MEDIAN REGRESSION MODELS FOR LONGITUDINAL EXPONENTIAL DATA

VARATHAN NAGARAJAH

Median Regression Models for Longitudinal Exponential Data

by

©Varathan Nagarajah

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirement for the Degree of

Master of Science in Statistics

Department of Mathematics and Statistics

Memorial University of Newfoundland

St. John's Newfoundland, Canada December 2012

Abstract

In independence setup, the quasi-likelihood estimation of the regression parameters involved in the mean function requires only the specification of the mean and variance function of the responses. In the longitudinal setup, one also has to accommodate the underlying correlation structure in order to obtain consistent and efficient regression estimates. Under the independence setup, when the responses follow an asymmetric distribution with a heavy tail, it has been argued in the literature that the regression estimates using mean regression model can be inefficient as compared to those obtained using a median regression model. Subsequently, the median regression models have been extended to study the asymmetric longitudinal data, but the longitudinal correlation of this type of asymmetric data have been computed using the moment estimates for all pairwise correlations. By considering an autoregressive order 1 (AR(1)) model for longitudinal exponential responses, in this thesis, it is 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. It is also argued in the thesis that the quasi-likelihood approach for median regression models may perform the same or worse than the mean regression models unless the data are highly asymmetric such as involving outliers. We illustrate the inference techniques discussed in the thesis by re-analysing the well-known labor pain data.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Professor Brajendra Sutradhar for the continuous support of my M.Sc study and research. To work with him has been a real pleasure to me, with heaps of fun and excitement. He has been a steady influence throughout my M.Sc program; and has oriented and supported me with promptness and care, and has always been patient and encouraging in times of new ideas and difficulties. His ability to select and to approach compelling research problems, his high scientific standards, and his hard work set an example. I admire his ability to balance research interests and personal pursuits. I have been extremely lucky to have a supervisor who cared so much about my work. Above all, I am especially grateful to Prof. Brajendra Sutradhar for his devotion to his students education and success.

I would also like to thank my co-supervisor Dr. Zhao Zhi Fan for his care and support during the program.

I am also grateful to the School of Graduate Studies and the Department of Mathematics and Statistics for their financial support in the form of Graduate Fellowship and Teaching Assistantship.

It is my great pleasure to thank all of my friends and well - wishers for making my stay in St.John's a pleasant and memorable one.

Finally, I take this opportunity to express the profound gratitude from my deep heart to my beloved parents, grandparents, and my siblings for their love and continuous support both spiritually and materially.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of the problem

In the Independence setup, when the responses follow an asymmetric distribution with a heavy tail, it has been argued in the literature (Bassett and Koenker (1978) and Koenker and Bassett (1978)) that the regression estimates using mean regression model can be inefficient as compared to those obtained using a median regression model. Subsequently many researchers such as Morgenthaler (1992), Kottas and Gelfand (2001), Kottas and Krnjajic (2009) and Reich, Bondell and Wang (2010), have studied various properties of the quantile regression estimates in the independence setup.

Suppose that a scalar response y_i and a p-dimensional vector of covariates x_i are observed for individuals $i = 1, 2, ..., K$. Let β denote the $p \times 1$ vector of regression parameters which measures the effects of x_i on y_i for all $i = 1, 2, ..., K$. When the responses are continuous and their distributions are asymmetric, fitting a median regression model (Morgenthaler (1992), Section 3) to this type of data is equivalent to use the relationship

$$y_i = g(m_i) + \varepsilon_i, \tag{1.1}$$

where m_i is the median of y_i, $g(.)$ is a suitable known link function, and ε_i is a model error component. Denote by f_i the density of the distribution of the ith individual such that $f_i(m_i) > 0$.

For

$$g(m_i) = x_i'\beta, \tag{1.2}$$

Morgenthaler (1992) proposed an absolute-deviation quasi-likelihood (ADQL) approach to estimate β consistently. To be specific, Morgenthaler (1992) suggested to solve the estimating equation

$$D^T\{\text{diag}[S_1, ..., S_i, ..., S_K]\}^{-1}\{\text{sgn}(y - m)\} = 0, \tag{1.3}$$

where

$$\begin{aligned} y &= (y_1, ..., y_i, ..., y_K)' \\ m &= (m_1, ..., m_i, ..., m_K)' \end{aligned}$$

with

$$\text{sgn}(y_i - m_i) = \begin{cases} 1 & \text{if } y_i \geq m_i \\ -1 & \text{otherwise,} \end{cases}$$

fo all $i = 1, ..., K$. Furthermore, $S_i = S(m_i)$ is a user-supplied dispersion function that models the scatter of the responses as a function of the median, and $D = (D_{ij}) : K \times p$, with $D_{ij} = \partial \mu_i / \partial \beta_j$, where $\mu_i = x_i' \beta$.

Note that for any random variable y_i with density f_i

$$\begin{aligned} E\{\text{sgn}(y_i - m_i)\} &= 1 \,.\, Pr(y_i \geq m_i) + (-1) \,.\, Pr(y_i < m_i) \\ &= 1 \,.\, \frac{1}{2} + (-1) \,.\, \frac{1}{2} \\ &= 0 \end{aligned}$$

implying that the estimating equation in (1.3) produces consistent estimates of the parameter β. Under some regularity conditions, Morgenthaler (1992) computed the variance of $\hat{\beta}$ (obtained from (1.3)) as

$$\text{var}(\hat{\beta}) = \{D^T S(m)^{-\frac{1}{2}} F D\}^{-1} \{D^T S(m)^{-1} D\} \{D^T S(m)^{-\frac{1}{2}} F D\}^{-1}, \qquad (1.4)$$

where $F = \text{diag}[2f_1(m_1), ..., 2f_K(m_K)]$. If $S_i^{-\frac{1}{2}} = 2f_i(m_i)$, then the var$(\hat{\beta})$ in (1.4) reduces to

$$\text{var}(\hat{\beta}) = \{D^T S(m)^{-1} D\}^{-1}.$$

To illustrate the use of the estimating equation (1.3) Morgenthaler (1992) considered an example with the response y_i following a gamma (asymmetric) distribution given by

$$f_i(y_i) = \frac{e^{-\frac{\nu_i}{\mu_i} y_i} y_i^{\nu_i - 1}}{\left(\frac{\nu_i}{\mu_i}\right)^{-\nu_i} \Gamma(\nu_i)}. \tag{1.5}$$

where

$$E(Y_i) = \int_0^\infty y_i f(y_i)\, dy_i = \int_0^\infty y_i \frac{e^{-\frac{\nu_i}{\mu_i} y_i} y_i^{\nu_i - 1}}{\left(\frac{\nu_i}{\mu_i}\right)^{-\nu_i} \Gamma(\nu_i)}\, dy_i = \mu_i,$$

and

$$\int_0^{m_i} f(y_i)\, dy_i = \int_0^{m_i} \frac{e^{-\frac{\nu_i}{\mu_i} y_i} y_i^{\nu_i - 1}}{\left(\frac{\nu_i}{\mu_i}\right)^{-\nu_i} \Gamma(\nu_i)}\, dy_i = \frac{1}{2} \Longrightarrow m_i = \mu_i \left[\frac{3\nu_i - 0.8}{3\nu_i + 0.2}\right]$$
$$= \mu_i m_{0i},$$

(Bannerheka and Ekanayake (2009)). It is clear that the median m_i and the mean μ_i of the ith observation are linked by a proportionality relation $m_i = m_{0i}\mu_i$. Evaluating the density at the median they found

$$f_i(m_i) = \Gamma(\nu_i)^{-1} \nu_i^{\nu_i} m_{0i}^{\nu_i} \exp(-\nu_i m_{0i})(1/m_i) \propto 1/m_i. \tag{1.6}$$

This (1.6) shows that a user, as suggested by Morgenthaler (1992), can choose $S_i(m_i)$ as $S_i \propto m_i^2$ implying that, $S_i = k_{0i}m_i^2$, where k_{0i} can be computed easily. Consequently, the ADQL estimating equation (1.3) reduces to

$$D^T \text{diag}[(\hat{m_i}^2)^{-1}]\{\text{sgn}(y - \hat{m})\} = 0, \tag{1.7}$$

which may be solved for β involved in the median function. Note that this ADQL approach appears to have several limitations. First S_i is chosen as $S_i \propto m_i^2$ which may be an appropriate choice only if m_i holds a proportionality relation to μ_i so that $m_i = m_{0i}\mu_i$ for a suitable constant m_{0i}. Second, using a mean response based gradient matrix D in (1.7) is also dependent on such proportionality relation between means and medians, which may not hold in general.

We remark that if the density of the response y_i at median m_i were known, one could use suitable likelihood approach to estimate β involved in the median function. For example, Jung (1996) has considered a longitudinal setup, where the distribution of the responses at the median is assumed to be known at any given time, but in general the marginal density may not be known. This implies that there is no way to compute the joint distribution of the repeated responses. Thus, Jung (1996) has extended the QL approach to the longitudinal setup but with an added restriction on the distribution of the response at the median. We now describe this median

regression problem in the longitudinal setup as follows.

Suppose that for the ith individual, $y_i = (y_{i1}, ..., y_{it}, ..., y_{iT})'$ is the $T \times 1$ response vector where elements are collected repeatedly over time. Let $X_i = (x_{i1}, ..., x_{it}, ..., x_{iT})'$ be 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 $t = 1, 2, ..., T$ and for all $i = 1, 2, ..., K$. When the responses are continuous and their distributions are symmetric, one fits the linear model

$$y_i = X_i\beta + \varepsilon_i \tag{1.8}$$

to estimate β. Suppose that $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT})'$ in (1.8) has the mean vector 0 and covariance matrix $\Sigma_i = A_i^{\frac{1}{2}} C(\rho) A_i^{\frac{1}{2}}$ with $A_i = \text{diag}[\text{var}(\varepsilon_{i1}), ..., \text{var}(\varepsilon_{it}), ... \text{var}(\varepsilon_{iT})]$ 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

$$\sum_{i=1}^{K} X_i' \Sigma_i^{-1} (y_i - X_i\beta) = 0. \tag{1.9}$$

Note that the estimating equation (1.9) does not depend on the distribution of y_i. However, 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 (1.8) can be inefficient (Bassett and Koenker (1978)). As a remedy, for the asymmetric data in the logitudinal setup, 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} be the median of y_{it} and $m_i = (m_{i1}, ..., m_{iT})'$. Furthermore, similar to (1.2), for some link function g(.), let

$$g(m_{it}) = x'_{it}\beta. \tag{1.10}$$

For longitudinal data, Jung (1996), for example, suggests to solve an indicator function based quasi-likelihood estimating equation, where the indicator variable is defined as $I(y_{it} \geq m_{it})$ with m_{it} is the median of y_{it} as in (1.10). More specifically, in the cluster regression setup, Jung's quasi-likelihood (QL) estimating equation can be expressed as

$$\sum_{i=1}^{K} \frac{1}{\phi_i} B_i' \Gamma_i \Omega_i^{-1} \{I(y_i \geq m_i) - \frac{1}{2} 1_T\} = 0, \tag{1.11}$$

where $I(y_i \geq m_i) = [I(y_{i1} \geq m_{i1}), ..., I(y_{it} \geq m_{it}), ..., I(y_{iT} \geq m_{iT})]'$ 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, i.e., $B_i = \partial m_i / \partial \beta'$, where $m_i = (m_{i1}, ..., m_{iT})'$ with $g(m_{it}) = x'_{it}\beta$, and $\phi_i^{-1}\Gamma_i = \phi_i^{-1}\text{diag}[\gamma(m_{i1}), ..., \gamma(m_{it}), ..., \gamma(m_{iT})]$, 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 (1.11) for β as the maximum quasi-likelihood estimate. Note that the QL estimating equation (1.11) may be treated as a generalization of the ADQL estimating

equation (1.3), 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 pair-wise bivariate distributions of the elements of the asymmetric response vector $y_i = (y_{i1}, ..., y_{it}, ..., y_{iT})'$. 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 pair-wise 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. Furthermore, in some situations when the pair-wise covariances will be a function of individual specific non-stationary covariates (such as under binary or count data models), one cannot take the average over the individuals to estimate the covariance or correlation matrix.

Note that in exponential failure times setup, some authors such as Hasan, Sutradhar and Sneddon (2007) dealt with mean regression model under a class of autocorrelation structures for the repeated exponential responses. But, no median regression was considered by these authors.

1.2 Objective of the thesis

Our main objectives are as follows:

1. We develop the median regression model for longitudinal exponential failure time data considered by Hasan et al. (2007).

2. We develop median regression based estimating equations for the regression parameters involved in the median function. This will be done by computing the exact correlation matrix under the assumed correlation model for repeated responses.

3. We will also consider several 'working' correlation structures based estimating equations. For example, following Jung (1996), (a) we will consider a non-parametric correlation structure based estimating equation approach, where pair-wise correlations are estimated by simple method of moments; (b) Second an independence assumption based correlation matrix will be used to construct

the estimating equations; (c) Also, a lag correlation based auto-correlation structure will be used to develop the desired estimating equations.

To examine the performance of the above mentioned inference techniques, we aim at a two-fold empirical study. First, the median regression estimation will be compared with mean regression based estimation. This comparison will also be studied by generating a few percentage of outlying observations, where in the absence of outliers the data are assumed to follow longitudinal exponential model. We also would like to compare the relative performance of the median based approaches, where these approaches differ from each other because of the correlation structures used to construct the respective estimating equations.

Chapter 2

Mean Regression Based GQL Estimation for Exponential Longitudinal Models

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 shared 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 the individual specific random effect may remain the same

throughout the data collection period and hence cannot represent any time effects.

Some other authors such as Hasan et al. (2007, Section 2.1, p. 552) have considered a class of non-stationary auto-correlation models for longitudinal exponential failure time data, AR(1) [auto-regressive of order 1] model being an important special case. Suppose that y_{it} is the exponential response collected at time t $(t = 1, ..., T)$ for the ith $(i = 1, ..., K)$ individual and $\mu_{it}(\beta)$ is the mean of y_{it}. Let $y_i = (y_{i1}, ..., y_{it}, ..., y_{iT})'$ with its mean vector $\mu_i(\beta) = (\mu_{i1}(\beta), ..., \mu_{it}(\beta), ..., \mu_{iT}(\beta))'$. Next suppose that $\Sigma_i(\beta, \rho)$ is the covariance matrix of y_i which has the formula

$$\Sigma_i(\beta, \rho) \quad = \quad A_i^{1/2} C_i(\rho) A_i^{1/2} \tag{2.1}$$

where $A_i = \text{diag}\,[\sigma_{i11}(\beta), ..., \sigma_{itt}(\beta), ..., \sigma_{iTT}(\beta)]$ with $\sigma_{itt}(\beta) = \text{var}(Y_{it})$ and $C_i(\rho) = (c_{iut}(\rho))$, $c_{iut}(\rho)$ being the correlation between y_{iu} and y_{it}. In (2.1), $C_i(\rho)$ is the $T \times T$ correlation matrix of $y_i = (y_{i1}, ..., y_{it}, ..., y_{iT})'$. Note that under a class of auto-correlation structures, this $C_i(\rho)$ takes the form

$$C_i(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{T-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho_{T-1} & \rho_{T-2} & \cdot & \dots & 1 \end{bmatrix}, \tag{2.2}$$

(Sutradhar (2003)) where ρ_ℓ is known as the ℓth $(\ell = 1, ..., T-1)$ lag auto-correlation

between y_{it} and $y_{i,t+\ell}$. Hasan et al. (2007) have shown that the longitudinal exponential data following ARMA (p^*, q^*) [auto -regressive moving average of order p^* and q^*] models satisfy the correlation matrix form in (2.2). For convenience, following Hasan et al. (2007), we demonstrate the validity of the general auto-correlation structure (2.2) for low order correlation models, namely, AR(1), MA(1), and EQC.

2.1 Exponential AR(1) Model - EAR(1)

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(.). That is,

$$f(y_{i1}) = \lambda_{i1} \exp(-\lambda_{i1}\ y_{i1}). \tag{2.3}$$

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

$$y_{it} = \rho_i y_{i,t-1} + I_{it} a_{it}, \quad t = 2, ..., T\ ;\ i = 1, ..., K \tag{2.4}$$

where, $\{a_{it}, t = 1, ..., T;\ i = 1, ..., K\}$ 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 $(0 \leq \rho \leq 1)$. In (2.4), 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.

Next we compute the mean, variance and correlations of the responses for the exponential AR(1) model (2.3)-(2.4). To be specific, following Hasan et al. (2007), one may write

$$\begin{aligned}\mu_{il}(\beta) = E(Y_{il}) &= E_{y_{i,t-1}} E\left[\{\rho_i y_{i,t-1} + I_{it} a_{it}\} \mid y_{i,t-1}\right] \\ &= E\left[\rho_i y_{i,t-1} + E(I_{it}) E(a_{it})\right] \\ &= \rho_i \left\{\frac{1}{\lambda_{i,t-1}}\right\} + (1-\rho)\left\{\frac{1}{\lambda_{it}}\right\} = \frac{1}{\lambda_{it}}, \end{aligned} \tag{2.5}$$

$$\begin{aligned}\sigma_{itt}(\beta) = \mathrm{var}(Y_{it}) &= E_{y_{i,t-1}} V[\{\rho_i y_{i,t-1} + I_{it} a_{it}\} \mid y_{i,t-1}] \\ &\quad + V_{y_{i,t-1}} E\left[\{\rho_i y_{i,t-1} + I_{it} a_{it}\} \mid y_{i,t-1}\right] = \frac{1}{\lambda_{it}^2}, \end{aligned} \tag{2.6}$$

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

$$c_{iut}(\rho) = Corr(Y_{iu}, Y_{it}) = \rho^{|t-u|}. \tag{2.7}$$

2.2 Exponential Moving Average of Order 1 Model - EMA(1)

Suppose that the repeated responses $y_{i1}, ..., y_{it}, ..., y_{iT}$ follow an exponential moving average of order 1 (EMA(1)) process. The EMA(1) model may be represented as

$$y_{it} = \rho a_{it} + I_{it}\eta_{it}a_{i(t+1)}, \tag{2.8}$$

(Hasan et al. (2007)) where ρ, I_{it} and a_{it} are as defined under the exponential AR(1) model (2.4) and $\eta_{it} = \lambda_{i(t+1)}/\lambda_{it}$. Then the mean, variance and correlation for repeated responses under (2.8) can be found as,

$$\begin{aligned} E(Y_{it}) &= E[\rho a_{it} + I_{it}\eta_{it}a_{i(t+1)}] \\ &= \rho E(a_{it}) + E(I_{it})E(\eta_{it})E(a_{it}) \\ &= \rho\left\{\frac{1}{\lambda_{it}}\right\} + (1-\rho)\left\{\frac{\lambda_{i(t+1)}}{\lambda_{it}}\right\}\left\{\frac{1}{\lambda_{i(t+1)}}\right\} \\ &= \frac{1}{\lambda_{it}}, \end{aligned} \tag{2.9}$$

$$\begin{aligned} \text{var}(Y_{it}) &= \text{v}[\rho a_{it} + I_{it}\eta_{it}a_{i(t+1)}] \\ &= \rho^2\text{v}(a_{it}) + \eta_{it}^2\text{v}[I_{it}a_{i(t+1)}] \\ &= \rho^2\text{v}(a_{it}) + \eta_{it}^2\left\{E^2(I_{it})\text{v}(a_{i(t+1)}) + E(I_{it})\text{v}^2(a_{i(t+1)}) + \text{v}(I_{it})\text{v}(a_{i(t+1)})\right\} \\ &= \frac{1}{\lambda_{it}^2}, \end{aligned} \tag{2.10}$$

and

$$\begin{aligned}\text{cov}(Y_{it}, Y_{i,t-\ell}) &= \text{cov}[\rho a_{it} + I_{it}\eta_{it}a_{i,(t+1)}, \rho a_{i,t-\ell} + I_{i,t-\ell}\eta_{i,t-\ell}a_{i,(t+1)-\ell}] \\ &= \text{cov}[\rho a_{it}, I_{i,t-\ell}\eta_{i,t-\ell}a_{i,(t+1)-\ell}] \\ &= \frac{\rho(1-\rho)}{\lambda_{it}\lambda_{i,t-\ell}},\end{aligned}$$

yielding

$$\text{corr}(Y_{it}, Y_{iu}) = \begin{cases} \rho(1-\rho) & \text{if } |t-u| = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

2.3 Exponential Equi-correlation Model - EEQC

Suppose that the initial response y_{i0} follow an exponential distribution with time independent parameter λ_{i0} (= 1, say). Further suppose that the response y_{it} has a functional relationship with y_{i0} of the form

$$y_{it} = \rho_i y_{i0} + I_{it}a_{it} \tag{2.12}$$

(Hasan et al. (2007)) where for all $t = 1, 2, ..., T$, a_{it} follows exponential distribution with parameter λ_{it} and $\rho_i = \rho\lambda_{i0}/\lambda_{it}$. It can be shown under (2.12) that the mean,

variance and correlation have the form given by

$$\begin{aligned} E(Y_{it}) &= E[\rho_i y_{i0} + I_{it} a_{it}] \\ &= \rho_i E(Y_{i0}) + E(I_{it}) E(a_{it}) \\ &= \frac{1}{\lambda_{it}}, \end{aligned} \tag{2.13}$$

$$\begin{aligned} \mathrm{v}(Y_{it}) &= \mathrm{v}[\rho_i y_{i0} + I_{it} a_{it}] \\ &= \rho_i^2 \mathrm{v}(Y_{i0}) + \mathrm{v}(I_{it} a_{it}) \\ &= \rho_i^2 \mathrm{v}(Y_{i0}) + \left\{ E^2(I_{it}) \mathrm{v}(a_{it}) + E(I_{it}) \mathrm{v}^2(a_{it}) + \mathrm{v}(I_{it}) \mathrm{v}(a_{it}) \right\} \\ &= \frac{1}{\lambda_{it}^2}, \end{aligned} \tag{2.14}$$

and

$$\begin{aligned} \mathrm{cov}(Y_{it}, Y_{i,t-\ell}) &= \mathrm{cov}[\rho_i y_{i0} + I_{it} a_{it}, \rho_i y_{i0} + I_{i,t-\ell} a_{i,t-\ell}] \\ &= \mathrm{cov}[\rho_i y_{i0}, \rho_i y_{i0}] \\ &= \rho_i^2 \mathrm{v}(Y_{i0}) \\ &= \frac{\rho^2}{\lambda_{it}^2}, \text{ for all } \ell \neq 0 \end{aligned}$$

yielding

$$\mathrm{corr}(Y_{it}, Y_{iu}) = \rho^2. \tag{2.15}$$

It is now clear from (2.7), (2.11), and (2.15) that the correlations under all three processes, namely, AR(1), MA(1) and EQC models for exponential data, satisfy the general auto-correlation matrix $C_i(\rho)$ in (2.2).

2.4 Mean Regression Based GQL Estimating Equation

Turning back to the estimation of β involved in the mean function $\mu_{it}(\beta) = \frac{1}{\lambda_{it}} = \exp(x'_{it}\beta)$, we may follow Sutradhar (2003) and write the GQL (Generalized Quasi-likelihood) estimating equation for β as

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\beta, \rho) \ (y_i - \mu_i) = 0 \tag{2.16}$$

where $C_i(\rho)$ involved in $\Sigma_i(\beta, \rho)$ has the auto-correlation structure form (2.2). Note that for known ρ, this equation (2.16) may be solved iteratively using

$$\hat{\beta}(r+1) = \hat{\beta}(r) + \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\beta, \rho) \frac{\partial \mu_i}{\partial \beta'}\right]_r^{-1} \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(y_i - \mu_i)\right]_r \tag{2.17}$$

where $[.]_r$ is computed by evaluating the quantity in [ ] using $\beta = \hat{\beta}(r)$. Next, because ρ_ℓ $(\ell = 1, ..., T-1)$ is unknown in practice, it must be estimated. For the estimation of this lag correlation parameter, we may use the method of moments and solve the ℓth $(\ell = 1, ..., T-1)$ lag correlation based moment equation. To be specific, the

moment estimator for ρ_ℓ has the formula

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{K}\sum_{t=1}^{T-\ell} \tilde{y}_{it}\ \tilde{y}_{i,t+1}/K(T-\ell)}{\sum_{i=1}^{K}\sum_{t=1}^{T} \tilde{y}_{it}^2/KT} \tag{2.18}$$

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

Chapter 3

Median Regression Based Estimation

It is generally recognized that the mean regression based estimate for β obtained from (2.16) may be biased when the responses y_{it} exhibit an asymmetric pattern. See for example, Bassett and Koenker (1978) for a discussion on this issue in a cross-sectional setup (i.e, when $T = 1$). As pointed out in Chapter 1, to resolve this problem, Morgenthaler (1992) has proposed ADQL approach to obtain median regression based estimates in a cross-sectional setup, and Jung (1996), for example, has proposed a median regression based QL approach to obtain unbiased and consistent estimator of β involved in the median function in a longitudinal setup, where the pair-wise

longitudinal correlations are estimated non-parametrically. However, because of non-parametric estimation of the true correlation structure, the QL approach may produce less efficient regression estimates, especially when the true correlation follow a specific pattern, such as the correlation structure given in (2.2). Moreover, there is no simulation study in Jung (1996) to examine the performance of the QL estimation approach.

The purpose of this chapter is to consider a specific correlation model for exponential data and develop median based regression estimates, where the correlation matrix will be estimated in various ways. In Section 3.1, we consider the AR(1) exponential model discussed in Section 2.1 and develop a median regression model based GQL estimating equation. The construction of the estimating equation is given in details.

3.1 GQL Estimation for Median Regression Model with AR(1) (Known Correlation Structure) Exponential Data

To develop the median regression based estimation approach, we minimize the distance function

$$[\delta(y_{it} \geq m_{it}) - E\{\delta(y_{it} \geq m_{it})\}] \tag{3.1}$$

for all $i = 1, ..., K$ and $t = 1, ..., T$, whereas in the mean regression based GQL approach we have minimized the distance function

$$[y_{it} - E(y_{it})] \tag{3.2}$$

for all i and t, where $y_{i1}, ..., y_{it}, ..., y_{iT}$ are correlated following the AR(1) model, that is, with correlation structure given by (2.7).

In (3.1), $\delta(y_{it} \geq m_{it})$ is an indicator variable defined as

$$\delta(y_{it} \geq m_{it}) = \begin{cases} 1 & \text{if} \quad y_{it} \geq m_{it} \\ 0 & \text{if} \quad y_{it} < m_{it} \end{cases} \tag{3.3}$$

where the median m_{it} can be derived as

$$\int_0^{m_{it}} f(y_{it})\, dy_{it} = \int_0^{m_{it}} \frac{1}{\lambda_{it}} \exp(-\lambda_{it} y_{it})\, dy_{it} = \frac{1}{2} \Longrightarrow m_{it} = \frac{log\, 2}{\lambda_{it}}. \tag{3.4}$$

The mean and variance of this indicator variable are given by

$$\begin{aligned}\tilde{\mu}_{it} = E[\delta(y_{it} \geq m_{it})] &= 1 \,.\, Pr(y_{it} \geq m_{it}) + 0 \,.\, Pr(y_{it} < m_{it}) \\ &= 1 \,.\, \frac{1}{2} + 0 \,.\, \frac{1}{2} \\ &= \frac{1}{2} \end{aligned} \tag{3.5}$$

and

$$\begin{aligned}\tilde{\sigma}_{itt} = \text{var}[\delta(y_{it} \geq m_{it})] &= E[\{\delta(y_{it} \geq m_{it})\}^2] - [E\{\delta(y_{it} \geq m_{it})\}]^2 \\ &= [1^2 \,.\, Pr(y_{it} \geq m_{it}) + 0^2 \,.\, Pr(y_{it} < m_{it})] - \tilde{\mu}_{it}^2 \\ &= \frac{1}{2} - \frac{1}{4} \\ &= \frac{1}{4} \end{aligned} \tag{3.6}$$

Furthermore, let

$$\delta(y_i \geq m_i) = [\delta(y_{i1} \geq m_{i1}), ..., \delta(y_{it} \geq m_{it}), ..., \delta(y_{iT} \geq m_{iT})]' : T \times 1. \tag{3.7}$$

It then follows that

$$E[\delta(y_i \geq m_i)] = \frac{1}{2}\mathbf{1}_T' \tag{3.8}$$

Next, denote the covariance between $\delta(y_{iv} \geq m_{iv})$ and $\delta(y_{it} \geq m_{it})$ by $\tilde{\sigma}_{ivt}$. That is

$$\begin{aligned}\tilde{\sigma}_{ivt} = \text{cov}[\delta(y_{iv} \geq m_{iv}), \delta(y_{it} \geq m_{it})] &= E[\delta(y_{iv} \geq m_{iv})\ \delta(y_{it} \geq m_{it})] - \tilde{\mu}_{iv}\tilde{\mu}_{it} \\ &= Pr(y_{iv} \geq m_{iv}, y_{it} \geq m_{it}) - \frac{1}{4}\,. \end{aligned} \tag{3.9}$$

The computation of $Pr(y_{iv} \geq m_{iv}, y_{it} \geq m_{it})$ under the AR(1) model (2.4) is given in the next subsection 3.1.1.

3.1.1 Formula for Joint Probability

Under AR(1) structure, the joint probability involved in (3.9) has the formula

$$Pr(y_{iv} \geq m_{iv}, y_{i(v+j)} \geq m_{i(v+j)}) = \begin{cases} e^{-\lambda_{iv} m_{iv}}, & \text{for} \quad m_{i(v+j)} \leq \rho_{v+j}\rho_{v+j-1}\cdots\rho_{v+1} m_{iv} \\ e^{-\lambda_{i(v+j)} m_{i(v+j)}} e^{-\lambda_{iv}(1-\rho^j) m_{iv}}, & \text{for} \quad m_{i(v+j)} > \rho_{v+j}\rho_{v+j-1}\cdots\rho_{v+1} m_{iv} \end{cases} \tag{3.10}$$

[Hasan (2004, Section 4.1.1, p. 59)] where $\rho_v = \rho^{\frac{\lambda_{i(v-1)}}{\lambda_{iv}}}$ is the lag v correlation for all $i = 1, ..., K$. For convenience, we highlight its derivation in the Appendix.

Note that for stationary case, $\lambda_{i(v-1)} = \lambda_{iv}$, implying that $\rho_v = \rho$. It then follows that the bivariate probability function in (3.10) reduces to

$$Pr(y_{iv} \geq m_{iv}, y_{i(v+j)} \geq m_{i(v+j)}) = \begin{cases} e^{-\lambda_{iv} m_{iv}}, & \text{for} \quad m_{i(v+j)} \leq \rho^j m_{iv} \\ e^{-\lambda_{i(v+j)} m_{i(v+j)}} e^{-\lambda_{iv}(1-\rho^j) m_{iv}}, & \text{for} \quad m_{i(v+j)} > \rho^j m_{iv} \end{cases} \tag{3.11}$$

3.1.2 GQL Estimating Equation for β

Note that under the AR(1) model, we have computed $\tilde{\sigma}_{ivt} = \text{cov}[\delta(y_{iv} \geq m_{iv}), \delta(y_{it} \geq m_{it})]$ by (3.9) (see also (3.11)). In matrix notation, we write

$$\tilde{\Sigma}_{i,\delta}(\beta, \rho) = (\tilde{\sigma}_{ivt}) = (\text{cov}[\delta(y_{iv} \geq m_{iv}), \delta(y_{it} \geq m_{it})]). \tag{3.12}$$

Next, following Sutradhar (2003) (See also Jung (1996)), we write the median regression based GQL estimating equation for β as

$$\sum_{i=1}^{K} \frac{\partial \delta'(y_i \geq m_i)}{\partial \beta} \tilde{\Sigma}_{i,\delta}^{-1}(\beta, \rho) \, [\delta(y_i \geq m_i) - E[\delta(y_i \geq m_i)]] = 0. \tag{3.13}$$

The formula for the derivative $\frac{\partial \delta'(y_i \geq m_i)}{\partial \beta}$ in (3.13) is given in the following subsection.

3.1.3 Formula for the derivative $\frac{\partial \delta'(y_i \geq m_i)}{\partial \beta} : p \times T$

Consider

$$\frac{\partial \delta(y_i \geq m_i)}{\partial \beta'} = \frac{\partial \delta(y_i \geq m_i)}{\partial m_i'} \frac{\partial m_i}{\partial \beta'}. \tag{3.14}$$

Note that $\frac{\partial \delta(y_i \geq m_i)}{\partial m_i'}$ may be computed by computing the general element $\frac{\partial \delta(y_{it} \geq m_{it})}{\partial m_{it}}$. Because $\delta(y_{it} \geq m_{it}) = 1$ when $y_{it} \geq m_{it}$, and $\delta(y_{it} \geq m_{it}) = 0$ when $y_{it} < m_{it}$, it is equivalent to write $\delta(y_{it} \geq m_{it}) \equiv \text{sgn}(y_{it} - m_{it})$ with

$$\delta(y_{it} \geq m_{it}) = \begin{cases} 1 & \text{if} \quad \text{sgn}(y_{it} - m_{it}) = 1 \\ 0 & \text{if} \quad \text{sgn}(y_{it} - m_{it}) = -1 \end{cases} \tag{3.15}$$

Next, let $F_{it}(y_{it}) = \int_0^{y_{it}} f(y_{it}) dy_{it}$ be the cumulative distribution function. Then we may write

$$\begin{aligned} \delta(y_{it} \geq m_{it}) &\equiv \text{sgn}(y_{it} - m_{it}) \\ &= [2F_{it}(y_{it} - m_{it}) - 1]. \end{aligned} \tag{3.16}$$

This is because for $y_{it} \geq m_{it}$ one writes

$$\frac{1}{2} \leq F_{it}(y_{it}) \leq 1 \quad \text{where} \quad 0 < y_{it} < \infty.$$

This implies that for $z_{it} = y_{it} - m_{it}$

$$\frac{1}{2} \leq F_{it}(z_{it}) \leq 1 \quad \text{for} \quad -m_{it} < y_{it} < \infty.$$

That is

$$\frac{1}{2} \leq F_{it}(y_{it} - m_{it}) \leq 1.$$

By the same token, when $y_{it} < m_{it}$, we write

$$0 \leq F_{it}(y_{it} - m_{it}) < \frac{1}{2}.$$

Hence

$$\begin{aligned} 2F_{it}(y_{it} - m_{it}) - 1 &= \begin{cases} +\text{ve} & \text{for } y_{it} \geq m_{it} \\ -\text{ve} & \text{for } y_{it} < m_{it} \end{cases} \\ &= \text{sgn}(y_{it} - m_{it}). \end{aligned} \tag{3.17}$$

It then follows that

$$\frac{\partial \delta(y_{it} \geq m_{it})}{\partial m_{it}} = -2f_{it}(m_{it}), \tag{3.18}$$

yielding

$$\frac{\partial \delta(y_i \geq m_i)}{\partial m_i'} = -2f_i(m_i), \tag{3.19}$$

where

$$f_i(m_i) = \text{diag } [f_{i1}(m_{i1}), ..., f_{it}(m_{it}), ..., f_{iT}(m_{iT})] : T \times T, \tag{3.20}$$

with

$$f_{it}(m_{it}) = f(y_{it})|_{\, y_{it}=m_{it}}, \tag{3.21}$$

is the density of the repeated response y_{it} evaluated at the median m_{it}. Hence the derivative in (3.14) has the formula

$$\frac{\partial \delta(y_i \geq m_i)}{\partial \beta'} = -2f_i(m_i)D_i' \tag{3.22}$$

where

$$D_i = \frac{\partial m_i'}{\partial \beta} = \left[\frac{\partial m_{i1}}{\partial \beta}, ..., \frac{\partial m_{it}}{\partial \beta}, ..., \frac{\partial m_{iT}}{\partial \beta}\right] : p \times T \tag{3.23}$$

with

$$\frac{\partial m_{it}}{\partial \beta} = \frac{\partial (log2/\lambda_{it})}{\partial \beta}.$$

Consequently

$$\begin{aligned} \frac{\partial \delta'(y_i \geq m_i)}{\partial \beta} &= -[2f_i(m_i)D_i']' \\ &= -2D_i f_i(m_i) \end{aligned} \tag{3.24}$$

Turning back to (3.13), we solve this estimating equation for β by using the Newton-Raphson iterative equation

$$\hat{\beta}(r+1) = \hat{\beta}(r) - \frac{1}{2}\left[\sum_{i=1}^{K} D_i f_i(m_i)\ \hat{\tilde{\Sigma}}_{i,\delta}^{-1}\ f_i(m_i) D_i'\right]_r^{-1} \left[\sum_{i=1}^{K} D_i f_i(m_i)\ \hat{\tilde{\Sigma}}_{i,\delta}^{-1}\{\delta(y_i \geq m_i) - \frac{1}{2}\mathbf{1}_T\}\right]_r \quad (3.25)$$

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

3.2 GQL Estimation With Unknown Correlation Structure

Note that the GQL estimating equation in (3.13) is quite general, even though we have computed $\tilde{\Sigma}_{i,\delta}$ for this equation under the AR(1) model. More specifically, if the correlation structure for exponential data are known, one may attempt to compute the $\tilde{\Sigma}_{i,\delta}$ accordingly. However, because in practice one may not know the exact correlation structure, some authors such as Jung (1996) has used an unstructured correlation matrix and estimated pairwise correlations non-parametrically. But the performance of this approach was not adequately studied, for example, by comparing with any possible known parametric structure. As opposed to this type of 'working' unstructured correlation approach, there exist a GQL approach (Sutradhar (2003)) where it is assumed that the data follow a class of autocorrelation structures that accommodates AR(1), MA(1) and EQC types of correlations. But this and other

related studies were confined to longitudinal count and/or binary data. However, as explained in Chapter 2, because this auto-correlation class also holds for exponential data, it makes this GQL approach as a reasonable alternative estimation approach. In this thesis we consider the GQL approach for repeated exponential data and examine its performance with other approaches including Jung's QL approach. We also will consider a simple 'independence' assumption based 'working' GQL approach. For convenience, we summarize these three approaches as in the following sections.

3.2.1 Jung's Approach

To apply the QL estimating equation (1.11), Jung (1996) has estimated the pairwise elements of $\tilde{\Sigma}_{i,\delta}$ matrix in (3.13) 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

$$\hat{Pr}(y_{iv} \geq m_{iv}, y_{it} \geq m_{it}) = \frac{\sum_{i=1}^{K} \delta(y_{iv} \geq m_{iv})\delta(y_{it} \geq m_{it})}{K}. \tag{3.26}$$

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

$$\tilde{\sigma}_{ivt} = \frac{\sum_{i=1}^{K} \delta(y_{iv} \geq m_{iv})\delta(y_{it} \geq m_{it})}{K} - \frac{1}{4}. \tag{3.27}$$

The QL estimate of β is then obtained by solving (1.11) or equivalently using the iterative equation (3.25), where $\tilde{\Sigma}_{i,\delta}$ is computed by using (3.27).

3.2.2 Lag-Correlation Approach

Note that, when the repeated responses $y_{i1}, ..., y_{iT}$ follow EAR(1), EMA(1) or EEQC structures, their correlations become lag dependent as in the Gaussian case. Thus, $\text{corr}(y_{iv}, y_{it})$ depends on $|v - t|$ rather than individuals $v, t = 1, ..., T$. For example, in the EAR(1) case $\text{corr}(y_{iv}, y_{it}) = \rho^{|t-v|}$, in EMA(1) case

$$\text{corr}(y_{it}, y_{iu}) = \begin{cases} \rho(1-\rho) & \text{if } |t-u| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

and in EEQC case $\text{corr}(y_{iv,it}) = \rho^2$. All these correlations may be represented simply by $\rho_{|t-v|}$. Thus irrespective of the correlation structure in such an auto-correlation class, one may use the auto-correlation matrix from (2.2) as

$$\tilde{C}_i(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{T-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho_{T-1} & \rho_{T-2} & . & \dots & 1 \end{bmatrix}. \tag{3.28}$$

When $\tilde{\Sigma}_{i,\delta}$ computed using (3.28) is compared to the unstructured covariance matrix (3.27) used by Jung (1996), it is clear that $\tilde{C}_i(\rho)$ based $\tilde{\Sigma}_{i,\delta}$ matrix requires fewer

elements. For the estimation of ρ_ℓ ($\ell = 1, ..., T-1$), we use moment estimate as follows:

$$\hat{\rho}_\ell = \left[\sum_{i=1}^{K} \sum_{t=1}^{T-\ell} \delta(y_{it} \geq m_{it}, y_{i,t+\ell} \geq m_{i,t+\ell}) / K(T-\ell) - \frac{1}{4} \right] / (1/4). \tag{3.29}$$

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 (3.13) or equivalently using the iterative equation (3.25).

3.2.3 Using Independence among repeated responses

In this case, the correlation index parameter ρ is assumed to be zero. Consequently, the pairwise bivariate probabilities for $\delta(y_{iv} \geq m_{iv})$ and $\delta(y_{it} \geq m_{it})$ given in (3.11), for example, reduce to

$$Pr(y_{iv} \geq m_{iv}, y_{it} \geq m_{it}) = e^{-\lambda_{iv} m_{iv}} e^{-\lambda_{it} m_{it}} = \frac{1}{4}, \tag{3.30}$$

because the median $m_{it} = \frac{log2}{\lambda_{it}}$ under the exponential model. Applying (3.30) to (3.9), one obtains zero covariances or correlations, i.e, $\tilde{\sigma}_{ivt} = 0$. Thus, one may simply use $\tilde{\Sigma}_{i,\delta} = Cov(\delta(y_i \geq m_i)) = A_{i,\delta}^{1/2} \tilde{C}_{i,\delta}(\rho) A_{i,\delta}^{1/2} = \frac{1}{4}\tilde{C}_{i,\delta} = \frac{1}{4} I_T$ in (3.13) or (3.25) for the estimation of β.

Chapter 4

Simulations Based Empirical Study

Note that when responses are asymmetric, the mean based regression estimates may be inefficient (Bassett and Koenker (1978)) as compared to the median based regression estimates. For the purpose of obtaining efficient regression estimates, in Chapter 3, we have discussed various median based GQL approaches constructed by using different forms and/or estimates for the covariance matrix $\tilde{\Sigma}_{i,\delta}$. As far as the mean based regression estimation is concerned in Chapter 2, we have discussed the GQL estimating equations approach for such an estimation for the regression parameters. In this chapter, we conduct an extensive simulation study to compare all the above approaches for the estimation of the parameters involved in a regression model for longitudinal exponential (asymmetric) data.

4.1 Simulation Design

For the generation of the repeated responses $y_{i1}, ..., y_{it}, ..., y_{iT}$, we follow the EAR(1) model (2.4), i.e.,

$$y_{it} = \rho_i y_{i,t-1} + I_{it} a_{it} \quad i = 1, ..., K \; ; \quad t = 2, ..., T \tag{4.1}$$

where, $\{a_{it}\}$ is a sequence of exponential random variables with parameter $\lambda_{it} = \exp(-x'_{it}\beta)$, $\rho_i = \rho \frac{\lambda_{i,t-1}}{\lambda_{it}}$ with ρ as a probability parameter or correlation parameter $(0 \leq \rho \leq 1)$ and I_{it} is an indicator variable such that $I_{it} \sim \text{bin}(1 - \rho)$. For our purposes, we choose

- $K = 100$ individuals, $T = 4$ time points
- Consider scalar $\beta = 0.5, 0.7, 1.0$
- For covariates x_{it}, we assume that they are staionary, i.e., they are not time dependent and consider

  $$x_{it} = \tilde{x}_i \sim U(0, 1)$$

  where $U(0, 1)$ denotes the Uniform distribution in the interval 0 to 1.
- For correlation index parameter ρ, we choose $\rho = 0.0, 0.5$ and 0.7.

4.2 Steps For Data Generation

Step 1: Generate y_{i1} such that $y_{i1} \sim \text{Exp}(\lambda_{i1} = \exp(-x'_{i1}\beta))$. Note that in general marginally y_{it} for all $t = 1, ..., T$ will follow $\text{Exp}(\lambda_{it})$ with

$$f(y_{it}) = \lambda_{it} e^{-\lambda_{it} y_{it}}.$$

Step 2: Generate a_{it} for $i = 1, ..., K$; $t = 1, ..., T$ following

$$a_{it} \sim \text{Exp}(\lambda_{\text{it}}).$$

Step 3: Generate indicator variable I_{it} $(i = 1, ..., K$; $t = 1, ..., T)$ following binary distribution with probability 1 - ρ. that is, for I_{it} we follow

$$I_{it} \sim \text{bin}(1 - \rho).$$

Step 4: Using a_{i2}, I_{i2} and y_{i1}, generated by following Steps 1 to 3, we generate y_{i2} by using (4.1). Next using a_{i3}, I_{i3} and y_{i2}, we generate y_{i3}. This continues until we generate y_{it} for $t = T$.

To have some feelings about the asymmetriness, we have used a selected set of parameters and with $\rho = 0$, and computed the averages of population means (pm) and medians (pmd) and sample means (sm) and medians (smd), and present them in the Table 4.1 below.

Table 4.1: Mean and median comparison for selected T, and for $\beta = 0.5$, $\rho = 0.0$ and $K = 100$

	Average of			
T	pm	pmd	sm	smd
10	1.0993	0.7619	2.1182	2.0031
50	1.0993	0.7619	2.4325	2.1621
100	1.0993	0.7619	2.6528	2.3502

In Table 4.1, the difference between the average of the population medians and the average of the population means shows asymmetriness. But the degree of asymmetriness does not appear to be high for this selected design.

4.3 Simulation Results

In this study, we have estimated the model parameters in 500 simulations using various combinations of the parameter values of $\beta = 0.5$, 0.7, & 1.0 and $\rho = 0.0$, 0.5, & 0.7. For the estimation of β, we first use the mean regression based GQL estimating equation (2.16) constructed by exploiting a class of auto-correlation models and denote the estimate as GQL(AC) estimate, where AC stands for a general auto-correlation structure. Next, for the median regression based estimation for β, we use the GQL(TC) estimating equation (3.13) constructed by using true EAR(1) model, and three non-parametric correlation structures based estimating equations,

namely by using Jung's QL approach (JQL) from 3.2.1; lag-correlation (LC) based GQL (GQL(LC)) from 3.2.2; and by using independence among repeated responses (IND) from 3.2.3. The simulated means (SM), simulated standard errors (SSE), simulated mean square errors (SMSE), percentage efficiency (E_1) among median regression based approaches, and overall percentage efficiency (E_2) as compared to mean regression based approach for the estimates of β are reported in Tables 4.2, 4.3 and 4.4 for $\rho = 0.0$, 0.5, and 0.7, respectively.

Table 4.2: Comparison of mean regression and median regression approaches for the estimation of regression parameter ($\beta = 0.5$, 0.7, 1.0) involved in an EAR(1) model with a correlation value $\rho = 0.0$; based on 500 simulations.

				Statistic				
ρ	β	Regression	Estimation Approach	SM	SSE	SMSE	E_1	E_2
0.0	0.5	Median	GQL(TC)	0.4993	0.0748	0.0056	100	52
			IND	0.4995	0.0745	0.0055	98	53
			JQL	0.4949	0.0839	0.0071	79	41
			GQL(LC)	0.4992	0.0754	0.0055	98	53
		Mean	GQL(AC)	0.4968	0.0535	0.0029	-	100
	0.7	Median	GQL(TC)	0.6993	0.0746	0.0056	100	52
			IND	0.6992	0.0744	0.0055	98	53
			JQL	0.6947	0.0843	0.0071	79	41
			GQL(LC)	0.6993	0.0753	0.0055	98	53
		Mean	GQL(AC)	0.6968	0.0535	0.0029	-	100
	1.0	Median	GQL(TC)	0.9994	0.0745	0.0055	100	52
			IND	0.9994	0.0745	0.0055	98	53
			JQL	0.9948	0.0841	0.0071	79	41
			GQL(LC)	0.9993	0.0752	0.0055	98	53
		Mean	GQL(AC)	0.9968	0.0535	0.0029	-	100

Table 4.3: Comparison of mean regression and median regression approaches for the estimation of regression parameter ($\beta = 0.5, 0.7, 1.0$) involved in an EAR(1) model with a correlation value $\rho = 0.5$; based on 500 simulations.

				Statistic				
ρ	β	Regression	Estimation Approach	SM	SSE	SMSE	E_1	E_2
0.5	0.5	Median	GQL(TC)	0.5018	0.1003	0.0100	100	53
			IND	0.5021	0.0999	0.0100	100	53
			JQL	0.5014	0.1125	0.0126	79	42
			GQL(LC)	0.5016	0.1014	0.0102	98	52
		Mean	GQL(AC)	0.5016	0.0726	0.0053	-	100
	0.7	Median	GQL(TC)	0.7016	0.1005	0.0100	100	53
			IND	0.7022	0.0999	0.0100	100	53
			JQL	0.7015	0.1124	0.0126	79	42
			GQL(LC)	0.7018	0.1014	0.0102	98	52
		Mean	GQL(AC)	0.7016	0.0726	0.0053	-	100
	1.0	Median	GQL(TC)	1.0019	0.1002	0.0100	100	53
			IND	1.0021	0.1002	0.0101	100	53
			JQL	1.0008	0.1127	0.0127	79	42
			GQL(LC)	1.0016	0.1013	0.0102	98	52
		Mean	GQL(AC)	1.0016	0.0726	0.0053	-	100

Table 4.4: Comparison of mean regression and median regression approaches for the estimation of regression parameter ($\beta = 0.5, 0.7, 1.0$) involved in an EAR(1) model with a 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)	0.5048	0.1148	0.0132	100	51
			IND	0.5055	0.1153	0.0133	99	50
			JQL	0.4887	0.1784	0.0319	41	21
			GQL(LC)	0.5000	0.1190	0.0141	94	48
		Mean	GQL(AC)	0.5054	0.0818	0.0067	-	100
	0.7	Median	GQL(TC)	0.7047	0.1146	0.0132	100	51
			IND	0.7057	0.1155	0.0133	99	50
			JQL	0.6890	0.1716	0.0295	45	23
			GQL(LC)	0.7053	0.1242	0.0154	86	44
		Mean	GQL(AC)	0.7054	0.0818	0.0067	-	100
	1.0	Median	GQL(TC)	1.0047	0.1147	0.0132	100	51
			IND	1.0056	0.1153	0.0133	99	50
			JQL	0.9848	0.2125	0.0454	29	15
			GQL(LC)	1.0015	0.1248	0.0155	85	43
		Mean	GQL(AC)	1.0054	0.0818	0.0067	-	100

The results of Tables 4.2, 4.3 and 4.4 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 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 $E_2(M) = \{\text{SMSE(Mean Based)}\}/\{\text{SMSE(M)}\} \times 100$. These results therefore do not appear to support the classical result (Bassett and Koenker (1978)) that a median based estimate may be preferable to the mean based estimate when data are asymmetric. This perhaps has happened because of the degree of asymmetriness in the present exponential data which is not so strong as indicated earlier based on Table 4.1.

However, to further explore the above contradiction, we have also generated asymmetric exponential data, 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 Tables 4.5, 4.6 and 4.7 for $\rho = 0$, 0.5, and 0.7, respectively. These results show that mean regression based GQL estimates are now biased when compared to the corresponding estimates obtained in the outliers free case as in Tables 4.2, 4.3 and 4.4, whereas

the median regression based new estimates do not appear to be affected by outliers. This prompted us to compare the relative bias as opposed to MSE, for the mean and median regression based estimates. It is clear from the last columns of Tables 4.5, 4.6 and 4.7 that mean regression based GQL estimates have much larger relative bias, for example in Table 4.7, 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.

Table 4.5: Comparison of mean regression and median regression approaches for the estimation of regression parameter ($\beta = 0.5, 0.7, 1.0$) involved in an EAR(1) model with a correlation value $\rho = 0.0$, in the presence of 1% outliers through shifted covariate values; based on 500 simulations.

				Statistic			
ρ	β	Regression	Estimation Approach	SM	SSE	SMSE	RB
0.0	0.5	Median	GQL(TC)	0.5094	0.0761	0.0058	
			IND	0.5094	0.0758	0.0057	
			JQL	0.5062	0.0856	0.0073	7.3272
			GQL(LC)	0.5090	0.0769	0.0060	
		Mean	GQL(AC)	0.5147	0.0539		27.4021
	0.7	Median	GQL(TC)	0.7126	0.0758	0.0059	
			IND	0.7120	0.0756	0.0058	
			JQL	0.7089	0.0862	0.0075	10.3960
			GQL(LC)	0.7123	0.0769	0.0060	
		Mean	GQL(AC)	0.7258	0.0547		47.2810
	1.0	Median	GQL(TC)	1.0162	0.0764	0.0061	
			IND	1.0160	0.0764	0.0061	
			JQL	1.0126	0.0864	0.0076	14.6391
			GQL(LC)	1.0159	0.0776	0.0062	
		Mean	GQL(AC)	1.0481	0.0570		84.3990

Table 4.6: Comparison of mean regression and median regression approaches for the estimation of regression parameter ($\beta = 0.5, 0.7, 1.0$) involved in an EAR(1) model with a correlation value $\rho = 0.5$, in the presence of 1% outliers through shifted covariate values; based on 500 simulations.

				Statistic			
ρ	β	Regression	Estimation Approach	SM	SSE	SMSE	RB
0.5	0.5	Median	GQL(TC)	0.5124	0.1023	0.0106	
			IND	0.5130	0.1022	0.0106	
			JQL	0.5127	0.1137	0.0131	11.2250
			GQL(LC)	0.5127	0.1034	0.0108	
		Mean	GQL(AC)	0.5194	0.0736		26.4000
	0.7	Median	GQL(TC)	0.7152	0.1026	0.0107	
			IND	0.7161	0.1024	0.0107	
			JQL	0.7147	0.1153	0.0135	12.7520
			GQL(LC)	0.7155	0.1036	0.0109	
		Mean	GQL(AC)	0.7302	0.0750		40.2970
	1.0	Median	GQL(TC)	1.0186	0.1039	0.0111	
			IND	1.0195	0.1028	0.0109	
			JQL	1.0189	0.1152	0.0136	16.3980
			GQL(LC)	1.0192	0.1037	0.0111	
		Mean	GQL(AC)	1.0515	0.0791		65.1340

Table 4.7: Comparison of mean regression and median regression approaches for the estimation of regression parameter ($\beta = 0.5, 0.7, 1.0$) involved in an EAR(1) model with a 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)	0.5174	0.1156	0.0136	
			IND	0.5186	0.1169	0.0140	
			JQL	0.5083	0.1428	0.0204	5.8304
			GQL(LC)	0.5104	0.1259	0.0159	
		Mean	GQL(AC)	0.5235	0.0818		28.7700
	0.7	Median	GQL(TC)	0.7201	0.1162	0.0139	
			IND	0.7211	0.1172	0.0141	
			JQL	0.7007	0.1518	0.0230	0.5151
			GQL(LC)	0.7138	0.1304	0.0171	
		Mean	GQL(AC)	0.7342	0.0802		41.4510
	1.0	Median	GQL(TC)	1.0319	0.1165	0.0146	
			IND	1.0307	0.1171	0.0146	
			JQL	1.0158	0.1529	0.0234	3.8451
			GQL(LC)	1.0316	0.1186	0.0151	
		Mean	GQL(AC)	1.0565	0.0845		66.8533

We now turn back to the results of Tables 4.2, 4.3 and 4.4 and compare the relative performance among the median regression based estimation approaches. 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.

Chapter 5

Labor Pain Data Analysis : An Illustration of the estimation methods

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 5.1 and 5.2, respectively.

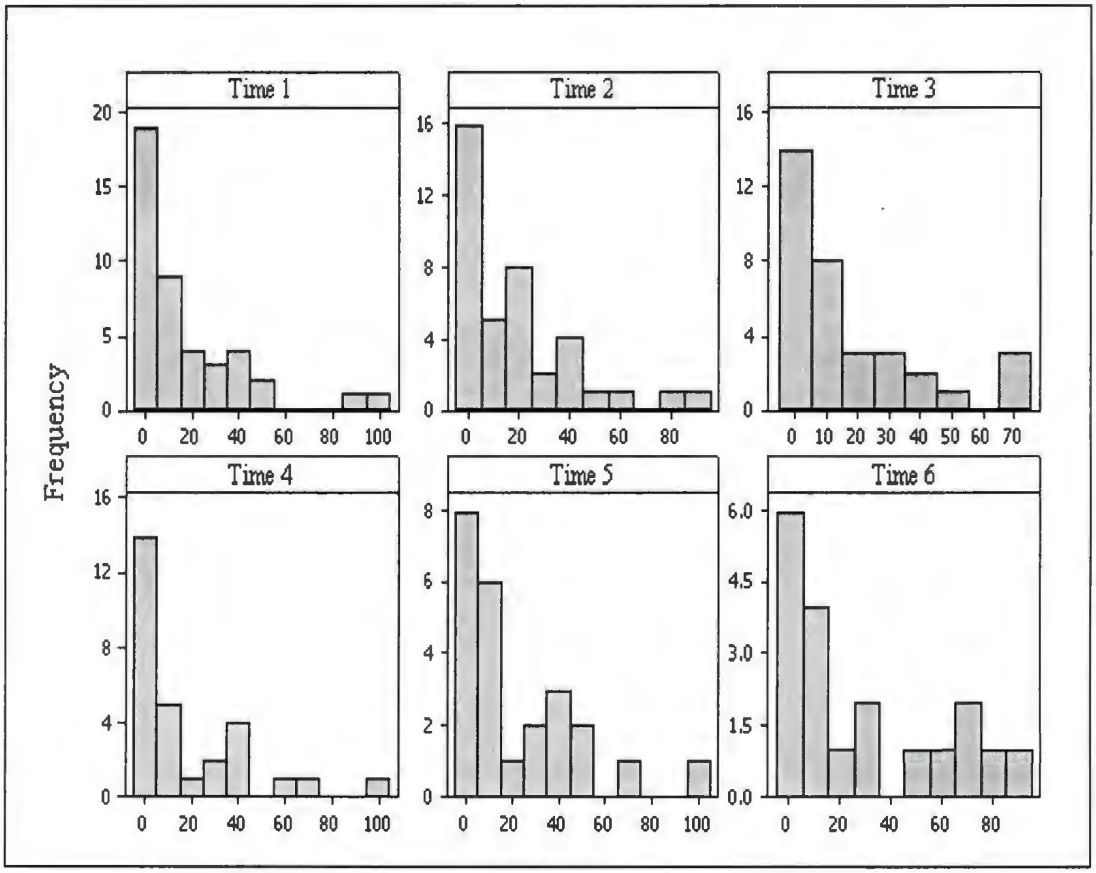

Figure 5.1: Labor pain observed data for treatment group

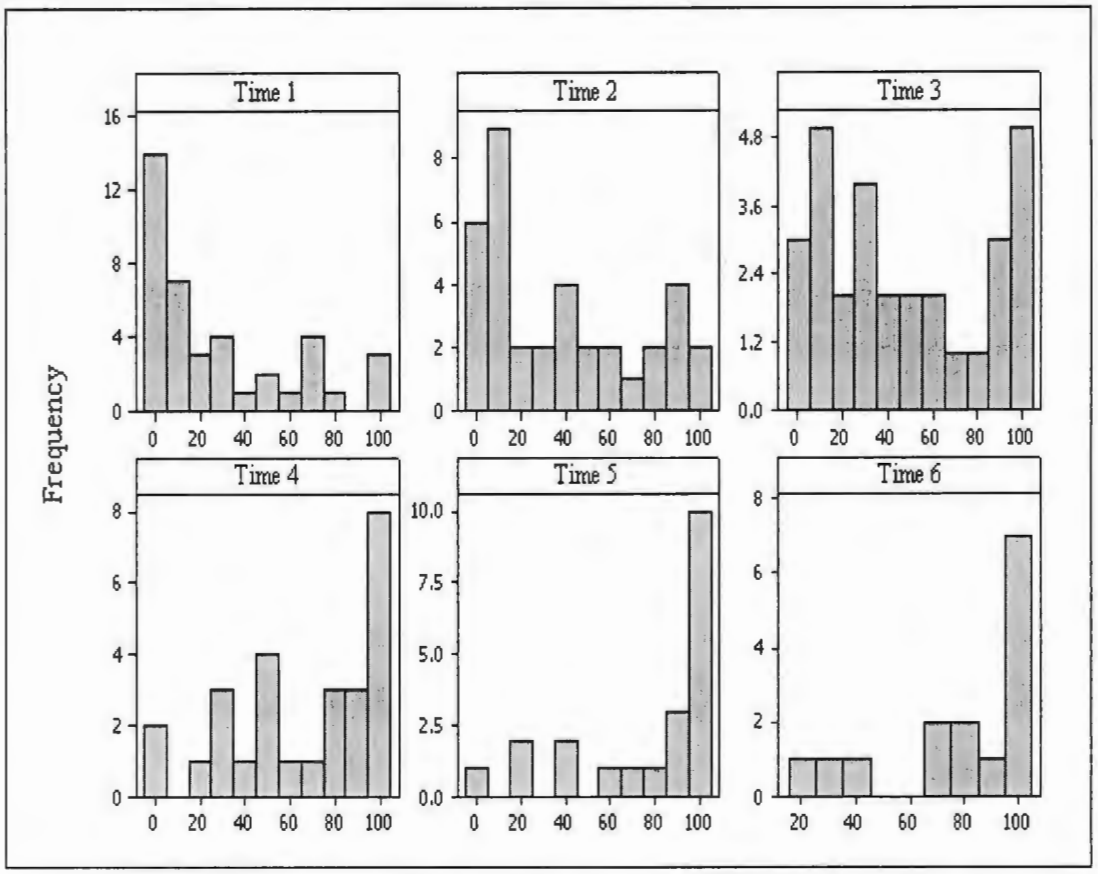

Figure 5.2: Labor pain observed data for placebo group

It appears from Figure 5.1 that under the treatment group, the labor pain at any given time (t = 1,...,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.

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 5.1, 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 5.3.

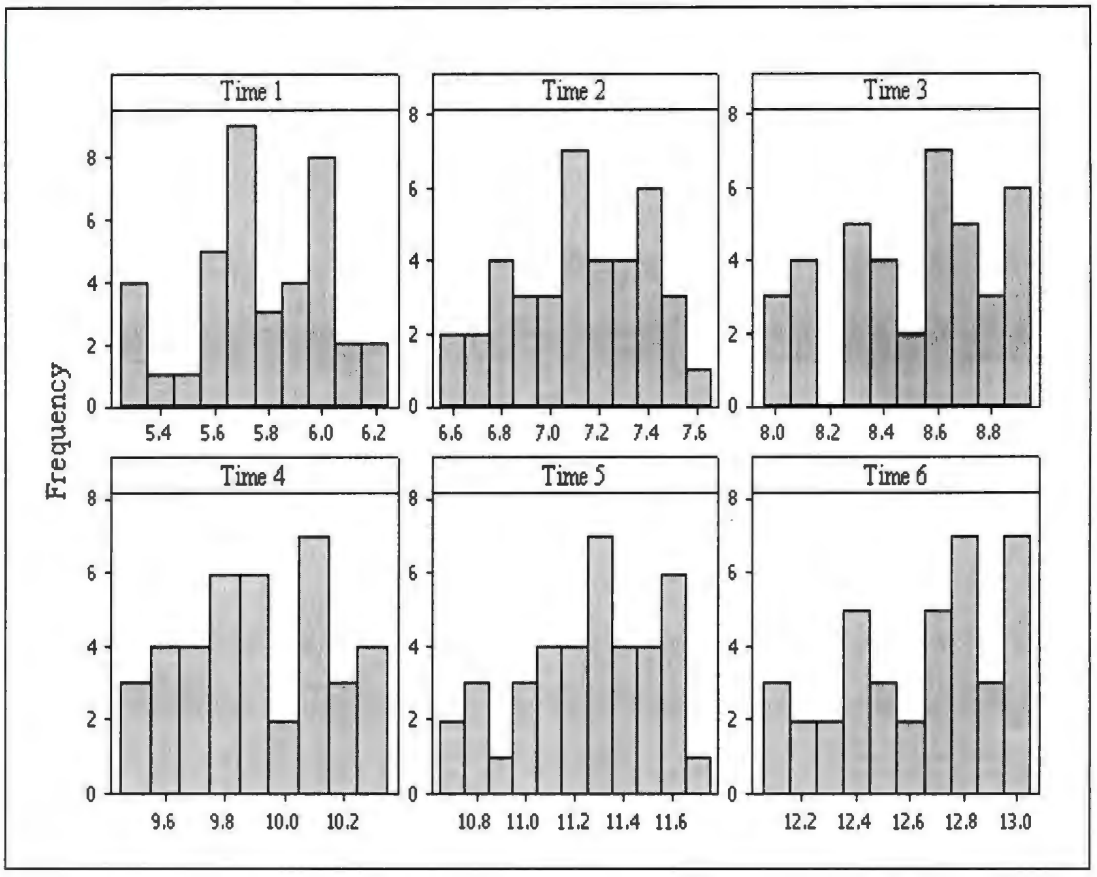

Figure 5.3: Linear median regression based fitted labor pain data for the treatment group with $U(-\frac{1}{2}, \frac{1}{2})$ error

It is however clear that the histogram in Figure 5.3 do not exhibit the exponential form exhibited by Figure 5.1. Thus, even though JQL approach fits the median well, the overall distribution fitting appears to be unsatisfactory. 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 Chapter 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 Chapter 4, it was found that IND followed by GQL(LC) produce more efficient regression estimates. As shown in Tables 4.2, 4.3, and 4.4, 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, $y_{it} = \rho_i y_{i,t-1} + I_{it} a_{it}$ following (2.4) with median m_{it}=(log2) $\exp(\beta_0 + \beta_1 t)$. 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 5.4, the fitted medians by these three approaches appear to agree well with the medians of the observed data (OBS).

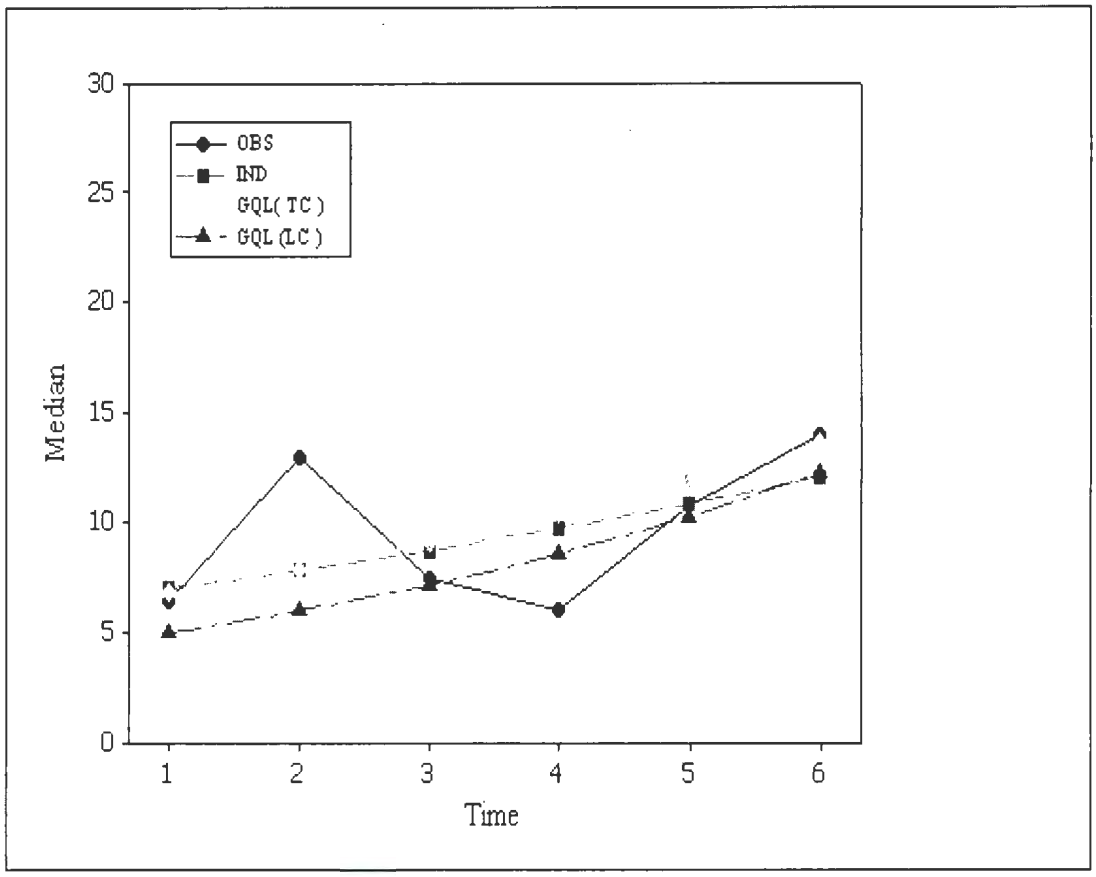

Figure 5.4: Observed versus various model based fitted medians for the treatment group

These approaches also appear to fit the over all data well. For example, using above mentioned $\hat{\beta}_0$, $\hat{\beta}_1$ in m_{it}=(log2) 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 5.5 and 5.6, respectively.

These distributions appear to agree well with seemingly exponential distribution for the observed data displayed in Figure 5.1.

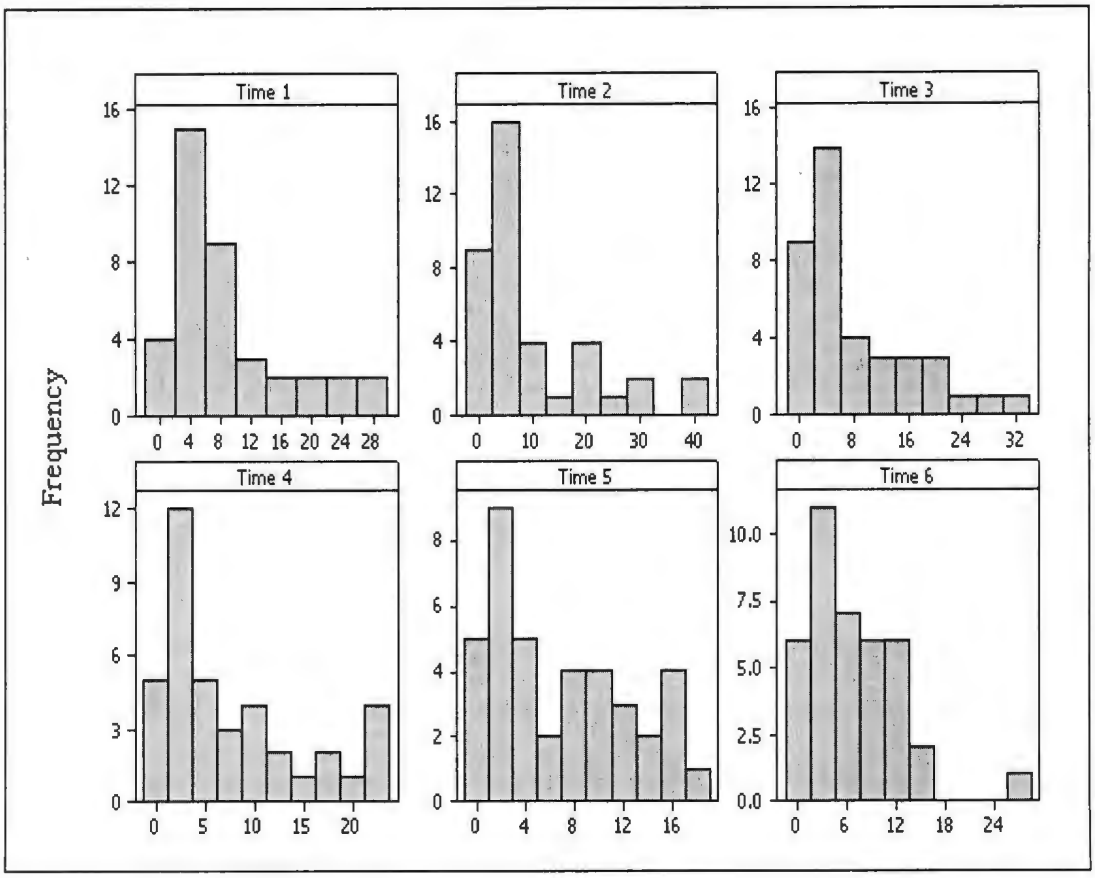

Figure 5.5: Exponential median regression based fitted labor pain data for the treatment group under GQL(TC)

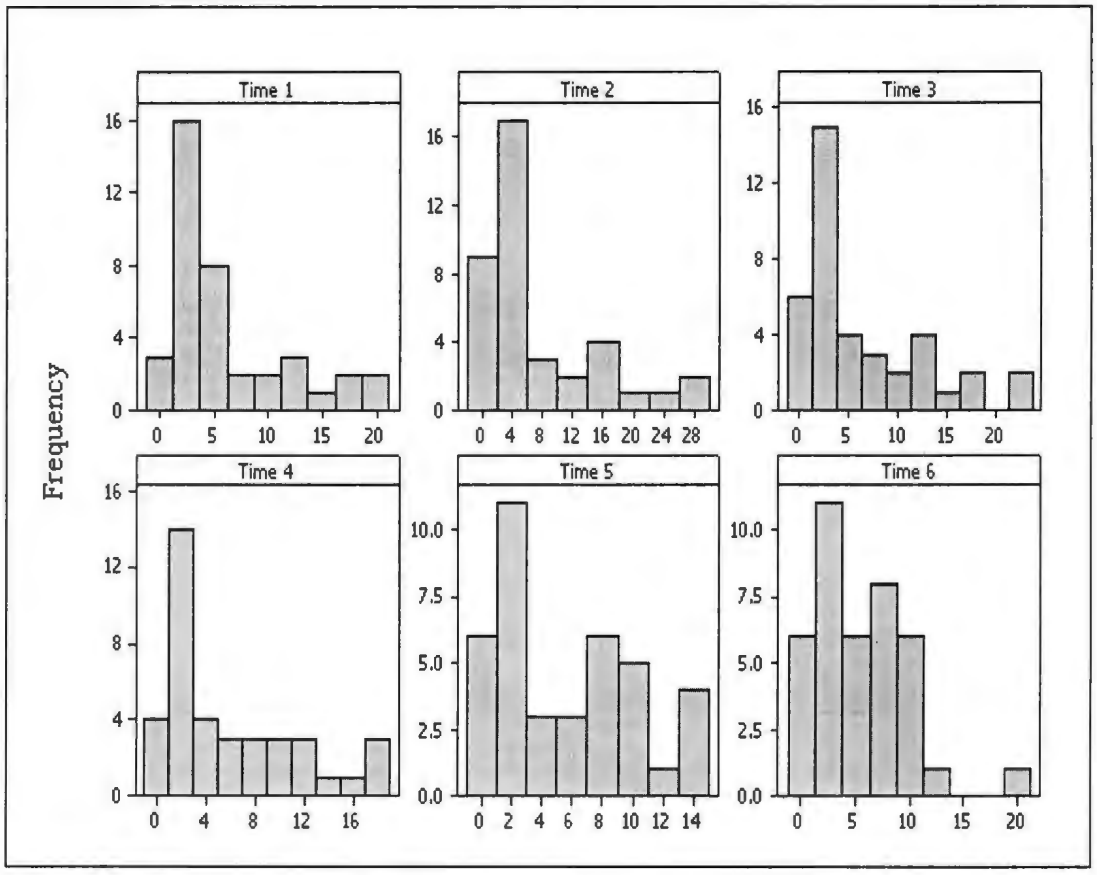

Figure 5.6: Exponential median regression based fitted labor pain data for the treatment group under GQL(LC)

We also have estimated the parameters of the exponential model (2.4) using the mean regression based GQL(AC) 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 correlation 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 Tables 4.2, 4.3, and 4.4. However, in view of the simulation results reported in Tables 4.5, 4.6, and 4.7 and because the observed data are highly asymmetric, the median based estimates are preferable to the mean based estimates.

Chapter 6

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 thesis, by using an AR(1) type correlation model for repeated exponential data, we have examined the performance of the simpler mean based GQL approach with several median based GQL approaches. The results of a simulation study indicates that the mean regression based GQL approach performs better than all median regression based QL approaches. This contradicts the classical result (Bassett and Koenker (1978)) that a median regression approach should be preferable to the mean regression based approach when the data are asymmetric. However, when the repeated exponential data

were contaminated through 1% outliers, the performances were reversed, that is, median regression based QL approaches were found to be better than the mean regression based QL approach. Thus, the study indicates that if the data are highly asymmetric then one should use median regression based approaches. Furthermore, when median regression based approach 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.

Appendix A

A.1 Derivation for Joint Probability in (3.10)

Following Hasan (2004, Section 4.1.1, p. 59), we first consider lag 1 apart repeated responses y_{iv} and $y_{i(v+1)}$, where y_{iv} and $y_{i(v+1)}$ must satisfy the relationship $y_{i(v+1)} \geq \rho_{v+1} y_{iv}$. Note that $m_{i(v+1)}$ (a realized value of the random variable $Y_{i(v+1)}$)) can be either $m_{i(v+1)} > \rho_{v+1} y_{iv}$ or $m_{i(v+1)} \leq \rho_{v+1} y_{iv}$. Therefore the bivariate probability function of y_{iv} and $y_{i(v+1)}$ may be computed as follows:

$$Pr(y_{iv} \geq m_{iv}, y_{i(v+1)} \geq m_{i(v+1)}) = \int_{m_{iv}}^{\infty} \int_{m_{i(v+1)}}^{\infty} f(y_{iv}, y_{i(v+1)})\, dy_{iv} dy_{i(v+1)}. \quad \text{(A.1)}$$

where $f(y_{iv}, y_{i(v+1)})$ is the bivariate probability density function of lag 1 apart non-stationary repeated responses y_{iv} and $y_{i(v+1)}$, which has the form

$$\begin{aligned} f(y_{iv}, y_{i(v+1)}) &= \lambda_{iv}\rho e^{-\lambda_{iv}y_{iv}}\delta(y_{i(v+1)} - \rho_{v+1}y_{iv}) \\ &\quad +\lambda_{iv}\lambda_{i(v+1)}(1-\rho)e^{-\lambda_{iv}y_{iv}}e^{-\lambda_{i(v+1)}(y_{i(v+1)}-\rho_{v+1}y_{iv})} \end{aligned} \tag{A.2}$$

where $\delta(x)$ is the discrete Dirac delta function, i.e. $\delta(x)$ is the distribution with atom of probability 1 at $x = 0$.

The lower limit of the integration with respect to $y_{i(v+1)}$ should be the maximum of $m_{i(v+1)}$ and $\rho_{v+1}y_{iv}$. It then follows that

$$\begin{aligned} Pr(y_{iv} \geq m_{iv}, y_{i(v+1)} \geq m_{i(v+1)}) &= \int_{m_{iv}}^{\infty}\int_{\max(m_{i(v+1)},\ \rho_{v+1}y_{iv})}^{\infty} f(y_{iv}, y_{i(v+1)})\, dy_{iv}dy_{i(v+1)} \\ &= \int_{m_{iv}}^{\infty} H(y_{iv})dy_{iv} \quad \text{(say)}. \end{aligned} \tag{A.3}$$

Note that there are two cases to consider to evaluate the integral $H(y_{iv})$ in (A.3): (a) $m_{i(v+1)} > \rho_{v+1}y_{iv}$ and (b) $m_{i(v+1)} \leq \rho_{v+1}y_{iv}$. For case (a) $m_{i(v+1)} > \rho_{v+1}y_{iv}$, $H(y_{iv})$ is

computed as

$$\begin{aligned}
H(y_{iv}) &= \int_{m_{i(v+1)}}^{\infty} f(y_{iv}, y_{i(v+1)})\, dy_{i(v+1)} \\
&= \int_{m_{i(v+1)}}^{\infty} \big[\lambda_{iv}\rho e^{-\lambda_{iv}y_{iv}}\delta(y_{i(v+1)} - \rho_{v+1}y_{iv}) \\
&\quad + \lambda_{iv}\lambda_{i(v+1)}(1-\rho)e^{-\lambda_{iv}y_{iv}}e^{-\lambda_{i(v+1)}(y_{i(v+1)}-\rho_{v+1}y_{iv})}\big]\, dy_{i(v+1)} \\
&= \int_{m_{i(v+1)}}^{\infty} \lambda_{iv}\lambda_{i(v+1)}(1-\rho)e^{-\lambda_{iv}y_{iv}}e^{-\lambda_{i(v+1)}(y_{i(v+1)}-\rho_{v+1}y_{iv})}dy_{i(v+1)} \\
&\qquad \big[\text{as } \delta(y_{i(v+1)} - \rho_{v+1}y_{iv}) = 0 \text{ for } m_{i(v+1)} > \rho_{v+1}y_{iv}\big] \\
&= \lambda_{iv}(1-\rho)e^{-\lambda_{iv}(1-\rho)y_{iv}}e^{-\lambda_{i(v+1)}m_{i(v+1)}}, \qquad \text{(A.4)}
\end{aligned}$$

whereas for case (b) $m_{i(v+1)} \leq \rho_{v+1}y_{iv}$, $H(y_{iv})$ is computed as

$$\begin{aligned}
H(y_{iv}) &= \int_{\rho_{v+1}y_{iv}}^{\infty} f(y_{iv}, y_{i(v+1)})\, dy_{i(v+1)} \\
&= \int_{\rho_{v+1}y_{iv}}^{\infty} \big[\lambda_{iv}\rho e^{-\lambda_{iv}y_{iv}}\delta(y_{i(v+1)} - \rho_{v+1}y_{iv}) \\
&\quad + \lambda_{iv}\lambda_{i(v+1)}(1-\rho)e^{-\lambda_{iv}y_{iv}}e^{-\lambda_{i(v+1)}(y_{i(v+1)}-\rho_{v+1}y_{iv})}\big]\, dy_{i(v+1)} \\
&= \int_{\rho_{v+1}y_{iv}}^{\infty} \lambda_{iv}\rho e^{-\lambda_{iv}y_{iv}}\delta(y_{i(v+1)} - \rho_{v+1}y_{iv})dy_{i(v+1)} \\
&\quad + \int_{\rho_{v+1}y_{iv}}^{\infty} \lambda_{iv}\lambda_{i(v+1)}(1-\rho)e^{-\lambda_{iv}y_{iv}}e^{-\lambda_{i(v+1)}(y_{i(v+1)}-\rho_{v+1}y_{iv})}dy_{i(v+1)} \\
&= \lambda_{iv}\rho e^{-\lambda_{iv}y_{iv}} + \int_{\rho_{v+1}y_{iv}}^{\infty} \lambda_{iv}\lambda_{i(v+1)}(1-\rho)e^{-\lambda_{iv}y_{iv}}e^{-\lambda_{i(v+1)}(y_{i(v+1)}-\rho_{v+1}y_{iv})}dy_{i(v+1)} \\
&\qquad \big[\text{as } \delta(y_{i(v+1)} - \rho_{v+1}y_{iv}) = 1 \text{ for } m_{i(v+1)} \leq \rho_{v+1}y_{iv}\big] \\
&= \lambda_{iv}\rho e^{-\lambda_{iv}y_{iv}} + \lambda_{iv}e^{-\lambda_{iv}y_{iv}} - \lambda_{iv}\rho e^{-\lambda_{iv}y_{iv}} \\
&= \lambda_{iv}e^{-\lambda_{iv}y_{iv}}, \qquad \text{(A.5)}
\end{aligned}$$

yielding

$$H(y_{iv}) = \begin{cases} \lambda_{iv}(1-\rho)e^{-\lambda_{iv}(1-\rho)y_{iv}}e^{-\lambda_{i(v+1)}m_{i(v+1)}} & \text{for} \quad m_{i(v+1)} > \rho_{v+1}y_{iv} \\ \lambda_{iv}e^{-\lambda_{iv}y_{iv}} & \text{for} \quad m_{i(v+1)} \leq \rho_{v+1}y_{iv} \end{cases} \tag{A.6}$$

Next, by using the above formula for $H(y_{iv})$ in (A.6) we evaluate the remaining integral in (A.3) over m_{iv} as follows. For case (a) $m_{i(v+1)} > \rho_{v+1}m_{iv}$, the integral in (A.3) is evaluated as

$$\begin{aligned} Pr(y_{iv} \geq m_{iv}, y_{i(v+1)} \geq m_{i(v+1)}) &= \int_{m_{iv}}^{\infty} \lambda_{iv}(1-\rho)e^{-\lambda_{iv}(1-\rho)y_{iv}}e^{-\lambda_{i(v+1)}m_{i(v+1)}}dy_{iv} \\ &= e^{-\lambda_{i(v+1)}m_{i(v+1)}}e^{-\lambda_{iv}(1-\rho)m_{iv}} \end{aligned} \tag{A.7}$$

whereas for case (b) $m_{i(v+1)} \leq \rho_{v+1}m_{iv}$, we evaluate the integral in (A.3) as

$$\begin{aligned} Pr(y_{iv} \geq m_{iv}, y_{i(v+1)} \geq m_{i(v+1)}) &= \int_{m_{iv}}^{\infty} \lambda_{iv}e^{-\lambda_{iv}y_{iv}}dy_{iv} \\ &= e^{-\lambda_{iv}m_{iv}} \end{aligned} \tag{A.8}$$

It then follows that the bivariate probability function of y_{iv} and $y_{i(v+1)}$ has the form given by

$$Pr(y_{iv} \geq m_{iv}, y_{i(v+1)} \geq m_{i(v+1)}) = \begin{cases} e^{-\lambda_{iv}m_{iv}} & \text{for} \quad m_{i(v+1)} \leq \rho_{v+1}m_{iv} \\ e^{-\lambda_{i(v+1)}m_{i(v+1)}}e^{-\lambda_{iv}(1-\rho)m_{iv}} & \text{for} \quad m_{i(v+1)} > \rho_{v+1}m_{iv} \end{cases} \tag{A.9}$$

Similarly, after some algebra it can be shown that the bivariate probability function of lag 2 apart repeated responses y_{iv} and $y_{i(v+2)}$ is

$$Pr(y_{iv} \geq m_{iv}, y_{i(v+2)} \geq m_{i(v+2)}) = \begin{cases} e^{-\lambda_{iv} m_{iv}} & \text{for} \quad m_{i(v+2)} \leq \rho_{v+2}\rho_{v+1} m_{iv} \\ e^{-\lambda_{i(v+2)} m_{i(v+2)}} e^{-\lambda_{iv}(1-\rho^2) m_{iv}} & \text{for} \quad m_{i(v+2)} > \rho_{v+2}\rho_{v+1} m_{iv} \end{cases} \tag{A.10}$$

Following (A.9) and (A.10), it can be shown that the bivariate probability function of lag j apart repeated responses y_{iv} and $y_{i(v+j)}$ is

$$Pr(y_{iv} \geq m_{iv}, y_{i(v+j)} \geq m_{i(v+j)}) = \begin{cases} e^{-\lambda_{iv} m_{iv}} & \text{for} \quad m_{i(v+j)} \leq \rho_{v+j}\rho_{v+j-1}\cdots\rho_{v+1} m_{iv} \\ e^{-\lambda_{i(v+j)} m_{i(v+j)}} e^{-\lambda_{iv}(1-\rho^j) m_{iv}} & \text{for} \quad m_{i(v+j)} > \rho_{v+j}\rho_{v+j-1}\cdots\rho_{v+1} m_{iv} \end{cases} \tag{A.11}$$

Bibliography

[1] Banneheka, B.M.S.G., and Ekanayake, G.E.M.U.P.D. (2009), "A new point estimator for the median of gamma distribution," *Journal of Science*, **14**, 095-103.

[2] Bassett, G. W., and Koenker, R. (1978), "Asymptotic theory of least absolute error regression," *Journal of the American Statistical Association*, **73**, 618-622.

[3] Buchinsky, M. (1995), "Estimating the asymptotic covariance matrix for quantile regression models: A Monte Carlo study", *Journal of Econometrics*, **68** , 303-338.

[4] 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.

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

[6] Gaver, D.P., and Lewis, P.A.W. (1980), "First order auto regressive gamma sequences and point processes," *Adv. Appl. Prob.*, **12**, 727-745.

[7] Geraci, M., and Bottai, M. (2007), "Quantile regression for longitudinal data using the asymmetric laplace distribution", *Biostatistics*, **8**, 1 , 140-154.

[8] Hasan, M.T., Sutradhar, B.C. and Sneddon, G. (2007), "On correlation models for longitudinal failure time data", *Sankhya*, **69**, 548-580.

[9] Hasan, M.T., (2004), Analysis of longitudinal failure time data, Ph.D. thesis, Memorial University of Newfoundland, Canada.

[10] He, X., Fu, B., and Fung, W.K. (2003), "Median regression for longitudinal data", *Statistics in Medicine*, **22(23)**, 3655-3669.

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

[12] Karlsson, A. (2008), "Nonlinear quantile regression estimation of longitudinal data", *Communications in Statistics- Simulation and computation*, **37**, 114-131.

[13] Knight, K. (2001), "Comparing conditional quantile estimators: first and second order considerations, ".

[14] Koenker, R. (1984), "A note on L-estimators for linear models", *Statist. Probab. Lett.*, **2** , 323-325.

[15] Koenker, R. (2004), "Quantile regression for longitudinal data", *Journal of Multivariate Analysis*, **91** . 74-89.

[16] Koenker, R. (2005), "Quantile Regression", *Cambridge University Press.*

[17] Koenker, R., and Bassett, G. W. (1978), "Regression Quantiles" *Econometrica*, **46**, 33-50.

[18] Koenker, R., and Park, B. J. (1996), "An interior point algorithm for non-linear quantile regression" *Journal of Econometrics*, **71**, 265-283.

[19] Kottas, A., and Gelfand, A. E. (2001), "Bayesian semiparametric median regression modeling," *Journal of American Statistical Association*, **96**, 456, 1458-1468.

[20] Kottas, A., and Krnjajic, M. (2009), "Bayesian semiparametric modelling in quantile regression," *Scandinavian Journal of Statistics*, **36**, 2, 297-319.

[21] Liu, Y., and Bottai, M. (2009), "Mixed-effects models for conditional quantiles with longitudinal data," *The International Journal of Biostatistics*, **5** : Iss. 1, Article 28.

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

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

[24] Reich, B.J., Bondell, H.D., and Wang, H. (2010), "Flexible bayesian quantile regression for independent and clustered data", *Biostatistics*, **11**, 2 , 337-352.

[25] Sutradhar, B.C. (2003), "An overview on regression models for discrete longitudinal responses", *Statist. Sci.*, **18**, 377-393.

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

[27] Sutradhar, B.C. (2011), "Dynamic Mixed Models for Familial Longitudinal Data", *Springer Series in Statistics.*

[28] Ying, Z., Jung, S.H., and Wei, L.J. (1995), "Survival analysis with median regression models", *Journal of American Statistical Association*, **90**, 429 , 178-184.