Com-Poisson Model with AR(1)-type Correlation Structure for Longitudinal Count Response Data

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

Mamodekhan, Naushad Ali
University of Mauritius, Department of Economics and Statistics
Reduit, Mauritius
E-mail: n.mamodekhan@uom.ac.mu

ABSTRACT

Analysis of longitudinal count responses, affected by one or more explanatory variables, requires adequate modeling of the correlations underlying the responses repeatedly collected over time. These correlation structures may be then used with the corresponding means and variances of the process to formulate quasi-likelihood estimating equations which can be solved to obtain reliable estimates of the regression parameters. True Gaussian ARMA-type correlation structures have been found to be more efficient against 'working' as well as 'random effects based' correlation structures in analyzing Poisson longitudinal counts. Quite often, count responses are under-dispersed or over-dispersed relative to Poisson distribution. Com-Poisson distribution, due to its convincing properties, has been widely used to describe under- or over- dispersed count data in cross-sectional set up. However, there exists no application of Com-Poisson model in longitudinal case. The challenge lies in modeling the correlation structure of repeated Com-Poisson counts. In this paper, we provide the framework for developing and analysing a Com-Poisson model with AR(1)-type correlation structure under the longitudinal set-up.

INTRODUCTION

Com-Poisson distribution[10] is a weighted Poisson distribution, capable of modeling the counts subject to over-, under- or equi- dispersion. This capability of Com-Poisson distribution makes it a very useful distribution in practice. Generalised linear models (GLM) based on this distribution have been developed to analyse the regression effects in cross-sectional data where the count responses are collected at a single time-point[3,6]. The estimation of regression parameters in cross-sectional models can be done using maximum likelihood or quasi-likelihood estimation procedures[1,6]. Applications of GLM's are demonstrated for the analysis of over- as well as under- dispersed cross-sectional counts arising in different studies[4,7,9]. However, in various fields of research, count responses from a group of independent subjects are collected repeatedly over several time points and one is interested in estimating the effects of some known factors on these longitudinal count responses. Modeling of such longitudinal data is quite challenging as the repeated counts of each subject are correlated and the correlation pattern is unknown in practice. This is because it is extremely difficult to describe the joint distribution of correlated count observations. The problem gets further mounted up when the counts are over- or under- dispersed.

For the analysis of over-dispersed longitudinal count data, models have been developed by some authors. For example, we refer to random effects models by Thall & Vail[13] and the marginal model by Jowaheer & Sutradhar[5]. Concerning the analysis of under-dispersed longitudinal count data, there exists no literature. In this paper,we construct longitudinal Com-Poisson model which can analyse any of the three types of count responses : under-, over- or equi-dispersed. This model is basically an extension of Com-Poisson cross-sectional model developed by Jowaheer & Mamodekhan[6]. However, in

extending a model from cross-sectional to longitudinal set-up, one has to adequately model the correlations underlying the responses that are repeatedly collected over time. The longitudinal correlations can be modeled using random effects [see Thall & Vail[13]], 'working' ARMA structures [see Liang & Zeger[8]] or 'true' Gaussian type ARMA structures [see Jowaheer & Sutradhar[5]; Sutradhar[11]]. It has been shown that 'true' Gaussian type ARMA correlation structures are the most efficient and reliable in the sense that they are able to model the true underlying correlations of the repeated counts (Crowder[2]; Sutradhar & Das[12]). Developing ARMA correlation structures of different orders for correlated random variables following a discrete distribution is quite challenging and Com-Poisson distribution in particular has a complicated probability mass function. Here, we provide a method for developing and analysing Com-Poisson model with true AR(1)-type correlation structure under the longitudinal set-up.

Com-Poisson Random Variable

Let y_t be the count response at time t $(i = 1, 2, \ldots, I; t = 1, 2, \ldots, T)$, following Com-Poisson distribution. Then

$$f(y_t) = \frac{\lambda^{y_t}}{(y_t!)^{\nu}} \frac{1}{Z(\lambda, \nu)}, \tag{1}$$

where

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

and the parameter ν is the dispersion index such that $\nu = 1$, $\nu < 1$and $\nu > 1$ correspond to equi-, over- and under- dispersion. Since equation (1) does not have closed form expressions, we use an asymptotic expression for $Z(\lambda, \nu)$ proposed by Shmueli et. al[10] given by

$$Z(\lambda, \nu) \simeq \frac{\exp(\nu \lambda^{\frac{1}{\nu}})}{\lambda^{\frac{\nu-1}{2\nu}} (2\pi)^{\frac{\nu-1}{2}} \sqrt{\nu}} \tag{3}$$

The approximation is particularly good for $\lambda > 10^{\nu}$. Hence,

$$E(Y_t) = \theta = \lambda^{1/\nu} - \frac{\nu - 1}{2\nu}, \quad Var(Y_t) = \frac{\lambda^{1/\nu}}{\nu} \tag{4}$$

We can say that

$$Y_t \sim CMP(\frac{\theta}{\nu}, \nu) \tag{5}$$

As the responses are repeatedly collected over T time points, $y_1, y_2, ..., y_T$ are likely to be correlated. Following Sutradhar & Das[12], these T repeated counts have Gaussian-type true but unknown autocorrelation structure given by

$$C_i(\rho_1, \ldots, \rho_{T-1}) = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \ldots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \ldots & \rho_{T-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \ldots & 1 \end{pmatrix} \tag{6}$$

yielding the autocovariance structure

$$\Sigma(\tilde{\rho}) = A^{\frac{1}{2}} C_(\rho_1, \ldots, \rho_{T-1}) A^{\frac{1}{2}} \tag{7}$$

with A=diag $var(Y_t)$. We now derive the components of matrix $C(\rho_1, \ldots, \rho_{T-1})$ under stationary AR(1) correlation pattern.

Generating a Sequence of AR(1)-Correlated Com-Poisson Responses

Let response random variable y_{t-1} at $(t-1)th$ time point follows $CMP(\frac{\theta}{\nu}, \nu)$ and the error term $d_t \sim CMP(\frac{(1-\rho)\theta}{\tilde{\nu}}, \tilde{\nu})$ where

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

with $q_0 = (1-\rho)\theta$, $q_1 = \frac{q_0}{\nu}[1+\rho(1-\nu)] + \frac{\nu-1}{2\nu^2}(1-\rho^2)$ and $0 < \rho < 1$. Then, the stationary AR(1) model for the sequence $\{y_t\}$ can be formulated as

$$y_t = \rho * y_{t-1} + d_t \tag{9}$$

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

$$\rho * y_{t-1} = \sum_{j=1}^{y_{t-1}} b_j(\rho) = g_t. \tag{10}$$

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

$$(g_t \mid y_{t-1}, \rho) \sim \text{Binomial}(y_{t-1}, \rho) \tag{11}$$

Hence, using equation (9), we obtain

$$\begin{aligned} E(Y_t) &= E_{Y_{t-1}} E(\rho * Y_{t-1} | Y_{t-1}, \rho) + E(d_t) \\ &= E(\rho Y_{t-1}) + E(d_t) \\ &= \rho E(Y_{t-1}) + (1-\rho)\theta \\ &= \theta \end{aligned} \tag{12}$$

$$\begin{aligned} Var(Y_t) &= Var(\rho * Y_{t-1}) + Var(d_t) \\ &= Var(E(\rho * Y_{t-1})) + E(Var(\rho * Y_{t-1} | Y_{t-1})) + Var(d_t) \\ &= Var(\rho Y_{t-1}) + E(Y_{t-1}\rho(1-\rho)) + Var(d_t) \\ &= \rho^2[\frac{\theta}{\nu} + \frac{\nu-1}{2\nu^2}] + \rho(1-\rho)\theta + (\frac{\theta}{\nu} + \frac{\nu-1}{2\nu^2})(1-\rho^2) - \rho(1-\rho)\theta \\ &= \frac{\theta}{\nu} + \frac{\nu-1}{2\nu^2} \end{aligned} \tag{13}$$

$$\begin{aligned} E(Y_t Y_{t-k}) &= E_{Y_{t-k}} E_{Y_{t-k+1}} \ldots E_{Y_t}(Y_t Y_{i,t-k} | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k}) \\ &= E_{Y_{t-k}}[\rho^k Y_{t-k}^2 + \rho^{k-1}(1-\rho)\theta^2 + \ldots + \rho(1-\rho)\theta^2 + (1-\rho)\theta | Y_{t-k}] \\ &= \rho^k(\frac{\theta}{\nu} + \frac{\nu-1}{2\nu^2} + \theta^2) + \theta^2(1-\rho)(\rho^k + \rho^{k-1} + \ldots + 1) \\ &= \rho^k[(\frac{\theta}{\nu} + \frac{\nu-1}{2\nu^2}) + \theta^2] + \theta^2(1-\rho^k) \end{aligned} \tag{14}$$

$$Cov(Y_t, Y_{t-k}) = E(Y_t Y_{t-k}) - E(Y_t)E(Y_{t-k}) = \rho^k(\frac{\theta}{\nu} + \frac{\nu-1}{2\nu^2}) \tag{15}$$

and

$$\rho_k = Corr(Y_t, Y_{t-k}) = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t)}\sqrt{Var(Y_{t-k})}} = \rho^k \tag{16}$$

Equation(16) justifies AR(1)-type stationary auto-correlations between longitudinal Com-Poisson counts modelled by equation (9).

Longitudinal Com-Poisson Regression Model

Let y_{it} be the count response of the ith individual at time t $(i = 1, 2, \ldots, I; t = 1, 2, \ldots, T)$. Let x_{it} be the p dimensional vector of covariates corresponding to y_{it}. Let β be the p dimensional vector of regression parameters. The Com-Poisson regression model can thus be formulated by replacing y_t with y_{it} and λ_t by λ_{it} in equations (1) and (2) where

$$\lambda_{it} = \exp(x_{it}^T\beta) \tag{17}$$

Hence,

$$f(y_{it}) = \frac{\exp(x_{it}^T\beta y_{it})[\exp(x_{it}^T\beta(\frac{\nu-1}{2\nu}))(2\pi)^{\frac{\nu-1}{2\nu}}\sqrt{\nu}]}{(y_{it}!)^\nu[\exp(\nu\exp(\frac{x_{it}^T\beta}{\nu}))]} \tag{18}$$

where

$$E(Y_{it}) = \theta_{it} = \lambda_{it}^{1/\nu} - \frac{\nu-1}{2\nu}, \quad Var(Y_{it}) = \frac{\lambda_{it}^{1/\nu}}{\nu} \tag{19}$$

and the autocovariance structure,

$$\Sigma_i(\tilde{\rho}) = A_i^{\frac{1}{2}}C_i(\rho_1, \ldots, \rho_{T-1})A_i^{\frac{1}{2}} \tag{20}$$

with A_i=diag $var(Y_{it})$. The correlation matrix is unique for all the I clusters ,i.e, $C_i(\rho) = C(\rho)$.

Generalized Quasi-likelihood Estimating Equations

The GQL equations to estimate the regression and dispersion parameters of the longitudinal model may be written as

$$\sum_{i=1}^{I} D_i^T\widetilde{\Sigma_i}^{-1}(f_i - \mu_i) = 0, \tag{21}$$

where $f_i = (f_{i1}^T, \ldots, f_{it}^T, \ldots, f_{iT}^T)^T$, $\mu_i = (\mu_{i1}^T, \ldots, \mu_{it}^T, \ldots, \mu_{iT}^T)^T$ are $2T \times 1$ vectors with $f_{it} = (y_{it}, y_{it}^2)$, $\mu_{it} = (\theta_{i1}, m_{i1})^T$. $\theta_{i1} = E(Y_{it})$ and $m_{i1} = E(Y_{it}^2)$
The components D_i,$\widetilde{\Sigma_i}$,f_i and μ_i are obtained in the same way as in Jowaheer & Sutradhar[5] such that

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

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 $t = 1, \ldots, T$ where

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

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

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

$$\partial m_{i1}/\partial\nu = \frac{1}{2\nu^3}[2\lambda_{i1}^{\frac{1}{\nu}}\nu\ln(\lambda_{i1}) + \nu - 1 - 4\lambda_{i1}^{\frac{2}{\nu}}\ln(\lambda_{i1})\nu - 4\lambda_{i1}^{\frac{1}{\nu}}\nu - 4\lambda_{i1}^{\frac{1}{\nu}}\ln(\lambda_{i1})].$$

The covariance matrix of f_i is expressed as

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

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 $t = 1, \ldots, T$ and $w = 1, \ldots, T$.

The GQL estimating equation (21) is solved by the Newton-Raphson iterative method to obtain the estimates $\hat{\beta}$ and $\hat{\nu}$ such that

$$\begin{pmatrix} \hat{\beta}_{r+1} \\ \hat{\nu}_{r+1} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_r \\ \hat{\nu}_r \end{pmatrix} + [\sum_{i=1}^{I} D_i^T \widetilde{\Sigma_i}^{-1} D_i]_r^{-1} [\sum_{i=1}^{I} D_i^T \widetilde{\Sigma_i}^{-1} (f_i - \mu_i)]_r \tag{22}$$

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. $\rho_{|t-w|}$ is consistently estimated using the method of moments

$$\hat{\rho}_l = \frac{\sum_{i=1}^{I} \sum_{t=1}^{T-l} \tilde{y_{it}} \tilde{y_{i,t+l}} / (T-l)}{\sum_{i=1}^{I} \sum_{t=1}^{T} \tilde{y_{it}}^2 / T} \tag{23}$$

for $(l = |t-w| = 1, \ldots, T-1)$ where $\tilde{y_{it}} = \frac{y_{it} - \theta_{i1}}{\sqrt{\frac{\lambda_{i1}^{\frac{1}{\nu}}}{\nu}}}$. The algorithm works as follows: For an initial value of $\hat{\beta}$ and $\hat{\nu}$, we calculate $\hat{\rho}_l$ using (23) 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. The estimators are consistent and under mild regularity conditions, for $I \to \infty$, it may be shown that $I^{\frac{1}{2}}((\hat{\beta}, \hat{\nu}) - (\beta, \nu))^T$ has an asymptotic normal distribution with mean 0 and covariance matrix $I[\sum_{i=1}^{I} D_i^T \widetilde{\Sigma_i}^{-1} D_i]^{-1} [\sum_{i=1}^{I} D_i^T \widetilde{\Sigma_i}^{-1} (f_i - \mu_i)(f_i - \mu_i)^T \widetilde{\Sigma_i}^{-1} D_i] [\sum_{i=1}^{I} D_i^T \widetilde{\Sigma_i}^{-1} D_i]^{-1}$

Conclusions

In this paper, we provide a method of generating a sequence of Com-Poisson counts exhibiting Gaussian AR(1)-type autocorrelations. AR(1) autocorrelation structure has been then used to develop Com-Poisson longitudinal model to estimate the effects of fixed covariates on the repeatedly observed under-, over- and equi-dispersed count responses obtained from a large number of independent individuals. The regression and dispersion parameters are estimated by generalised quasi-likelihood estimation approach, whereas correlation parameters are estimated using method of moments.

REFERENCES

1. Consul, P. and Shoukri, M.(1984). Maximum likelihood estimation for generalized Poisson distribution. Communication in Statistics, Theory and Methodology, 17, 1533-1547.
2. Crowder, M. (1995) On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika, 82.
3. Guikema, S. (2007). Formulating informative, data-based priors for failure probability estimation in reliability analysis. Reliability Engineering and System Safety, 92(4), 490-502.
4. Guikema, S. and Gofelt, J. (2008). A flexible count data regression model for risk analysis. Risk Analysis, 28, 213-223.
5. Jowaheer, V. and Sutradhar, B.C. (2002). Analysing longitudinal count data with overdispersion. Biometrika, 89, 389-399.
6. Jowaheer, V. and Mamodekhan, N. (2009). Estimating regresion effects in Com-Poisson generalized linear model. International Journal of Mathematical and Statistical Sciences. 1:2, 59-63.
7. Kadane, G., Shmueli, G., Minka, G., Borle, T. and Boatwright, P. (2006). Conjugate analysis of the Conway Maxwell Poisson distribution. Bayesian Analysis, 1, 363-374.
8. Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalised linear models. Biometrika 73, 13-22.
9. Lord, S., Guikema, S., Geedipally, S. (2008). Application of the Conway-Maxwell-Poisson generalised linear model for analysing motor vehicle crashes, Accident Analysis and Prevention 40(3), 1123-1134.
10. Shmuelli, G., Minka, T.P., Kadane, J., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Com-Poisson. Applied Statistics: Journal of Royal Statistical Society, Series C., 54(1), 127-142.
11. Sutradhar, B. C. (2003). An overview on regression models for discrete longitudinal responses. Statistical Science, 18(3), 377-393.
12. Sutradhar, B.C. and Das, K. (1999). On the efficiency of regression estimators in generalised linear models for longitudinal data, Biometrika, 86, 459-465.
13. Thall, P. and Vail, S. (1990). Some covariance models for longitudinal count data with overdispersion. Biometrika 46, 657-671.