International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:9, No:12, 2015

A Survey on Quasi-Likelihood Estimation Approaches for Longitudinal Set-ups

Naushad Mamode Khan

***Abstract*—**The Com-Poisson (CMP) model is one of the most popular discrete generalized linear models (GLMS) that handles both equi-, over- and under-dispersed data. In longitudinal context, an integer-valued autoregressive (INAR(1)) process that incorporates covariate specification has been developed to model longitudinal CMP counts. However, the joint likelihood CMP function is difficult to specify and thus restricts the likelihood-based estimating methodology. The joint generalized quasi-likelihood approach (GQL-I) was instead considered but is rather computationally intensive and may not even estimate the regression effects due to a complex and frequently ill-conditioned covariance structure. This paper proposes a new GQL approach for estimating the regression parameters (GQL-III) that is based on a single score vector representation. The performance of GQL-III is compared with GQL-I and separate marginal GQLs (GQL-II) through some simulation experiments and is proved to yield equally efficient estimates as GQL-I and is far more computationally stable.

***Keywords*—**Longitudinal, Com-Poisson, Ill-conditioned, INAR(1), GLMS, GQL.

I. Introduction

IN several real-life medical studies, count data are collected repeatedly over time for different subjects or individuals. Some common examples include the monitoring of CD-4 counts in HIV patients and the number of seizures among epilepsy patients. In such data set-ups, the subjects are treated as independent while the repeated observations for a particular individual are most likely time correlated and are influenced by some explanatory variables like age, gender, treatment, baseline measurements and among others. One more feature is these observations may be equi-, over- or under-dispersed relative to their means. Thus, CMP is most appropriate for modelling such data. Some interesting facts about CMP is that it is a member of the family of GLMs and satisfies the moment properties of the exponential dispersion family. In this respect, Iterative Reweighted Least Squares (IRWLS) or quasi-likelihood estimating equations [6], [13] can be used to estimate the regression effects.

Until now, CMP has been largely explored in cross-sectional studies [9], [3], [4]; but, only few literature is available on longitudinal set-ups. Mamode Khan and Jowaheer [5] formulated an INAR(1) process([1], [7], [2]) to model longitudinal CMP counts based on time-independent covariates. In this process, the marginal moments and joint covariances were derived. These authors considered the GQL approach [10] to estimate the regression and dispersion effects while the serial correlations were obtained through a robust moment estimating equation approach [11].

N. Mamode Khan is a Senior Lecturer in Statistics in the Department of Economics and Statistics, University of Mauritius, Mauritius, (e-mail: n.mamodekhan@uom.ac.mu).

Since CMP is a two-parameter model, the classical GQL was modified to accommodate a score vector constituting of paired responses. However, the challenge in this joint GQL (GQL-I) approach was the modelling of the joint longitudinal covariance structure. This structure required the computation of high-order moments which were derived under the multivariate normality assumption [8]. Under these conditions, the regression and dispersion estimators were shown to be asymptotically normal and consistent. The Newton-Raphson iterative procedure was used to estimate the regression and dispersion parameters. However, some computational difficulties were noted here. Firstly, in simulation and real-life studies, the longitudinal covariance matrix and the Hessian component in the iterative process were often ill-conditioned and this leads to unreliable estimates or an entire blockage of the estimation system. The pseudo-inverse was considered to be an alternative to the inverse problem but still the problem of non-convergence persists.

In literature, Prentice and Zhao [8] and Sutradhar and Farrell [12] have considered the option of setting two separate GQLs (GQL-II) to estimate different sets of parameters which has yielded consistent and slightly less efficient estimates as GQL-I. However, GQL-II yielded very few convergence problems and very few cases of singular covariance or Hessian matrix were reported. In this paper, we propose a GQL approach (GQL-III) which estimates jointly the regression and dispersion parameters using a single estimating equation and a score vector of single observations rather than paired observations. This is possible since the mean and variance functions of the CMP model constitutes of both the regression and dispersion effects in their formulations. Such a technique has not yet been explored and we are proposing here to compare the performance of these three GQL approaches. Thus the organization of this paper is as follows: In the next section, the longitudinal CMP model, the GQL-I and GQL-II approaches are presented. In the same section, the components of the derivative and longitudinal covariance matrices are shown. In The GQL-III approach is also introduced here followed by a section on simulation experiment and results. The conclusion is presented in the last section.

II. Methodology

Let y_{it} be a count response and x_{it} be a p-dimensional vector of covariates for subject $i(i = 1, \ldots, I)$ observed at time t $(t = 1, \ldots T)$. Assume β is the $p \times 1$ regression vector. For the ith subject, let $y_i = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT})^T$ be the $T \times 1$ response vector and $X_i = (x_{i1}, \ldots, x_{iT})^T$ be the

$T \times p$ matrix of time-independent covariates. The Com-Poisson density function of $y_{it} \sim CMP(\frac{\theta_{it}}{\nu}, \nu)$ is written as

$$f(y_{it}) = \frac{\lambda_{it}^{y_{it}}}{(y_{it}!)^{\nu}} \frac{1}{Z(\lambda_{it}, \nu)}, \tag{1}$$

where

$$Z(\lambda_{it}, \nu) = \sum_{j=0}^{\infty} \frac{\lambda_{it}^{j}}{(j!)^{\nu}} \tag{2}$$

and

$$\lambda_{it} = \exp(x_{it}^T \beta) \tag{3}$$

where the parameter ν corresponds the dispersion index. Values of $\nu = 1$, $\nu < 1$and $\nu > 1$ correspond to equi-, over- and under- dispersion respectively. To facilitate computation, [9] considered an approximation for $Z(\lambda_{it}, \nu)$ which yielded

$$E(Y_{it}) = \theta_{it} = \lambda_{it}^{1/\nu} - \frac{\nu - 1}{2\nu} \tag{4}$$

and

$$Var(Y_{it}) = \frac{\theta_{it}}{\nu} + \frac{\nu - 1}{2\nu^2} \tag{5}$$

Note that under the time-independent covariates, that is stationarity, the correlations within same lags are same for all individuals such that the general autocorrelation structure can be expressed as

$$C_i(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{T-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \dots & 1 \end{bmatrix} \tag{6}$$

where further to findings of [11]

$$\hat{\rho}_l = \frac{\sum_{i=1}^{I} \sum_{t=1}^{T-l} \tilde{y_{it}} y_{i(\tilde{t+l})} / (T-l)}{\sum_{i=1}^{I} \sum_{t=1}^{T} \tilde{y_{it}}^2 / T} \tag{7}$$

where $\tilde{y_{it}} = \frac{y_{it} - \theta_{i1}}{\sqrt{Var(Y_{it}}}$.

A. A Review of the GQL-I and GQL-II and Introducing GQL-III

This section provides an overview of the GQL-I approach developed by [5] for the CMP longitudinal model.

$$\sum_{i=1}^{I} D_i^T \widetilde{\Sigma_i}^{-1} (f_i - \mu_i) = 0, \tag{8}$$

where $f_i = (f_{i1}^T, \dots, f_{it}^T, \dots, f_{iT}^T)^T$, $\mu_i = (\mu_{i1}^T, \dots, \mu_{it}^T, \dots, \mu_{iT}^T)^T$ are $2T \times 1$ vectors with f_{it} a pair of the response score(y_{it}, y_{it}^2), $\mu_{it} = (\theta_{i1}, m_{i1})^T$ for $t = 1, \dots, T$. $\theta_{i1} = E(Y_{it})$ and $m_{i1} = E(Y_{it}^2)$ where

$$m_{i1} = \frac{\lambda_{i1}^{\frac{1}{\nu}}}{\nu} + \theta_{i1}^2 \tag{9}$$

with $\lambda_{i1} = \exp(x_{i1}^T \beta)$. $\widetilde{\Sigma_i}$ is the covariance matrix of the score vector f_i and D_i is the $2T \times (p+1)$ derivative matrix consisting of

$$D_i = [\partial \mu_i / \partial \beta^T, \partial \mu_i / \partial \nu] = [D_{i1}^T, \dots, D_{it}^T, \dots, D_{iT}^T]^T,$$

with

$$D_{it} = \begin{pmatrix} \partial \theta_{i1} / \partial \beta^T & \partial \theta_{i1} / \partial \nu \\ \partial m_{i1} / \partial \beta^T & \partial m_{i1} / \partial \nu \end{pmatrix}$$

for $t = 1, \dots, T$ where

$$\partial \theta_{i1} / \partial \beta^T = \frac{\lambda_{i1}^{\frac{1}{\nu}}}{\nu} x_{i1}^T \tag{10}$$

$$\partial \theta_{i1} / \partial \nu = \frac{\nu - 1}{2\nu^2} - \frac{1}{2\nu} - \frac{\lambda_{i1}^{\frac{1}{\nu}} x_{i1}^T \beta}{\nu^2} \tag{11}$$

$$\partial m_{i1} / \partial \beta^T = x_{i1}^T \left(\frac{2\lambda_{i1}^{\frac{1}{\nu}} + 2\nu \lambda_{i1}^{\frac{2}{\nu}} - \nu \lambda_{i1}^{\frac{1}{\nu}}}{\nu^2} \right) \tag{12}$$

$$\tag{13}$$

The longitudinal joint covariance matrix of f_i is expressed as

$$\widetilde{\Sigma_i} = \begin{pmatrix} \widetilde{\Sigma_{i1}} & \widetilde{\Omega_{i12}} & \widetilde{\Omega_{i13}} & \cdots & \widetilde{\Omega_{i1T}} \\ & \widetilde{\Sigma_{i2}} & \widetilde{\Omega_{i23}} & \cdots & \widetilde{\Omega_{i2T}} \\ & & \widetilde{\Sigma_{i3}} & \cdots & \widetilde{\Omega_{i3T}} \\ & & & \ddots & \\ & & & & \widetilde{\Sigma_{iT}} \end{pmatrix} \tag{14}$$

where the diagonal submatrix

$$\widetilde{\Sigma_{it}} = \begin{pmatrix} var(Y_{it}) & cov(Y_{it}, Y_{it}^2) \\ & var(Y_{it}^2) \end{pmatrix}$$

and for $t \neq w$, the off-diagonal submatrix

$$\widetilde{\Omega_{itw}} = \begin{pmatrix} cov(Y_{it}, Y_{iw}) & cov(Y_{it}, Y_{iw}^2) \\ cov(Y_{it}^2, Y_{iw}) & cov(Y_{it}^2, Y_{iw}^2) \end{pmatrix}$$

for $t = 1, \dots, T$ and $w = 1, \dots, T$ [Refer to [5] for more details on these components]. The GQL-I estimating equation is solved by the Newton-Raphson iterative process

$$\begin{bmatrix} \hat{\beta}_{r+1} \\ \hat{\nu}_{r+1} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_r \\ \hat{\nu}_r \end{bmatrix} + \left[\sum_{i=1}^{I} D_i^T \widetilde{\Sigma_i}^{-1} D_i \right]_r^{-1} \left[\sum_{i=1}^{I} D_i^T \widetilde{\Sigma_i}^{-1} (f_i - \mu_i) \right]_r \tag{15}$$

where $\hat{\beta}_r$ is the value of $\hat{\beta}$ at the r^{th} iteration. $[.]_r$ is the value of the expression at the r^{th} iteration. The algorithm works as follows: For an initial value of $\hat{\beta}$ and $\hat{\nu}$, we calculate $\hat{\rho}$ to obtain the correlation structure and then use these two sets of parameters to update the values of $\hat{\beta}$ and $\hat{\nu}$. Then the new set of parameters is used to calculate $\hat{\rho}_l$ and the iteration continues in this way until convergence. However, these authors reported that few simulations converge under the GQL-I approach because the longitudinal covariance matrix $\widetilde{\Sigma_i}$ and the Hessian component are close to singularity. A possible solution to this problem is to adopt the GQL-II approach of [8] ad-hoc estimation and [12] separate marginal estimating equations. The GQL-II is split into

$$\sum_{i=1}^{I} [\partial \theta_{i1} / \partial \beta^T]^T [Cov(Y_i)]^{-1} [y_i - \theta_i] = 0 \tag{16}$$

$$\sum_{i=1}^{I} [\partial m_{i1} / \partial \nu]^T [Cov(Y_i^2)]^{-1} [y_i^2 - m_i] = 0 \tag{17}$$

where $[Cov(Y_i)]$ and $[Cov(Y_i^2)]$ comprise of the covariance matrix of y_i and y_i^2 respectively and are $T \times T$ dimensional matrices.

B. GQL-III

GQL-III is a simpler version of GQL-I where only the score vector y_i is considered instead of y_i and y_i^2. Since the first moment of y_{it} is a function of both β and ν, we may consider only a $(p+1)\times T$ derivative matrix, that is,

$$\sum_{i=1}^{I} \widetilde{D}_i^T [Cov(Y_i)]^{-1}[y_i - \theta_i] = 0 \tag{18}$$

where

$$\widetilde{D}_{it} = \left(\partial\theta_{i1}/\partial\beta^T \quad \partial\theta_{i1}/\partial\nu\right) \tag{19}$$

The Newton-Rapshon iterative procedure is implemented in a similar way as in GQL-I but with the modified derivative and covariance matrix.

III. Simulation Study and Results

A. The Generating Procedure

A simulation experiment is run 10,000 times with sample sizes $I = 60, 100$ and 500 and with $T = 4$ and correlation coefficient $\rho = 0.9$ along with $\beta_1 = \beta_2 = 1$ and $\nu = 0.5$ and 1. The two-dimensional covariate design is

$$x_{it1} = \begin{cases} -1+i & (i = 1,\ldots,I/4), \\ 0 & (i = (I/4)+1,\ldots,3I/4), \\ 1+i & (i = (3I/4)+1,\ldots,I), \end{cases}$$

and x_{it2} is a Poisson distributed with mean parameter 2. This implies $\lambda_{it} = \exp(x_{it1}\beta 1 + x_{it2}\beta_2)$ for $t = 1, 2, \ldots, T = 4$. Moreover, the INAR(1) sequence of the Com-Poisson counts is

$$y_{it} = \rho * y_{i,t-1} + d_{it} \tag{20}$$

where $0 < \rho < 1$. The symbol $*$ indicates the binomial convolution thinning operation such that

$$\rho * y_{i,t-1} = \sum_{j=1}^{y_{i,t-1}} b_j(\rho) = g_{it}. \tag{21}$$

where $prob[b_j(\rho) = 1] = \rho$ and $prob[b_j(\rho) = 0] = 1 - \rho$. therefore

$$(g_{it} \mid y_{i,t-1}, \rho) \sim \text{Binomial}(y_{i,t-1}, \rho) \tag{22}$$

and the error term $d_{it} \sim CMP(\frac{(1-\rho)\theta_{i1}}{\tilde{\nu}}, \tilde{\nu})$ where

$$\tilde{\nu} = \frac{(2q_0+1) + \sqrt{(2q_0+1)^2 - 8q_1}}{4q_1} \tag{23}$$

with $q_0 = (1-\rho)\theta_{i1}$, $q_1 = \frac{q_0}{\nu}[1+\rho(1-\nu)] + \frac{\nu-1}{2\nu^2}(1-\rho^2)$ and $0 < \rho < 1$ [See [5] for further details]. The simulation results are shown in the next section.

B. Numerical Results

TABLE I
Regression and Dispersion Estimates under GQL-I and GQL-II

I	ν	Method		$\hat{\beta}_1$	$\hat{\nu}$	$\hat{\rho}_1$	$\hat{\rho}_2$
60	0.5	GQL-I	0.9992	0.9991	0.4830	0.8982	0.7885
			(0.0192)	(0.0230)	(0.0120)		
		GQL-II	0.9991	0.9991	0.4828	0.9014	0.7992
			(0.0220)	(0.0242)	(0.0134)		
		GQL-III	0.9991	0.9991	0.4832	0.8988	0.8201
			(0.0183)	(0.0225)	(0.0110)		
100		GQL-I	0.9998	0.9999	0.5091	0.9045	0.7956
			(0.0115)	(0.0199)	(0.0115)		
		GQL-II	0.9999	0.9997	0.4989	0.9026	0.8015
			(0.0120)	(0.0206)	(0.0120)		
		GQL-III	0.9997	0.9999	0.5020	0.8988	0.8112
			(0.0115)	(0.0200)	(0.0099)		
500		GQL-I	0.9997	0.9999	0.4990	0.9012	0.7999
			(0.0082)	(0.0074)	(0.0101)		
		GQL-II	1.0002	0.9999	0.4992	0.8972	0.8012
			(0.0083)	(0.0076)	(0.0108)		
		GQL-III	0.9997	0.9999	0.4995	0.8989	0.8121
			(0.0062)	(0.0056)	(0.0085)		
60	1	GQL-I	1.0101	1.0002	0.9982	0.8952	0.8182
			(0.0198)	(0.0252)	(0.0152)		
		GQL-II	1.1002	0.9999	0.9989	0.9010	0.8231
			(0.0201)	(0.0265)	(0.0155)		
		GQL-III	1.0102	1.0001	0.9985	0.8991	0.8042
			(0.0130)	(0.0230)	(0.0130)		
100		GQL-I	1.0001	1.0001	1.0031	0.8925	0.8201
			(0.0138)	(0.0192)	(0.0136)		
		GQL-II	0.9999	0.9999	0.9995	0.9001	0.8182
			(0.0141)	(0.0210)	(0.0142)		
		GQL-III	1.0001	1.0001	0.9995	0.8997	0.8096
			(0.0123)	(0.0160)	(0.0101)		
500		GQL-I	0.9991	0.9999	0.9990	0.9021	0.8315
			(0.0101)	(0.0086)	(0.0128)		
		GQL-II	0.9991	0.9999	1.0012	0.8991	0.8159
			(0.0102)	(0.0090)	(0.0135)		
		GQL-III	1.0001	1.0001	1.0012	0.9101	0.8201
			(0.0098)	(0.0074)	(0.0087)		

The results demonstrate that for the different values of ν, the estimates of β converge to the true values and the correlation estimates under the moment estimating equation are close to the autoregressive structure. As the cluster size increases, the standard errors in the GQL approaches decrease. However, the standard errors in GQL-I are slightly superior than GQL-II but when compared with GQL-III, there are some significant gaps. In fact, in the majority of the simulations, the standard error of GQL-III regression estimates are comparable to GQL-I but GQL-III yields highly efficient estimates for the dispersion parameter. The same trend has been noted across the different clusters and the different ν values. However, we need to add that the number of non-convergent simulations has significantly decreased in the GQL-III approach. In fact, for $I = 60$ and $\nu = 0.5$, GQL-I fails in 3516 simulations while GQL-II and GQL-III flop in only 320 simulations. For $\nu = 1$ and $I = 100$, GQL-I survives in only 1000 simulations while the other two algorithms yield around 4300 simulations. The failures in GQL-I were due mainly to the ill-conditioned longitudinal covariance matrix and in some simulations due to the Hessian matrix in the Newton-Raphson iteration. The

pseudo-inverse was used in GQL-I but the converged estimates deviate from the true values and in some cases NaN or Inf expressions were obtained. Based on the above simulation study, GQL-III yields satisfactory results.

IV. Conclusion

This paper introduces a new GQL-III approach to estimate the regression and dispersion parameters in a longitudinal Com-Poisson set-up. As compared to the existing GQL-I and GQL-II, this approach is more computationally feasible and yields slightly more efficient estimates than GQL-I and GQL-II under both dispersed data set-ups.

References

[1] M.A.Al-Osh and A.A.Alzaid, First-order integer-valued autoregressive (INAR(1))process, *Journal of Time Series Analysis*, 8, 261-275, (1987).
[2] K. Brannas, Explanatory variable in the AR(1) count data model,*Umea University, Department of Economics*, 381, (1995).
[3] ,J.B. Kadane and G. Shmueli and G. Minka and T.P. Borle and P. Boatwright, Conjugate analysis of the Conway-Maxwell Poisson distribution,*Bayesian Analysis*, 1, 363-374, (2006).
[4] D. Lord and S.D. Guikema and S. Geedipally, Application of the Conway-Maxwell-Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes, *Accident Analysis and Prevention*,40,1123-1134, (2008).
[5] N. Mamode khan and V. Jowaheer, Comparing joint GQL estimation and GMM adaptive estimation in COM-Poisson longitudinal regression model, *Communication in Statistics- Simulation and Computation*, 42, 755-770, (2013).
[6] P. McCullagh and J.A. Nelder, Generalized Linear Models, *Chapman and Hall*, (1999).
[7] E. McKenzie, Autoregressive moving-average processes with negative binomial and geometric marginal distrbutions, *Advanced Applied Probability*, 18, 679-705,(1986).
[8] R.L Prentice and L.P Zhao,Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics*, 47,825-839, (1989).
[9] G. Shmueli and T. Minka and J.B. Borle and P. Boatwright, A useful distribution for fitting discrete data, *Applied Statistics,Journal of Royal Statistical Society*, (2005).
[10] B.C. Sutradhar, An Overview on Regression Models for Discrete Longitudinal Responses, *Statistical Science*,18, 377-393, (2003).
[11] B.C Sutradhar and K. Das, On the efficiency of regression estimators in generalised linear models for longitudinal data, *Biometrika*, 86, 459-465, (1999).
[12] B.C. Sutradhar and P. Farrell, Analyzing Multivariate Longitudinal Binary Data: A Generalized Estimating Equations Approach, *Canadian Journal of Statistics*, 32, 39-55, (2004).
[13] R. Wedderburn, Quasi-likelihood functions, generalized linear models, and the GaussNewton method, *Biometrika*, 61, 439-447, (1974).