Generalized Quasilikelihood Inference for Zero Inflated Longitudinal Count Data

Jannatul Ferdous Antu, Sabina Sharmin and Taslim Sazzad Mallick[*]

Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh

Abstract

In this paper, we extend an observation-driven model for time series of zero inflated count data to longitudinal data setup. Basic properties of the models are discussed. For statistical inference of the proposed model, a generalized quasilikelihood (GQL) estimating equation has been derived for the regression parameter. A pharmaceutical data has been reanalyzed using the proposed approach and results are compared. The proposed approach produces similar estimates as given in the earlier work with much smaller standard errors.

I. Introduction

Regression analysis of count data has become a common interest in almost every discipline. For example, in public health, it may be of interest to analyze the frequency of antenatal care visits or hospital visits. Similar applications are also found in finance, insurance and criminology, where data on frequency of failure of a financial institution, frequency of accidents and the number of criminal offenses are analyzed. Analyses of such count data are routinely carried out using generalized linear model (GLM)[1] with Poisson or negative binomial distribution. There are, however, situations where count data exhibit excess zeros than the standard count model can assume. GLM without accommodating this 'excess zeros' may result in lack of fit[2]. The modified count model that incorporates the probability of such excess zeros is termed as zero inflated or two-part model. This type of models is in fact mixture of two distributions; one of them is essentially a binary distribution that puts extra mass for the zero counts. In the count data literature, zero-inflated (ZI)[3] and Hurdle[4] models are two such models that are commonly used depending on the nature of zeros. The difference between ZI and hurdle models lies in the fact that ZI assumes the population consists of both 'at risk' and 'not at risk' respondents, while hurdle assumes only 'at risk' respondents in the population. Consequently, ZI models incorporates excess zero from both binary and Poisson processes and hurdle models uses zero truncated count model assuming zeros originated only from the binary components. Cohen[5] proposed a general method for estimating the mixing mixtures of known distributions. Applications of these modified count models are found for analyzing number of recreational boating trips[6], number of non-payments on credit scoring applications[7], modelling abundance of rare species[8], accident counts[9], the number of decayed and filled tooth surfaces[10] among others in the literature.

Recently, analyses of ZI correlated count data have been used in social and health related studies[11]. Because of the correlation among repeated responses in longitudinal setting, the behavior of counts as well as zero inflation is expected to change over time. Therefore, modelling such correlated ZI count data is much more challenging compared to the models used in longitudinal count data literature. There exist few studies that deal with modelling such ZI longitudinal count data along with its inference procedures. These studies mostly extended cross-section models by introducing random effects through which the longitudinal ZI counts are assumed to be correlated. Assuming independence among binary responses, Hall used subject specific random effects to accommodate correlation among longitudinal count responses[12]. On the other hand, some authors used random effects for both binary and Poisson processes under ZIP or hurdle setup. Using Gaussian random effects for each of the two components of a hurdle model, Yau & Lee[13] discussed Penalized Quasiikelihood (PQL) approach for the inference procedure. ZIP mixed model has also been used in a study[14] for modelling clustered length of hospital stay data (LOS) using independent random effects for binary and count portion of the model. This study[14] used residual maximum likelihood (REML) method through expectation maximization (EM) algorithm for parameter estimation. In another study[15], LOS data was analyzed using zero inflated negative binomial mixed regression. Min & Agresti[2] used correlated Gaussian random effects in Poisson hurdle model for a longitudinal pharmaceutical study, where two treatments were compared in terms of number of episodes of a certain side effect. For the estimation of parameters, they have used both ML and nonparametric maximum likelihood (NPML) estimation approaches. Some authors[16] used two independent random effects, one for within-cluster and the other for within-individual correlations for both binary and count parts of ZIP to model clustered longitudinal data. For the estimation of the parameters, they have used restricted ML approach. Recently, Long *et al.*[17] used marginalized ZI Poisson model for longitudinal data using certain random effects. They employed adaptive Gauss-Hermite quadrature to obtain ML estimates of the parameters associated with their proposed model. For a nice review of recent development on two-part models for longitudinal data, see Farewell *et al.*[11].

Although modelling correlated count data is convenient using random effect models, interpretation of correlation

---

[*]Author for correspondence. e-mail: *tsmallick@du.ac.bd*

among repeated responses is not clear as it is modelled as function of unobserved random effects. There exists alternative observation-driven models in time series of count data, where correlations are modelled through past count responses by allowing correlation structures to have autoregressive or moving average type structures which are widely used in continuous time series data[18-21]. For a comprehensive review of such models, we refer to Weiß[22]. Since longitudinal data on n individuals can be thought of as n independent time series, several authors[23,24] have used those models under longitudinal data setup. Hasan and Sneddon[25] proposed an observation-driven ZIP model for analyzing longitudinal data of annual number of visits to physician. For statistical inference, they have used generalized quasilikelihood (GQL) approach[26]. Their proposed model provides AR(1) type correlation structure, where the lag correlations among repeated ZIP responses are function of the product of inflation and correlation parameters of count responses. Since the domains of inflation and correlation parameters lie between 0 and 1, this model is useful for longitudinal ZIP responses with low correlations.

Recently, Jazi *et al.*[27] have introduced an observation-driven ZIP model for time series data. This approach of modelling is appealing and can be extended for the longitudinal data. Note that Jazi *et al.*[27] did not consider their model under regression setup. The aim of this paper is to extend the ZIP model of Jazi *et al.*[27] for longitudinal data under regression setup, derive simpler GQL estimating equation for estimating regression effects and reanalyze the pharmaceutical data previously analyzed by Min & Agresti[2]. In Section 2, the model is discussed along with its basic properties, GQL estimating equation is derived in Section 3, application of the proposed methods is given in Section 4 and the paper concludes in Section 5.

II. Proposed ZIP Autoregressive Model for Longitudinal Data

Let, y_{it} be the count response for i-th$(i = 1, 2, ..., n)$ respondent at t-th$(t = 1, 2, ..., T)$ time point. Following Jazi *et al.*[27], we assume that the longitudinal zero-inflated count data have been generated from the following process

$$y_{i1} \sim ZIP\,(\alpha_{i1}, \mu_{i1})$$
$$y_{it} = \rho * y_{i,t-1} + d_{it};\ t > 1, \tag{1}$$

Where $d_{it} \sim ZIP\,(\alpha_{it}, \mu_{it})$; $ZIP\,(\alpha_{it}, \mu_{it})$ being the probability mass function (pmf) of a mixture of a distribution degenerate at zero and a Poisson distribution with mean μ_{it} having the form

$$P(d_{it} = k) = \alpha_{it}I_{\{0\}}(k) + \frac{(1 - \alpha_{it})e^{-\mu_{it}}\mu_{it}^k}{k!}, k \tag{2}$$
$$= 0, 1, 2, \cdots.$$

We denote the pmf in (2) as $d_{it} \sim ZIP\,(\alpha_{it}, \mu_{it})$ with $I_A(x)$ as an indicator function that equals to 1, if $x \in A$ else equal

to zero. The term $\rho * y_{i,t-1}$ in (1) is computed through a binomial thinning operation[20] defined as

$$\rho * y_{i,t-1} = \sum_{j=1}^{y_{i,t-1}} B_j(\rho), \tag{3}$$

where the counting series $B_j(\rho)$ is a sequence of independent identically distributed binary random variables with $P[B_j(\rho) = 1] = 1 - P[B_j(\rho) = 0] = \rho, \rho \in (0,1)$. For $t = 1$, $y_{i1} \sim ZIP(\alpha_{i1}, \mu_{i1})$. Note that the parameters $\alpha_{it} = P(Y_{it} = 0)$ and ρ in (1)-(3) are referred to as inflation and correlation parameters, respectively. For a p-dimensional vector of covariates associated with i-th respondent at t-th time, $x_{it} = (x_{it1}, x_{it2}, \cdots, x_{itp})'$ and the corresponding effect parameters $\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$, under regression setup, we define $\mu_{it} = \exp(x_{it}'\beta)$ in (2). It follows from (1)-(3) that

$$\lambda_{it} = \mathrm{E}(Y_{it}) = \frac{(1 - \alpha_{it})\mu_{it}}{1 - \rho}, \tag{4}$$

$$\sigma_{i,tt}^2 = Var(Y_{it}) = \frac{\mu_{it}(1 - \alpha_{it})(1 + \rho + \alpha_{it}\mu_{it})}{1 - \rho^2}, \tag{5}$$

$$Cov(Y_{it}, Y_{i,t+l}) = \sigma_{i,tt}\sigma_{i,(t+l)(t+l)}\rho^l. \tag{6}$$

For $\rho = 0$, it is easy to verify that the model reduces to ZIP model under cross-section setup. Assuming stationary process, $x_{i1} = x_{i2} = \cdots = x_{iT}$, implying $\mu_{it} = \mu_i$ and $\alpha_{it} = \alpha_i$ for all $i = 1, 2, \cdots, n;\ t = 1, 2, \cdots, T$, the mean and variance of Y_{it} is time invariant, i.e., $\lambda_{it} = \lambda_i$ and $\sigma_{i,tt} = \sigma_i$. As opposed to the random effect models, model (1) provides AR(1) correlation structure, i.e., the lag-l correlation, $Cor(Y_{it}, Y_{i,t+l}) = Cor(Y_{it}, Y_{i,t+l}) = \rho^l$. Note that the assumption of AR(1) correlation structure is appealing from practical point of view.

III. Statistical Inference: Generalized Quasilikelihood Approach

Let $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{iT})'$ denote the mean vector and $(\Sigma_i)_{u,t} = Cov(Y_{iu}, Y_{it}) = \sigma_{i,uu}\sigma_{i,tt}\rho^l$ denote the (u, t)th component of the $T \times T$ covariance matrix for i-th respondent. For the purpose of estimating regression parameter β associated with model (1), we use GQL estimating equation[29] given by

$$\sum_{i=1}^n \frac{\partial \lambda_i'}{\partial \beta} \Sigma_i^{-1}(Y_i - \lambda_i) = 0, \tag{7}$$

Where $\partial \lambda_i / \partial \beta_j = (\partial \lambda_{i1}/\partial \beta_j, \partial \lambda_{i2}/\partial \beta_j, \cdots, \partial \lambda_{iT}/\partial \beta_j)'$ with $\partial \lambda_{it}/\partial \beta_j = x_{itj}(1 - \alpha_{it})\mu_{it}/(1 - \rho)$. The covariance matrix Σ_i in (7) is constructed based on (5)-(6). Note that the covariance matrix Σ_i under generalized estimating equation (GEE)[28] is constructed based on some 'working' correlation matrix, but in GQL it is constructed based on model based true correlation structure (5)-(6). Several authors[29-31] have used GEE under longitudinal ZIP data by

considering 'working' correlation in conjunction with the cross-sectional ZIP model. The use of 'working' correlation has been shown to have many pitfalls for standard longitudinal count data[32,33]; therefore this approach may not be appropriate under more complex ZI longitudinal data. Under some mild regularity conditions[34], it may be shown that the GQL estimator $\hat{\beta}$ obtained from (7) asymptotically follows a Gaussian distribution with mean β and covariance matrix given by

$$Cov(\hat{\beta}) = \left[\sum_{i=1}^{n} \frac{\partial \lambda_i'}{\partial \beta} \Sigma_i^{-1} \frac{\partial \lambda_i}{\partial \beta} \right]^{-1}.$$

Note that the GQL estimating equation (7) is a function of other unknown parameters α_{it} and ρ which should be estimated from the data.

Assuming suitable 'working' correlation structure and a set of covariates $z_{it} = (z_{it1}, z_{it2}, \cdots, z_{itq})'$ with corresponding effect parameters $\eta = (\eta_1, \eta_2, \cdots, \eta_q)'$, a GEE with logit link $\text{logit}(1 - \alpha_{it}) = z_{it}'\eta$ can be used to estimate η which is then used to obtain $\hat{\alpha}_{it}$. As far as the estimation of correlation parameter ρ is concerned, following, we propose a consistent estimator using method of moments[35]. The moment estimator of ρ can be written as

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T} \tilde{y}_{it}\tilde{y}_{i,t-1}}{\sum_{i=1}^{n}\sum_{t=2}^{T} \tilde{y}_{it}^2} \frac{T \times n}{n \times (T-1)}, \quad (8)$$

where $\tilde{y}_{it} = (y_{it} - \lambda_{it})/\sigma_{i,tt}$.

The proposed estimation approach is carried out in the following sequence:

1. Using GEE with suitable 'working' correlation matrix, obtain $\hat{\alpha}_{it}$.

2. With an initial value of ρ and $\hat{\alpha}_{it}$, the GQL estimate $\hat{\beta}$ is obtained using Newton-Raphson iterative procedure.

3. The converged estimate $\hat{\beta}$ from step 2 is then used to compute $\hat{\rho}$.

4. The estimation of β and ρ are recursively executed until convergence.

IV. Analysis of Pharmaceutical Data

The proposed ZIP model for longitudinal data and GQL estimation approach will be illustrated by reanalyzing a pharmaceutical data[2]. The purpose of the data analysis is to compare two treatments for a particular disease, where the number of episodes of a certain side effect are the count responses. A total of $n = 118$ patients were randomly allocated to receive either one of two treatments, where 59 of them received treatment A and the other 59 received B. The count response of side effect episodes was measured at each of six visits. Figure-1 shows the frequencies of the side effects at each follow-up, approximately 80% of the cases are observed without any side effect at each visits. This motivates one to use zero inflated count models for the

analysis. Overall, the average number of episodes of the side effect was found as 0.158 and 0.415 for treatment A and B, respectively (p-value < 0.01 for standard t-test).
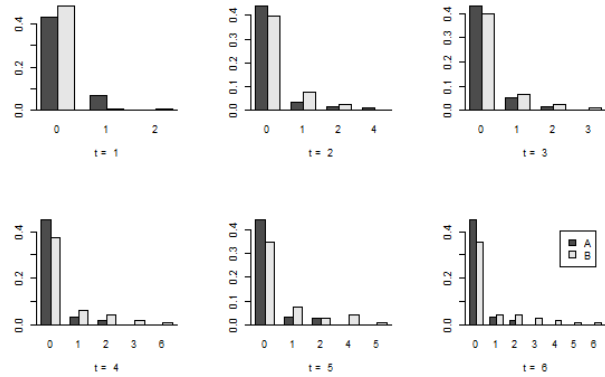

Fig. 1. Episodes of side effects at each visits by treatment

We next fit our proposed longitudinal ZIP regression model to the data. For the purpose of comparison, we consider the same setup as of Min and Agresti[2]. To be specific, for i-th respondent at the t-th visit, we consider

$$\text{logit}(1 - \alpha_{it}) = \eta_0 + \eta_1 TrtB + \eta_2 \log(\text{Time}),$$

$$\log(\mu_{it}) = \beta_0 + \beta_1 TrtB + \beta_2 \log(\text{Time}).$$

Recall that, the proposed model assumes $E(Y_{it}) = \mu_{it}(1 - \alpha_{it})/(1 - \rho)$. Unlike Min and Agresti[2], we have estimated η_j's $(j = 0, 1, 2)$ using generalized estimating equation GEE[30] with unstructured correlation structure. The columns 6-7 of Table-I under the heading 'Full data GQL' represent the GQL estimates of β, GEE estimates of η and the moment estimate of ρ, along with their standard errors (SE). For comparison, we also report the ML and NPML estimates of Min & Agresti[2] in the same table. Overall, the parameter estimates obtained by our proposed method indicate similar effect as in Min & Agresti[2] with much smaller standard errors, but indicate different strength of association between the effects of treatment and time with episodes of side effect counts.

Table 1. Comparison of parameter estimates under proposed GQL approach with ML and NPML of Min and Agresti[2] for side effect count data

Parameter	ML Estimate	SE	NPML Estimate	SE	Full data GQL Estimate	SE
η_0	-2.874	0.622	-2.813	0.576	-2.091	0.518
η_1	0.895*	0.417	0.958*	0.335	0.483	0.302
η_2	0.021	0.186	0.022	0.185	0.040	0.165
β_0	-2.844	0.735	-2.880	0.619	-0.629	0.197
β_1	0.963*	0.352	0.898*	0.294	0.488*	0.063
β_2	0.540*	0.192	0.494*	0.188	0.152*	0.016
ρ	-	-	-	-	0.422	-

*p-value < 0.05

For example, for the binary part, ML estimate[2] of treatment effect indicates that the odds of positive side effect count for treatment B is 145% higher as compared to treatment A [Odds ratio = $\exp(0.895) = 2.45$], while GEE estimated it

to be about 62% higher for treatment B. Note that, although ML estimate found B to have significantly higher odds of positive side effect counts, GEE shows no significant effect of treatment. The effect of time appears to be insignificant in all approaches for the binary part. As far as the estimates of the count model are concerned, all the estimation approaches indicate that treatment B has significantly higher number of side effect counts and an increase in time between follow-ups results in significantly higher number of side effect counts. For example, the ML estimates[2] shows that patients receiving treatment B experienced side effect episodes 162% more often, while our proposed GQL estimate reveals that it is only 63% more for treatment B.

Note that in the data, there are 64 respondents for whom no side effect counts were observed throughout the study period. This raises a question about the existence of a separate population who may never show such side effects. If this is the case, we can drop them from the study as they have no chance of experience that particular side effect. Figure-2 depicts side effect counts for the reduced data of $n = 54$ respondents. The data still contains a large proportion of zeros and thus modified count regression model is still appropriate to analyze it.
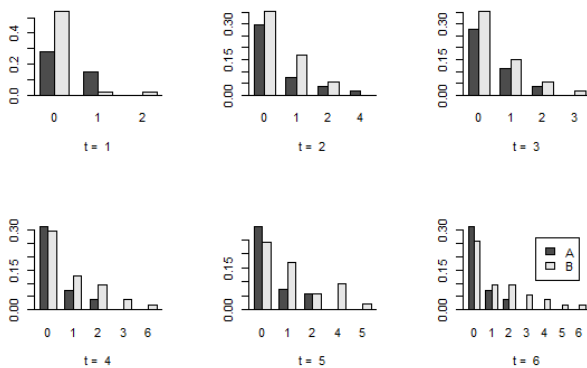

Fig. 2. Episodes of side effects at each visits by treatment for 'reduced' data

The GQL estimates of β, GEE estimates of η and the moment estimate of ρ for reduced data are reported in Table-II. The estimates are very close to what we have found by our proposed approach for the full dataset.

Table 2. GQL estimates for 'reduced' side effect count data

Parameter	Estimate	SE	Parameter	Estimate	SE
η_0	-0.824	0.579	β_0	-0.554	0.200
η_1	0.264	0.230	β_1	0.454*	0.057
η_2	0.009	0.203	β_2	0.178*	0.018
			ρ	0.311	-

V.   Conclusion

In this paper, we have extended an observation-driven ZIP time series model [1] to the longitudinal data with regression formulation. As opposed to the existing parameter-driven models, where correlations are modelled using certain random effects, the proposed stochastic model

accommodates AR(1) correlation structure among repeated stationary zero inflated counts. Since the repeated count response on certain visit depends on both binary and count response of the past, it is expected that the average count on that visit cannot be the same as in the baseline. The model we propose for the longitudinal ZI counts not only provides different mean on follow-ups, but also provides stochastic AR(1) correlation structure that is more practical and appealing to accommodate temporal association. Moreover, the proposed GQL estimation approach is computationally simpler as compared to other existing estimation approaches.

When we reanalyzed the pharmaceutical data, similar to Min & Agresti[2], the proposed method found the treatment B has significantly higher side effect count, however the rate of occurrence was much smaller in our approach.

References

1.   Nelder J.A., R.W. Wedderburn, 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General).* **135(3)**, 370-384.

2.   Min Y., A. Agresti, 2005. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling.* , **5(1)**, 1-19.

3.   Lambert D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics,* **34(1)**, 1-14.

4.   Mullahy J., 1986. Specification and testing of some modified count data models. *Journal of Econometrics,* **33(3)**, 341-365.

5.   Cohen A.C., 1963. *Estimation in mixtures of discrete distributions.* Statistical Pub Society

6.   Gurmu S., P.K. Trivedi, 1996. Excess zeros in count models for recreational trips. *Journal of Business & Economic Statistics,* **14(4)**, 469-477.

7.   Dionne G., M. Artís, M. Guillén, 1996. Count data models for a credit scoring system. *Journal of Empirical Finance.* **3(3)**, 303-325.

8.   Welsh A.H., R.B. Cunningham, C. Donnelly , D.B. Lindenmayer, 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling,* **88(1-3)**, 297-308.

9.   Shankar V., J. Milton, F. Mannering, 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention,* **29(6)**, 829-837.

10. Preisser J.S., K. Das, D.L. Long, K. Divaris, 2016. Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in medicine,* **35(10)**, 1722-1735.

11. Farewell V., D. Long, B. Tom, S. Yiu, L. Su, 2017. Two-part and related regression models for longitudinal data. *Annual Review of Statistics and its Application,* **4**, 283-315.

12. Hall D.B., 2000.   Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics,* **56(4)**, 1030-1039.

13. Yau K.K., A.H. Lee, 2001. Zero-inflated Poisson regression with random effects to evaluate an occupational injury

prevention programme. *Statistics in medicine,* **20(19)**, 2907-2920.

14. Wang K., K. K. Yau, A.H. Lee, 2002. A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. *Computer Methods and Programs in Biomedicine,* **68(3)**, 195-203.

15. Yau K.K., K. Wang, A.H. Lee, 2003. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: Journal of Mathematical Methods in Biosciences,* **45(4)**, 437-452.

16. Lee A.H., K. Wang, J.A. Scott, K. K. Yau, G. J. McLachlan, 2006. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research,* **15(1)**, 47-61.

17. Long D. L., J. S. Preisser, A. H. Herring, C.E. Golin, 2015. A marginalized zero-inflated Poisson regression model with random effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* **64(5)**, 815-830.

18. McKenzie E., 1986. Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Advances in Applied Probability,* **18(3)**, 679-705.

19. McKenzie E., 1988. Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability,* **20(4)**, 822-835.

20. Al-Osh M.A., E-E.A. Aly, 1992. First order autoregressive time series with negative binomial and geometric marginals. *Communications in Statistics-Theory and Methods* **21(9)**, 2483-2492.

21. Al-Osh M., A.A. Alzaid, 1987. First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis,* **8(3)**, 261-275.

22. Weiß C.H., 2018.*An introduction to discrete-valued time series.* John Wiley & Sons.

23. Sutradhar B.C., 2003. An overview on regression models for discrete longitudinal responses. *Statistical Science,* **18(3)**, 377-393.

24. Qaqish B.F., 2003. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika,* **90(2)**, 455-463.

25. Hasan M.T. , G. Sneddon. Zero-inflated Poisson regression for longitudinal data. *Communications in Statistics-Simulation and Computation,* **38(3)**, 638-653.

26. Sutradhar B.C., R.P. Rao. On quasi-likelihood inference in generalized linear mixed models with two components of dispersion. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique,* 2003, 415-435.

27. Jazi M.A.,G. Jones, C.D. Lai, 2012. First-order integer valued AR processes with zero inflated poisson innovations. *Journal of Time Series Analysis,* **33(6),** 954-963.

28. Liang K.Y., S.L. Zeger, 1986. Longitudinal data analysis using generalized linear models. *Biometrika,* **73(1)**, 13-22.

29. Dobbie M.J., A.H. Welsh, 2001. Theory & Methods: Modelling Correlated Zero-inflated Count Data. *Australian & New Zealand Journal of Statistics.* **43(4)**, 431-444.

30. Hall D.B., Z. Zhang, 2004. Marginal models for zero inflated clustered data. *Statistical Modelling.* **4(3)**, 161-180.

31. Yang Y., D. Simpson, 2010. Unified computational methods for regression analysis of zero-inflated and bound-inflated data. *Computational Statistics & Data Analysis,* **54(6)**, 1525-1534.

32. Crowder M.,1995. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika,* **82(2)**, 407-410.

33. Sutradhar B.C., K. Das, 1999. On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika,* **86(2)**, 459-465.

34. Zeger S.L., K.Y. Liang, P.S. Albert . Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049-1060.

35. Mallick T.S, B.C. Sutradhar, 2008. GQL Versus Conditional GQL Inferences for Non-Stationary Time Series of Counts with Overdispersion. *Journal of Time Series Analysis,* **29(2)**, 402-420.