Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Proposed Score Test of Homogeneity of Groups in the Multilevel Poisson Regression Model for Clustered Discrete Data

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

[1]Department of Statistics, College of Science and Technology, Andhra University, Visakhapatnam, India. Email: munistats@gmail.com

[2]Department of Statistics, College of Science and Technology, Andhra University, Visakhapatnam, India.Email: aragaw2018@gmail.com.Mobile: +919063200433.

ABSTRACT

The NB model is useful to analyze discrete data. For clustered discrete data when the observations are correlated from individual subjects, the NB parameter estimates can be severely biased. A score test for testing the random parameter in the over-dispersed clustered discrete data are developed and analyzed for the NB model with the assumption that it is used for model fit under the null hypothesis. Here, we derived the procedure of the fitted multilevel NB model and developed a score test for testing the random parameter; the likelihood ratio tests are used as an alternative test to select the best test statistic in terms of power. To demonstrate our proposed method, a simulation study and an illustrative example are used. The results showed that when the dataset has heterogeneous groups in the clustered discrete data, the multilevel NB model gives a good approximation and correct result in the analysis while NB is clearly not adequate for handling heterogeneous data since it gives wrong results and it is only appropriate and reliable for homogenous groups. From the simulation study, for fixed values of the random and dispersed parameters, when the sample size is increasing the power of the score test is increasing. For large values of the sample size and random parameters, the difference among different tests become trivial in terms of its power. For other cases, the proposed score test is more appropriate for general use because of its high Power.

KEYWORDS: Clustered discretedata; multilevel NB model; Homogeneity; Random effects.

***Corresponding author**

Aragaw Eshetie Aguade

Department of Statistics,

College of Science and Technology,

Andhra University, Visakhapatnam, India.

Email: aragaw2018@gmail.com. Mobile: +919063200433.

1. INTRODUCTION

The conventional regression methods, including multiple linear regression, logistic regression and generalized linear models, assume independence of the observations. In medical sciences, as well as in many other fields, data are hierarchical and independence assumption is conceptually violated. In fact in this sense, a standard negative binomial regression model is a purely fixed effects model. The random coefficients models when the model coefficients are also allowed to vary, more complex mixed effects models may be constructed by nesting levels within one another.

Gaussian distribution is typically used to describe the intercept randomness. The intercept only model sometimes referred to as an empty model or null model; this is the simplest case of the multilevel regression model. This model only contains random groups and random variation within groups. It can be expressed as a model where the dependent is the sum of a general mean, the random effect at the group level, and a random effect at the individual level. The addition of an extra parameter to indicate a randomly distributed intercept classifies the models as a random intercept. It is understood that random effects are the same within each cluster, but they differ between clusters.

The multilevel regression model has become known in the research literature under a different name, such as, "random coefficient model"(de Lecuw & Kreft, 1986 ;Long ford, 1993)[1], "variance component model"(Long ford, 1987)[2], "hierarchical linear model"(Raudenbush & Bryk, 1986, 1988)[3], "mixed effects or mixed model"(Little, Milliken, Straup & Wolfing, 1996)[4] all are very similar and jointly as multilevel regression models. All of them are assuming that there is a hierarchical data set with one single outcome variable that is measured at the lowest level and explanatory at all existing levels (Goldsten, 2003)[5]. Until recently, nearly all discussion and application of multilevel models have been of continuous response models. Binary response models, especially logistic models were introduced about more than a decade. The multilevel mixed models are a comparatively new area of research, with multilevel count models being the most recent, J.M. Hilbe (2007)[6].Comprehensive studies of mixed models are given by Searle, casella, and MC-Culloch (1992)[7], Verbeke and Molenberghs (2000)[8], Raudenbush and Bryk (2002)[9], Demidenko (2004)[10], Hedeker and Gibbons (2006)[11], McCulloch, Searle, and Neuhaus (2008)[12], Rabe- Hesketh , and Skrondal (2012)[13] are good review on applied multilevel count data.

The basic idea of the multilevel analysis is that data sets with a nesting structure that include unexplained variability at each level of nesting are usually not adequately represented by the multiple regression analysis. The reason is that the unexplained variability in single level regression analysis is only the variance of the residual term. The multilevel data has a more complicated structure related to the fact that many populations are concerned with modeling in such data which

embrace one population for every level (Snijders and Bosker, 1999)[14]. The objectives of our study wereto develop a multilevelNB model for clustered discrete data.

2.STATISTICAL METHODS

2.1. The Multilevel Negative Binomial Regression Model

The effects of adding a random component to the linear predictor are to add extra correlation to the model, which in turn induces over dispersion, the level of dependency must be adjusted by the models if the resulting levels of over dispersion are to be accommodated.In this study, the clustering of the data points within geographical regions offers a natural 2- level hierarchical structure of the data, that is, children are nested within regions. The random effects model begin with the same notation as fixed effects models in that heterogeneity parameters is added to the linear predictor. Moreover, the fixed effects parameters β , is now considered to be an iid random parameter rather than a fixed parameter and it is derived from a known probability distribution.

When the outcome variable is a count denoting the number of time that an incident occurred, a Poisson regression model can be accustomed relate the mean number of events to a group of explanatory variables employing a logarithmic link function.The NB regression model be used as

$$log\ (\mu_i) = X_i\beta\ ,\ Y_i \sim NB\ (\mu_i, c) \text{----------(1)}$$

Where X_i denotes a p×1 column matrix of covariates measured with the i^{th} subject, Y_i denotes the count outcome variable measured with the i^{th} subject, β denotes a 1×p row matrix of the regression coefficients and the parameter μ_i denotes the expected or mean number of events for the i^{th} subject given their set of observed covariates. Let Y_{ij} be the response variable for the observation j of group i, i =1, 2… k; j = 1, 2… n_i . Without lose of generalization, consider a two level 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 over dispersion parametersc in a mean over dispersion parameterization, is

$$f(y_{ij}/\mu_{ij}, c_i) = exp\left[\sum_{i=1}^{n_i}\{log\ \Gamma(y_{ij} + c_i^{-1}) - log\ \Gamma(y_{ij} + 1) - log\ \Gamma(c_i^{-1})\} + C(y_{ij}, c_i)\right] \ldots (2)$$

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 with$\alpha_i = \alpha + D^{\frac{1}{2}}u_i$,$c_i$ is the dispersion parameter for group i, and x_{ij} is a $p\times1$ vector of time 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 meanα and variance *D*. Model (2) is an extension of the NB regression model to include normally distributed random effects at different group levels. The standard NB 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$H_0: D = 0$ against the alternative$H_0: D > 0$. This implies that testing homogeneity across groups as well as testing homogeneity with in groups as the intra cluster or within group correlation coefficient assuming that common over dispersion parameter c overall groups or individuals (see carrasco and Jover, 2005)[15].

2.2. Derivations of the Score Test Based on the Multilevel NB Model

We consider a multilevel NB model in (2), our purpose is to develop a score test of homogeneity between and within groups for over dispersed count data. Let Y_{ij} be the response variable for the observation j of group i, i =1, 2… k; j = 1, 2… n_i from the NB distribution denoted by $f(y_{ij}/\mu_{ij}, c_i)$ and given by

$$\frac{\Gamma(y_{ij} + c_i^{-1})}{\Gamma(y_{ij} + 1)\,\Gamma(c_i^{-1})}\left(\frac{1}{1 + c_i\mu_{ij}}\right)^{c^{-1}}\left(\frac{c_i\mu_{ij}}{1 + c_i\mu_{ij}}\right)^{y_{ij}}$$

The mean and the variance of the outcome variable Y_{ij} are

$\mu_{ij} = E\left(Y_{ij}\right) = g`(\theta_{ij}) = exp\left(x_{ij}\beta + z_{ij}\alpha_i\right)$ and $Var(y_{ij}) = \sigma_{ij}^2 = c_i g``(\theta_{ij}) = \mu_{ij}(1 + c\mu_{ij})$

The i^{th} term in the log- likelihood of the multilevel NBmodel in (1) can be written as

$$L_i(\beta, \alpha, c) = \sum_{j=1}^{n_i}\left[\sum_{l=1}^{y_{ij}} log(1 + cl) + y_{ij}\,(x_{ij}`\beta + \left(\alpha + D^{\frac{1}{2}}u_i\right)\right.$$

$$\left. - \left(y_{ij} + c^{-1}\right)log\left(1 + cexp\left(x_{ij}`\beta + \left(\alpha + D^{\frac{1}{2}}u_i\right)\right)\right)\right] \ldots\ldots\ldots (3)$$

To determine the score function, we follow the procedures in the above equation. Then, the first and second derivatives of the log likelihood equation with respect to$\sqrt{D}$ are

$$\frac{\partial}{\partial\sqrt{D}} log\prod_{j=1}^{n_i} f_{ij}\left(y_{ij},\beta,\alpha,c\right) = u_i\sum_{j=1}^{n_i}\frac{\left(y_{ij}-\mu_{ij}\right)}{\left(1+c\mu_{ij}\right)}$$

$$and$$

$$\frac{\partial^2}{\partial^2\sqrt{D}} log\prod_{j=1}^{n_i} f_{ij}\left(y_{ij},\beta,\alpha,c\right) = -u_i^2\sum_{j=1}^{n_i}\frac{\mu_{ij}\left(1+c\mu_{ij}\right)}{\left(1+c\mu_{ij}\right)^2}$$

Therefore, at $D = 0$ the score statistic becomes

$$S_N(\beta,\alpha,c) = \sum_{i=1}^{k}\frac{\partial l_i(\beta,\alpha,c)}{\partial D}\,\Big|\, D = 0$$

As in Jacqmin- Gadda and commenges (1995), using Liang (1987) and Chesher (1984) and after evaluation of the partial derivatives, we obtain the score function evaluated under the null hypothesis of homogeneity. The score test statistic for testing $H_0: D = 0$ for the known nuisance parameters γ and c are as follows.

$$S_N(\beta,\alpha,c) = \frac{1}{2}\sum_{i=1}^{k}\left\{\left[\sum_{j=1}^{n_i}\frac{\partial logf_{ij}}{\partial\alpha_i}(\beta,\alpha,c)\right]^2 + \sum_{j=1}^{n_i}\frac{\partial^2 logf_{ij}}{\partial\alpha_i^2}(\beta,\alpha,c)\right\}$$

Where

$$\frac{\partial logf_{ij}}{\partial\alpha_i}(\beta,\alpha,c) = \sum_{j=1}^{n_i}\left(\frac{y_{ij}-\mu_{ij}}{1+c\mu_{ij}}\right) \; and \; \frac{\partial^2 logf_{ij}}{\partial\alpha_i^2}(\beta,\alpha,c) = -\sum_{j=1}^{n_i}\left[\frac{\mu_{ij}\left(1+cy_{ij}\right)}{\left(1+c\mu_{ij}\right)^2}\right]$$

Therefore, the score test statistic will be

$$S_N(\beta,\alpha,c) = \frac{1}{2}\sum_{i=1}^{k}\left\{\left[\sum_{j=1}^{n_i}\frac{\left(y_{ij}-\mu_{ij}\right)}{\left(1+c\mu_{ij}\right)}\right]^2 - \sum_{j=1}^{n_i}\frac{\mu_{ij}\left(1+cy_{ij}\right)}{\left(1+c\mu_{ij}\right)^2}\right\}$$

Then the score test statistic for testing $H_0: D = 0$ for the known nuisance parameters γ and c is

$H_{NB} = S_N^2(\beta,\alpha,c)/(I_{DD} - AB^{-1}A^T)$.. *(3)*

Now, the asymptotic variance as $k \to \infty$, of $S_N(\beta,\alpha,c)$ under H_0 (Cox and Hinkley, 1974) is

$I = I_{DD} - AB^{-1}A^T$ Where $I_{DD} = \sum_{i=1}^{k} E\left[\frac{\partial l_i}{\partial D}\,\big|\, D = 0\right]^2$ is a scalar and $A = [A_1, A_2]$,

$$A_1 = \sum_{i=1}^{k} E\left[\left(\frac{\partial l_i}{\partial D}\right)\left(\frac{\partial l_i}{\partial\gamma}\right)^{`}\Big|\, D = 0\right] \text{ is a } 1\times(p+1) \text{ vector,}$$

$$A_2 = \sum_{i=1}^{k} E\left(\frac{-\partial^2 l_i}{\partial D\partial c}\,\Big|\, D = 0\right) \text{ is a scalar,}$$

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, B_{11} = \sum_{i=1}^{k} E\left[\left(\frac{\partial l_i}{\partial \gamma_s}\right)\left(\frac{\partial l_i}{\partial \gamma_r}\right)` \mid D = 0\right] isa(p+1) \times (p+)\text{matrix}$$

$$B_{12} = B_{12} = \sum_{i=1}^{k} E\left[\left(\frac{\partial l_i}{\partial \gamma}\right)\left(\frac{\partial l_i}{\partial c}\right)` \mid D = 0\right] \text{ is a } (p+1) \times 1 \text{ vector}$$

$$\text{and} B_{22} = \sum_{i=1}^{k} E\left(\frac{-\partial^2 l_i}{\partial c^2} \mid D = 0\right) \text{isa scalar.}$$

2.3. *Parametric Estimation of the Score test Based on the Multilevel Negative Binomial Model*

Now we need to evaluate the variance of the score function (S) defined as

$\text{Var}(S) = I_{DD} - AB^{-1}A^T$. The i^{th} summand of I_{DD} can be written as

$$E\left(\frac{\partial l_i}{\partial D} \mid D = 0\right)^2 = \frac{1}{4} E\left\{\left[\sum_{j=1}^{n_i} \frac{(y_{ij} - \mu_{ij})}{(1 + c\mu_{ij})}\right]^2 - \sum_{j=1}^{n_i} \frac{\mu_{ij}(1 + cy_{ij})}{(1 + c\mu_{ij})^2}\right\}^2$$

Where

$$a_i = \left[\sum_{j=1}^{n_i} \frac{(y_{ij} - \mu_{ij})}{(1 + c\mu_{ij})}\right]^2 \text{ and} b_i = \sum_{j=1}^{n_i} \frac{\mu_{ij}(1 + cy_{ij})}{(1 + c\mu_{ij})^2}$$

To derive quantitative such as$E(a_i)^2$, we need some basic moment results from the

$\text{NB}\left(\mu_{ij}, c\right)$distribution. Let$U = \frac{(y_{ij} - \mu_{ij})}{(1 + c\mu_{ij})}$. Then it can be shown that the first four cumulates of U are

$$k_1 = 0, \quad k_2 = \frac{\mu_{ij}}{(1 + c\mu_{ij})}, \quad k_3 = \frac{(\mu_{ij} + 2c\mu_{ij}^2)}{(1 + c\mu_{ij})^2}, \quad k_4 = \frac{(\mu_{ij} + 6c\mu_{ij}^2 + 6c^2\mu_{ij}^3)}{(1 + c\mu_{ij})^3}$$

Appling the first four cumulates results we obtain

$$E(a_i^2) = \sum_{j=1}^{n_i} \left[\frac{\mu_{ij} + 6c\mu_{ij}^2 + 6c^2\mu_{ij}^3 + 6c^3\mu_{ij}^4}{(1 + c\mu_{ij})^3} + 3\left(\frac{\mu_{ij}}{1 + c\mu_{ij}}\right)^2\right]$$

$$E(a_i b_i) = \sum_{j=1}^{n_i} \frac{\mu_{ij}(\mu_{ij} + c\mu_{ij} + 2c\mu_{ij}^2 + 3c^2\mu_{ij}^2 + 2c^3\mu_{ij}^3 + c^2\mu_{ij}^3)}{(1 + c\mu_{ij})^4}$$

and

$$E(b_i^2) = \sum_{j=1}^{n_i} \left[\frac{(\mu_{ij}^2 + c^2\mu_{ij}^4 + c^2\mu_{ij}^3 + c^3\mu_{ij}^4 + 2c\mu_{ij}^3)}{(1 + c\mu_{ij})^4}\right] + \sum_{j=1}^{n_i} \sum_{j` \neq j}^{n_i} \left[\frac{\mu_{ij}\mu_{ij}`}{(1 + c\mu_{ij})(1 + c\mu_{ij}`)}\right]$$

Finally, we obtain the variance of the score function(S) as follows

$$E\left(\frac{\partial l_i}{\partial D} \mid D = 0\right)^2$$

$$= \frac{1}{4}\left\{\sum_{j=1}^{n_i}\frac{\left(\mu_{ij} + 6c\mu_{ij}^2 + 6c^2\mu_{ij}^3\right)}{\left(1 + c\mu_{ij}\right)^3} + 3\sum_{j=1}^{n_i}\left(\frac{\mu_{ij}}{1 + c\mu_{ij}}\right)^2\right.$$

$$- 2\sum_{j=1}^{n_i}\left[\frac{\mu_{ij}\left(\mu_{ij} + c\mu_{ij} + 2c\mu_{ij}^2 + 3c^2\mu_{ij}^2 + 2c^3\mu_{ij}^3 + c^2\mu_{ij}^3\right)}{\left(1 + c\mu_{ij}\right)^4}\right]$$

$$\left. + \sum_{j=1}^{n_i}\left[\frac{\left(\mu_{ij}^2 + c^2\mu_{ij}^4 + c^2\mu_{ij}^3 + c^3\mu_{ij}^4 + 2c\mu_{ij}^3\right)}{\left(1 + c\mu_{ij}\right)^4}\right] + \sum_{j=1}^{n_i}\sum_{j`\neq j}^{n_i}\left[\frac{\mu_{ij}\mu_{ij}`}{\left(1 + c\mu_{ij}\right)\left(1 + c\mu_{ij}`\right)}\right]\right\}$$

Now,

$$\frac{\partial l_i(\beta, \alpha, c)}{\partial \gamma}\Big|_{D=0} = \sum_{j=1}^{n_i}\left[\left(\frac{y_{ij} - \mu_{ij}}{1 + c\mu_{ij}}\right)W_{ij}`\right]$$

Then

$$E\left[\left(\frac{\partial l_i}{\partial D}\right)\left(\frac{\partial l_i}{\partial \gamma}\right)` \mid D = 0\right] = \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\mu_{ij}}{\left(1 + c\mu_{ij}\right)}w_{ij}`$$

hence we obtain

$$A_1 = \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\mu_{ij}}{\left(1 + c\mu_{ij}\right)}w_{ij}`$$

And we need also to evaluate A_2

$$\frac{\partial l_i(\beta, \alpha, c)}{\partial D}\Big|_{D=0} = \frac{1}{2}\sum_{i=1}^{k}\left\{\sum_{j=1}^{n_i}\left(\frac{y_{ij} - \mu_{ij}}{1 + c\mu_{ij}}\right)^2 + \sum_{j=1}^{n_i}\sum_{j`\neq j}^{n_i}\frac{\left(y_{ij} - \mu_{ij}\right)\left(y_{ij}` - \mu_{ij}`\right)}{\left(1 + c\mu_{ij}\right)\left(1 + c\mu_{ij}`\right)} - \sum_{j=1}^{n_i}\frac{\mu_{ij}\left(1 + cy_{ij}\right)}{\left(1 + c\mu_{ij}\right)^2}\right\}$$

Then

$$\left(\frac{-\partial^2 l_i(\beta, \alpha, c)}{\partial D\partial c}\Big|_{D=0}\right)$$

$$= \frac{1}{2}\sum_{i=1}^{k}\left\{-\sum_{j=1}^{n_i}\frac{2\mu_{ij}\left(y_{ij} - \mu_{ij}\right)^2}{\left(1 + c\mu_{ij}\right)^3} - \sum_{j=1}^{n_i}\sum_{j`\neq j}^{n_i}\frac{\mu_{ij}\mu_{ij}`\left(y_{ij} - \mu_{ij}\right)\left(y_{ij}` - \mu_{ij}`\right)}{\left(1 + c\mu_{ij}\right)^2\left(1 + c\mu_{ij}`\right)^2}\right.$$

$$\left. - \left[\sum_{j=1}^{n_i}\frac{\mu_{ij}y_{ij}\left(1 + c\mu_{ij}\right) - 2\mu_{ij}^2\left(1 + cy_{ij}\right)}{\left(1 + c\mu_{ij}\right)^3}\right]\right\}$$

Therefore

$$A_2 = \frac{1}{2}\sum_{i=1}^{k}\left(\frac{\mu_{ij}}{1 + c\mu_{ij}}\right)^2 \; and \; B_{11} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{\mu_{ij}}{1 + c\mu_{ij}}\right)x_{ij}`x_{ij}$$

The diagonal elements of B_{12} and B_{12} will be zero. We now obtain B_{22}. The partial derivatives of the log likelihood function of the negative binomial with respect to c is given by

$$\left(\frac{\partial l_i}{\partial c} \mid D = 0\right) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\sum_{l=0}^{y_{ij}-1}\left(\frac{l}{1+cl}\right) - \left(y_{ij} + c^{-1}\right)\left(\frac{\mu_{ij}}{1+c\mu_{ij}}\right) + c^{-2}log\left(1+c\mu_{ij}\right)\right]$$

$$\left(\frac{\partial^2 l_i}{\partial c^2} \mid D = 0\right) = -\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\sum_{l=0}^{y_{ij}-1}\left(\frac{l}{1+cl}\right)^2 + 2c^{-3}log\left(1+c\mu_{ij}\right) - \frac{2c^{-2}\mu_{ij}}{\left(1+c\mu_{ij}\right)} - \frac{\mu_{ij}^2\left(y_{ij}+c^{-1}\right)}{\left(1+c\mu_{ij}\right)^2}\right]$$

Following Fisher (1941) and collings (1981) the above equations can be simplified as

$$E\left(-\frac{\partial^2 l_i}{\partial c^2}\right) = c^{-4}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\sum_{l=0}^{\infty}\frac{l!\left(cq_{ij}\right)^{l+1}}{(l+1)d_l} + \frac{c^2\mu_{ij}}{1+c\mu_{ij}} - \frac{c^2\mu_{ij}}{1+c\mu_{ij}}\right]$$

Thus

$$E\left(-\frac{\partial^2 l_i}{\partial c^2}\right) = c^{-4}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\sum_{l=0}^{\infty}\frac{l!\left(cq_{ij}\right)^{l+1}}{(l+1)d_l}\right]$$

Where $q_{ij} = \frac{c\mu_{ij}}{(1+c\mu_{ij})}$ and $d_l = \prod_{j=1}^{l}(1+jc)$.

The regression (γ) parameter and dispersion(c) parameter in H_{NB} given equation (4.4) are replaced by their maximum likelihood estimates, obtained from the negative binomial regression model under the null hypothesis (see Lawless (1987)). The score test statistic H_{NB} then reduce to

$$\frac{\left(\sum_{i=1}^{k}\left\{\left[\sum_{j=1}^{n_i}\frac{(y_{ij}-\mu_{ij})}{(1+c\mu_{ij})}\right]^2 - \sum_{j=1}^{n_i}\frac{\mu_{ij}\left(1+cy_{ij}\right)}{\left(1+c\mu_{ij}\right)^2}\right\}\right)^2}{\left\{\frac{1}{4}\sum_{i=1}^{k}\left[\begin{array}{l}\sum_{j=1}^{n_i}\frac{\left(\mu_{ij}+6c\mu_{ij}^2+6c^2\mu_{ij}^3\right)}{\left(1+c\mu_{ij}\right)^3} + 3\sum_{j=1}^{n_i}\left(\frac{\mu_{ij}}{1+c\mu_{ij}}\right)^2 - \sum_{j=1}^{n_i}\left[\frac{\left(\mu_{ij}^2+c^2\mu_{ij}^4+c^2\mu_{ij}^3+c^3\mu_{ij}^4+2c\mu_{ij}^3\right)}{\left(1+c\mu_{ij}\right)^4}\right] \\ 2\sum_{j=1}^{n_i}\left[\frac{\mu_{ij}\left(\mu_{ij}+c\mu_{ij}+2c\mu_{ij}^2+3c^2\mu_{ij}^2+2c^3\mu_{ij}^3+c^2\mu_{ij}^3\right)}{\left(1+c\mu_{ij}\right)^4}\right] + \sum_{j=1}^{n_i}\sum_{j`\neq j}^{n_i}\left(\frac{\mu_{ij}\mu_{ij`}}{(1+c\mu_{ij})(1+c\mu_{ij`})}\right)\end{array}\right]\right.$$
$$\left. - \left[\begin{array}{c}\left(\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\mu_{ij}}{(1+c\mu_{ij})}W_{ij}`\right)\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{\mu_{ij}}{1+c\mu_{ij}}\right)x_{ij}`x_{ij}\right)^{-1}\left(\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\mu_{ij}}{(1+c\mu_{ij})}W_{ij}`\right)` \\ + \left(\frac{1}{2}\sum_{i=1}^{k}\left(\frac{\mu_{ij}}{1+c\mu_{ij}}\right)^2\right)^2\left(c^{-4}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\sum_{l=0}^{\infty}\frac{l!(cq_{ij})^{l+1}}{(l+1)d_l}\right]\right)^{-1}\end{array}\right]\right\} \quad ..(4)$$

Now the maximum likelihood estimate of γ can be estimated iteratively by fisher`s scoring method from the following equations.

$\gamma^{(t+1)} = [(WQ^{-1}W^T)^{-1}WQZ]^{(t)}$, $t = 1, 2,3,\ldots$, where $Q = diag\left(\frac{\hat{\mu}}{1+c\hat{\mu}}\right)$ is an $N\times N$ matrix and $Z = W^T\hat{\gamma}^{(t+1)} + \frac{Y-\hat{\mu}}{\hat{\mu}}$, t = 1, 2, 3,…, is an $N\times 1$ vector. Fisher`s scoring equation to estimate c is given by $c^{(t+1)} = c^{(t)} + \left[B_{22}^{-1}\left(\frac{\partial l}{\partial c}\right)\right]^{(t)}$, where l is the log likelihood function given by H_{NB} with $D = 0$. Note

that these two equations must be solved simultaneously to get the maximum likelihood estimates of the parameters γ and c under the null hypothesis.

2.4. Fisher Scoring Method for the Estimation of β and the Dispersion Parameter C

Define the function $\eta_{ij} = log\mu_{ij}$. Then, $\frac{\partial \eta_{ij}}{\partial \mu_{ij}} = \frac{1}{\mu_{ij}}$, also let $V_{ij} = Var(y_{ij}) = \mu_{ij} + c\mu_{ij}^2$. Further we define $w_{ij} = \left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right)^2 V_{ij}^{-1} = \frac{\mu_{ij}}{(1+c\mu_{ij})}$. Then the score equation for βs, s=1, 2…p. can be written as $u_s = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{\mu_{ij}}{(1+c\mu_{ij})}\right)\left(\frac{y_{ij}-\mu_{ij}}{\mu_{ij}}\right)x_{ijs}$ And the fisher information matrix for βis $I = E\left(-\frac{\partial^2 l_i}{\partial\beta_S\beta_r}\right) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij}x_{ijs}x_{ijr}$`. Nowwe define $u = (u_1, u_2, u_3, \dots, u_p)$`,

$y = (y_{11}, \dots, y_{1n_1}, \dots, y_{k1}, \dots, y_{kn_k})$`. $\mu = (\mu_{11}, \dots, \mu_{1n_1}, \dots, \mu_{k1}, \dots, \mu_{kn_k})$` and $N = \sum_{I=1}^{K} n_i$. Further, let x be a $N \times N$ diagonal matrix with elements w_{ij}. Then the score equations in vector notation can be written as $u = X`W\left(\frac{Y-\mu}{\mu}\right)$ and the fisher information matrix can be written as $I = (X`WX)$. The fisher scoring equations for solving for the regression parameters β become

$I^{(t)}\beta^{(t+1)} = I^{(t)}\beta^{(t)} + u^{(t)}, (X`WX)\beta^{(t+1)} = X`WZ.$

Thus $\beta^{(t+1)} = \left((X`WX)\right)^{-1}(X`WZ)$` … … … … … … … … … … … … … … … … … … (5)

Where *t=0, 1, 2…* and $Z = X`W\left(\frac{Y-\mu}{\mu}\right)$. The Fisher scoring equation to solving for c is

$$c^{(t+1)} = c^{(t)} + I_c^{-1}u_{(t)} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (6)$$

Where $I_{CC} = E\left(-\frac{\partial^2 l_i}{\partial c^2}\right)$ *and* $v = \frac{\partial l}{\partial c}$ as defined before, the maximum likelihood estimates of β and c are obtained by iterating between equations *(5) and (6)* after putting in initial values.

3. SIMULATION STUDY

In this section we consider a simulation study and assumed thatthe random effect is the intercept($z_{ij} = 1$). Wegeneratedsets of count data viathe multilevel NB distribution of the response variable with different number of groups and individuals according to the variance (D) of the group-specific random effects and for different values of the over dispersion parameter c. The sample comprised k = 5,10,20,50 groups with $n = 10,20,50\ and\ 100$ observations. Each simulated experiment for level and power was based on 10000 simulated samples. The following log-linear model for the response variable is considered.

$$\log(\mu_{ij}) = 0.8x_{1ij} + 0.5u_i - 1.5 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.8)$$

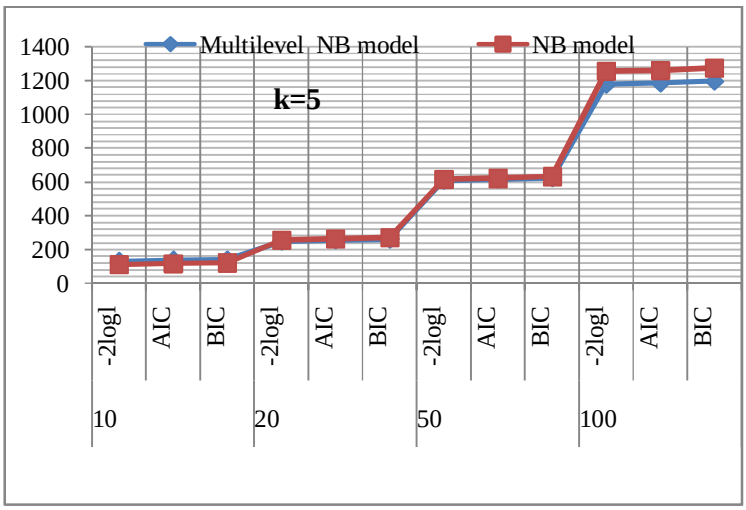
For i =1, 2… k and j = 1, 2… n_i .The variable x_1 is subjected-specific effect and $u_i$is the group specific effect simulated from a standard normal distribution. To simulate correlated data, we added a group specific random intercept in the model for the response variable which is$\alpha_{ij} = \alpha + D^{\frac{1}{2}}u_i$, where, $u_i$is a random effect. Therefore, the random effects are normally distributed with mean α and variance D. Our objective is to test the power of the proposed score test and to compare the proposed model with the standard NB regression model. Table 1 displays the goodness of fit test for the proposed model when the data are simulated from theNB distribution under the hypothesis of homogeneity with common overdispersion parameter c.

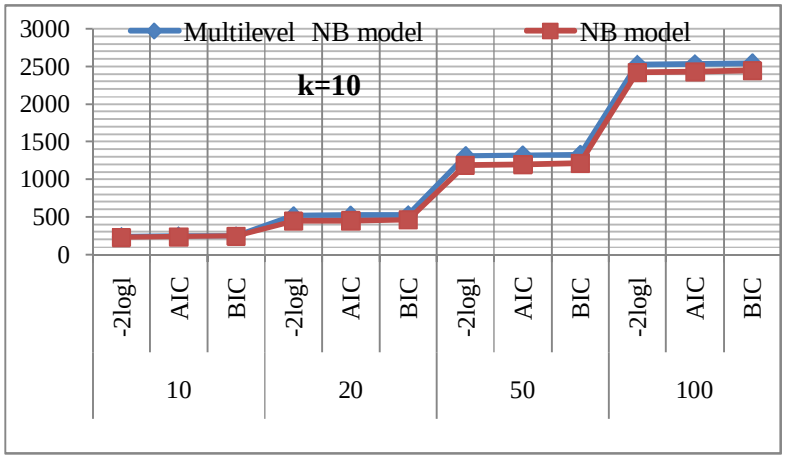
Table1: The goodness of fit tests of the multilevel NB and NB models via simulated data

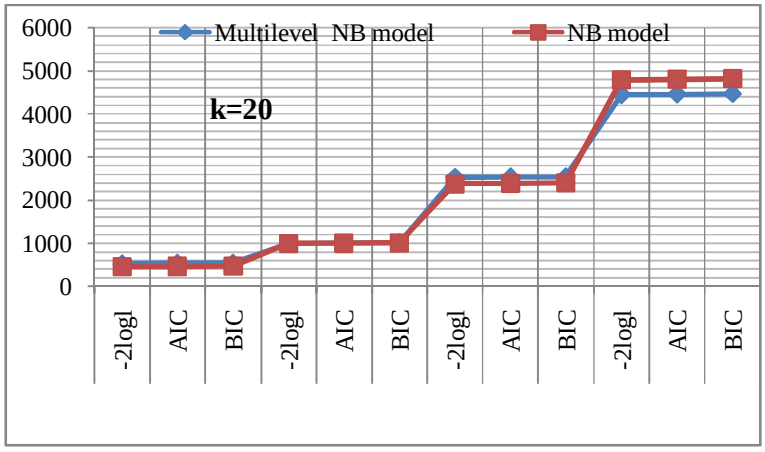
Group(k)	Observation(n)	Model	-2loglikelihood	AIC	BIC
5	10	Multilevel NB model	129.2	137.2	138.4
		NB model	111.1	117.1	122.9
	20	Multilevel NB model	248.1	256.1	260.1
		NB model	256.1	262.1	270
	50	Multilevel NB model	608	616	623.7
		NB model	614.8	620.8	631.3
	100	Multilevel NB model	1175.5	1183.5	1194
		NB model	1253.4	1259.4	1272.1
10	10	Multilevel NB model	243.4	251.4	252.6
		NB model	234.9	240.9	248.8
	20	Multilevel NB model	521.7	529.7	533.6
		NB model	447.2	453.2	463.1
	50	Multilevel NB model	1315	1323	1330.4
		NB model	1194.8	1200.8	1213.4
	100	Multilevel NB model	2524.7	2532.7	2542.9
		NB model	2424.5	2430.5	2445.2
20	10	Multilevel NB model	531.5	539.5	540.7
		NB model	460.9	466.9	476.8
	20	Multilevel NB model	999.3	1007.3	1011.3
		NB model	1003.7	1009.7	1021.6
	50	Multilevel NB model	2533.8	2541.8	2549.5
		NB model	2383.8	2389.8	2404.5
	100	Multilevel NB model	4453.7	4461.7	4472
		NB model	4794.3	4800.3	4817.1
50	10	Multilevel NB model	1336.2	1344.2	1345.4
		NB model	1250.2	1256.2	1268.8
	20	Multilevel NB model	2374.6	2382.6	2386.6
		NB model	2324.2	2330.2	2344.9
	50	Multilevel NB model	6566.5	6574.5	6582
		NB model	6648.2	6654.2	6671.6
	100	Multilevel NB model	11898	11906	11916
		NB model	11984	11990	12010

In the Table 1 we investigated how the information criteria perform the model selection problems via simulations. The result revealed that AIC was superior to BIC in all the sample data set.

When the sample size was large, except (k=10) the multilevel NB regression model is better than the NB regression model, where as in small sample size the NB regression model is better than the multilevel NB regression model. From the simulations results we found that the performance of the criteria depends on sample size and model complexity.







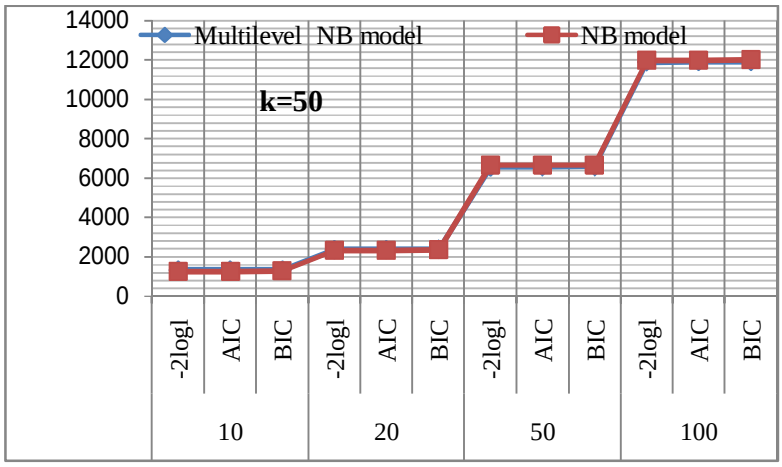


Figure 1. Goodness of fit tests of the NB and MNB models via simulated data

From Table 2 note that as type I error probability increases power increases and also as the number of observations in each group increases power increases. When the values of D increases from 0.15to 0.40 the power decreases.

Table 2: Empirical power of the score test based on 1000 replications generated from the multilevel ZINB model under the hypothesis of homogeneity

Cluster (k)	Subject (n)	D=0.05			D=0.10			D=0.15		
		Significance level			Significance level			Significance level		
		α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
5	10	0.99889	0.99696	0.98330	0.99487	0.99487	0.97448	0.93093	0.87852	0.70945
	20	0.99968	0.99903	0.99348	0.98300	0.96446	0.88280	0.91172	0.84997	0.66291
	50	0.98788	0.97371	0.90699	0.99748	0.99361	0.96954	0.97386	0.94800	0.84372
	100	0.99786	0.99448	0.97294	0.98512	0.96842	0.89292	0.98068	0.96020	0.87226
10	10	0.99327	0.98447	0.93826	0.99956	0.99870	0.99168	0.97943	0.95791	0.86673
	20	0.98980	0.97745	0.91741	0.98763	0.97323	0.90568	0.98525	0.96867	0.89356
	50	0.99261	0.98310	0.93403	0.99892	0.99703	0.98364	0.99343	0.98481	0.93933
	100	0.99761	0.99391	0.97067	0.98276	0.96401	0.88168	0.97592	0.95162	0.85197
20	10	0.99972	0.99914	0.99413	0.99994	0.99978	0.99816	0.99940	0.99828	0.98955
	20	0.99622	0.99076	0.95908	0.99911	0.99751	0.98583	0.99999	0.99995	0.99950
	50	0.99268	0.98326	0.93453	0.98622	0.97050	0.89838	0.98468	0.96760	0.89080
	100	0.99439	0.98682	0.94574	0.99324	0.98441	0.93808	0.97896	0.95706	0.86471
50	10	0.9688	0.93931	0.82463	1.00000	0.99998	0.99980	0.99751	0.99366	0.96973
	20	0.99346	0.98486	0.93948	0.99993	0.99977	0.99808	0.99779	0.99433	0.97232
	50	0.99742	0.99347	0.96899	0.98938	0.97663	0.91510	0.97439	0.94893	0.84583
	100	0.99617	0.99065	0.95867	0.98731	0.97259	0.90397	0.97162	0.94408	0.83500

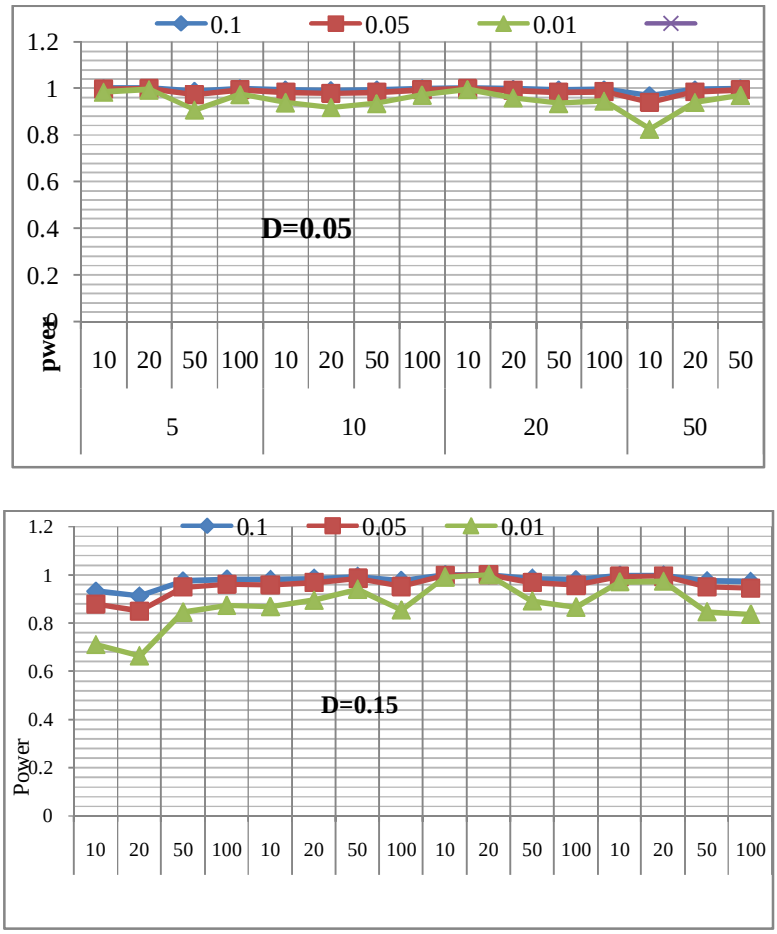


Figure 2. Empirical power of the score test of the MNB model via simulation

4. APPLICATION STUDY

In this chapter, we focused on multilevel negative binomial regression model to take account of the coefficient of regression and random parameters in NB counts with overdispersion. For estimating the parameters of a score test, we used an EM algorithm in the multilevel NB regression model against the standard NB regression model, and for testing the significance of regression coefficients. The demographic and health related survey of Ethiopia (EDHS) data is used to illustrate the proposed score test. The number of children deaths is an outcomevariableand the others nine are predictor's variables.

Table 3: FittedMultilevel NB Model with covariates via EDHS

Number of children	Negative Binomial				Multilevel negative binomial			
	Estimate	S.E	Z-value	p-value	Estimate	S.E.	Z-value	P-value
Residence	.4442592	.0303513	14.64	0.000 *	.2629343	.0341172	7.71	0.000 *
Educ. level	.0057544	.0033045	1.74	0.082	.004331	.0033312	1.30	0.194
Toiletfac	.0619004	.001014	61.05	0.000*	.0621577	.0010159	61.19	0.000 *
Religion	-.049073	.0025395	-19.32	0.000*	-.0483643	.0026752	-18.08	0.000 *
HHSMembers	-.6220319	.0279866	-22.23	0.000 *	-.612779	.028402	-21.58	0.000 *
Age mother	.4740397	.0168476	28.14	0.000 *	.4060037	.0201374	20.16	0.000 *
Current Mari	-.0154326	.098995	-0.16	0.876	.0183504	.0995902	0.18	0.854
Agemarriage	.0615666	.0163336	3.77	0.000 *	.0012938	.0166597	0.08	0.938
Sourcdrinkwat	-.0129167	.0231555	-0.56	0.577	-.0179391	.0237627	-0.75	0.450
Constant	-2.173634	.0830942	-26.16	0.000 *	-1.788503	.1096869	-16.31	0.000*
/lnalpha	-.6424236	.0263217			-.7008252	.0272103	-25.76	0.000*
alpha	.526016	1.026671			0.4961757	1.027584		
Region|var(cons)					.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 0.01levels.

From Table3 explained those predictor variables significantly association with the outcome variable. It is observed that there exist significance differences between the β coefficients of these two models for each of the explanatory variables except education level, current marital status of women, age of mothers for first marriage, and sources of drinking water were found to be significant variation in the deaths of children among regions. In the multilevel NB analysis, a two-level structure is used with regions as the second-level unit and children as the first-level unit. The nesting structure is children within regions that resulted in a set of 11 regions with a total of 25420 mothers. A chi-square test statistic was applied to assess heterogeneity between regions. The test yield LRT = 184.55, P < 0.000. Thus, there is evidence for heterogeneity among regions with respect to deaths.

The difference in β coefficients estimated from a multilevel negative binomial model and the standard negative binomial model arises because of the addition of the random effects. The analysis of the random intercept parameter result revealed that deaths of children varied among regions.In the estimate of σ_u^2 is 0.05 with standard error 0.02. The estimate of the variance component σ_u^2drops down to 0.05 compared with multilevel Poisson model, which is not surprising given that now we have additional parameters that control the variability of the data. Because the conditional over dispersion α is assumed to be greater than 0, it is parameterized on the log scale, and its log estimate is reported as/ lnalpha in the output. In our model $\hat{\alpha} = \exp(-0.64) = 0.53$.we can also compute the conditional over dispersion in this model by using the corresponding formula $\exp(0.05) * (1 + 0.53) - 1 = 0.61$. The reported likelihood ratio test showed that there is enough variability between regions to favour the a multilevel negative binomial model over NB model without random effects.

Table 4: Goodness of fit tests of the NB and multilevel NB models via EDHS, 2005

Model	Obsll(null	ll(model)	df	AIC	BIC
Multilevel NB	-	-34673.53	12	69371.05	69468.77
NB	-38532.49	-34855.01	11	69732.03	69821.6

The multilevel negative binomial model can be considered as a parametricversion of assessing heterogeneity among regions with respect to Death`s of children.Moreover, based on the AIC values,themultilevel NB model is better than NB model.

Table 5: The observed and predicted probabilities of the fitted models via EDHS

Number of Children Death	Observed frequency	Observed probability	Predicted probability	
			Negative Binomial Model	Multilevel Negative Binomial Model
0	12000	0.472069237	0.31454695	0.311111111
1	5680	0.223446105	0.221168436	0.224713805
2	3560	0.140047207	0.198307401	0.200808081
3	1880	0.073957514	0.109837762	0.111245791
4	1140	0.044846577	0.070427066	0.070841749
5	500	0.019669552	0.033427988	0.033670034
6	320	0.012588513	0.024511177	0.024579125
7	80	0.003147128	0.005419803	0.004915825
8	120	0.004720692	0.007900621	0.007676768
9	40	0.001573564	0.00283033	0.002424242
10	40	0.001573564	0.005210895	0.004040404
11	20	0.000786782	0.001710156	0.001818182
13	20	0.000786782	0.000530449	0.00047138
18	20	0.000786782	0.004170959	0.003232323

The predicted probabilities for NB and the multilevel NB regression model are presented in Table 5 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. From Fig 3, since the predicted probabilities from multilevel NB model is closer to the observed probabilities for each count. Then we conclude that the multilevel NB model is essentially more appropriate than the NB model for predicting the number of deaths of children in EDHS, Ethiopia.

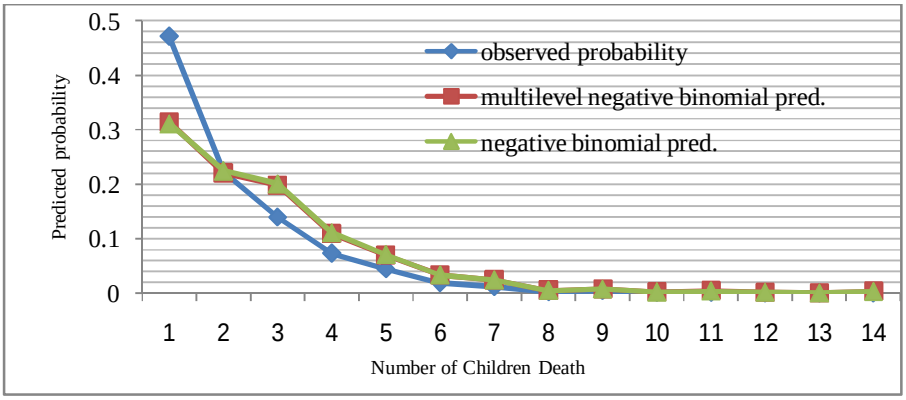


Fig 3. Compressions of Models with covariates via Predicted Probabilities

5. DISCUSSION AND CONCLUSION

In the Table3, the multilevel NB regression model identified the variation among eleven regions in the deaths of children less than eighteen years.The results revealed that place of residences of mother, toilet facility, Religions, size of household members, age of mothers were found to bethe main determining factors for deaths of children under eighteen in Ethiopia. The likelihood ratio result showed that the predictors were significantly associated with the outcome variable ($p<0.000$).

The level and power properties of the statistics, in general, remainsimilarirrespectiveofwhichmechanismofover- dispersion is use dtogeneratecountdata. Thisalsoseems tobetrueirrespectiveofwhethertheover-dispersionparameter c is varyingorconstant. For testinghomogeneitybetweenand within individuals forclusteredcountdatawithover-dispersion, our recommendation,then,istousethe multilevel NB model, so, the proposed model is more preferablethan the standard NB model.

6. REFERENCE

1. DeLecuw, J., & Kreft, I. Random coefficient models for multilevel analysis. Journal of Educational Statistics. 1986; 11: 57-85.
2. Longford, N.T. Random Coefficient Models: Oxford Clarendon Press; 1993.
3. Longford, N.T. A fast scoring algorism for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika. 1987; 74: 817 – 827.
4. Raudenbush, S. W., &Bryk, A. S. A hierarchical model for studying school effects. Sociology of Education.1986; 59: 1-17.
5. Raudenbush, S. W., &Bryk, A. S. Methodological advances in analyzing the effects of schools and classrooms on student learning. Washington, DC: American Educational Research Association.1988;15: 423-475.
6. Littell, R. C., G. A. Milliken, W. W. Stroup, & R. D. Wolfinger. SAS System for Mixed Models. Cary, C: SAS Institute, Inc; 1996.
7. H. Goldsten. Multilevel Statistical Models, 3rd editions. Edward Arnold, London; 2003.
8. Hilbe, J. M. Negative binomial regression. Cambridge, UK: Cambridge University Press; 2007.
9. Searle S.R., Casella, G. and McCulloch, C.E. Variance components. Wiley, New York;1992.
10. Verbeke, G. and Molenberghs, G. Linear mixed models for longitudinal data, Springer Series in Statistics: Springer-Verlag, New-York ;2000.
11. Raudenbush, S.W., &Bryk, A.S. Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage ;2002.
12. Demidenko,E. Mixed Models: Theory and applications, Wiley ; 2004.

13. Hedeker D, Gibbons RD. Longitudinal data analysis. New Jersey: John Wiley & Sons. 2006; 337: ISBN 978-0-471-42027-9
14. McCulloch, C. E., Searle, S. R. &Neuhaus, J. M. Generalized, Linear, and Mixed Models. Hoboken, New Jersey: John Wiley & Sons, Inc., 2nd ed ;2008.
15. Rabe-Hesketh,S., Skrondal, A. and Zheng, X., Generalized multilevel structural equation modeling. In Hoyle, R. (Ed.). Handbook of Structural Equation Modelling. Guilford Press. 2012; 512-531.
16. Snijders, T., &Bosker, R., Multilevel analysis: An introduction to basic and advanced multilevel modeling. London/Thousand Oaks/New Delhi: SAGE Publications Ltd ;1999.
17. Josep L. Carrasco and Lluis Jover, Concordancecorrelationcoefficientappliedtodiscretedata. Bioestadstica; Departament de SalutPublica; Facultat de Medicina; Universitat de Barcelona; Casanova.2005; 143: E-08036.
18. Jacqmin- Gadda and commenges,H. Tests of Homogeneity for generalized linear Models. Journal of the American Statistical Association. 1995;90: 1237 – 1246.
19. Liang , K.Y., A locally most powerful test for homogeneity with many strata, Biometrica. 1987 ;74: 259 – 264.
20. Chesher. A. Testing for Neglected Heterogeneity. Econometrical. 1984;52: 865 – 872.
21. Cox , D.R. and Hinkley, D.V. Theoretical statistic, Chapman and Hall, London ;1974.
22. Fisher, R. A., Average excess and average effect of a gene substitution. Annals of Eugenics. 1941; 11: 53–63.
23. Collings, B. The negative binomial distribution: an alternative to the Poisson. Unpublished Ph.D. thesis. University of North Carolina at Chapel Hill (1981).
24. Lawless,J.F.Negative binomial and mixed Poisson regression. The Canadian Journal of statistics, 1987; 15: 209 – 225.
25. Rabe-Hesketh, S, and Skrondal, A. Multilevel and Longitudinal Modelling using Stata, Stata Press, Stata Corp, College Station, Texas ;2005.