Modelling Truncated and Clustered Count Data

Ayoub Saei, Ray Chambers

Abstract

Count response data often exhibit departures from the assumptions of standard Poisson generalized linear models (McCullagh & Nelder 1989). In particular, cluster level correlation of the data and truncation at zero are two common characteristics of such data. In this paper we describe a random components truncated Poisson model that can be applied to clustered and zero-truncated count data. Residual maximum likelihood method estimators for the parameters of this model are developed and their use illustrated using a data set of non-zero counts of sheets with edge strain defects in iron sheets produced by the Mobarekeh Steel Complex, Iran. We also report on a small scale simulation study that supports the estimation procedure.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/10

Modelling Truncated and Clustered Count Data

Ayoub Saei and Ray Chambers

Southampton Statistical Sciences Research Institute, University of Southampton,
Highfield, Southampton, SO17 1BJ, United Kingdom

Summary

Count response data often exhibit departures from the assumptions of standard Poisson generalized linear models (McCullagh & Nelder 1989). In particular, cluster level correlation of the data and truncation at zero are two common characteristics of such data. In this paper we describe a random components truncated Poisson model that can be applied to clustered and zero-truncated count data. Residual maximum likelihood method estimators for the parameters of this model are developed and their use illustrated using a data set of non-zero counts of sheets with edge strain defects in iron sheets produced by the Mobarekeh Steel Complex, Iran. We also report on a small scale simulation study that supports the estimation procedure.

Key words: Cluster, Poisson, Random components, REML, Truncated.

1. Introduction

Edge strain defects in iron sheets are a major production problem at the Mobarekeh Steel Complex (MSC), Iran. Data on numbers of coils with edge defect for iron sheets produced by the tandem-mill and skin-pass processes at MSC were recorded over a period of 9 months, from March to December 2000. At the same time as these counts were recorded, information was also collected on other characteristics of these sheets, including average weight, thickness and chemical composition, with the aim of investigating the impact of these factors on the number of sheets with edge strain defects observed. Unfortunately, no information was collected on sheets with no edge defects, nor was there a count of the number of such sheets. Consequently information on numbers of sheets with edge defects was only available for days on which there was at least one sheet with such defects.

A flexible and the most widely used model for analysing count response data is the Poisson generalized linear model (McCullagh & Nelder, 1989 Chap. 6). In our case however these data are truncated at zero. The distribution of strictly positive Poisson count data is called the zero-truncated Poisson and has a long history, dating back to the papers of David & Johnson (1952) and Plackett (1953). See also Johnson, Kotz & Kemp (1992). Shaw (1988) extends the Poisson generalized linear model (PGLM) of McCullagh & Nelder (1989) to deal with truncated count data. Alternatively, zero truncated count data can be modelled via the negative binomial generalized linear model (NBGLM), see Gurmu (1991) and Grogger & Carson (1991). Gurmu & Trivedi (1992) present tests for overdispersion in the truncated count model. The truncated Poisson generalized linear model (TPGLM) has been applied to adenomatous polyps data by Xie & Aicken (1997). Examples of economic applications of the TPGLM are given in Cameron & Trivedi (1998 Chap. 1).

In many applications, however, the dependence structure is more complex than the independent observations assumed by the TPGLM. In particular, count data often exhibit clustering, with observations within clusters correlated with one another. A typical example of this is where repeated measurements are taken of a single subject and a cluster consists of the set of observations for this subject. Ignoring such clustering effects in the TPGLM leads to overdipersion and can result in biased and inconsistent estimates of the regression coefficients in the model (Grogger & Carson, 1991).

This paper extends the TPGLM to zero-truncated longitudinal and repeated measurements Poisson count data. Two models are developed; in the first a subject random effect is introduced into the linear predictor to account for overdispersion. The second then extends the linear predictor to include an additional correlated random effect that allows for a possible increase in correlation in the second and subsequent observations made on a subject. In the context of the iron sheets example given above, subject corresponds to the day of observation, and the additional random effect allows for within day effects due to repetition of the manufacturing process.

Sections 2 and 3 define the random effects TPGLM and estimation procedures for model parameters respectively. Section 4 then sets out results from a small simulation study that explores the behaviour of these estimators, while Sections 5 contains a description of their application to the MSC data set on edge strain defects for iron sheets. The results of the paper are discussed in Section 6.

2. Model Specification

In what follows i = 1, 2, ..., N denotes a cluster and t = 1,2, ... $, n_i$ denotes an observation within a cluster. Let Y_{it} represent the value that a count response variable Y takes at observation t within cluster i. In our application, i denotes day and t denotes time within day, we will refer to these indices as day and time from now on. We assume that $Y_{it} \overset{\mathrm{d}}{=} \mathrm{Pn}(\eta_{it})$, while the data collection process is such that only strictly positive values of Y_{it} are observed. We consider two different models for λ_{it}. In the first, η_{it} = ln(λ_{it}) is assumed to be a linear function of a vector $\boldsymbol{x}_{it}$ of p covariates as well as a random day effect u_{1i} to account for variation not explained by the values in $\boldsymbol{x}_{it}$. That is,

$$\eta_{it} = \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + u_{1i} \qquad (1)$$

where $\boldsymbol{\beta}$ is an unknown vector of regression coefficients. The random effects u_{1i} are assumed to be realisations of independent N(0, ϕ_1) random variables.

In the second model η_{it} includes an extra random effect u_{2i}, allowing a possible change in variance and pattern of association in the second and following observations within a day. This is consistent with the idea that the any day departures from the model in the first time period (first observation in a day) are likely to be carried over into subsequent time periods(in subsequent observations in a day) and to be augmented further errors. This model is

$$\eta_{it} = \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + u_{1i} + \Delta_t u_{2i} \qquad (2)$$

where Δ_t = I(t >1), $\boldsymbol{u}_i = (u_{1i}, u_{2i}) \overset{\mathrm{d}}{=} \mathrm{N}(\mathbf{0}, \boldsymbol{\Phi})$ and

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1 & \phi_3 \\ \phi_3 & \phi_2 \end{bmatrix}. \qquad (3)$$

The u_{1i} is the day effect at baseline (first observation in a day), and u_{2i} measures the average deviation from that value on the second and following observations. It is important to realise that a negative value for $\hat{u}_{2i}$ implies a larger decline in η_{it}, so that small $\boldsymbol{Y}$ observations are likely (an improvement in the production lines).

Let $\boldsymbol{u}_1 = (u_{11}, u_{12}, ..., u_{1N_1})$, $\boldsymbol{u}_2 = (u_{21}, u_{22}, ..., u_{2N_2})$ and let $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ denote the incidence matrices for the random effects vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ respectively. Let N_1 denote the number of the clusters N_2 the number of the clusters with more than one observation. In general the model for $\boldsymbol{\eta} = (\eta_{it},\ i = 1,2,...,N,\ t = 1,2,...,n_i)$ can be expressed as $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zu}$, where $\boldsymbol{X}$ is a known matrix of regression variables, $\boldsymbol{Z} = \boldsymbol{Z}_1$ and $\boldsymbol{u} = \boldsymbol{u}_1$ under (1) and $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$ and $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2)$ under (2). The random vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}$ are distributed as multivariate normal with zero mean vectors and variance-covariance matrices given by

$$\boldsymbol{A} = \phi_1 \boldsymbol{I}_{N_1} \text{ and } \boldsymbol{A} = \begin{bmatrix} \phi_1 \boldsymbol{I}_{N_1} & \phi_3 J \\ \phi_3 J^{\mathrm{T}} & \phi_2 \boldsymbol{I}_{N_2} \end{bmatrix} \text{ respectively,}$$

where J is given by deleting the columns of the identity matrix $\boldsymbol{I}_{N_1}$ that correspond to clusters with a single observation.

3. Model Estimation

Henderson (1963, 1973a, and 1975) develops best linear unbiased predictors for linear mixed models. These ideas have been extended to obtain maximum likelihood (ML) and restricted or residual maximum likelihood (REML) estimators in Harville (1977), Thompson (1980), Fellner (1986, 1987) and Speed (1991). McGilchrist (1994) extends this approach to generalized linear mixed models. This method has elements in common with Schall (1991), Breslow & Clayton (1993), Wolfinger (1993), Nelder & Lee (1996) and Saei & McGilchrist (1998). Lee & Nelder (2001a, 2001b) further extend the work of Nelder & Lee (1996) to correlated non-normal data. Below we outline the extension of this approach to the random component truncated Poisson model.

Let l_1 be the loglikelihood function of truncated Poison observations conditional on the value of the random component vector $\boldsymbol{u}$ and let l_2 be the logarithm of the probability density function of $\boldsymbol{u}$. For the model (2) the functions l_1 and l_2 are

$$l_1 = \sum_{i=1}^{N} \sum_{t=1}^{n_i} (y_{it}\eta_{it} - \exp(\eta_{it}) - \ln(1 - \exp(-\exp(\eta_{it}))) - \ln(y_{it}!))$$

$$l_2 = -\frac{1}{2}(const. + \ln|\boldsymbol{A}| + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{u})$$

respectively. Penalised likelihood (PL) estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$ are obtained by maximising l = l_1 + l_2 with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ respectively. These estimates are then used as an initial step in finding ML and REML estimates of ϕ_j via Anderson (1973) and Henderson (1973b) algorithm. The iterative procedure used to obtain the ML and REML estimators and their approximate variance-covariance matrices can be specified as follows:

(a) Starting from initial values $\boldsymbol{\beta}_0$, $\boldsymbol{u}_0$ and ϕ_{j0} (hence $\boldsymbol{A}_0$) successive iterations are obtained by finding changes $\Delta\boldsymbol{\beta}$ and $\Delta\boldsymbol{u}$ to the current estimates from the equations

$$\boldsymbol{V}\begin{bmatrix}\Delta\boldsymbol{\beta}\\ \Delta\boldsymbol{u}\end{bmatrix} = \boldsymbol{W}^{\mathrm{T}}\frac{\partial l_1}{\partial\boldsymbol{\eta}_0} - \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{A}_0^{-1}\boldsymbol{u}_0\end{bmatrix} \tag{4}$$

where $\boldsymbol{W}$ = [$\boldsymbol{X}$, $\boldsymbol{Z}$], $\boldsymbol{\eta}_0 = \boldsymbol{X\beta}_0 + \boldsymbol{Zu}_0$ and $\boldsymbol{V} = \boldsymbol{W}^{\mathrm{T}}(-\frac{\partial^2 l_1}{\partial\boldsymbol{\eta}_0\partial\boldsymbol{\eta}_0^{\mathrm{T}}})\boldsymbol{W} + \begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}_0^{-1}\end{bmatrix}$; $\partial l_1/\partial\boldsymbol{\eta}_0$ and $\partial^2 l_1/\partial\boldsymbol{\eta}_0\partial\boldsymbol{\eta}_0^{\mathrm{T}}$ are first and second order derivatives of l_1 with respect to $\boldsymbol{\eta}$ and evaluated at initial value $\boldsymbol{\eta}_0$.

(b) Once iterations of (4) have converged to $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$, let $\tilde{\boldsymbol{\eta}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}$, $\boldsymbol{B} = -\partial^2 l_1/\partial\tilde{\boldsymbol{\eta}}\partial\tilde{\boldsymbol{\eta}}^{\mathrm{T}}$, $\boldsymbol{T}^* = [\boldsymbol{T}_{jk}^*] = [\boldsymbol{A}_0^{-1} + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{BZ}]^{-1}$, a_1 = $\mathrm{tr}(\boldsymbol{B}_1\boldsymbol{T}_{11}^*) + \tilde{\boldsymbol{u}}_1^{\mathrm{T}}\boldsymbol{B}_1\tilde{\boldsymbol{u}}_1$, a_2 = $\mathrm{tr}(\boldsymbol{T}_{22}^*) + \tilde{\boldsymbol{u}}_2^{\mathrm{T}}\tilde{\boldsymbol{u}}_2$, a_3 = $\mathrm{tr}(\boldsymbol{B}_2\boldsymbol{T}_{11}^*) + \tilde{\boldsymbol{u}}_1^{\mathrm{T}}\boldsymbol{B}_2\tilde{\boldsymbol{u}}_1$ and $a_4 = 2(\mathrm{tr}(\boldsymbol{T}_{12}^* J^{\mathrm{T}}) + \tilde{\boldsymbol{u}}_1^{\mathrm{T}} J\tilde{\boldsymbol{u}}_2)$. $\boldsymbol{B}_1$ is a diagonal matrix with entry $(\phi_{10}\phi_{20} - \phi_{30}^2)^2/(\phi_{10}\phi_{20})^2$ if the corresponding cluster has a single observation and 1 otherwise, while $\boldsymbol{B}_2$ is a diagonal matrix with diagonal element zero if the corresponding cluster has a single observation and 1 otherwise. The estimates of ϕ_j are then given by

$$\begin{aligned}
\hat{\phi}_1 &= \frac{a_1 + (\phi_{30}/\phi_{20})^2 a_2 + (N_2\phi_{30}^2/\phi_{20}) - (\phi_{30}/\phi_{20})a_4}{\mathrm{tr}(\boldsymbol{B}_1)}\\
\hat{\phi}_2 &= \frac{a_2 + (\phi_{30}/\phi_{10})^2 a_3 + (N_2\phi_{30}^2/\phi_{10}) - (\phi_{30}/\phi_{10})a_4}{\mathrm{tr}(\boldsymbol{B}_2)}\\
\hat{\phi}_3 &= \frac{\phi_{10}\phi_{20}a_4}{N_2\phi_{30}^2 \text{ - } a_4\phi_{30} + \phi_{20}a_3 + \phi_{10}a_2 - \phi_{10}\phi_{20}N_2}
\end{aligned} \tag{5}$$

(c) The preceding two steps are then repeated, with initial values set to $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{u}}$ and $\hat{\phi}_j$.

At convergence, $\hat{\phi}_j$ is the ML estimate $\hat{\phi}_{j(\text{ML})}$ of ϕ_j. The asymptotic variance-covariance matrix for $\hat{\boldsymbol{\phi}}_{(\text{ML})} = (\hat{\phi}_{1(\text{ML})}, \hat{\phi}_{2(\text{ML})}, \hat{\phi}_{3(\text{ML})})$ is

$$\text{var}(\hat{\boldsymbol{\phi}}_{(\text{ML})}) = 2\begin{bmatrix} r_{111}+r_{211}-2r_{311} & r_{112}+r_{212}-2r_{312} & r_{113}+r_{213}-2r_{313} \\ & r_{122}+r_{222}-2r_{322} & r_{123}+r_{223}-2r_{323} \\ & & r_{133}+r_{233}-2r_{333} \end{bmatrix}^{-1} \tag{6}$$

where

$$r_{1jk} = \text{tr}(\boldsymbol{A}(\frac{\partial \boldsymbol{A}^{-1}}{\partial \phi_j})\boldsymbol{A}(\frac{\partial \boldsymbol{A}^{-1}}{\partial \phi_k}))$$

$$r_{2jk} = \text{tr}(\boldsymbol{T}^*(\frac{\partial \boldsymbol{A}^{-1}}{\partial \phi_j})\boldsymbol{T}^*(\frac{\partial \boldsymbol{A}^{-1}}{\partial \phi_k}))$$

$$r_{3jk} = \text{tr}(\boldsymbol{A}(\frac{\partial \boldsymbol{A}^{-1}}{\partial \phi_j})\boldsymbol{T}^*(\frac{\partial \boldsymbol{A}^{-1}}{\partial \phi_k}))$$

for j, k = 1, 2, 3.

Let $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}$ and $\boldsymbol{V}^{-1} = \begin{bmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} \\ \boldsymbol{T}_{21} & \boldsymbol{T}_{22} \end{bmatrix}$ denote the partitions of the matrix $\boldsymbol{V}$ and its inverse corresponding to the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{u}$. Replacing $\boldsymbol{T}^*$ by $\boldsymbol{T}_{22}$ in (5) and (6) yields the REML estimate $\hat{\phi}_{j(\text{REML})}$ of ϕ_j and the variance-covariance matrix for the REML estimators $\hat{\boldsymbol{\phi}}_{(\text{REML})} = (\hat{\phi}_{1(\text{REML})}, \hat{\phi}_{2(\text{REML})}, \hat{\phi}_{3(\text{REML})})$ respectively.

4. Simulations

A limited simulation study was undertaken to examine the performance of the method. Truncated count observations were generated from the one random component truncated Poisson model, $\eta_{it} = \beta_0 + x_{it}\beta_1 + u_{1i}$ and the two random component truncated Poisson model, $\eta_{it} = \beta_0 + x_{it}\beta_1 + u_{1i} + \Delta_t u_{2i}$. The values of u_{1i} were independently generated from normal distribution with zero mean and variance ϕ_1 in the first model. The values of $\boldsymbol{u}_i = (u_{1i}, u_{2i})$ were independently generated from bivariate normal distribution with zero mean and variance-covariance matrix $\boldsymbol{\Sigma}$ is given by (3). The x_{it}

were randomly assigned to values of zero or one. Observations were generated for times $t = 1, 2, \ldots, 5$ and for subjects $i = 1, 2, \ldots, 30$. The TPGLM was then fitted to these data and estimates of ϕ_1, ϕ_2, ϕ_3, β_0, and β_1 obtained, as well as the mean deviance (MD) and lower and upper confidence intervals for β_0 and β_1 (with $\alpha = 0.05$). This process was then independently replicated 10000 times. Table 1 defines the quantities reported in simulation results set out in tables 2 – 6. Tables 3 – 6 show the results by both ML (in brackets) and REML methods.

Table 1 about here

Table 2 shows the results from fitting the TPGLM via ML to truncated observations generated by one and two random components TPGLMs. Corresponding results from REML and ML (in brackets) fits of a one random component TPGLM are set out in Table 3.

Tables 2 and 3 about here

Tables 4 and 5 explores the impact of model misspecification, showing the results from REML and ML fits of a one and a two random components TPGLM to data generated by a two correlated random components model. In comparison, Table 6 shows what happens when the correct two component model is fitted to the two component data.

Tables 4 – 6 about here

The results set out in Table 2 show that the ML estimators of β_0 and β_1 under the standard TPGLM are seriously biased when applied to observations generated under random components TPGLM. This bias increases from one component random effect to the two correlated random components. It is also increases with the size of the variance component in the random components TPGLM. The average mean deviances (amd) are very far from 1 in all cases and the average 95% confidence limits (alc, auc) exclude the true values of the parameters β_0 and β_1. The average estimated standard errors (av.se) are also much smaller than their corresponding simulation standard errors (sd.est) of the estimates of these parameters.

The results in Tables 3 and 6 show that the REML estimates of the variance components and the regression slope are biased downwards. This bias is even more pronounced for the ML estimates of the variance components. Increasing the size of the variance components increases the biases of the estimators of ϕ_1, ϕ_2 and ϕ_3, but not that of the regression parameters. The simulation standard errors sd.est of the parameter estimators agrees with their average estimated standard errors av.se under both ML and REML. This is also true for the $\phi_{1,}$ ϕ_2 and ϕ_3. Although not shown here, increasing the number of observations for each subject from 5 to 15 significantly decreased the biases of the parameter estimators and their estimated variances. We also see that in all cases the average 95% confidence limits (alc, auc) contain the true values of β_0, β_1, $\phi_{1,}$ ϕ_2 and ϕ_3.

Finally, in Tables 4 and 5 we show that the effect of model misspecification when fitting a random components TPGLM. In these cases we fitted a one random component TPGLM and a two independent random components to data generated from a two correlated random components TPGLM. This led to positively biased variance component estimators and increased bias for the estimators of β_0 and β_1 for the one random component model. These biases are small when we ignore the covariance between the two random components. The "average" confidence interval does not include the true value for ϕ_1 in Table 4.

5. Application to Iron Sheet Data

In this section we illustrate the preceding theory by applying it to data obtained in an investigation of the impact of different factors on edge strain defects in iron sheet production at the Mobarekeh Steel Complex (MSC). Information on edge strain defects was recorded over a period of 122 days from March 2000 to December 2000, with the number of coils with edge strain defects recorded at equal time intervals on each day. No data were collected on coils with no edge strain defects. The level of operation of tandem-mill and skin-pass processes that led to the defective sheets (A, B, C and D) was also recorded, as were the average sheet width and thickness for the sheets with edge strain defects and their average carbon (C), selenium (Se), magnesium (Mg), aluminium (Al), nitrogen (N), tension strength, elongation and hardness measurements and also coil weight. Figure 1(a) shows the number of observations over the 122 days. Note the decreasing trend in both average and variability evident in the plot. Unfortunately, there is no information to link this behaviour to increasing numbers of zero observations. If that was the case then a more complex model that included a time series structure for the random effects would have been appropriate. However, in the absence of any information in this regard, we decided to persist with the simple two component random effects model for these data. Figure 1(b) shows number of coils with edge strain defect.

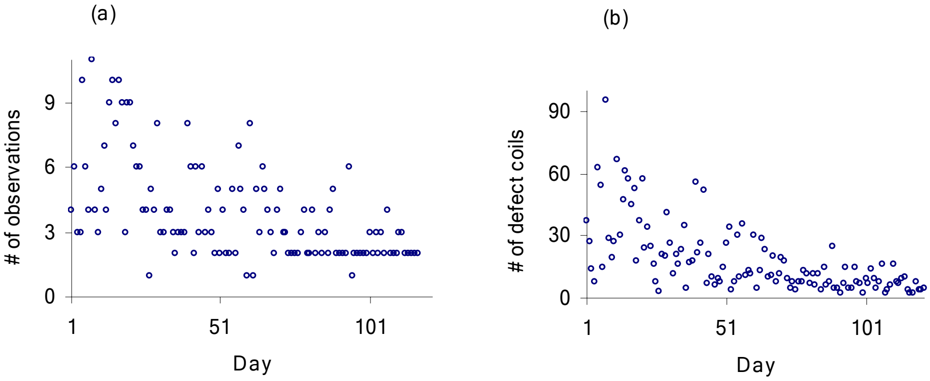

Figure 1. (a) Number of observations on each 122 days; (b) number of coils with edge strain defect over 122 days.

Table 7 sets out the parameter estimates and associated standard errors for four zero-truncated Poisson GLMs fitted to these data, with the response variable in each case corresponding to the number of sheets with edge strain defects observed on each day. Model 1 is the standard TPGLM, fitted via ML. Model 2 is a one random effect TPGLM, fitted via REML, with the random effect corresponding to day of observation. Model 3 is a two independent random effects TPGLM, with an extra random effect introduced to allow for within day heterogeneity. Finally, Model 4 relaxes the independent assumption in model 3 and includes an extra parameter to account for dependence between two random effects. Models 3 and 4 are also fitted using REML.

Examination of the results for Model 4 in Table 7 show that there is statistically significant variation between days ($\hat{\phi}_1$ = 0.34 with an estimated standard error of 0.07). This day effect varies significantly within a day ($\hat{\phi}_2$ = 0.46 with an estimated standard error of 0.11). Furthermore the two random effects are correlated with estimated covariance of $\hat{\phi}_3$ = -0.26 and estimated standard error of 0.08. The estimated correlation between two random effects is $\hat{\rho}$ = -0.66. The predicted values of the second random component ($\hat{u}_{2i}$) are used to identify days that show greater declines, and 13 days are identified for further study. Under Models 2, 3 and 4, neither sheet thickness nor sheet width effects are statistically significant. In contrast, under Model 1 (TPGLM) both sheet width and sheet thickness are highly significant. There are also significant tandem-mill and skin-pass process effects, with Wald statistics ($\hat{\boldsymbol{\beta}}^{\mathrm{T}}[\mathrm{var}(\hat{\boldsymbol{\beta}})]^{-1}\hat{\boldsymbol{\beta}}$) of 9.96 and 23.2 respectively. However, with the exception of magnesium and selenium, the conclusions are the same for the remaining covariates under all four models. The selenium does not have significant effect under both models 3 and 4 whereas magnesium does have a significant effect under only model 4. In contrast, under both Model 1 (TPGLM) and Model 2 (single random component TPGLM) selenium and magnesium (marginally for model 2) are significant. The aluminium, elongation and tension strength are other significant covariates. This conclusion is supported by all four models.

Table 7 about here

6. Summary and Discussion

In this paper we introduce a simple method for analysing longitudinal or repeated measurement count response data that are truncated at zero. Two types of models, one involving a single random component to account for between subject heterogeneity, and a second involving two random components, with the second component used to account for within subject heterogeneity, have been investigated. These random components are allowed to be correlated. When applied to zero truncated data on counts of coils with edge defects in iron sheets produced at the Mobarekeh Steel Complex, the two random components model indicates the existence of statistically significant day to day and within day heterogeneity in these data. The two random components are also significantly correlated. Furthermore, misleading inferences are obtained if the within day heterogeneity and the dependence between random components are ignored in model specification.

Results from a small scale simulation (Tables 2 - 6) support this conclusion, in the sense that parameter estimates are biased if no account is taken of extra-Poisson variability when modelling data that includes such effects. These results also show the REML method leads to regression parameter estimates that are both unbiased and efficient. However, the REML estimators of the variance components in the model tend to be negatively biased. Increasing number of observation per cluster and clusters decreases the bias and improves the asymptotic performance of the REML estimators. Our inference is based on the asymptotic properties of the REML estimators. In general, the development of asymptotic properties is difficult for models with random effects. This difficulty is not specific to generalized linear mixed models (the approach taken in this paper). Jiang (1996) has investigated the asymptotic behaviour of REML estimators for linear mixed models. Breslow and Clayton (1993) have showed that for Poisson data the accuracy of these estimators improves as the mean increases. Lin (1997) has pointed out that REML estimator of variance components are not normally distributed unless the number of clusters is large and the estimators bounded away from zero. Using a Laplace approximation, Lin & Breslow (1996) show that the REML estimators are biased for large variance components and introduced a bias correction to improve the asymptotic performance of the REML estimators.

Turning now to ML estimation, we note that in theory ML estimates can be calculated via numerical integration using Gauss-Hermite quadrature. However, this is an impractical method for the models with high dimension random effects. More recently McCulloch (1997) and Booth & Hobert (1999) have explored the use of Markov Chain Monte Carlo (MCMC) and Monte Carlo EM in this context. These are computer intensive approaches and their application to the complex models is very difficult (if not impossible). They also require fairly sophisticated computer programming since there is no generally available software.

Finally, we note that the translated Poisson distribution, defined by $\Pr(Y_{it} = y_{it})$ = $\exp(-\lambda_{it})\lambda_{it}^{y_{it}-1}/(y_{it} - 1)!$ is an alternative to the truncated Poisson distribution for positive response values. The application of this model via the specification $\ln(\mu_{it} - 1)$ = $\eta_{it} = \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}$, i.e., a fixed effect model, to the iron sheet data yielded a mean deviance of 3.1 whereas the TPGLM without day random effect (TPGLM fixed effect model) had a mean deviance of 2.994. Clearly, further research needs to be done to explore alternative models for positive valued data.

Acknowledgments

The first author's research was partially supported by research council of Isfahan University of Technology. The authors thank the associate Editor and a referee for their helpful comments.

References

Anderson, T.W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1**, 135 - 141.

Booth, J.G. & Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. Roy. Statist. Soc. Ser. B* **61**, 265 - 285.

Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalised linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9 - 25.

Cameron, A.C. & Trivedi, P. (1998). *Regression analysis of count data*. Cambridge University Press, Cambridge.

David, F.N. & Johnson, N.L. (1952). The truncated Poisson. *Biometrics* **8**, 275 - 285.

Gurmu, S. (1991) Tests for detecting overdispersion in the positive Poisson regression model. *J. Bus. Econom. Statist.* **9**, 215 - 222.

Gurmu, S. & Trivedi (1992). Overdispersion tests for truncated Poisson regression models. *J. Econometrics* **54**, 347 - 370.

Grogger, J.T. & Carson, R.T. (1991). Models for truncated counts. *J. Appl. Econometrics* **6**, 225 - 238.

Henderson, C.R. (1963). Selection index and expected genetic advance. *In Statistical Genetics and Plant Breeding* (W.D. Hanson and H.F. Robinson, eds.), 141 - 163. National Academy of Sciences and National Research Council Publication No. 982, Washington, D.C.

Henderson, C.R. (1973a). Sire evaluation and genetic trends. *In Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. Jay L. Lush* 10-41. Amer. Soc. Animal Sci.-Amer. Dairy Sci. Assn.-Poultry Sci. Assn. Champaigne, Illinois.

Henderson, C.R. (1973b). Maximum likelihood estimation of variance components. unpublished manuscript.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under selection model. *Biometrics* **31**, 423 - 447.

Jiang, J. (1996). REML estimation: asymptotic behaviour and related topic. *Ann. Statist.* **24**, 255 - 286.

Johnson, N.L., Kotz, S. & Kemp, A.W. (1992). *Univariate discrete distributions*. Second Ed. John Wiley & Sons, Inc.

Lee, Y. & Nelder, J.A. (2001a). Hierarchical generalized linear models: A synthesis of generalised linear models, random-effect model and structured dispersions. *Biometrika* **88**, 987 - 1006.

Lee, Y. & Nelder, J.A. (2001b). Modelling and analysing correlated non-normal data. *Statistical Modelling* **1**, 3 - 16

Lin, X. & Breslow, N.E. (1996). Bias correction in generalized linear mixed model with multiple components of dispersion. *J. Amer. Statist. Assoc.* **91**, 1007 - 1016.

McCullagh, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B*, **42**, 109 - 142.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. Second Ed. London: Chapman and Hall.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92**, 162 - 170.

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *J. Roy. Statist. Soc. Ser. B* **56**, 61 - 69.

Nelder, J.A. & Lee, Y. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58**, 619-678.

Plackett, R.L. (1953). The truncated Poisson distributions. *Biometrics* **9**, 485 - 488.

Saei, A. & McGilchrist, C. (1998). Longitudinal threshold models with random components. *The Statistician* (JRSS Series D) **47**, 365 - 375.

Schall, R (1991). Estimation in generalised linear models with random effects. *Biometrika* **78**, 719 - 727.

Shaw, D. (1988). One-site samples’ regression problems of non-negative integers, truncation, and endogenous stratification. *J. Econometrics* **37**, 211 - 223.

Speed, T. (1991). Comment on Robinson: Estimation of random effect. *Statistical Science* **6**, 42 - 44.

Wolfinger, R. (1993). Laplace’s approximation for nonlinear mixed models. *Biometrika* **80**, 791 - 796.

Xie, T. & Aicken, M. (1997). A truncated Poisson regression model with applications to occurrence of adenomatous polyps. *Statistics in Medicine* **16**, 1845 - 1857

Table 1 Definitions of quantities reported in simulation results set out in tables 2 – 6.

Quantity	Description
tv	true parameter values
ae	average error
sd.est	actual standard error over 10000 simulations
av.se	average estimated standard error
alc	average lower 95% confidence limit
auc	average upper 95% confidence limit
amd	average mean deviance

Table 2 Simulation results for the ML fit of the TPGLM to data generated from one and two random component truncated Poisson models; see Table 1 for definitions.

	1 Component Model			2 Components Model									
Quantity	ϕ_1	β_0	β_1	ϕ_1	ϕ_2	ϕ_3	β_0	β_1	ϕ_1	ϕ_2	ϕ_3	β_0	β_1
				Parameter set 1					Parameter set 2				
tv	1.00	2.00	0.50	1.00	1.00	0.00	2.00	0.50	1.00	1.00	0.50	2.00	0.50
ae		0.45	-0.03				0.65	-0.13				0.74	-0.19
sd.est		0.30	0.41				0.35	0.46				0.37	0.51
av.se		0.03	0.05				0.03	0.04				0.03	0.04
alc		2.37	0.38				2.59	0.28				2.67	0.23
auc		2.51	0.56				2.71	0.45				2.80	0.39
amd		13.45				20.80					24.50		

Table 3 Simulation results for the REML (ML) fit to data generated from a one random component truncated Poisson model; see Table 1 for definitions.

Quantity	ϕ_1	β_0	β_1	ϕ_1	β_0	β_1
	Parameter set 1			Parameter set 2		
tv	1.00	2.00	0.50	2.00	2.00	0.50
ae	-0.03 (-0.10)	0.02 (0.02)	-0.01 (-0.01)	-0.11 (-0.25)	0.04 (0.04)	-0.02 (0.03)
sd.est	0.27 (0.25)	0.26 (0.26)	0.37 (0.37)	0.53 (0.49)	0.37 (0.37)	0.52 (0.52)
av.se	0.27 (0.24)	0.26 (0.25)	0.37 (0.36)	0.52 (0.47)	0.36 (0.35)	0.51 (0.50)
alc	0.44 (0.43)	1.51 (1.53)	-0.23 (-0.21)	0.87 (0.84)	1.33 (1.36)	-0.53 (-0.50)
auc	1.50 (1.37)	2.53 (2.52)	1.21 (1.21)	2.91 (2.91)	2.75(2.73)	1.50 (1.45)

Table 4 Simulation results for REML (ML) fit of a one random component truncated Poisson model to data generated from a two correlated random components truncated Poisson model; see Table 1 for definitions.

Quantity	ϕ_1	ϕ_2	ϕ_3	β_0	β_1
tv	1.00	1.00	0.50	2.00	0.50
ae	0.97 (1.03)			0.10 (0.11)	-0.02 (-0.02)
sd.est	0.57 (0.55)			0.40 (0.39)	0.56 (0.56)
av.se	0.58 (0.54)			0.39 (0.38)	0.55 (0.53)
alc	0.83 (0.98)			1.34 (1.37)	-0.60 (-0.56)
auc	3.11 (3.10)			2.87 (2.84)	1.57 (1.52)

Table 5 Simulation results for REML (ML) fit of a two independent random components truncated Poisson model to data generated from a two correlated random components truncated Poisson model; see Table 1 for definitions.

Quantit y	ϕ_1	ϕ_2	ϕ_3	β_0	β_1
tv	1.00	1.00	0.50	2.00	0.50
ae	0.03 (-0.05)	-0.04 (-0.04)		0.02 (0.03)	-0.02 (-0.02)
sd.est	0.32 (0.29)	0.29 (0.29)		0.29 (0.29)	0.40 (0.40)
av.se	0.31 (0.28)	0.29 (0.29)		0.28 (0.27)	0.40 (0.38)
alc	0.43 (0.41)	0.38 (0.39)		1.46 (1.49)	-0.30 (-0.27)
auc	1.64 (1.49)	1.53 (1.53)		2.57 (2.56)	1.26 (1.23)

Table 6 Simulation results for REML (ML) fit to data generated from a two correlated random components truncated Poisson model; see Table 1 for definitions.

Quantity	ϕ_1	ϕ_2	ϕ_3	β_0	β_1
tv	1.00	1.00	0.50	2.00	0.50
ae	-0.07 (-0.11)	-0.04 (-0.04)	-0.03 (0.03)	0.05 (0.06)	-0.02 (-0.02)
sd.est	0.22 (0.21)	0.23 (0.23)	0.16 (0.16)	0.21 (0.21)	0.28 (0.28)
av.se	0.21 (0.20)	0.22 (0.22)	0.16 (16)	0.20 (0.19)	0.27 (0.27)
alc	0.51 (0.5)	0.53 (0.53)	0.15 (0.16)	1.66 (1.68)	-0.05 (-0.02)
auc	1.35 (1.29)	1.40 (1.40)	0.79 (0.78)	2.44 (2.44)	1.02 (1.00)

Table 7 Parameter estimates (est), standard errors (se) and *t* (*t* = est/se) for four models fitted to the MSC iron sheet data.

Effect	Model 1			Model 2			Model 3			Model 4		
	est	se	*t*	est	se	*t*	est	se	*t*	est	se	*t*
ϕ_1				0.20	0.04	5.36	0.19	0.05	4.08	0.34	0.07	4.57
ϕ_2							0.24	0.07	3.69	0.46	0.11	4.13
ϕ_3										-0.26	0.08	-3.26
Intercept	10.0	4.38	2.29	7.33	5.06	1.45	6.08	5.32	1.14	7.38	5.54	1.33
Thickness	-0.48	0.20	-2.42	-0.34	0.26	-1.32	-0.31	0.28	-1.13	-0.44	0.28	-1.54
Width	0.38	0.12	3.13	-0.06	0.15	-0.37	-0.05	0.17	-0.28	-0.03	0.17	-0.19
Tandem-mill A	0.08	0.07	1.18	0.11	0.07	1.44	0.14	0.08	1.72	0.06	0.09	0.72
Tandem-mill B	0.18	0.06	2.88	0.18	0.07	2.58	0.21	0.07	2.82	0.21	0.08	2.74
Tandem-mill C	0.09	0.07	1.36	0.06	0.07	0.81	0.05	0.08	0.61	0.04	0.08	0.47
Skin-pass A	-0.04	0.06	-0.69	0.02	0.07	0.36	0.08	0.07	1.18	0.07	0.07	0.99
Skin-pass B	-0.08	0.06	-1.37	-0.10	0.07	-1.43	-0.07	0.07	-1.03	-0.09	0.07	-1.25
Skin-pass C	-0.24	0.07	-3.65	-0.35	0.07	-4.67	-0.30	0.08	-3.82	-0.31	0.08	-3.88
Carbon	0.00	0.14	0.03	0.19	0.16	1.22	0.07	0.17	0.41	0.02	0.17	0.10
Selenium	0.21	0.07	3.06	0.18	0.08	2.24	0.10	0.08	1.14	0.05	0.09	0.60
Magnesium	-1.42	0.36	-4.00	-0.79	0.41	-1.91	-0.80	0.43	-1.85	-1.02	0.45	-2.26
Aluminium	0.44	0.15	2.89	0.43	0.17	2.49	0.47	0.18	2.52	0.51	0.19	2.66
Nitrogen	0.00	0.09	-0.04	0.13	0.11	1.21	0.17	0.12	1.44	0.15	0.12	1.20
Tension strength	-1.83	0.68	-2.68	-2.05	0.83	-2.46	-1.87	0.87	-2.14	-1.79	0.91	-1.98
Elongation	1.26	0.30	4.18	1.51	0.37	4.11	1.47	0.40	3.72	1.52	0.41	3.74
Hardness	0.00	0.28	-0.01	0.29	0.40	0.74	0.53	0.43	1.24	0.42	0.43	0.98
Weight	0.08	0.04	1.80	0.10	0.06	1.84	0.08	0.06	1.40	0.07	0.06	1.11

- Tandem-mill D and Skin-pass D are fixed at zero for identifiability
- Model 1 = TPGLM
- Model 2 = A one random component truncated Poisson model
- Model 3 = A two independent random components truncated Poisson model
- Model 4 = A two random components truncated Poisson model