*Chaotic Modeling and Simulation (CMSIM)* **3: 235-241, 2015**

# Simulation of MA(1) Longitudinal Negative Binomial Counts and Using a Generalized Methods of Moment Equation Approach

**Naushad Mamode Khan**
Department of Economics and Statistics,University of Mauritius,Reduit Mauritius
(E-mail:n.mamodekhan@uom.ac.mu)

**Abstract**
Longitudinal count data often arise in financial and medical studies. In such applications, the data exhibit more variability and thus the variance to mean ratio is greater than one. Under such circumstances, the negative binomial is more convenient to be used for modeling these longitudinal responses. Since these responses are collected over time for the same subject, it is more likely that they will be correlated. In literature, various correlation models have been proposed and among them the most popular are the autoregressive and the moving average structures. Besides, these responses are often subject to multiple covariates that may be time-independent or time-dependent. In the event of time-independence, it is relatively easy to simulate and model the longitudinal negative binomial counts following the MA(1) structures but as for the case of time-dependence, the simulation of the MA(1) longitudinal count responses is a challenging problem. In this paper, we will use the binomial thinning operation to generate sets of MA(1) non-stationary longitudinal negative binomial counts and the efficiency of the simulation results are assessed via a generalized method of moments approach.
Keywords:
Negative Binomial, Longitudinal, Moving Average ,Binomial thinning, Stationary, Non-stationary, Generalized method of moments

## 1 Introduction

In today's era, longitudinal data has become extremely useful in applications related to the health and financial sectors. It constitutes of a number of subjects that are measured over a specified number of time points. Since these measurements are collected for a particular subject on a repetitive basis, it is more likely that the data will be correlated. The correlation structures may be following autoregressive, moving average, equi-correlation, unstructured or any other general autocorrelation structures[4][5]. Moreover, in longitudinal studies, the responses are influenced by many factors such as in the analysis of CD4 counts, the influential factors are the treatment, age, gender and many others. In order to estimate the contribution and the significance of each of these factors

Received: 28 July 2014 / Accepted: 5 December 2014

ISSN 2241-0503

towards the response variable, it is important to transform the data set-up into a regression framework. In literature, the regression parameters have been estimated by various approaches. Initially, the method of Generalized Estimating equations (GEE) were developed but it fails under misspecified correlation structure particularly under the independence correlation structure [5]. Thereafter, Prentice and Zhao [2] developed a Joint Estimation approach to estimate jointly the regression and correlation parameters and yielded more efficient regression estimates than the GEE approach but the joint estimation is based on higher order moments. Their approach is also based on the working correlation structure but the presence of these high order moments dilute the misspecification effect and boost the efficiency of the estimates. On the other hand, Qu and Lindsay [3] developed an adaptive quadratic inference based Generalized Method of Moments (GMM) approach where they assumed powers of the empirical covariance matrices as the bases. These bases are then used to form score vectors or moment estimating equations and thereafter, they were combined to form a quadratic function in a similar way as the GMM approach. This approach of analyzing longitudinal regression models has so far been tested on normal, Poisson data [3] but has not yet been explored in negative binomial correlated counts data. In this paper, our objectives are to develop the moment estimating equations based negative binomial model, construct the quadratic inference function and then obtain the regression estimates by maximizing the function. However, one challenging issue is that since the negative binomial model is a two parameter model (that is, depending on the mean and over-dispersion parameter), it implies that we will require higher order moments. This estimation approach will be tested via simulations on MA(1) stationary and non-stationary negative binomial counts. The organization of the paper is as follows: In the next section, we will review the negative binomial model along with its MA(1) Gaussian autocorrelation structure and the adaptive GMM approach following Qu and Lindsay [3]. In section 3, we will develop the estimating equations for the negative binomial model followed by simulation results.

## 2 Negative Binomial model

Longitudinal data comprise of data that are collected repeatedly over $t = 1,2,3,\ldots T$ time points for subjects $i = 1,2,3,\ldots I$ . Thus any $i^{th}$ random observation at $t^{th}$ time point will have a representation of the form $y_{it}$ . The negative binomial model for $y_{it}$ is given by

$$f(y_{it}) = \frac{\Gamma(c^{-1} + y_{it})}{\Gamma(c^{-1}) y_{it}!} \left( \frac{1}{1 + c\theta_{it}} \right)^{c^{-1}} \left( \frac{c\theta_{it}}{1 + c\theta_{it}} \right)^{y_{it}}$$

with $E(y_{it}) = \theta_{it} = \exp(x_{it}^T \beta)$ and $Var(y_{it}) = \theta_{it} + c\theta_{it}^2$ , $c > 0$ where in notation form ,

$$y_{it} \sim NeBin(\frac{1}{c}, c\theta_{it})$$

given a $p \times 1$ vector of covariates $x_{it}^T$ and vector of regression parameters $\beta$ of the form $\beta = [\beta_1, \beta_2, ..., \beta_p]^T$, $y_i = [y_{i1}, y_{i2}, ..., y_{it}, ....., y_{iT}]^T$ and $\theta_i = [\theta_{i1}, \theta_{i2}, ..., \theta_{it}, ....., \theta_{iT}]^T$.

Since these counts $y_{it}$ are collected repeatedly over time, it is more likely that $y_{it}$ will be correlated over time. In this paper, we will assume that the simulated $y_{it}$ set of response variables come from the family of MA(1) Gaussian autocorrelation structure. The derivation of the MA(1) stationary negative binomial counts follows from McKenzie binomial thinning process[1]. However, the derivation of the MA(1) non-stationary correlation structure has not yet appeared in statistical literature. In the next section, we provide an in-depth derivation of the MA(1) non-stationary Gaussian autocorrelation structure.

## 3 MA(1) Non-Stationary Gaussian autocorrelation Structures

In the non-stationary set-up, the mean parameter at each time point will differ as the covariates are time-dependent, that

$$\theta_{i1} \neq \theta_{i2} \neq ...\theta_{it} \neq ....\theta_{iT}$$

Following McKenzie[1], we set up the framework to generate the MA(1) non-stationary Gaussian autocorrelation structure. Tthe binomial thinning process assumes that

$$y_{it} = \alpha_{it} * d_{i,t-1} + d_{it}$$

where

$$d_{it} \sim NeBin(\frac{1}{\dot{c}}, \dot{c}\theta_{i1}), \alpha_{it} \sim Beta(\frac{\rho}{c}, \frac{1-\rho}{c}) \quad \text{and,}$$

$$\alpha_{it} * y_{i,t-1} = \sum_{j=1}^{y_{i,t-1}} b_j(\alpha_{it}) = z_{it},$$

prob[$b_j(\alpha_{it})$=1]= $\alpha_{it}$, prob[$b_j(\alpha_{it})$=0]=1- $\alpha_{it}$ and

$$\dot{c} = \frac{c(1+\rho+2\rho^2+c+2c\rho+c\rho^2)}{1+\rho^2+c+c\rho}$$

That is the conditional distribution of $\alpha_{it} * d_{i,t-1}$ follows the binomial distribution with parameters $d_{it-1}$ and $\alpha_{it}$. Under these assumptions, it can be proved that $y_{it} \sim NeBin(\frac{1}{c}, c\theta_{it})$ and the set of $y_i = [y_{i1}, y_{i2}, ..., y_{it}, ...y_{iT}]^T$ follows the MA(1) structure.. Under these distributional assumptions, we note that the covariance

between $y_{it}$ and $y_{it-k}$ is given by $\frac{\rho\theta_{i,t-k}}{1+\rho}+\dot{c}\frac{\rho\theta_{i,t-k}^2}{(1+\rho)^2}$ for $k=1$ and for other lags, the covariance does not exist.

## 4. Simulation of MA(1) Non –Stationary NB counts

The simulation process will follow from the binomial thinning operation explained in the previous section with $\theta_{it}=\exp(x_{it}^T\beta)$, that is we need to provide a given set of covariate designs and a set of regression vector $\beta$ that respects the dimension of the covariate matrix. Note that for the stationary case, the covariate matrix will be time independent while for the non-stationary, the covariate design will be time-dependent. As such, we assume for the non-stationary case the following designs,

Design A

$$x_{it1}=\left\{\begin{array}{c}-0.5t\times rbinom(3,0.2), t=1...\frac{I}{4}\\ t\times rpois(2), t=\frac{I}{4}+1,...,\frac{3I}{4}\\ 1.5+t, t=\frac{3I}{4}+1,...,I\end{array}\right\}$$

Design B

$$x_{it1}=\left\{\begin{array}{c}-0.5t\times \sin t, t=1...\frac{I}{4}\\ \exp(t), t=\frac{I}{4}+1,...,\frac{3I}{4}\\ \cos t, t=\frac{3I}{4}+1,...,I\end{array}\right\}$$

Design C

$$x_{it1}=\left\{\begin{array}{c}t, t=1...\frac{I}{4}\\ \ln(t), t=\frac{I}{4}+1,...,\frac{3I}{4}\\ t-1, t=\frac{3I}{4}+1,...,I\end{array}\right\}$$

and $x_{it2}$ is generated from the Poisson distribution with mean parameter 2. In this way, the mean parameter for each subject $i$ will vary. Thus, for these set of covariates and initial estimate of the regression vector, dispersion parameter and correlation parameter, we generate MA(1) Negative Binomial random variables by first simulating the error

components $d_{it}$, $y_{it-1}$ and the thinning operation random variables $\alpha_{it} * y_{i,t-1}$. For our simulation process, we will assume the values of $\beta = [1,1]^T$.

## 5. Estimation Methodology

Qu and Lindsay [3] have developed an estimation approach based Generalized Methods of Moments that do not require any assumption in the underlying correlation structure and do not require any estimation of the correlation parameter. In fact, Qu and Linsday [3] assumed a score vector that only needs the empirical covariance estimation matrix

$$V = \frac{1}{I}\sum_{i=1}^{I}(y_i - \theta_i)(y_i - \theta_i)^T ,$$

$$g = \begin{pmatrix} \sum_{i=1}^{I} D_i^T (y_i - \theta_i) \\ \sum_{i=1}^{I} \alpha^T D_i^T V (y_i - \theta_i) \end{pmatrix}$$

where $D_i$ is the gradient matrix: $D_{it} = \dfrac{\partial \theta_{it}}{\partial \beta^T}$ and $\alpha$ is an orthogonal vector. The calculation of the parameter $\alpha$ requires the conjugate gradient method [see Qu and Lindsay [3]]. In the context of the negative binomial model, the score vector $g$ is defined as:

$$g = \begin{pmatrix} \sum_{i=1}^{I} D_i^T (f_i - \theta^*{}_i) \\ \sum_{i=1}^{I} \alpha^T D_i^T V (f_i - \theta^*{}_i) \end{pmatrix}$$

where the vectors $f_i = [y_i, y_i^2]^T$, $\theta^*{}_i = E[f_i] = [\theta_i, \theta_i + (c+1)\theta_i^2]^T$ ,

$V = \dfrac{1}{I}\sum_{i=1}^{I}(f_i - \theta^*{}_i)(f_i - \theta^*{}_i)^T$ and

$D_i = [\dfrac{\partial \theta^*{}_i}{\partial \beta^T}, \dfrac{\partial \theta^*{}_i}{\partial c}] = [D_{i1}, D_{i2}, ..., D_{it}, ...D_{iT}]^T$ where

$$D_{it} = \begin{pmatrix} \dfrac{\partial \theta_{it}}{\partial \beta^T} & 0 \\ \dfrac{\partial [\theta_i + (c+1)\theta_i^2]}{\partial \beta^T} & \dfrac{\partial [\theta_i + (c+1)\theta_i^2]}{\partial c} \end{pmatrix}$$

Using the score vector $g$, Qu and Lindsay [3] defined the objective function

$$Q(\beta, c) = g^T C^{-1} g$$

where $C$ is the sample variance of $g$

$$\begin{pmatrix} \sum_{i=1}^{I} D_i^T V D_i & [\sum_{i=1}^{I} D_i^T V^2 D_i]\alpha \\ [\sum_{i=1}^{I} D_i^T V^2 D_i]\alpha & \alpha^T [\sum_{i=1}^{I} D_i^T V^3 D_i]\alpha \end{pmatrix}$$

By maximizing the objective function with respect to the unknown set of parameters, we obtain the estimating equation

$$\dot{Q}(\beta, c) = 2\ \dot{g}^T C^{-1} g$$

with $\dot{g} = [\frac{\partial g}{\partial \beta^T}, \frac{\partial g}{\partial c}]^T$. Since the above estimating equation is non-linear, we solve the equation using the Newton-Raphson procedure that yields an iterative equation of the form

$$\begin{pmatrix} \hat{\beta}_{r+1} \\ \hat{c}_{r+1} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_r \\ \hat{c}_r \end{pmatrix} - [\dot{\dot{Q}}(\beta, c)]_r^{-1} [\ \dot{Q}(\beta, c)]_r$$

where $[\dot{\dot{Q}}(\beta, c)] = 2\,\dot{g}^T C^{-1} \dot{g}$ is the double derivative hessian part of the score function and this is being used for calculating the variance of the regression and over-dispersion parameters. As illustrated by Qu and Lindsay [3], this method yields consistent and efficient estimators and tends towards asymptotic normality for large sample size.

## 6. Results and Conclusion

Following the previous sections, we have run 10,000 simulations for each of the sample sizes $I = 20,50,100,200,500$ based on the different covariate designs for the non-stationary set-ups. Note that for the stationary case, the mean is held constant at all time points whilst for non-stationary, the mean varies with the time points given the time-dependent covariates. The table provides the simulated mean estimates of the regression parameters along with the standard errors in brackets.

| I | Design A | Design B | Design C |
|---|---|---|---|
| 20 | 0.9919;1.0010<br>(0.1351;0.2120) | 1.0121;0.9987<br>(0.1401;0.1971) | 0.9956;1.0013<br>(0.2212;0.1898) |
| 50 | 1.0110;0.9978<br>(0.1022;0.1762) | 0.9919;0.9995<br>(0.1211;0.1881) | 0.9982;1.0121<br>(0.1580;0.1) |
| 100 | 0.9982;0.9995<br>(0.0812;0.1120) | 1.0101;0.9961<br>(0.0754;0.1052) | 0.9988;1.0015<br>(0.0889;0.1010) |
| 200 | 1.0012;1.0005<br>(0.0661;0.0991) | 0.9992;0.9992<br>(0.0762;0.0975) | 1.0042;1.0141<br>(0.0562;0.0888) |
| 500 | 0.9999;1.0001<br>(0.0552;0.0808) | 0.9992;0.9993<br>(0.0432;0.0652) | 0.9978;1.0010<br>(0.0466;0.0762) |

Based on the simulation results, we note that the estimates of the regression parameters are close to the population values and as the sample size increases, the standard errors of the regression parameters decrease which indicates that the estimates are consistent and

efficient. However, we have remarked a significant number of failures in the simulations as we increase the sample size. These failures were mainly due to ill-conditioned nature of the double derivative Hessian matrix. To overcome this problem in some simulations, we have used the Moore Penrose generalized inverse method in R (ginv in Library MASS) to perform the iterative procedures. Overall, the generalized method of moments estimation technique is a statistically sound technique but in terms of computation, it may not always be reliable.

## References

1. E. McKenzie. Autoregressive moving-average processes with negative binomial and geometric marginal distrbutions. Advanced Applied Probability 18, 679–705, 1986.
2. R. Prentice, R. & L. Zhao (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47, 825–39,1991.
3. A.Qu & B. Lindsay (2003). Building adaptive estimating equations when inverse of covariance estimation is difficult. Journal of Royal Statistical Society 65, 127–142,2003.
4. B. Sutradhar. An overview on regression models for discrete longitudinal responses. Statistical Science 18(3), 377–393, 2003.
5. B. Sutradhar, B. & K. Das. On the efficiency of regression estimators in generalized linear models for longitudinal data. Biometrika 86, 459–65, 1999.