Communications of the Korean Statistical Society
2012, Vol. 19, No. 6, 761–770

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.6.761

Modelling Count Responses with Overdispersion

Kwang Mo Jeong[1,a]

[a]Department of Statistics, Pusan National University

Abstract

We frequently encounter outcomes of count that have extra variation. This paper considers several alternative models for overdispersed count responses such as a quasi-Poisson model, zero-inflated Poisson model and a negative binomial model with a special focus on a generalized linear mixed model. We also explain various goodness-of-fit criteria by discussing their appropriateness of applicability and cautions on misuses according to the patterns of response categories. The overdispersion models for counts data have been explained through two examples with different response patterns.

Keywords: Clustered data, overdispersion, quasi-likelihood, dispersion parameter, zero-inflated Poisson, negative binomial, generalized linear mixed model.

1. Introduction

Poisson distribution is the standard model for count responses, for which the variance of responses is expected to equal mean. But violations in the assumptions characterizing a Poisson distribution may lead to the presence of extra variation in the analysis of counts and rates in longitudinal studies. There may exist a dependence between the elemental units in a study where experimental units are clusters. When responses are observed in clusters they frequently exhibit extra variation than the permitted variance of the assumed model. We often encounter this extra variation in a real data of count or binomial responses. Overdispersion may sometimes be observed when the residual variation obtained is greater than which can be attributed to the sampling variation assumed by the model.

McCullagh and Nelder (1989) showed that overdispersion is not uncommon in practice. Overdispersion should be considered deliberately in modelling count responses. The ordinary Poisson general linear model(GLM) cannot be fitted well in the presence of overdispersion. In the analysis of clustered or longitudinal data the generalized linear mixed model(GLMM) and the generalized estimating equations(GEE) approaches are most popular. The GLMM incorporates random subject effects into a GLM by allowing subjects variability; however, the GEE method solves equations that include the correlation structure of repeated responses to estimate regression coefficients.

Thall and Vail (1990) discussed covariance models for longitudinal count data with overdispersion, and Jowaheer and Sutradhar (2002) applied a GEE method to analyze longitudinal count data; however, Sutradhar *et al.* (2007) suggested three kinds of mixture models to account for the overdispersion in binomial data. The beta-binomial, the finite mixture, and the zero-inflated binomial model all belong to the same class of mixture models. They suggested a chi-squared type goodness-of-fit statistic to test the assumed null model against the alternative three kinds of models without considering covariate variables. Recently, Morel and Neerchal (2012) have extensively studied various overdispersion models using SAS.

This work was supported for two years by Pusan National University Research Grant.

[1] Professor, Department of Statistics, Pusan National University, Jangjeon-Dong, Kumjung-Gu, Pusan 609-735, Korea.
E-mail: kmjung@pusan.ac.kr

It is necessary to assess the goodness-of-fit of fitted model. There are many criteria such as the chi-squared statistic based on Pearson residuals, and the deviance using the likelihood based statistics. Furthermore, Pan and Lin (2005) suggested test statistics using cumulative sums of residuals in GLMM. Recently, Xu and Lu (2009) suggested a nonparametric Monte Carlo test based on the cumulative sums of residuals for the longitudinal count data with overdispersion.

This paper considers several overdispersion models for count data in the framework of mixture models; the quasi-Poisson model, the zero-inflated Poisson(ZIP) model, the Poisson-normal GLMM, and the negative binomial(NB) model. The NB model and the Poisson-normal GLMM belong to the GLMM family in the sense that Poisson responses are mixed with random effects distribution. In Section 2, we review the quasi-likelihood method for count responses by allowing variance function to include a dispersion parameter. The commonly used goodness-of-fit criteria such as chi-squared statistic, deviance statistic, and Akaike information criterion(AIC) have been discussed in the respect of appropriateness or misuses according to the patterns of response categories. As a motivation for overdispersion, we introduce a real data which is not well fitted by the ordinary Poisson GLM. Section 3 discusses several overdispersion models in the context of mixture models. The NB model and the Poisson-normal GLMM are the two competing models for the overdispersion of count data. The NB model also belongs to a class of GLMM but with nonnormal random effects in contrast to the Poisson-normal GLMM which has normal random effects. In Section 4 the overdispersion models will be compared in detail through two examples of different response patterns, the one is contingency table data, and the other is ungrouped count responses. In particular, we compare the Poisson-normal GLMM with the NB model to emphasize the normality assumption of random effects. Finally we summarize the paper with comments on future research areas.

2. Quasi-Poisson Model for Counts

2.1. Quasi-likelihood function

Let Y denote count response and $x_1, x_2, \ldots, x_p$ be p covariate variables. Given a data $(y_i, \boldsymbol{x}_i)$ with $\boldsymbol{x}_i = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{ip})'$, where $x_{i0} = 1$, and $i = 1, 2, \ldots, n$, the Poisson GLM for count responses is represented in terms of the log link and the mean $\mu_i = E(Y|\boldsymbol{x}_i)$ as

$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$. We simply let $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ be the linear predictor of (2.1). Before we introduce the quasi-likelihood method that allows for the extra variation we consider an exponential density for Y

$$f(y, \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right].$$

The relationships $\mu = E(Y) = b'(\theta)$, $\mathrm{Var}(Y) = a(\phi)b''(\theta)$ hold in general. Thus $b(\theta)$ determines the first and second moments of Y. The log-likelihood function can be written as

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi). \tag{2.2}$$

The $L(\boldsymbol{\beta})$ depends on $\boldsymbol{\beta}$ through the assumed model (2.1). The maximum likelihood estimator (MLE)

of $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}$, can be solved from the score equations

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1, \ldots, p. \tag{2.3}$$

We note that the score equations depend on the distribution of Y_i only through μ_i and $\mathrm{Var}(Y_i)$. The quasi-likelihood method simply uses the relationship of a mean and variance function instead of a Poisson assumption of Y_i. The variance itself also depends on the mean through a particular functional relationship $\mathrm{Var}(Y_i) = v(\mu_i)$, where $v(\mu_i)$ is a function of μ_i. In particular, for the Poisson GLM $v(\mu_i) = \mu_i$. However, in the quasi-Poisson model the $\mathrm{Var}(Y_i)$ replaced by a certain variance function $v(\mu_i) = \phi\mu_i$ multiplied by a dispersion parameter to account for the extra variation of count responses. When $\phi > 1$ it means overdispersion. The quasi-likelihood MLE $\hat{\beta}_i$ from the quasi-score equations of (2.3) with $\mathrm{Var}(Y_i) = \phi\mu_i$ coincides with that of ordinary Poisson GLM but has a larger variance multiplied by $\hat{\phi}$. We may refer to Wedderburn (1974) for detailed discussions on the quasi-likelihood method.

2.2. Goodness-of-fit criteria

By substituting the ML estimator $\hat{\boldsymbol{\beta}}$ into (2.1) we obtain the estimated mean $\hat{\mu}_i = \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$. To assess the goodness-of-fit(GOF) of a fitted model there are several statistics such as a Pearson chi-squared like statistic, deviance, and AIC measures. Firstly, the Pearson chi-squared statistic is the sum of squares of Pearson residuals $(y_i - \hat{\mu}_i)/\sqrt{\widehat{\mathrm{Var}}(Y_i)}$ given by

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{\mathrm{Var}}(Y_i)}. \tag{2.4}$$

The X^2 of the quasi-Poisson model is given by multiplying $1/\hat{\phi}$ to the chi-squared statistic of ordinary Poisson GLM because $\widehat{\mathrm{Var}}(Y_i) = \hat{\phi}\hat{\mu}_i$ holds for the quasi-likelihood method. The X^2 has an approximate chi-squared distribution for large $\{\hat{\mu}_i\}$. But we should be cautious on using the X^2 statistic when $\hat{\mu}_i$'s are small compared to 5, as discussed by Wood (2002). The deviance statistic, denoted by $D(\boldsymbol{y};\hat{\boldsymbol{\mu}})$, is formally defined as the difference of the log-likelihood between the saturated model and the fitted model. The deviance statistic is

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right\}. \tag{2.5}$$

Deviance $D(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ has also an approximate chi-squared distribution for large $\{\hat{\mu}_i\}$. The AIC measure is defined to be $-2L(\hat{\boldsymbol{\beta}}) + 2r$, where r is the number of estimated parameters.

Before we discuss other overdispersion models we briefly explain the lack-of-fit of Poisson GLM through an example. Table 1 shows responses of 1308 subjects by race about the number of homicide victims they have known within the past 12 months. The data comes from Agresti (2002), and originally given in a General Social Survey of 1990 by National Opinion Research Center. The sample mean of 159 blacks is 0.522 with a variance of 1.150, and the sample mean of 1149 whites is 0.092 with a variance of 0.155. The sample variances are larger than the sample means and denote extra variations than that expected under Poisson distribution. We consider a Poisson GLM

$$\log(\mu_i) = \beta_0 + \beta_1 x_i,$$

Table 1: Number of victims known in past year, by race, with fit of Poisson GLM

Response	Data		Poisson GLM	
	Black	White	Black	White
0	119	1070	94.3	1047.7
1	16	60	49.2	96.7
2	12	14	12.9	4.5
3	7	4	2.2	0.1
4	3	0	0.3	0
5	2	0	0	0
6	0	1	0	0
Total	159	1149	159	1149

where x_i denotes race for subject i taking values 1 or 0 according to black or white. The MLEs of Poisson GLM are $\hat{\beta}_0 = -2.38$ (SE = 0.097), $\hat{\beta}_1 = 1.73$ (SE = 0.147). However, $\hat{\phi} = 1.75$ for the quasi-Poisson model and denotes overdispersion. The regression coefficients are the same for the quasi-Poisson model but their standard errors are multiplied by $\sqrt{\hat{\phi}} = \sqrt{1.75}$ to those of ordinary Poisson GLM. The expected frequencies of Poisson GLM greatly deviate from the observed frequencies, in particular at response count 0 and 1. The given data has excessively many 0's, and this type of excess of 0 counts can be improved by fitting the ZIP model which will be explained in later sections. The estimated means is $\hat{\mu}_i = \exp(-2.38+1.73) = 0.522$ for blacks and $\exp(-2.38) = 0.092$ for whites. The $X^2 = 2279.9$ with degrees of freedom 1306 for the assumed Poisson GLM. The GOF of quasi-Poisson has been improved to be $X^2/\hat{\phi} = 1302.8$ with the same degrees of freedom. However, we need to be cautious on using this goodness-of-fit measure since the approximate chi-squared distribution is not guaranteed for small estimated means such as 0.522 and 0.092.

3. Mixture Models for Overdispersed Counts

3.1. Zero-inflated Poisson model

When there occurs excess of zeroes it can be modelled by the ZIP distribution. The count response Y has a probability distribution mixed with degenerate distribution at zero and the Poisson distribution with mean μ; therefore, the probability density of Y can be written as

$$f(y) = \omega I(y = 0) + (1 - \omega)\frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots, \ 0 < \omega < 1,$$

where $I(y = 0)$ is an indicator and ω is a weight at $y = 0$.

We can easily compute the mean and variance of ZIP distribution as

$$E(Y) = (1 - \omega)\mu, \qquad \mathrm{Var}(Y) = (1 - \omega)\left(\mu + \omega\mu^2\right).$$

Therefore $\mathrm{Var}(Y) > E(Y)$ and means that the ZIP distribution is overdispersed in the sense that the variance is larger than its mean. But the variance is smaller than the usual Poisson distribution having variance μ. If ω approximates to 0 then the ZIP is close to the usual Poisson model. We comment that the ω can be estimated using *gamlss* library in R software by additional modeling of ω using logit link. The ZIP has been introduced as one candidate model for overdispersed counts data with an extraordinarily large proportion of zeros; however, we should be cautious that it cannot be used for the overdispersed count outcomes that do not have an excess of zeros, for which the ZIP provides similar results with the ordinary Poisson GLM.

3.2. Poisson-normal GLMM

In this section y_{ij} be responses within subject i, where $j = 1, 2, \ldots, t_i$ and $i = 1, 2, \ldots, n$. When cluster sizes all equal one, that is, $t_i \equiv 1$ for all clusters, the y_{ij} coincides with y_i defined before. Hence to be consistent with notations we use y_i instead of y_{ij}. Given a subject effect $\boldsymbol{u}_i$ we assume that count responses are Poisson with mean $\mu_i = E(y_i|\boldsymbol{u}_i)$. Furthermore the subject effects $\boldsymbol{u}_i$ is usually regarded as independent random variables compared to the fixed coefficient $\boldsymbol{\beta}$. The GLMM for count responses is defined as

$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u}_i, \tag{3.1}$$

where $\boldsymbol{z}_i$ is a q dimensional covariate vector related with the random effects $\boldsymbol{u}_i$. To simplify the problem we only consider the random intercept model by taking $\boldsymbol{z}_i = 1$, that is, the Poisson-normal GLMM is the form

$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \tag{3.2}$$

where u_i is assumed to be $N(0, \sigma^2)$. Under the Poisson-normal GLMM the likelihood function $L(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma)'$ is obtained by integrating out u_i with respect to the density of random effects.

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int f(y_i|u_i) f(u_i)\, du_i, \tag{3.3}$$

where $f(y_i|u_i)$ is the Poisson density with mean μ_i characterized by (3.2), and $f(u_i)$ is a normal density with mean 0 and variance σ^2. Numerical approximation of (3.3) and then maximization steps are needed to find the MLEs. The adaptive Gaussian quadrature or Laplacian method is commonly applied to approximate the marginal integral numerically. The random effects u_i are usually predicted by the empirical Bayes method. Under some regularity conditions, the MLE $\hat{\boldsymbol{\theta}}$ satisfies $\sqrt{n}$-consistency and the asymptotic normality. We may refer to Breslow and Clayton (1993) for details of inference under GLMM.

Now we are to discuss the Poisson-normal GLMM in the respect of overdispersion. Firstly, the marginal mean of Y_i can be written as

$$E(Y_i) = \int \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i\right) f(u_i)\, du_i. \tag{3.4}$$

From the moment generating function of normal density we obtain $E(Y_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma^2/2)$. Next, to derive the marginal variance of Y_i we use the relationship

$$\mathrm{Var}(Y_i) = E[\mathrm{Var}(Y_i|u_i)] + \mathrm{Var}[E(Y_i|u_i)]. \tag{3.5}$$

Since the conditional distribution of Y_i given u_i has been assumed to be Poisson with mean μ_i given by (3.2), therefore $E(Y_i|u_i) = \mathrm{Var}(Y_i|u_i) = e^{\boldsymbol{x}_i'\boldsymbol{\beta}} e^{u_i}$. Similarly to finding marginal mean we finally obtain the relationship

$$\mathrm{Var}(Y_i) = E(Y_i) + \{E(Y_i)\}^2 \left(e^{\sigma^2} - 1\right). \tag{3.6}$$

Clearly, $\mathrm{Var}(Y_i) > E(Y_i)$, and σ denotes the amount of overdispersion in GLMM.

3.3. Negative binomial model

The most popular model to account for overdispersion on count responses is negative binomial distribution. The NB distribution is usually interpreted as the distribution of the number of failures before a pre-determined number of successes occur in a sequence of Bernoulli trials. However, in the context of overdispersion, the NB distribution can be applied to account for extra variation of count responses. To formulation the NB distribution as a mixture model we assume that the conditional distribution of count responses is Poisson. Further, the mean rates vary according to a gamma distribution. It turns out that the unconditional distribution of counts is NB distribution. Formally we let Y_i be Poisson with mean $\mu_i\gamma_i$ given γ_i, and further assume that the γ_i follows gamma distribution with shape κ and scale $1/\kappa$ for $\kappa > 0$. As $1/\kappa \to 0$ the gamma distribution has $\mathrm{Var}(\gamma_i) = 1/\kappa \to 0$, and it converges to a degenerate distribution at μ_i. In this case the NB distribution converges to the Poisson distribution with mean μ_i. The marginal density of Y_i can be derived as

$$f(y) = \frac{\Gamma(y+\kappa)}{\Gamma(\kappa)\Gamma(y+1)}\left(\frac{\kappa}{\mu_i+\kappa}\right)^{\kappa}\left(1-\frac{\kappa}{\mu_i+\kappa}\right)^{y}, \quad y = 0, 1, 2, \ldots.$$

Since the gamma distribution is a conjugate family of Poisson distribution the marginal mean and variance of NB distribution can be easily obtained by using the variance relationship (3.5). Thus we find $E(Y_i) = \mu_i$ and

$$\mathrm{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\kappa}. \tag{3.7}$$

The greater $1/\kappa$, the greater the overdispersion compared to the Poisson distribution. The index $1/\kappa$ is called the dispersion parameter. From (3.6) and (3.7) we see the same form for the variances of Poisson-normal GLMM and NB model.

We next discuss the NB model in the context of GLMM. The Poisson-gamma mixture density can be written in GLMM by taking $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ and $u_i = \log(\gamma_i)$ in (3.2). In this case the random effects u_i is nonnormal but $\exp(u_i)$ is gamma distributed as mentioned before. This means that NB model can be formulated as GLMM with a nonnormal random effect. Most of the commonly used software to fit GLMM allows only normal distribution for the random effect, and therefore Poisson-gamma GLMM cannot be fitted in a routine way without writing special programming codes and restricts the use of Poisson-gamma GLMM of NB model. But NB model can alternatively be fitted using ordinary GLM software by regarding the NB distribution for count responses. Until now we discussed three kinds of overdispersion models with the GOF criteria. The ZIP model can be used to overcome overdispersion due to excess of zeros but useless for other cause of overdispersion; however, the NB model and GLMM can generally be used for overdispersed counts data. The GLMM is computationally more intensive than the NB model, and furthermore most statistical packages fit GLMM under normal assumption for random effects. The violation of normality may cause inferential problems for the variance component of random effects. The NB model can be fitted simply without any restrictions but the choice between two competing overdispersion models should be considered in several respects that include GOF criteria.

4. Implementation and Examples

In this section we further explain the overdispersion models through two kinds of real data with appropriate GOF criteria. Various software can fit the models discussed for overdispersed count data. For

Table 2: Number of victims known in past year, by race, predicted by mixture models

Responses	Data		ZIP		NB		GLMM	
	Black	White	Black	White	Black	White	Black	White
0	119	1070	128.3	1059.3	122.8	1064.9	119.2	1101.7
1	16	60	10.8	72.3	17.9	67.5	24.6	40.9
2	12	14	9.8	15	7.8	12.7	7.8	4.2
3	7	4	6	2.1	4.1	2.9	3.8	1.18
4	3	0	2.7	0.2	2.4	0.7	1.9	0.45
5	2	0	1	0	1.4	0.2	1	0.22
6	0	1	0.3	0	0.9	0.1	0.4	0.12
X^2 (df)			965.3 (1305)		1424 (1305)		278.63 (1305)	
$D(\mathbf{y};\hat{\boldsymbol{\mu}})$ (df)			999.9 (1305)		412.6 (1305)		728.1 (1305)	
AIC			1005.9		1001.8		734.1	

Table 3: Number of victims collapsed over responses and race in Table 2

Category	Observed	Poisson	ZIP	NB	GLMM
0	1189	1142	1187.6	1187.7	1220.9
1	76	145.9	83.1	85.4	65.5
2	26	17.4	24.8	20.5	12
3	11	2.3	8.1	7	5
4+	6	0.3	6.8	10.9	4.1
GOF (df)		180.9 (3)	1.80 (2)	7.0 (2)	27.2 (2)
P-value		0.000	0.407	0.030	0.000

example, *NLMIXED* in SAS and *glmmML* in R can be used to fit Poisson-normal GLMM; however, *GENMOD* in SAS and *glm.nb* in R are useful to fit NB model. In these examples we used R packages mentioned above to obtain the result.

Example 1. We revisit the data of Table 1 to overcome the overdispersion problems in count responses. The sample mean of blacks is 0.522 with a variance of 1.150, and the sample mean of whites is 0.092 with a variance of 0.155 as noted before. This data set also has excess of zeros with proportion of 0.909; subsequently, the ZIP model can be a natural selection. Table 2 summarizes the expected count responses predicted by the ZIP model, NB model, and Poisson-normal GLMM with their assessment criteria. The ordinary Poisson GLM was already fitted in Section 2. When we fit the ZIP model the estimates are $\hat{\beta}_0 = -0.88\,(\text{SE} = 0.175)$, $\hat{\beta}_1 = 1.48\,(\text{SE} = 0.199)$ and $\hat{\omega} = 0.77\,(\text{SE} = 0.178)$. But under NB model $\hat{\beta}_0 = -2.38\,(\text{SE} = 0.117)$, $\hat{\beta}_1 = 1.73\,(\text{SE} = 0.239)$, and $\kappa = 4.94\,(\text{SE} = 0.041)$. The estimated coefficients of the NB model are similar to the ordinary Poisson GLM of Section 2; however, when Poisson-normal GLMM are fitted we obtain $\hat{\beta}_0 = -3.69\,(\text{SE} = 0.244)$, $\hat{\beta}_1 = 1.90\,(\text{SE} = 0.246)$ and $\hat{\sigma} = 1.63\,(\text{SE} = 0.155)$.

The goodness-of-fit of various models can be assessed by the usual Pearson chi-squared statistic, denoted as GOF in Table 3 that compares the observed frequencies and the expected frequencies of grouped data given by contingency table. Original response categories 4 through 6 are combined so that the expected frequencies are properly large compared to 5. Table 3 shows the expected frequencies merged with respect to race and also the P-values of four kinds of models, among which the ZIP model is the best and the NB model is the next alternative; however, the ordinary Poisson GLM and the Poisson-normal GLMM do not fit. We doubt the violation of normality of random effects in the fit of Poisson-normal GLMM as shown in the Q-Q plot of Figure 1.

Example 2. As a second example of overdispersed count responses Figure 2 shows a time series plot of 534 monthly counts of mumps cases in New York City, 1928–1972 (Waagepetersen, 2006).

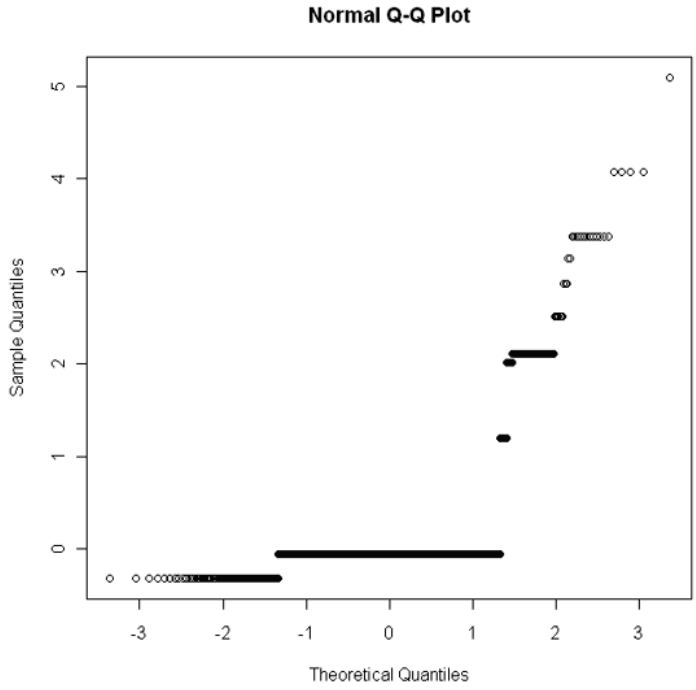

Figure 1: *Q-Q plot of predicted random effects*

In Figure 2 we see a pronounced seasonal variation that varies from the smallest count of 20 to the maximum count of 1956. The graph of the autocorrelation function, omitted for space, denotes strong autocorrelations with the significant periodic variation of mumps. The mean and variance of mumps is 487.7 and 147721.5, respectively, thus there exists an expected large extra variation versus the variance of the Poisson distribution.

We first consider an ordinary Poisson model having several covariates such as a month and time variable measured in the unit of month given by

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i2})^2, \quad i = 1, 2, \ldots, 534,$$

where x_{i1} denotes the categorical covariate month (1∼12) and x_{i2} is the log transformed time variable. The month is included to explain the seasonal variation, and the time variable is the elapsed time measured in units of month. The MLEs for each fitted model are listed in Table 4. The coefficient $\hat{\beta}_1$ varies according to the month, for example, in a Poisson-normal GLMM fit, the estimates are 0.00, 0.13, 0.52, 0.59, 0.59, 0.48, −0.10, −0.80, −1.20, −1.14, −0.79, −0.40. These estimates are similar under the Poisson, quasi-Poisson, NB models. The signs of $\hat{\beta}_1$ are negative from month 7 to month 12, particularly small in the season of month 8 to month 11. Finally the ZIP model closely coincides with the ordinary Poisson GLM with $\hat{\omega} \approx 0$ because the observed counts are positively large as shown in Figure 2.

The GOF criteria are summarized in Table 4 with the parameter estimates and their standard errors in parentheses. The Poisson-normal GLMM fits best among the four models and then the quasi-Poisson and the NB models are the next in the sense of AIC and X^2 criteria. We checked the validity of normal random effects through a Q-Q plot of predicted random effects, which are not as violated as in Example 1. This fact seems to improve the fit of the Poisson-normal GLMM compared

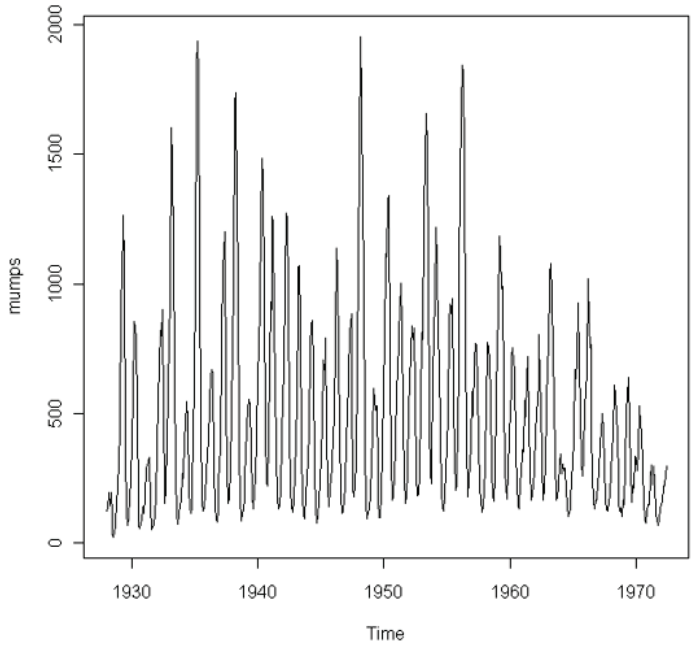

Figure 2: *Number of mumps against monthly times between 1928 and 1972*

Table 4: Estimates for the mumps data fitted by four types of models

Parameter	Poisson	Quasi-Poisson	NB	GLMM
$\hat{\beta}_0$	2.74	2.74	3.83	3.35
	(0.045)	(0.447)	(0.247)	(0.255)
$\hat{\beta}_1$	1.51	1.51	0.97	1.12
	(0.019)	(0.186)	(0.109)	(0.113)
$\hat{\beta}_2$	−0.16	−0.16	−0.10	−0.11
	(0.002)	(0.019)	(0.012)	(0.013)
Dispersion	-	$\hat{\phi} = 99.9$	$\hat{\kappa} = 0.19$	$\hat{\sigma} = 0.46$
X^2 (df)	51940 (520)	520 (519)	541.1 (519)	475.9 (519)
$D(\mathbf{y};\hat{\boldsymbol{\mu}})$ (df)	49605 (520)	49605 (519)	550.8 (519)	2848 (519)
AIC	53755	-	6993.4	2878

to the alternative NB model; however, the NB model is best in the sense of $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ statistic.

Several overdispersion models have been discussed with their GOF criteria for counts data through two examples of different response types. Most responses in Example 1 are smaller than three with a very large proportion of zeros; however, Example 2 denotes time series outcomes with strong autocorrelations that denote significant seasonal variations and the violation of independence. We emphasize that an appropriate overdispersion model should be chosen with proper assessment criteria for the model GOF.

5. Discussion and Further Research

The various models for count responses with overdispersion has been discussed in the context of mixture models. The quasi-likelihood method using only the relationship between mean and variance incorporates a dispersion parameter to account for extra variation. The GLMM is more elaborate than

the quasi-likelihood method; subsequently, we have two competing candidate models, the NB model versus the Poisson-normal GLMM. The NB model has been used as standard for the overdispersion model of count data. We show that the NB model belongs to a class of GLMM but has nonnormal random effects. The goodness-of-fit of GLMM may depend on the random effects distribution that is usually assumed to be normal. Commonly used software to fit GLMM is restricted to normal random effects; therefore, we need to check the normality. The NB model in the framework of GLMM cannot be directly fitted via commonly used software without special programming codes.

Goodness-of-fit for each overdispersed model can be assessed by usual criteria such as the chi-squared statistic, deviance, and AIC measure. These statistics should be deliberately applied according to the structure of data because the approximate distribution depends on the magnitude of estimated means. Further the preference between the NB model and the Poisson-normal GLMM also may be determined according to the validity of assumption for random effects.

Two examples have been introduced to explain several overdispersion models with their goodness-of-fit criteria. We finally recommend that the Poisson-normal GLMM and the NB model are the most competing models for count responses with overdispersion. The choice depends on assessment criteria and assumptions for random effects. This paper has a limitation in that the selection method of an appropriate model has not been justified through an empirical study. A Monte Carlo study that compares models in terms of mean squared errors of estimates and GOF criteria will be a justifiable method. These remain future research topics and include GOF statistics that use cumulative sums of residuals.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed., Wiley, New York.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9–25.

Jowaheer, V. and Sutradhar, B. C. (2002). Analysing longitudinal count data with overdispersion, *Biometrika*, **89**, 389–399.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Ed., Chapman and Halls, London.

Morel, J. G. and Neerchal, N. K. (2012). *Overdispersion Models in SAS*, SAS Institute Inc.

Pan, Z. and Lin, D. Y. (2005). Goodness-of-Fit methods for generalized linear mixed models, *Biometrics*, **61**, 1000–1009.

Sutradhar, S. C., Neerchal, N. K. and Morel, J. G. (2007). A goodness-of-fit test for overdispersed binomial or multinomial models, *Journal of Statistical Planning and Inference*, **138**, 1459–1471.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics*, **46**, 657–671.

Waagepetersen, R. (2006). A simulation-based goodness-of-fit test for random effects in generalized linear mixed models, *Scandinavian Journal of Statistics*, **33**, 721–731.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika*, **61**, 439–447.

Wood, G. R. (2002). Assessing goodness-of-fit for Poisson and negative binomial models with low means, *Communications in Statistics*, *Theory and Methods*, **31**, 1977–2001.

Xu, W. and Lu, Y. (2009). Goodness-of-fit for longitudinal count data with overdispersion, *Communications in Statistics*, *Theory and Methods*, **38**, 3745–3754.

Received August 9, 2012; Revised September 5, 2012; Accepted September 24, 2012