Proposed Score Test for Over-dispersion Parameter in the Multilevel Negative Binomial Regression Model

[1] Aragaw Eshetie Aguade , [2*] B.Muniswamy

[1*] Department of Statistics, College of Science and Technology, Andhra University, Visakhapatnam, India

[2*] Departments of Statistics, College of Science and Technology, Andhra University, Visakhapatnam, India

Abstract: Overdispersion is familiar in count data models particularly in the area of ecological and biological science because of non- independent, aggregations of data and an excess frequency of zeros. Every cluster levels received a singular level of a random effect that models the additional Poisson variation given within the data, are usually utilized to address overdispersion in count data. However, studies investigating that the power of cluster level random effects as a way to data with overdispersion is scarce. A situation where the variance of the response variable exceeds the mean, and hence, both overdispersion and heterogeneity between groups parameters occur. In the appropriate imposition of the multilevel Poisson model may underestimate the standard error and overestimate the significance of the regression parameters, and consequently, giving misleading inference about the regression parameters. This paper suggested that the multilevel negative binomial models as alternatives for handling overdispersion; an algorithm is developed for the residual maximum likelihood estimate (REML) of the regression coefficients and variance component parameters. In addition, the predicted random effects can provide information on the interregional variation after adjustment for children characteristic and death features. In this paper we used an application and simulation study, a simulation study showed that the estimators obtained from multilevel negative binomial model perform well in all the setting considered. The simulations reveal that failing to account for overdispersion in mixed models can erroneously inflate measures of explained variance ( ), which may lead to researchers overestimating the predictive power of variables of interest. Application to a set of deaths of children under 18 year's data is illustrated. The result revealed that the proposed model is better than the multilevel Poisson regression model and hence, there is a variation among regions in the deaths of children less than 18 years. Both the predicted probability and the information criteria indicated that the multilevel negative binomial model is better than the multilevel Poisson regression model. This work suggests the use of group level random effects provides a simple and robust means to a count data, but also that this ability to minimize bias is not uniform across all types of overdispersion and must be applied thoughtfully.

***Key Words*:** Clustered Count Data, Overdispersion, Homogeneity, Random Effects, Multilevel ZIP and ZINB models.

I. INTRODUCTION

Count data are tremendously familiar in the field of biological and ecological sciences; researchers are often interested in quantifying the factors affecting variables such as how many offspring an individual produces, counts of parasite load, abundance of species within and between habitats, or the frequency with which individually perform certain behaviours perhaps the most known method employed to model count data is to assume the data approximate a Poisson distribution, and specify statistical models accordingly (Bolker et al., 2009). However, a persistent problem with Poisson model is that they often exhibit overdispersion, where the variance of the response variable is greater than the mean (Hilbe., 2011), resulting in a poor fit to the data. Accounting for overdispersion when it is present is critical; failing to do so can lead to biased parameter estimates (Crawley, 2007; Hilbe 2011); and causes researchers to erroneously conclude that covariates have a significant effect when in fact they do not (Crawley, 2007; Richards, 2008; Zuur et al.; 2009).

The need for accurate biological inference, therefore, we use tool to both identify and adequately deal with overdispersion to minimize the risk of type I error (Hilbe, 2011). Primarily overdispersion occurs for two reasons" apparent over-dispersion'' (Hilbe, 2011) arises when models are poorly specified such as by failing to include important predictors, an interaction between predictors that have already been measured or by specifying the incorrect links function (Hilbe, 2011). Conversely, "real over-dispersion" can arise when there is an excess number of zeros in the data (Zuur et al.; 2009), or when the variance of the response truly is greater than the mean. In cases of the real overdispersion, the fit of the model to the data will be poor, even if the model uncovers significant results.

The dependence of clustered data (children nested within regions) may result in spurious associations and misleading inferences (Leving et al., 1998). There has been little research in regional discrepancies as a relevant attribute of deaths of children variations. The aim of this paper is to compare the multilevel negative binomial model and the multilevel Poisson models that accommodate simultaneously the overdispersion and heterogeneity of the outcome variables. The approach is based on the generalized linear mixed model formulation (McGilchrist,1994), where the random effects are incorporated in the linear predictor of each component. Residual maximum likelihood estimation of the regression coefficients and the random component parameters is achieved via an EM algorithm (Lee et al., 2005). The predicted random effects from fitting the mixture model provided information on inter-hospital variations after adjustment for patient characteristics and relevant risk factors (Yau, Luang, and Lee, 2003; Ng et al., 2004).

In this chapter, we develop a score test for overdispersion in generalized linearity mixed effects models, Cox (1983), based on the multilevel negative binomial models. A simulation is conducted to assess the performance of the REML estimators of the models parameters. An empirical datasets in deaths of children in Ethiopia is used to illustrate the practical applications of the methodology. If the observed variance is larger than the assumed variance which is known as overdispersion and if the dispersion is ignored, statistical inference results in an inaccurate conclusion by underestimating the variability of the data.

II. FORMULATIONS OF STATISTICAL MODEL

2.1. THE MULTILEVEL NEGATIVE BINOMIAL MODEL

Let Y_{ij} (i =1, 2,…, k; j = 1, 2,…, n_i) represents the response variable for the j^{th} observation in the i^{th} group, where k is the number of groups, n_i is the number of individual observation i , the total number of observations being $n = \sum_{i=1}^{k} n_i$. The probability density function of Y_{ij} is assumed to be a negative binomial model for cluster i, the conditional distributions of the outcome variable $y_i = (y_{i1}, y_{i2}, ..., y_{in_i})^`$, given a set of cluster level random effect α_i and the conditional overdispersion parameters c in a mean overdispersion parameterization, is

$$exp\left[\sum_{i=1}^{n_i}\{log\,\Gamma(y_{ij}+c_i^{-1}) - log\,\Gamma(y_{ij}+1) - log\,\Gamma(c_i^{-1})\} + C(y_{ij}, c_i)\right] \ldots.(5.1)$$

Where $C(y_{ij}, c_i)$ is defined as $-\frac{1}{c_i}log\left\{1+exp\left(\eta_{ij} + logc_i\right)\right\} - y_{ij}log\left\{1+exp\left(-\eta_{ij} - logc_i\right)\right\}$

Where $log(\mu_{ij}) = \eta_{ij} = x_{ij\beta} + z_{ij}\alpha_i$ the mixed effect model for the mean response is $\alpha_i = \alpha + D^{\frac{1}{2}}u_i$, c_i is the dispersion parameter for group i, and x_{ij} is a $p\times1$ vector of independent covariates. Where the u_i`s are independently and identically distributed with normal distribution with zero mean and unit variance. Since we want to test homogeneity across and within groups we consider the random intercept model in which $z_{ij} = 1$for all i,j. Therefore $\alpha_i$`s are independently and identically distributed with mean α and variance *D*. Model (5.1) is an extension of the negative binomial regression model to include normally distributed random effects at different group levels. The standard negative binomial model is used to model over dispersed count data for which the variance is greater than that of a Poisson model. In a Poisson model, the variance is equal to the mean, and thus over dispersions are defined as the extra variability compared with the mean. Our interest is to test the null hypothesis $H_0: C = 0$ against the alternative $H_0: C > 0$. This implies that testing the dispersion parameters assuming that heterogeneity of individuals across groups (see carrasco and Jover, 2005).

2.2. DERIVATION OF THE SCORE TEST BASED ON THE DISPERSION PARAMETER OF THE MULTILEVEL NEGATIVE BINOMIAL MODEL

Now we assumed that the outcome variable y_{ij} , j=1, 2,…,n_i, i=1, 2,…, k comes from the negative binomial model, then the probability distribution functions of y_{ij} is

$$f_{ij}(y_{ij}, \mu_{ij}, c_i) = \frac{\Gamma(y_{ij}+c_i^{-1})}{y_{ij}!\,\Gamma(c_i^{-1})}\left(\frac{1}{1+c_i\mu_{ij}(x)}\right)^{c_i^{-1}}\left(\frac{c_i\mu_{ij}(x)}{\tau+\mu_{ij}(x)}\right)^{y_{ij}}$$

We assume a common over dispersion parameter c, that is, $c_1 = c_2 = \cdots = c_k = c.$
Further, we assume the mixed effects model approach.

$$\theta_{ij} = log(\mu_{ij}) = X_{ij}{}^`\beta + \alpha_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..\ldots.(5.2$$

Where β is a vector of p unknown regression parameters and α_i`s are identically and independently distributed random variables having a normal distribution with mean α and variance D. The log-likelihood function for the i^{th} group is given by

$$l_i(\beta, c) = log\int\prod_{j=1}^{n_i} f_{ij}(y_{ij}, \beta, c/\alpha_i)f_{ij}(\alpha_i)\,d\alpha_i \ldots\ldots\ldots(5.3)$$

Where

$$logf_{ij}(y_{ij}, \beta, c/\alpha_i) = \left[\sum_{l=1}^{y_{ij}-1}(1+cl) + y_{ij}(x_{ij}{}^`\beta + \alpha_i) - (y_{ij}+c^{-1})log(1+ce^{x_{ij}{}^`\beta+\alpha_i})\right]\ldots.(5.4)$$

Our purpose is to test the null hypothesis $H_0: C = 0$ against the alternative $H_A: C > 0$. To obtain the score function we need to integrate out α_i from (5.3). However, in practice, it is difficult to carry out the integration. So, instead, we use (5.4) to obtain the likelihood and develop the score test for given α_i. That is, for the development of the score test. We consider β to be a nuisance parameter and α_i to be known. We deal with the issue of α_i being random. Later in this chapter, the resulting log-likelihood of β and *c* for given α_i is

$$l = \sum_{i=1}^{k}\sum_{j=1}^{n_i} logf_{ij}\left(y_{ij}, \beta, c/\alpha_i\right)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\sum_{l=0}^{y_{ij}-1} log(1+cl) + y_{ij}\left(x_{ij}`\beta + \alpha_i\right) - \left(y_{ij} + c^{-1}\right)log\left(1+ \ ce^{x_{ij}`\beta+\alpha_i}\right)\right] \dots (5.5)$$

Then the score function for testing the null hypothesis $\mathrm{H_0}: \mathrm{C} = 0$ is obtained as (see also collings and Margoline, 1985)

$$S_c = \frac{\partial l}{\partial c} \mid_{c=0} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{\partial l}{\partial c} logf_{ij}\left(y_{ij}, \beta, \frac{c}{\alpha_i} \mid c = 0\right)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\frac{\left(\mu_{ij}^2 - 2\mu_{ij}y_{ij}\right)}{2} + \sum_{l=0}^{y_{ij}-1} l\right\} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\left(y_{ij} - e^{x_{ij}`\beta+\alpha_i}\right)^2 - y_{ij}\right\} \dots \dots \dots . (5.6)$$

So the score test statistic for testing $\mathrm{H_0}: \mathrm{C} = 0$ is

$$\mathrm{T_C} = \frac{\mathrm{S_c^2}}{\mathrm{I}}$$

where $I = I_{CC} - I_{C\beta}I_{\beta\beta}^{-1}I_{C\beta}`$, $I_{CC} = E\left(\frac{\partial l_i}{\partial c} \mid_{c=0}\right)^2$, $I_{C\beta} = E\left(-\frac{\partial^2 l_i}{\partial c \partial \beta} \mid_{c=0}\right)$ *is a 1×p* vector and $I_{\beta\beta} = E\left(-\frac{\partial^2 l_i}{\partial \beta_s \partial \beta_r} \mid_{c=0}\right)$ is p×p matrix.

Now

$$E\left(\frac{\partial l}{\partial c}\sum_{i=1}^{n_i} logf_{ij}\right)^2 = \frac{1}{4}E\left[\sum_{j=1}^{n_i}\left(y_{ij} - \mu_{ij}\right)^2 - \sum_{j=1}^{n_i} y_{ij}\right]^2$$

$$= \frac{1}{4}E\left\{\sum_{j=1}^{n_i}\left(y_{ij} - \mu_{ij}\right)^4 + \sum_{j=1}^{n_i}\sum_{j\neq j`}^{n_i}\left(y_{ij} - \mu_{ij}\right)^2\left(y_{ij}` - \mu_{ij}`\right)^2 + \sum_{j=1}^{n_i} y_{ij}^2 + \sum_{j=1}^{n_i}\sum_{j\neq j`}^{n_i} y_{ij}y_{ij}` - 2\sum_{j=1}^{n_i}\left(y_{ij} - \mu_{ij}\right)^2 y_{ij}\right\}$$

Using the first four moments of the Poisson distributions in the above expression

$$E\left(y_{ij} - \mu_{ij}\right)^4 = \mu_4 = k_4 + 3k_2^2, \ E\left(y_{ij} - \mu_{ij}\right)^2 = \mu_2 = \sigma_{ij}^2 = k_2 = k_4 = \mu_{ij}$$

Where $E\left[\left(y_{ij} - \mu_{ij}\right)^2 y_{ij}\right] = \mu_{ij} + \mu_{ij}^2$

After simplification it can be shown that

$$I_{CC} = \frac{1}{2}\sum_{i=1}^{k}\left(\sum_{j=1}^{n_i}\mu_{ij}^2 + \sum_{j=1}^{n_i}\sum_{j\neq j`}^{n_i}\mu_{ij}\mu_{ij}`\right)$$

$$I_{C\beta} = I_{\beta C} = E\left(-\frac{\partial^2 l}{\partial c \partial \beta}\sum_{i=1}^{n_i} logf_{ij}\right) = E\left(\sum_{i=1}^{n_i}\mu_{ij}\left(y_{ij} - \mu_{ij}\right)x_{ij}\right) = 0.$$

And

$$\frac{\partial l}{\partial \beta}\sum_{i=1}^{n_i} logf_{ij} = \sum_{i=1}^{n_i}\left(y_{ij}x_{ij} - \mu_{ij}x_{ij}\right)$$

$$\frac{\partial^2 l}{\partial \beta_s \partial \beta_r}\sum_{i=1}^{n_i} logf_{ij} = -\sum_{i=1}^{n_i}\mu_{ij}x_{ij}x_{ij}`$$

So

$$I_{\beta\beta} = E\left(-\frac{\partial^2 l}{\partial \beta_s \partial \beta_s}\sum_{i=1}^{n_i} logf_{ij}\right) = \sum_{i=1}^{n_i}\mu_{ij}x_{ij}x_{ij}`$$

Then, the score test statistic $\mathrm{T_c}$ can be written as

$$T_c = \frac{\hat{s}_c^2}{\hat{I}_{CC}} = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\left(y_{ij} - \hat{\mu}_{ij}\right)^2 - y_{ij}\right\}\right)^2}{\frac{1}{2}\sum_{i=1}^{k}\left(\sum_{j=1}^{n_i}\hat{\mu}_{ij}^2 + \sum_{j=1}^{n_i}\sum_{j\neq j`}^{n_i}\hat{\mu}_{ij}\hat{\mu}_{ij}`\right)}$$

Where $\hat{\mu}_{ij}^2 = e^{x_{ij}`\hat{\beta}+\alpha_i}$

$\hat{\beta}$ is the maximum likelihood estimation of β under the null hypothesis. Asymptotically, as $k \to \infty$, T_c has a chi-square distribution with one degree of freedom. Note that the above results are based on α_i being known. However, since $\alpha_i`s$ are random effects these should have been integrated out of (5.3). As indicated earlier such integration is difficult to carry out. So, we replace these by their estimates. One way of obtaining estimates of the random effects is through using an empirical Bayes procedure (see collet, 2003). The Maximum likelihood estimates of β and the empirical Bayes estimates of α_i under the null hypothesis are given as follows.

2.3. PARAMETRIC ESTIMATION OF THE PARAMETERS $\boldsymbol{\beta}$ UNDER THE NULL HYPOTHESIS.

Note that the mixed affects models (5.2) can be written as

$$log(\mu_{ij}) = x_{ij}`\beta + \sqrt{D}u_i \ldots\ldots\ldots\ldots..(5.7)$$

Where u_i has a standard normal distribution and now define $\eta_{ij} = x_{ij}`\beta$ for the linear component of the model obtained from the fixed effects, then (5.7) becomes $log(\mu_{ij}) = \eta_{ij} + \sqrt{D}u_i$. The kernel of the likelihood for β, D and u_i, i= 1, 2,…, k, for Poisson data is given by

$$L = L(\beta, D, u_1, u_2, u_3, \ldots, u_k) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\left[exp\{y_{ij}log(\mu_{ij}) - \mu_{ij}\}\right]$$

$$= \prod_{i=1}^{k}\prod_{j=1}^{n_i}\left[exp\left\{y_{ij}\left(\eta_{ij} + \sqrt{D}u_i\right) - exp\left(\eta_{ij} + \sqrt{D}u_i\right)\right\}\right] \ldots\ldots(5.8)$$

Further, since u_i is a random variables, it needs to be integrated out. Then the likelihood function for β and D can be written as

$$L(\beta, D) = \prod_{i=1}^{k}\int_{-\infty}^{\infty}\prod_{j=1}^{n_i}\left[exp\left\{y_{ij}\left(\eta_{ij} + \sqrt{D}u_i\right) - exp\left(\eta_{ij} + \sqrt{D}u_i\right)\right\}\right]\frac{exp\left(-\frac{u_i^2}{2}\right)}{\sqrt{2\pi}}du_i. (5.9)$$

The likelihood function (5.9) has (p+1) unknown parameters $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ and D. maximum likelihood estimates of the parameters β and D are obtained by maximizing (5.9). The integration in (5.9) is difficult to carry out. However, this can be evaluated approximately by using Gauss-Hermite formula for numerical integration. Therefore, the marginal likelihood function (5.9) becomes

$$\prod_{i=1}^{k}\prod_{j=1}^{n_i}\prod_{r=1}^{m}W_r\left[exp\left\{y_{ij}\left(\eta_{ij} + \sqrt{D}s_r\sqrt{2}\right) - exp\left(\eta_{ij} + \sqrt{D}s_r\sqrt{2}\right)\right\}\right] \ldots(5.10)$$

Where $w_1, w_2, w_3, \ldots, w_m$ are the weights with

$$W_r = \frac{2^{m-1}m!\sqrt{2}}{m^2[H_{m-1}(s_r)]^2}$$

Where m is the number of quadrature the points and $s_1, s_2, s_3, \ldots., s_m$ are the roots of Hermite polynomial

$$H_{m(s)} = (-1)^m e^{-s^2/2}\frac{d^m}{ds^m}e^{-s^2/2}.$$

The evaluation points s_r (abscissas) and w_r (weights) (Abramowitz and Stegun 1972).The values $\hat{\beta}$ and $\widehat{D}$, which maximize (5.10) or its logarithm can then be determined numerically. The computer package SAS procedure GLIMMIX or NLMIXED or R function glmm ML can be used to evaluate equation (5.10).

2.4. PARAMETRIC ESTIMATIONS OF THE RANDOM EFFECTS $\boldsymbol{\alpha_i}$

From equation (5.9) the joint posterior density of $u_1, u_2, u_3, \ldots, u_k$, given $\hat{\beta}$and $\widehat{D}$, the maximum likelihood estimates of β and D will be obtained.

$$\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left[exp\left\{y_{ij}\left(\widehat{\eta}_{ij} + \sqrt{\widehat{D}}u_i\right) - exp\left(\widehat{\eta}_{ij} + \sqrt{\widehat{D}}u_i\right)\right\}\right]\frac{exp(-u_i^2/2)}{\sqrt{2\pi}} \ldots(5.11)$$

$$\text{Where } \widehat{\eta}_{ij} = x_{ij}`\hat{\beta}.$$

Now, the log of the j^{th} term of (5.11) is given by

$$L_i(\hat{\beta}, \widehat{D}, u_i) = constant + \prod_{j=1}^{n_i}\left[y_{ij}\left(\widehat{\eta}_{ij} + \sqrt{\widehat{D}}u_i\right) - exp\left(\widehat{\eta}_{ij} + \sqrt{\widehat{D}}u_i\right)\right] - \frac{u_i^2}{2} \ldots.(5.12)$$

The empirical Bayes estimate $\widehat{u_i}$ of u_i is obtained by solving $\frac{\partial L_i(\hat{\beta}, \widehat{D}, u_i)}{\partial u_i} = 0$. This is equivalent to obtain $\widehat{u_i}$ by solving

$$\frac{\partial L_i(\hat{\beta}, \widehat{D}, u_i)}{\partial u_i} = 0 + \sqrt{\widehat{D}}\sum_{j=1}^{n_i}y_{ij} - exp\left(\widehat{\eta}_{ij} + \sqrt{\widehat{D}}u_i\right)\left(\widehat{\eta}_{ij} + \sqrt{\widehat{D}}u_i\right)` - \widehat{u_i} = 0$$

$$\sqrt{\widehat{D}}\sum_{j=1}^{n_i}exp\left(\widehat{\eta}_{ij} + \sqrt{\widehat{D}}u_i\right) + \widehat{u_i} = \sqrt{\widehat{D}}\sum_{j=1}^{n_i}y_{ij} \ldots\ldots\ldots\ldots\ldots\ldots..(5.13)$$

Equation (9.13) is a non-linear equation to be solved by using a numerical method. The empirical Bayes estimate of α_ι then $\widehat{\alpha_\iota} = \sqrt{\widehat{D}}\widehat{u_\iota}$.

III. SIMULATION STUDY

The simulation is conducted to assess the performance of the proposed score test estimates obtained via the EM algorithm. Data are simulated under the multilevel negative binomial and multi level Poisson regression. Empirical power of the score test based on 1000 replications generated from the multilevel negative binomial regression model under the hypothesis of homogeneity. Levels considered are 10%; 5% and 1%. The sample size scenarios for each test groups = 5, 10, 20, 50, 100 and number of observation=10; 20; 50; 100 and the dispersion parameters= 0.15, 0.25, 0.4, are considered and the simulated results are shown as in Table 1.

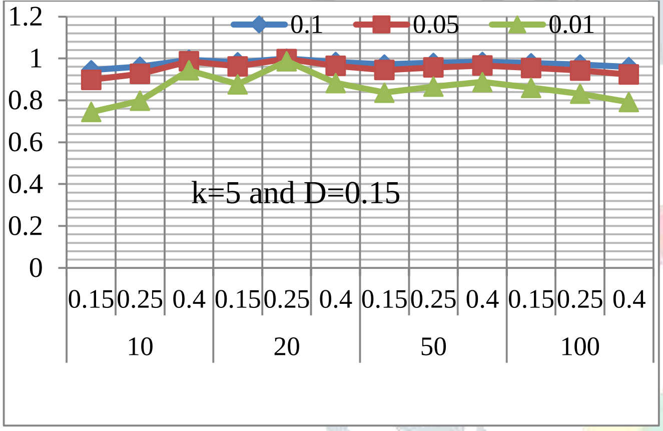
Table 1. TestS of power and size effects of the over dispersion parameter in the multilevel NB model via simulated data

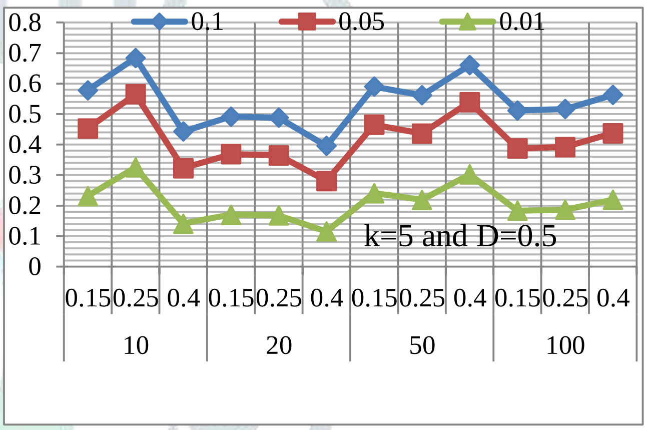
Cluster (k)	Observation (n)	Dispersion parameter (C)	D=0.15			D=0.5		
			Significance level			Significance level		
			α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
5	10	0.15	0.94386	0.89848	0.74444	0.57799	0.45272	0.23122
		0.25	0.96127	0.92654	0.79804	0.68384	0.56481	0.32537
		0.40	0.99409	0.98618	0.94368	0.44304	0.32285	0.14091
	20	0.15	0.98176	0.96217	0.87710	0.49181	0.36823	0.17034
		0.25	0.99922	0.99779	0.98715	0.48815	0.36477	0.16802
		0.40	0.98303	0.96452	0.88296	0.39598	0.28066	0.11550
	50	0.15	0.97207	0.94488	0.83676	0.59008	0.46503	0.24079
		0.25	0.97916	0.95743	0.86558	0.56147	0.43608	0.21858
		0.40	0.98395	0.96622	0.88726	0.66029	0.53901	0.30223
	100	0.15	0.97784	0.95505	0.85994	0.51193	0.38745	0.18348
		0.25	0.97084	0.94274	0.83207	0.51670	0.39205	0.18669
		0.40	0.95951	0.92363	0.79219	0.56276	0.43737	0.21955
10	10	0.15	0.99429	0.98661	0.94507	0.60536	0.48077	0.25328
		0.25	1.00000	1.00000	0.99998	0.77512	0.67026	0.43044
		0.40	0.98805	0.97402	0.90786	0.51173	0.38726	0.18335
	20	0.15	0.99081	0.97946	0.92320	0.64739	0.52509	0.29013
		0.25	0.92889	0.87543	0.70423	0.57594	0.45064	0.22962
		0.40	0.99836	0.99567	0.97776	0.67649	0.55670	0.31799
	50	0.15	0.98212	0.96284	0.87875	0.64321	0.52061	0.28629
		0.25	0.93991	0.89231	0.73337	0.60967	0.48525	0.25689
		0.40	0.94340	0.89775	0.74312	0.53620	0.41103	0.20017
	100	0.15	0.98822	0.97436	0.90877	0.39166	0.27686	0.11330
		0.25	0.98345	0.96529	0.88489	0.59670	0.47183	0.24615
		0.40	0.96442	0.93181	0.80881	0.46626	0.34424	0.15451
20	10	0.15	1.00000	1.00000	1.00000	0.27433	0.17860	0.061706
		0.25	0.96698	0.93612	0.81783	0.52728	0.40232	0.19393
		0.40	0.99852	0.99606	0.97941	0.30693	0.20497	0.074563
	20	0.15	0.97750	0.95445	0.85853	0.43662	0.31701	0.13728
		0.25	0.99700	0.99251	0.96541	0.57201	0.44667	0.22659
		0.40	0.99076	0.97935	0.92288	0.36990	0.25794	0.10258
	50	0.15	0.94200	0.89556	0.73917	0.52100	0.39621	0.18961
		0.25	0.96629	0.93495	0.81537	0.33652	0.22952	0.087190
		0.40	0.97385	0.94798	0.84369	0.54490	0.41960	0.20639
	100	0.15	0.97651	0.95267	0.85439	0.55584	0.43045	0.21439
		0.25	0.99174	0.98134	0.92874	0.54156	0.41631	0.20399
		0.40	0.97054	0.94223	0.83094	0.51646	0.39182	0.18653
50	10	0.15	0.99998	0.99992	0.99923	0.62811	0.50457	0.27275
		0.25	0.99724	0.99306	0.96744	0.62372	0.49995	0.26891
		0.40	0.99570	0.98963	0.95515	0.55190	0.42654	0.21149
	20	0.15	0.99408	0.98616	0.94363	0.74956	0.63979	0.39820
		0.25	0.99982	0.99942	0.99578	0.30600	0.20421	0.074183
		0.40	0.98739	0.97275	0.90438	0.39950	0.28376	0.11730
	50	0.15	0.99361	0.98517	0.94049	0.58210	0.45689	0.23444
		0.25	0.96191	0.92760	0.80019	0.57141	0.44606	0.22613
		0.40	0.99198	0.98181	0.93014	0.69056	0.57227	0.33223

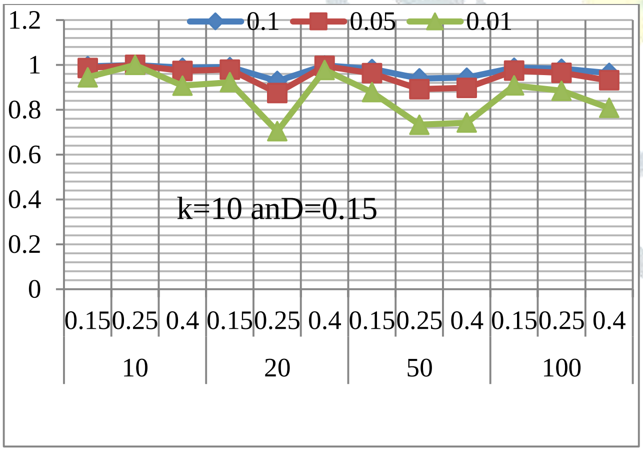
	100	0.15	0.97706	0.95365	0.85668	0.48103	0.35805	0.16355
		0.25	0.97722	0.95395	0.85736	0.52909	0.40408	0.19519
		0.40	0.96702	0.93620	0.81800	0.52140	0.39660	0.18988

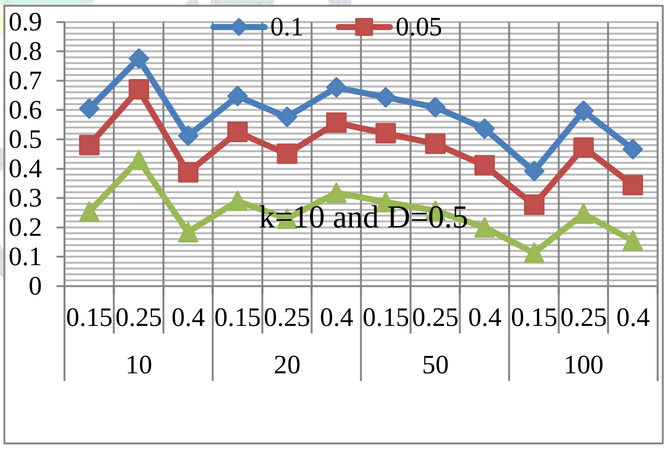
From Table 1, for large sample size, the multilevel Poisson model is efficient than the multilevel negative binomial model whereas in small sample size the multilevel negative binomial model is better than the multilevel Poisson regression model. In a small sample case, the dispersion parameter varies with the level of significance whereas in a large sample the level of significance has no effects on the dispersion parameter. The power study was extended for the situation in which data are generated from heterogeneous lognormal negative binomial distribution, the results which are not given here show similar behaviour to those when data are generated from the heterogeneous lognormal Poisson distribution. In summary, as k,c and ni increase, the power increases for all the statistics.

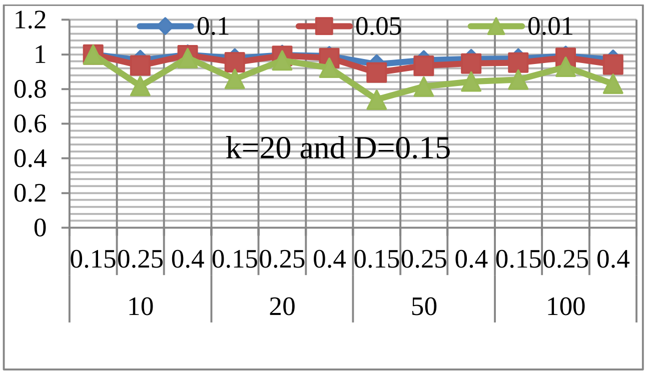
The statistics, the multilevel negative binomial model, in general, shows highly inflated level properties. The statistics the multilevel Poisson model show some conservative level properties, however, as the values of c and k increase, empirical levels become closer to the nominal level. The power of the statistic multilevel negative binomial model is, in general, larger than those of multilevel Poisson model. The powers of both statistics are not similar in all the cases studied. We extended the simulation study of the properties of the two statistic interms of empirical size and power to situations where the over-dispersion parameter c is not the same for all groups. For this, we generated data from the heterogeneous negative binomial and lognormal Poisson distributions with heterogeneous over-dispersion parameter c. The results for size and power are given in Table 1 only for data that are generated from the heterogeneous negative binomial distribution. The level and power properties of all the statistics, in general, remain similar irrespective of which mechanism of over-dispersion is used to generate count data. This also seems to be true irrespective of whether the over-dispersion parameter c is varying or constant.

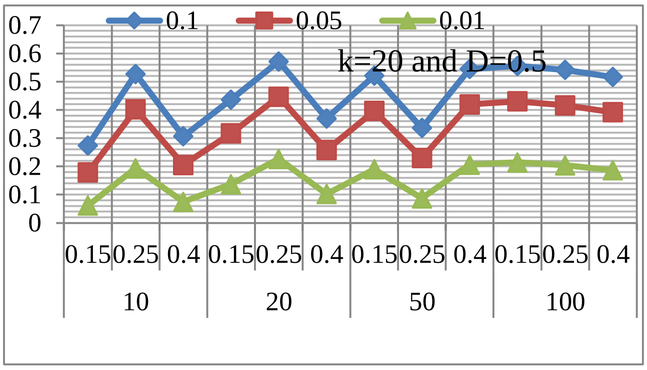












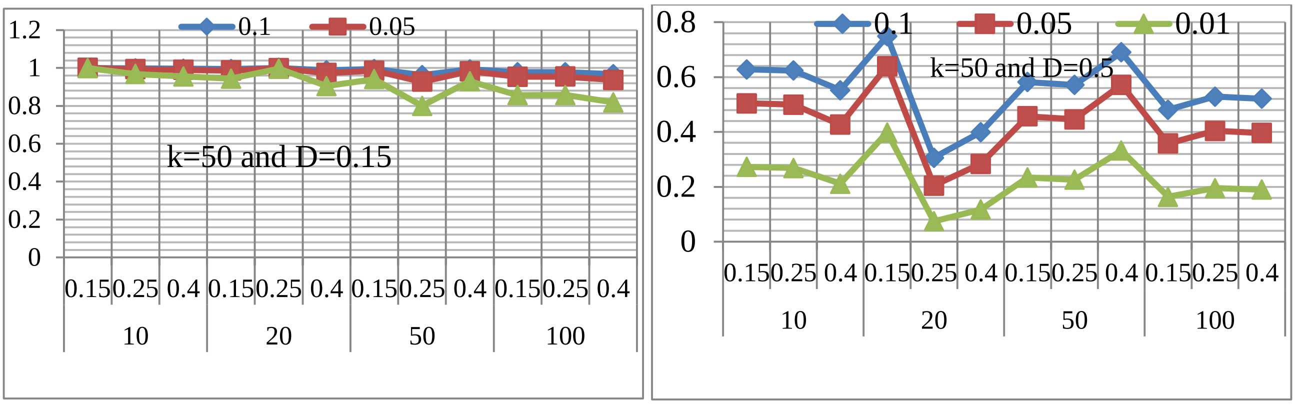


Figure 1. Simulated power and size effects of the overdispersion parameter in the multilevel NB models

Table 2 .Comparisons of goodness of fit Statistic between the multilevel NEGBIN and Poisson models based on simulated data.

Group(k)	Observation(n)	Model	-2loglikelihood	AIC	BIC
5	10	Multilevel Poisson model	140.6	146.6	147.5
		Multilevel NEGBIN model	129.2	137.2	138.4
	20	Multilevel Poisson model	225.5	228.5	231.5
		Multilevel NEGBIN model	248.1	256.1	260.1
	50	Multilevel Poisson model	647.2	653.2	658.9
		Multilevel NEGBIN model	608	616	623.7
	100	Multilevel Poisson model	1453.7	1459.7	1467.5
		Multilevel NEGBIN model	1175.5	1183.5	1194
10	10	Multilevel Poisson model	298.7	304.7	305.6
		Multilevel NEGBIN model	243.4	251.4	252.6
	20	Multilevel Poisson model	525	531	533.9
		Multilevel NEGBIN model	521.7	529.7	533.6
	50	Multilevel Poisson model	1292.9	1298.9	1304.6
		Multilevel NEGBIN model	1315	1323	1330.4
	100	Multilevel Poisson model	2720.5	2726.5	2734.4
		Multilevel NEGBIN model	2524.7	2532.7	2542.9
20	10	Multilevel Poisson model	477.9	483.9	484.9
		Multilevel NEGBIN model	531.5	539.5	540.7
	20	Multilevel Poisson model	1039	1045	1048
		Multilevel NEGBIN model	999.3	1007.3	1011.3
	50	Multilevel Poisson model	2466.7	2472.7	2478.5
		Multilevel NEGBIN model	2533.8	2541.8	2549.5
	100	Multilevel Poisson model	5244.8	5250.8	5258.6
		Multilevel NEGBIN model	4453.7	4461.7	4472
50	10	Multilevel Poisson model	1394.7	1400.7	1401.6
		Multilevel NEGBIN model	1336.2	1344.2	1345.4
	20	Multilevel Poisson model	2377.1	2383.1	2386.1
		Multilevel NEGBIN model	2374.6	2382.6	2386.6
	50	Multilevel Poisson model	6433.4	6439.4	6445.2
		Multilevel NEGBIN model	6566.5	6574.5	6582
	100	Multilevel Poisson model	12993	12999	13007
		Multilevel NEGBIN model	11898	11906	11916

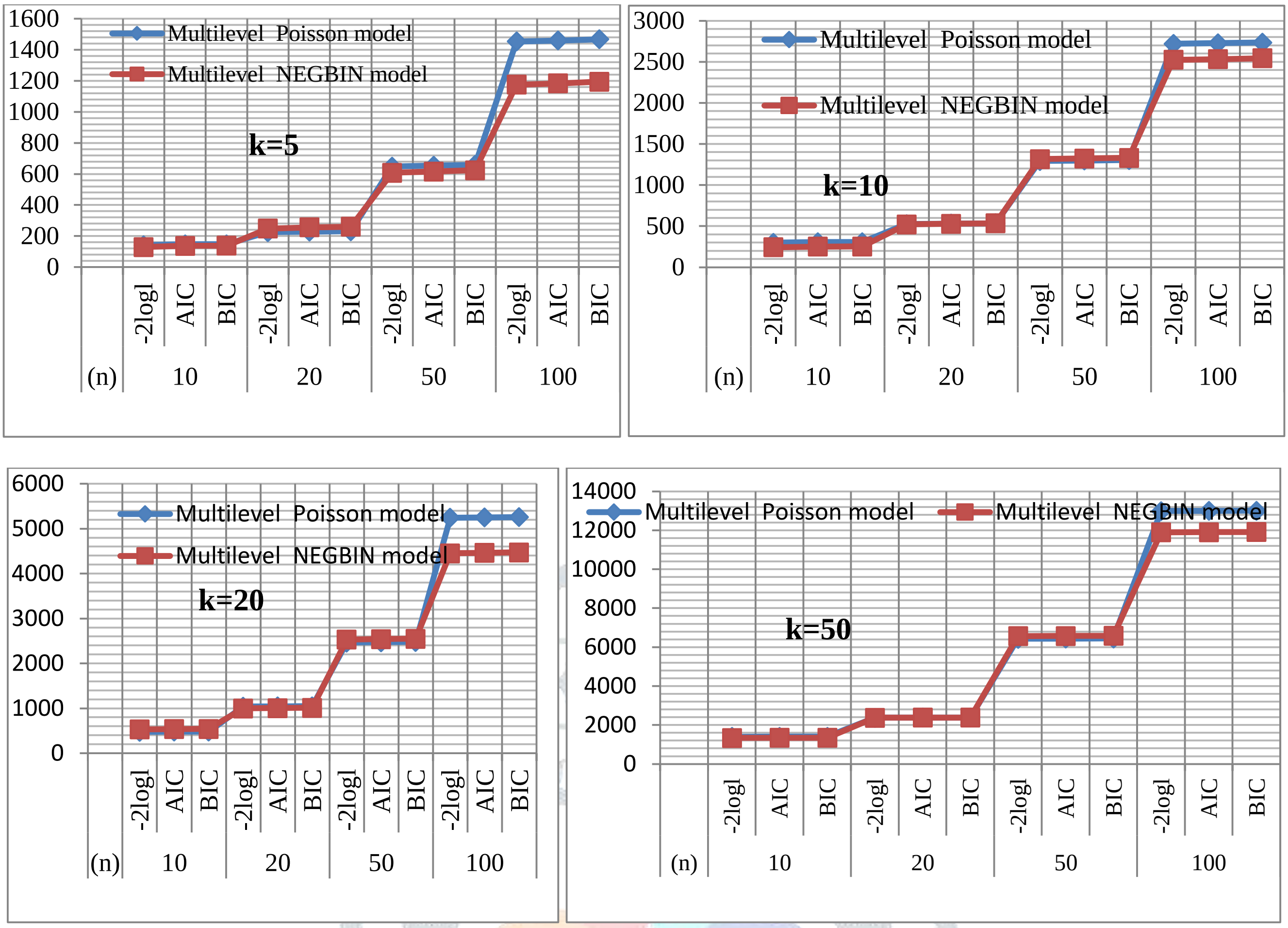


Figure 2. Comparisons of the multilevel Poisson and multilevel NB models

The power study was extended for the situation in which data are generated from heterogeneous lognormal negative binomial distribution. The results which are not given here shows similar behaviour to those when data are generated from the heterogeneous lognormal Poisson distribution, in summary, as k,c and ni increase, the power increases for all the statistics. The statistic MNEBIN, in general, shows highly inflated level properties. The statistics MNEGBIN show some conservative level properties, however, as the values of c and k increase, empirical levels become closer to the nominal level. The power of the statistic MNEGBIN is, in general, larger than those of Poisson. The powers of both statistics are not similar in all the cases studied. We extended the simulation study of the properties of the two statistic interms of empirical size and power to situations where the over-dispersion parameter c is not the same for all groups. For this, we generated data from the heterogeneous negative binomial and lognormal–Poisson distributions with heterogeneous over-dispersion parameter c. The results for size and power are given in Table 2 only for data that are generated from the heterogeneous negative binomial distribution. The level and power properties of all the statistics, in general, remain similar irrespective of which mechanism of over-dispersion is used to generate count data. This also seems to be true irrespective of whether the over-dispersion parameter c is varying or constant.

Table 3. Estimated Empirical power and size of the score test between the multilevel Poisson regression model and the multilevel negative binomial regression models under the hypothesis of homogeneity based on 1000 replications. Levels considered are 10%; 5% and 1%. The sample size scenarios for each test = 10, 20, 50,100 and the number of groups = 5, 10, 20 and 50 are considered.

Cluster (k)	Observation (n)	Multilevel Poisson Model			Multilevel Negative Binomial Model		
		Power at D=0.15			Power at D=0.15		
		α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
5	10	0.97221	0.94512	0.83729	0.93093	0.87852	0.70945
	20	0.99307	0.98406	0.93701	0.91172	0.84997	0.66291
	50	0.97598	0.95174	0.85224	0.97386	0.94800	0.84372
	100	0.97808	0.95548	0.86097	0.98068	0.96020	0.87226
10	10	0.99926	0.99790	0.98769	0.97943	0.95791	0.86673
	20	0.99136	0.98057	0.92644	0.98525	0.96867	0.89356
	50	0.97843	0.95611	0.86245	0.99343	0.98481	0.93933
	100	0.98096	0.96070	0.87348	0.97592	0.95162	0.85197
	10	0.94187	0.89536	0.73881	0.99940	0.99828	0.98955

	20	0.94689	0.90326	0.75316	0.99999	0.99995	0.99950
	50	0.98101	0.96080	0.87372	0.98468	0.96760	0.89080
	100	0.98768	0.97332	0.90592	0.97896	0.95706	0.86471
50	10	0.96108	0.92623	0.79741	0.99751	0.99366	0.96973
	20	0.91279	0.85153	0.66537	0.99779	0.99433	0.97232
	50	0.99780	0.99433	0.97234	0.97439	0.94893	0.84583
	100	0.98751	0.97298	0.90501	0.97162	0.94408	0.83500

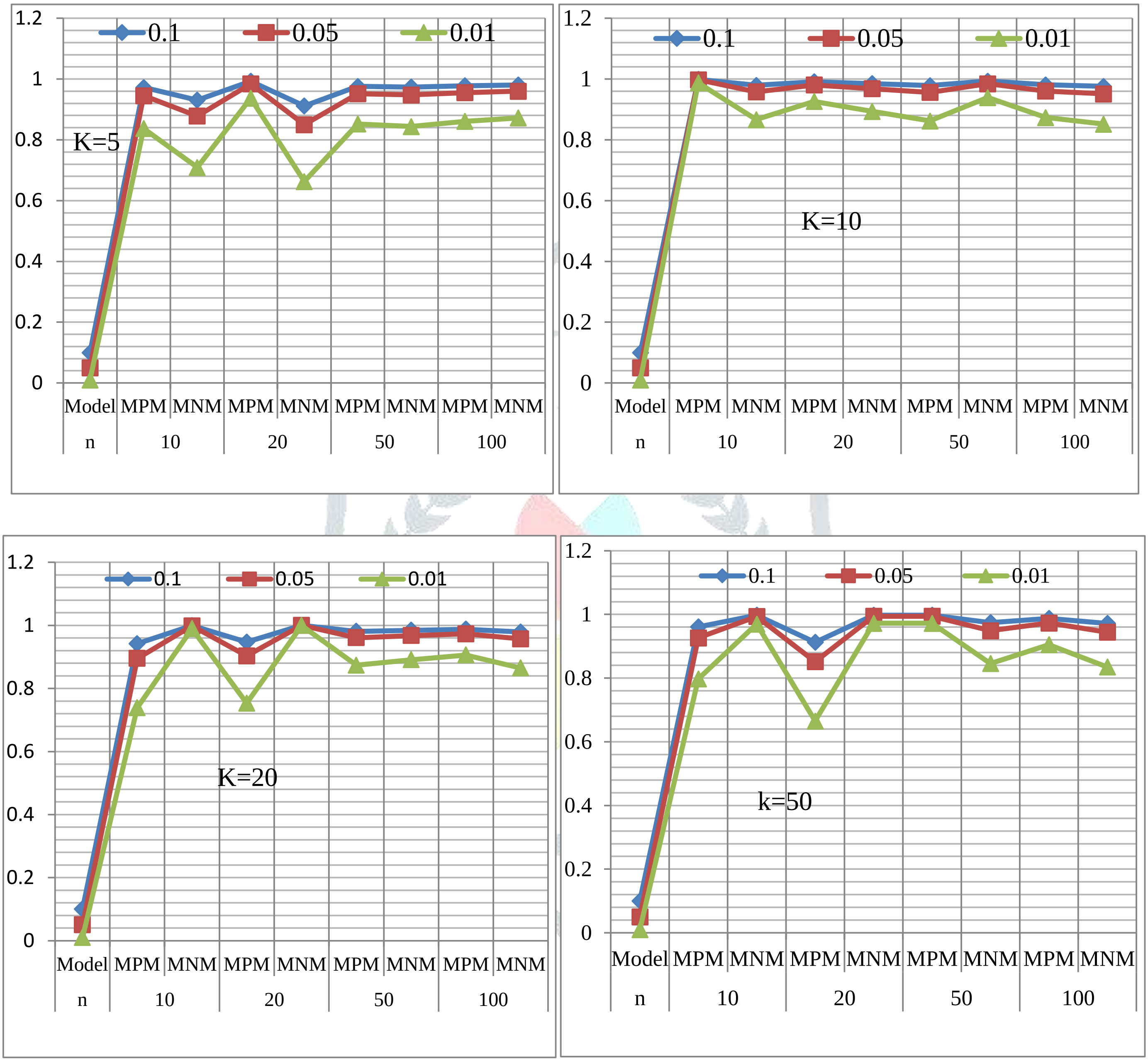


Figure 1. Comparisons of the power and size effects of the multilevel Poisson and the multilevel NB models

IV. APPLICATION STUDY

In this chapter, we focus on multilevel negative binomial regression model to take account of the coefficient of regression and random parameters in negative binomial counts with overdispersion. An algorithm for estimating parameters was obtained and a score test was presented for testing the multilevel negative binomial regression model against the multilevel Poisson regression model, and for testing the significance of the dispersion parameter. In the application study, the Ethiopian demographic and health-related survey under eighteen years Children death rate data is used to illustrate the proposed score tests.

Table 4. Fitting the dispersion parameters in Multilevel Negative Binomial Model with covariates

Number of children	Multilevel Poisson				Multilevel negative binomial			
	Estimate	S.E	Z-value	p-value	Estimate	S.E.	Z-value	P-value
Residence	0.4479238	0.0251482	17.81	0.000**	.2629343	.0341172	7.71	0.000**
Educ. level	-0.6493431	0.0245608	-26.44	0.000**	.004331	.0033312	1.30	0.194
Toiletfac	0.027909	0.0180485	1.55	0.122	.0621577	.0010159	61.19	0.000**
Religion	0.377617	0.0132241	28.56	0.000**	-.0483643	.0026752	-18.08	0.000**
HHSMembers	0 .0086175	0 .0024202	3.56	0.004**	-.612779	.028402	-21.58	0.000**
Age mother	0.056305	0 .0007035	80.04	0.000**	.4060037	.0201374	20.16	0.000**
Current Mari	0.2137659	0.0802338	3.66	0.008**	.0183504	.0995902	0.18	0.854
Agemarriage	-0.038131	0.0019082	-19.98	0.000**	.0012938	.0166597	0.08	0.938
Sourcdrinkwat	0.0311697	0.0121839	2.56	0.011*	-.0179391	.0237627	-0.75	0.450
Constant	-2.156525	0 .0662962	-32.53	0.000**	-1.788503	.1096869	-16.31	0.000**
/lnalpha					-.7008252	.0272103	-25.76	0.000**
alpha								
Region(var)	0 .1012529	0.1017809			.0479054	.0212992		
LR test vs. Poisson model: chibar2(01) = 184.55 Prob >= chibar2 = 0.0000 LR test of alpha=0: chibar2(01) = 3816.91 Prob >= chibar2 = 0.000								

*: Significant at 0.05 level. Alpha is clearly > 0! Over dispersion is evident; LR test p<.05.

From Table 4, the result showed that the multilevel negative binomial regression model is more efficient than multilevel Poisson regression model. Hence, overdispersion is evident and the proposed score test is appropriate to handle overdispersion.

Table 5. Fitting statistic of Poisson and multilevel Poisson models for the number of death notice of children with covariates in EDHS, 2005

Model	-2l	AIC	BIC
Multilevel Negative binomial model	69347.06	69371.05	69468.77
Multilevel Poisson model	73240	73095	73185.06

The multilevel negative binomial model can be considered as a parametric version of assessing heterogeneity among regions with respect to the Deaths of children. Moreover, the AIC and BIC value of multilevel negative binomial model is smaller than the multilevel Poisson model. This indicates that the multilevel negative binomial model is better than the multilevel Poisson model.

Table 6. The observed and predicted probabilities of Multilevel Poisson model and multilevel negative binomial models

Number of Children Death	Observed frequency	Observed probability	Predicted probability	
			Multilevel Poisson Model	Multilevel Negative Binomial Model
0	12000	0.472069237	0.327120871	0.311111111
1	5680	0.223446105	0.220180056	0.224713805
2	3560	0.140047207	0.19380382	0.200808081
3	1880	0.073957514	0.108047166	0.111245791
4	1140	0.044846577	0.068320502	0.070841749
5	500	0.019669552	0.032645033	0.033670034
6	320	0.012588513	0.02318702	0.024579125
7	80	0.003147128	0.005151118	0.004915825
8	120	0.004720692	0.007569642	0.007676768
9	40	0.001573564	0.002869634	0.002424242
10	40	0.001573564	0.005065108	0.004040404
11	20	0.000786782	0.001566147	0.001818182
13	20	0.000786782	0.000606783	0.00047138
18	20	0.000786782	0.003857189	0.003232323

The predicted probabilities for the multilevel Poisson and the multilevel negative binomial regression model are presented in Table 4 and Fig 3. To check the analysis, whether the negative binomial and the multilevel negative binomial regression model would fit the data better, we fitted the maximum likelihood of the parameters and the maximized log likelihoods for them. The fitted statistic for multilevel Poisson model and the multilevel negative binomial regression model are shown in Table 4, we note that the AIC and BIC values for both models, the multilevel negative binomial model is better than the multilevel Poisson model, this can also be noticed from Fig 4, since the predicted probabilities from the multilevel negative binomial model is closer to the observed probabilities for each count. Then we can make a conclusion that the multilevel negative binomial model is essentially more appropriate than the multilevel Poisson model for the number of children of Death in EDHS, Ethiopia.

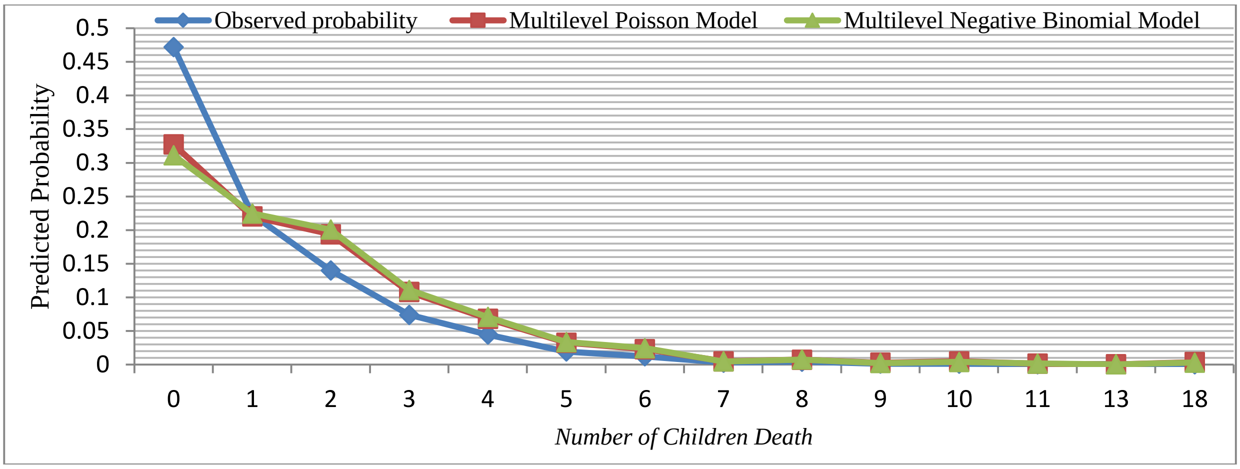


Figure 2. Comparisons of the multilevel Poisson and multilevel NB models using predicted Probabilities

V. DISCUSSION AND CONCLUSION

This paper proposed the multilevel negative binomial model and the multilevel Poisson model as an alternative for handling overdispersion specifically; two types of dispersions for two types of regression models, the multilevel Poisson regression model, and the multilevel negative binomial regression model, the fitting procedures can be carried out easily by using the EM algorithms. The estimations of the dispersion parameter, k, can be implemented by the maximum likelihood method. In this paper, we briefly discussed the goodness of fit measures which were already familiar to those who used multilevel negative binomial model and Multilevel Poisson model. The measures which are applicable to the two model likelihood ratio test, Akaike information criteria (AIC) and Bayesian Schwartz information criteria (BIC).

In this paper the multilevel negative binomial model and the multilevel Poisson regression model were fitted, and compared using illustrative examples of an application datasets from the Ethiopian Demographic and health related survey, and simulation data were used.This paper showed that multilevel Poisson and multilevel negative binomial procedures is similar estimates for the regression parameters, the standard errors for the negative binomial model is larger than the multilevel Poisson model, therefore, the multilevel Poisson overestimates the significance of the regression parameters in the presence of overdispersion. This paper also showed that in the presence of overdispersion, the multilevel Poisson overestimates the significant of the rating factors. The variance of the multilevel negative binomial model is larger than the multilevel Poisson, and this allows the multilevel negative binomial models to handle overdispersion. This study demonstrates the application of the multilevel negative binomial and the multilevel Poisson to death of children; it highlights significance heterogeneity in the regions for children deaths. The method avoids arbitrary trimming and transformation of the data for a single component analysis (Quantin et al., 1999), that is, the assumption of a homogenous children population is no longer required.

This paper used a simulation approach to test the ability of group level random effects to minimize overdispersion and thus recover unbiased parameter estimates in multilevel Poisson models. Simulation result revealed that the appropriateness of employing observations level random effects in mixed models depends on the process generating the overdispersion in the data.
The likelihood ratio result showed that the predictors were significantly associated with the outcome variable ($p<0.000$). We derived a score test statistics MNB for testing homogeneity between and within individuals for clustered count data with over dispersed. We then compared these two statistics using simulations with the two statistics.

The statistic the multilevel negative binomial model is based on a specific over-dispersion model, namely the multilevel negative binomial model and the multilevel Poisson model. The statistic multilevel negative binomial model, in general, shows highly inflated level properties. The statistics multilevel Poisson show some conservative level properties, however, as the values of c and k increase, empirical levels become closer to the nominal level. The power of the statistics multilevel negative binomial model is, in general, larger than the multilevel Poisson model.

In terms of both level and power, there does not seem to be much difference in the performance of the two statistics. The level and power properties of all the statistics, in general, remain similar irrespective of which mechanism of overdispersion is used to generate count data. This also seems to be true irrespective of whether the over-dispersion parameter c is varying or constant. For testing homogeneity between and within individuals for clustered count data with over-dispersion, our recommendation, then, is to use multilevel negative binomial model might be preferable.

REFERENCES

[**1**] Abramowitz and stegun,I.A. 1972. Handbook of Mathematical Functions, New York: Dover Publications, Inc.
[**2**] Bolker BM, Brooks ME, Clark CJ, Geanges SW, Poulsens MHH, WhiteJSS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24: 127-135.

[3] Carrasco , J.L . and Jover, L .2005. Concordance Correlation Coefficient Applied to Discrete Data. Statistics in Medicine, 24: 4021 - 4034.

[4] Collet, D. 2003. Modeling Binary Data. Second edition, Chapaman and Hall.

[5] Collings, J., and Margoline, H. 1985. Testing Goodness of fit test for the Poisson Assumption when Observations are not Identically Distributed. Journal of the American Statistical Association, 80: 411-418.

[6] Cox, D.R. 1983. Some remarks on overdispersion. Biometrika, 70: 269-274.

[7] Crawley MJ.2007. The R book. United Kingdom: John Wiley & Sons Ltd.

[8] Hilbe JM.2011. Negative binomial regression (2nd edition). Cambridge: Cambridge University Press.

[9] Harrison XA.2014. Using observation level random effects to model overdispersion in count data in ecology and evolution. Peer J2:E616 http://doi.org/10.7717/peerj.616

[10] Lee, A.H., Wang, K., Yau, K.K.W., Carrivick, P.J.W. and Stevenson, M.R.2005. Modeling bivariate count series with excess zeros. Mathematical Biosciences 196: 226 –232.

[11] Leung, K.M., Elashoff, R.M., Rees, K.S., Hasan, M.M. and Legorreta, A.P.1998. Hospital and patient related characteristics determining maternity length of stay: a hierarchical linear model approach. American Journal of Public Health 88:377 – 381.

[12] McGilchrist,C.A. 1994. Estimation in generalized mixed models. Journal of Royal Statistical Society B(56): 61 – 69.

[13] Ng, S.K., McLachlan, G.J., Yau,K.K.W. and Lee, A.H.2004. Modeling the distribution of ischemic stroke specification survival time using an EM based mixture approach with random effects adjustment. Statistics in Medicine 23: 2729-2744.

[14] Quantin, C., Sauleau, E., Bolard, P., Mousson,C.,Kerkri, M., Brinet Lecomte, P., Moreau, T. and Dusserie, L.1999. Modeling of high cost patient distribution within renal failure diagnosis related group. Journal of Clinical Epidemiology 52: 251 – 258.

[15] Richards SA.2008. Dealing with over dispersed count data in applied ecology. Journal of Applied Ecology 45: 218-227.

[16] Yau, K.K.,Lee, A.H. and Ng, A.S.K. 2003. Finite mixture regression model with random effects: Application to neonatal hospital length of stay. Computational Statistics & Data Analysis 41: 359 – 366.

[17] Zuur AF, leno EN, Walker NJ, Saveliev AA, Smith GM. 2009 . Mixed effects models and extensions in ecology with R. New York: Springer.