Analysis of The Mixed Effects Regression Model for Clustered Count Response Data

Aragaw Eshetie Aguade (aragaw2018@gmail.com)
University of Gondar College of Natural and Computational Sciences
B.Muniswamy Begari
Andhra University

Research

Keywords: Clustered count response data, Score test, homogeneity, MPM, and GLMM

Posted Date: February 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-157733/v1

Analysis of the Mixed Effects Regression Model for Clustered Count Response Data

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

[1*] Department of Statistics, College of Natural and Computational Sciences, University of Gondar, Gondar, Ethiopia.

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

[*]Corresponding Author: Aragaw Eshetie Aguade (Ph.D.), the University of Gondar, Gondar, Ethiopia. aragaw2018@gmail.com.

Email Addresses:

Aragaw Eshetie Aguade: aragaw2018@gmail.com;

B.Muniswamy Begari: munistats@gmail.com

Abstract

Background

The Poisson regression model is useful for analyse count data, but, when the observations are correlated the Poisson estimate will be biased. Whereas, when the over-dispersion and heterogeneity problems occur the imposition of the Poisson model underestimate the standard error and overestimate the significance of the regression parameters. Therefore, the objective of this paper was to develop a test statistic to model and predict clustered count response data via the application and simulation data.

Methods

This paper concentrated on the clustered count data model to take into account heterogeneity. Accordingly, we developed a score test based on the multilevel Poisson model for testing heterogeneity with the alternative Poisson regression model. In addition, for the model application, we used the EDHS children`s data. Therefore, to evaluate the proposed model, we used both simulation and application data.

Results

Simulation results showed that the proposed score test has high power to predict and used to control heterogeneity between groups. Oromia, Amhara, and SNNPR are among the regions with the highest child mortality rates (Table 1). The results indicated that women who made marriage a mean age of 16 years and gave birth to the first child a mean age of 18 years and 8 months. Table 1 showed that 81% of all child deaths have recorded in rural areas. 78% of child families were illiterate, as a result, 75% of children don't have access to latrines and drinking water. Rivers and open-source waters are the common sources of drinking water, which comprised 79% of the total water supply. Therefore, from the research finding, it is possible to conclude that most child mortality is due to scarcity of water.

Conclusion

The Power of test estimates indicated that the proposed method was better than the existing models. All covariant and dummy explanatory variables have a significant effect on the deaths of children. Hence, the multilevel Poisson model results indicated that there exists high variability among regions for the deaths of children. Therefore, this work suggested that the applications of the random-effects model provided a simple and robust means to predict the count response data model.

Keywords: Clustered count response data, Score test, homogeneity, MPM, and GLMM.

Background

Modelling is the heart of applied Mathematical and Statistical sciences. Model is the key component in any Mathematical and Statistical analysis. Hence, by introducing an error term, the mathematical model becomes a statistical model or the stochastic model. The applications of statistical linear regression models had made by [1] [2] [3], [4], [5], [6] [7], [8], [9], [10] [11], [12], [13], [14], [15] and [16] among others.

One of the classical assumptions of the regression model is: the error follows a normal distribution with mean zero and variance unit. Hence, when a distribution of the continuous response variable is asymmetric, to meet the assumption of normality, a simple transformation of the response variable can produce normally distributed errors. Nevertheless, if the response variable of interest is discreet, then, a simple transformation of the skewed distribution cannot produce normally distributed errors. Besides, nested data are common in the social and health sciences, hence, the assumption of independent observation is violating, and resulting in an incorrect standard error and inefficient estimate [17].

Fitting separate regressions for each group doesn`t allow examination of what group characteristics may be important in explaining the outcome. Classical regression methods, including multiple linear regression, logistic regression, and generalized linear modeling, assume independence of the

observations. However, the dependence of clustered data can lead to the imprecision of coefficient estimates which affects the statistical significance of risk factors. Testing homogeneity among clustered data adjusting for the effects of covariates [18]. Hence, ignoring such inter-cluster correlations may result in overlooking the importance of cluster effects and call into question the validity of traditional statistical techniques used for studying data relationships consequently, this analysis has lower statistical power.

Data with a multilevel nature happens in public health, health service research, behavioural sciences, and medicine. For clustered count data, the observations are correlated, the main approaches for correlated count data are conditional models, random-effects model, generalized estimating equations (GEE's) [19]. Conditional models are convenient for particular cases, such as data with small group sizes or with order structure [20]. The multilevel regression models incorporate cluster-specific random effects that account for the interdependence observations. This paper focused on a multilevel Poisson regression model approach, the distribution of the response variables is conditionally modelled in a group-specific parameter itself a random variable [21], [22].

This paper developed a score test to test heterogeneity of variance among groups, and estimating the regression parameters and the cluster-specific random effects. The power of the proposed score test has been evaluated through simulation and application data.

Objectives

- To develop a score test for testing heterogeneity of groups for discrete distribution.
- To evaluate the efficiency of the score test and compare it with an alternative model by using simulation and application data.
- To assess the various problems of the GLM for clustered discrete response data.

Significance of the Study

- To highlight the appropriate models for clustered count response data.
- Clustering information provides a correct standard error, confidence intervals, and significance tests, [23]

Methods

Multilevel Poisson Regression Model

Suppose Y_{ij} stands for the outcome variable for the case j of cluster i, i =1, 2… k; j = 1, 2… n_i, $N = \sum_{i=1}^{k} n_i$ the distribution is conditional, the response variable of interest $y_i = (y_{i1}, y_{i2}, ..., y_{in_i})`$, assumed group-specific random effects α_i having a probability density function of Y_{ij} is defined as follows $f\left(Y_{ij}|\mu_{ij}\right) = exp\left[\sum_{i=1}^{n_i}\left\{\theta_i y_{ij} - g\left(\theta_{ij}\right)\right\} + \quad C\left(y_{ij}, c_i\right)\right]$ ... (1).

Defining $C\left(y_{ij}, c_i\right) = \sum_{i=1}^{n_i} log\left(y_{ij}!\right)$, where $C\left(y_{ij}, c_i\right)$ does not depend on the model parameters. The mean and the variance of Y_{ij} are $\mu_{ij} = g'\left(\theta_{ij}\right)$ and$\sigma_{ij}^2 = c_i g''\left(\theta_{ij}\right) = \mu_{ij} = exp\left(x_{ij}\beta + z_{ij}\alpha_i\right)$, where$'$denotes the differentiation with respect to the parameter θ_{ij}. The mixed effects model considers at least one regression coefficient to be random is $\theta_{ij} = log(\mu_{ij}) = \eta_{ij} = x_{ij}\beta + z_{ij}\alpha_i$---- (2)

In the model the mean μ_{ij} and the variance σ_{ij}^2 are conditional on α_i, now, to examine the model under the null hypothesis, we assumed that all of the variables and the correlation between the random effects are zero in a generalized linear mixed model. Therefore, the parameter α_i, will be written as $\alpha_i = \alpha + D^{1/2}u_i$ [24].

Score Test of the Multilevel Poisson Model

In the model defined in (1) and (3), the conditional probability distribution function of Y_{ij} is denoted by $f_{ij}(y_{ij}, \beta, \alpha + D^{\frac{1}{2}}u_i)$, then the log likelihood function for the group i, is $l_i((\beta, \alpha, , D) = \frac{1}{2}\sum_{i=1}^{k}\left\{\sum_{j=1}^{n_i} y_{ij}\left(x_{ij}\beta + \alpha + D^{\frac{1}{2}}u_i\right) - exp\left(x_{ij}\beta + \alpha + D^{\frac{1}{2}}u_i\right)\right\} \ldots\ldots (3)$. Using L` Hopital`s rule following [25] and [26]the score function for testing the hypothesis of homogeneity can be written as $S(\beta, \alpha, D) = \frac{1}{2}\sum_{i=1}^{k}\left\{\left[\sum_{j=1}^{n_i}\frac{\partial log f_{ij}}{\partial \alpha_i}(\beta, \alpha, D = 0)\right]^2 + \sum_{j=1}^{n_i}\frac{\partial^2 log f_{ij}(\beta,\alpha,D=0}{\partial {\alpha_i}^2}\right\}$.

The first and the second partial derivatives of the log likelihood function with respect to the scalar random subject parameter α_i ,for the multilevel Poisson regression model is given by

$$\left[\frac{\partial}{\partial \alpha_i} log f_{ij}(\beta, \alpha, , D = 0)\right] = \sum_{j=1}^{n_i}(y_{ij} - \mu_{ij})z_j \text{ and } \left[\frac{\partial^2}{\partial \alpha^2} log f_{ij}(\beta, \alpha, , D = 0)\right] = -\sum_{j=1}^{n_i} z_{ij}^2\sigma_{ij}^2.$$

Using the first and second partial derivation of the log likelihood equation with respect to α_i, the score statistic is given by $S(\beta, \alpha,) = \frac{1}{2}\sum_{i=1}^{k}\left\{\left[\sum_{j=1}^{n_i}(y_{ij} - \mu_{ij})z_{ij}\right]^2 - \sum_{j=1}^{n_i}\left[z_{ij}^2\sigma_{ij}^2\right]\right\}$.Under homogeneity test, $H_0: D = 0$ for the known nuisance parameters β can be written as $H_{PD} = \frac{S_{PD}^2(\beta,\alpha)}{\left(I_{DD} - I_{D\beta}I_{\beta\beta}^{-1}I_{\beta D}\right)} \ldots\ldots\ldots\ldots\ldots\ldots (4)$. Now, the asymptotic variance function as the group size $k \to \infty$ of $S_{PD}(\beta, \alpha, D)$ und

er H_0 [27]can be expressed as a function of the information matrix. Then the asymptotic variance function for the score test is $D(\beta) = I_{DD} - I_{D\beta} I_{\beta\beta}^{-1} I_{\beta D}$ where $I_{DD} = \sum_{i=1}^{k} E\left[\frac{\partial l_i}{\partial D} \mid D = 0\right]^2$ a scalar is, $I_{\beta\beta} = \sum_{i=1}^{k} E\left\{\frac{-\partial^2 l_i}{\partial\beta\partial\beta`}\right\}$, $I_{\beta D} = I_{D\beta} = \sum_{i=1}^{k} E\left\{\frac{-\partial^2 l_i}{\partial\beta\partial D}\right\}$.

Parametric estimations of the Multilevel Poisson model

The function of the proposed test is $\mathrm{Var(D)} = \mathrm{I_{DD}} - I_{D\beta} I_{\beta\beta}^{-1} I_{\beta D}$. Moreover, the information matrix $(\mathrm{I_{DD}})$ for the proposed score test is defined as $\left(\frac{\partial l_i}{\partial D} \mid D = 0\right)^2 = \frac{1}{2}\sum_{i=1}^{k} E\left\{\left[\sum_{j=1}^{n_i}(y_{ij} - \mu_{ij})\right]^2 - \sum_{j=1}^{n_i}\sigma_{ij}^2\right\}^2$. Now, to solve the variance of the score function, let us to define $U_{ij} = y_{ij} - \mu_{ij}$, the i^{th} term of the score test can be written as $\frac{\partial l_i}{\partial D}\Big|_{D=0} = \frac{1}{2}\left[\sum_{j=1}^{n_i} U_{ij}\right]^2 - \sum_{j=1}^{n_i}\mu_{ij}$. Thus, $E\left(\frac{\partial l_i}{\partial D} \mid D = 0\right)^2 = \frac{1}{4}E\left(U_i^2 - \sum_{j=1}^{n_i}\sigma_{ij}^2\right)^2$ where $U_i = \sum_{j=1}^{n_i} U_{ij}$, since $E(U_{ij}) = 0$, then $E(U_i^2) = E\left(\sum_{j=1}^{n_i} U_{ij}^2\right)^2 = \sum_{j=1}^{n_i}\sigma_{ij}^2$. Therefore, $E\left(\frac{\partial l_i}{\partial D}\Big|_{D=0}\right)^2 = \frac{1}{4}(\mu_4 - \mu_2^2)$.

In the above equation, the second and the fourth central moments of the outcome variable Y_{ij} are μ_2 and μ_4, respectively, it can be written as a function of the fourth and the second cumulates, cumulates K_4 and K_2 of Y_{ij} ([28] , $\mu_4 = K_4 + 3K_2^2$ $then$ $\mu_4 = \mu_{ij} + 3\mu_{ij}^2$ since $\mu_2 = K_2 = \sigma^2 = \mu_{ij}$ [29]. Subsequently, the simplified form of the information matrix (I_{DD}) is defined as $\frac{1}{4}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mu_{ij} + 2\mu_{ij}^2)$.

The variance of the score functions can be derived from the Fisher information. $I_{(\beta)} = \begin{pmatrix} I_{\beta\beta} & I_{\beta D} \\ I_{D\beta} & I_{DD} \end{pmatrix}$.

Where $\frac{\partial l_i}{\partial\beta} = \sum_{j=1}^{n_i} y_{ij}x_{ij} - \sum_{j=1}^{n_i}\mu_{ij}x_{ij}$ and $\frac{\partial^2 l_i}{\partial\beta\partial\beta`} = -\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mu_{ij}x_{ij}\,x_{ij}'$. Therefore $I_{\beta\beta} = \sum_{i=1}^{k} E\left\{\frac{-\partial^2 l_i}{\partial\beta\partial\beta`}\right\} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\mu_{ij}x_{ij}\,x_{ij}'$. $I_{\beta D} = I_{D\beta} = \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mu_{ij}x_{ij}$. Where $x_{ij} = \left(1, x_{1ij}, \ldots . x_{n_i ij}\right)`$,

the parameter β in the score test statistic (H_{PD}) is given equation (1), which is replaced by their ML estimates obtained from the Poisson regression model under the null hypothesis, [30]. Then, the score test statistic H_{PD} will be reduced to $H_{PD} = \frac{\widehat{S_P^2}(\widehat{\beta},\widehat{\alpha})}{\left(\hat{I}_{DD} - \hat{I}_{D\beta}\hat{I}_{\beta\beta}^{-1}\hat{I}_{\beta D}\right)}$. Therefore, the proposed score test statistic, $H_{PD,}$ will be

$$\left\{\frac{\left[\sum_{j=1}^{n_i}(y_{ij}-\mu_{ij})^2 + \sum_{j=1}^{n_i}\sum_{j\neq j`}^{n_i}(y_{ij}-\mu_{ij})(y_{ij`}-\mu_{ij`}) - \sum_{j=1}^{n_i}\mu_{ij}\right]^2}{\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\hat{\mu}_{ij}+2\hat{\mu}_{ij}^2) - \frac{\left(\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\hat{\mu}_{ij}x_{ij}\right)\left(\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\hat{\mu}_{ij}x_{ij}\right)'}{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\hat{\mu}_{ij}x_{ij}\,x_{ij}'\right)}\right]}\right\} \quad (5)$$

In large samples, the proposed test statistic follows a chi-square distribution with unit degrees of freedom, now the maximum likelihood estimate of β can be estimated iteratively by using Fisher`s scoring method from the following equations. $\beta^{(t+1)} = [(XW^{-1}X^T)^{-1}XWZ]^{(t)}$, *t = 1, 2, 3…, where* $W = diag(\hat{\mu})$ is an *N*×*N* matrix and $Z = X^T\hat{\beta}^{(t+1)} + \frac{Y-\hat{\mu}}{\hat{\mu}}$, t = 1, 2, 3… is an *N*×*1* vector.

Method of Estimation of the Regression parameters

Define the function $\eta_{ij} = log\mu_{ij}$, then, $\frac{\partial \eta_{ij}}{\partial \mu_{ij}} = \frac{1}{\mu_{ij}}$, also let $V_{ij} = Var(y_{ij}) = \mu_{ij}$. Further, let us define $w_{ij} = \left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right)^2 V_{ij}^{-1} = \mu_{ij}$. Then the score equation for the regression parameters $\beta`s$, s=1, 2…p. Can be written as $u_s = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_{ij})x_{ijs}$. And the fisher information matrix for β is $I = \sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij}x_{ijs}x_{ijr}'$. Now 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 $N = \sum_{I=1}^{K} n_i$. Additionally, let X be a squared diagonal matrix with weighting variable w_{ij}. Then, the score function is defined as $u = x'w(\frac{y-\mu}{\mu})$ and the information matrix is also defined as $I = (X'WX)$. The scoring function for the regression parameters β

become $I^{(t)}\beta^{(t+1)} = I^{(t)}\beta^{(t)} + u^{(t)}$. Thus $\beta^{(t+1)} = \left((X`WX)\right)^{-1}(X`WZ)'$, *t=0, 1, 2...* with $Z = X'W\left(\frac{Y-\mu}{\mu}\right)$ ……(6)

Alternative Tests of Homogeneity of Clustered Count Data

The multilevel Poisson regression model is a special case of the Poisson regression model, the model is used for testing if a Poisson model is adequate corresponding to testing homogeneity $H_0 : D = 0$ against the alternative $H_0 : D > 0$, where one possible test statistic is the likelihood ratio test, $LRT = -2 \times \left[l(\hat{\mu}) - l\left(\hat{\mu}, \widehat{D}\right)\right]$ where $l(\hat{\mu})$ and $l\left(\hat{\mu}, \widehat{D}\right)$ are the maximum likelihood estimation under the Poisson and multilevel Poisson models, respectively.

Alternative Tests for Coefficient of Regression Parameters

To test the significance of the coefficient of the regression parameter (β_k), the hypothesis denoted as $H_0: \beta_k = 0 \; vs \; H_0: \beta_k \neq 0$ and the LRT test statistics for testing the null hypothesis (6) is given by $\text{LRT}\beta_k = 2 \times \left[l\left(\hat{\beta}_0\right) - l\left(\hat{\beta}\right)\right]$, where $l\left(\hat{\beta}_0\right)$ and $l\left(\hat{\beta}\right)$ are the log-likelihood estimate of the parameters for the restricted and unrestricted parameter, respectively. The associated Wald test statistic is $W_{\hat{\beta}} = \widehat{\beta}`\left(Cov(\hat{\beta})\right)^{-1}\hat{\beta}$ where $\hat{\beta}$ is the maximum likelihood estimate of X_{lij}, $Cov(\hat{\beta})$ is the variance-covariance matrix of these estimations, determined from the estimation of the variance-covariance matrix $\widehat{\mathrm{I}^{-1}}(\beta, \mathrm{D})$. For an adequate model, the LRT also has an asymptotic chi-square distribution with the degree of freedom equals to the number of restricted parameters in the null hypothesis. There are several applications by using the likelihood ratio test or the Wald test for the count regression model, such as [31].

Model Selection

Information criteria used to select the best model which fits the data instead of using the likelihood ratio test by using the Akaike information criteria (AIC) and Bayesian information criteria (BIC).

Akaike Information Criteria

AIC is the most common means of identifying the model which fits the data well by comparing two or more than two models. The goodness of fit test against the complexity of the model is similar to that the coefficient of multiple determination(R^2); however, it penalized by the number of parameters included in the complexity of the model. Unlike the R^2, the good model is the one that has the minimum AIC value. It is given by the following formula AIC=−2ℓ+2k, where ℓ is the log-likelihood function of a model that will compare with the other models and k is the number of parameters in the model including the intercept [32].

Bayesian Information Criteria

Unlike the Akaike information criteria, the Bayesian information criteria take into account the size of the data under consideration. It is given by BIC=−2ℓ+$k$log (n) where ℓ is the log-likelihood of a model that will compare with the other models, n is the sample size of the data and k is the number of parameters in the model including the intercept.

Result and Discussion

Simulation study

In this section, a simulation study is conducted to compare the proposed and existing models in terms of sizes and powers. For studying the properties of the statistic in terms of empirical size, generating count data from a Poisson distribution under the null hypothesis of homogeneity and assume that the random effect's parameter is once$(z_{ij} = 1)$ and the samples are comprised of 10; 20; 50; 100

observations and 5; 10; 20; 50 groups with simulated data, the multilevel Poisson distribution is simulated for the distribution of the response variable assuming that under the null hypothesis there is homogeneity within a different number of groups. Each simulation experiment for level and power was based on 1000 simulated samples. The model used under the simulation study is $\log(\mu_{ij}) = 0.8x_{1ij} + 0.5u_i - 0.5$---------- (7). $y_{ij} \sim \text{Poisson}(\mu_{ij})$, i =1, 2… k; j = 1, 2… n_i . For the variable x_{1ij} is a subject-specific effect which is simulated from a uniform distribution with mean zero and unit variance and u_i is group specific effects and simulated from a standard normal distribution and generate a set of random numbers from

a uniform distribution in the interval (0,1) as the values of x_{1ij}, $\alpha_i = \alpha + D^{1/2}u_i$.

To simulate correlated data the u_i`s is identical and independently distributed with a standard normal distribution with mean zero and unit variance. Therefore, $\alpha_i$`s are identical and independently distributed with mean α and variance D. For each set of generating data, a multilevel Poisson model is fitted for calculating the score test and the existing tests followed by the powers of the score tests. Results from the simulation study are presented in Table- 2 and 3.

In Table 2 the results investigated that how the information criteria, perform in the multilevel Poisson regression model selection problems via simulations. When the number of clusters is large and the number observation is small and the multilevel Poisson regression model is better than the standard Poisson regression model. Simulated data results indicated that the performance of the true model generally involved an increase of sample size, despite differences in performance between the information criteria. The simulated results found that the performance of the model depends on the sample size and the number of clusters.

In Table 2 noted that an error probability increases power increases. The power increases when α increases and for large sample groups and small variance of the group effect (k = 50, n = 10, D=0.05) the power increase quickly, and approaches to 1. For small sample groups, when the standard deviations of the group effects increase, the power increases slowly, whereas, in large sample groups, (k = 50, n = 50, D=0.15 and α =0.1) as standard deviations of the group effect increases, the power increases slowly, however, when the values of D increase from 0.05 to 0.15, the power decreases. Generally, as the number of groups increases the power increased. Therefore, the proposed score test is important to examine the heterogeneity of group effects and fixing the number of observations and groups due to its high power to predict the model.

Application Study

This paper to intake account of the random parameters of the Poisson model. Hence, to illustrate the proposed model, we used the Ethiopian Demographic and health survey children's data. Oromia, Amhara, and SNNPR are among the regions with the highest child mortality rates (Table 1). Whereas, the child mortality rates were significantly lower in Afar and Harari regional states. The results indicated that women who made marriage a mean age of 16 years and gave birth to the first child a mean age of 18 years and 8 months. Besides, the results from Table 1 showed that 81% of all child deaths have recorded in rural areas. 78% of child families were illiterate, as a result, 75% of children don't have access to latrines and drinking water. Rivers and open-source waters are the common sources of drinking water, which comprised 79% of the total water supply. Therefore, from the research finding, it is possible to conclude that most child mortality is due to scarcity of water.

To illustrate the proposed method for fitting a multilevel Poisson regression model, we considered the 2005 Ethiopian demographic and health survey children's deaths data, 25420 women whose ages 15 and 49 years had interviewed. This paper considered the number of child mortality aged lower than 18

years that each mother has experienced in her lifetime. The minimum and maximum values of the count of the response variable are lies in the interval zero and eighteen with a mean 1.2 and variance 1.7.

Estimation of the random and regression parameters for the two models have given in Table 4. The estimated values of the random effect parameter (σ_u^2) is 0.101 with standard error 0.102, the result indicated that a variation of child mortality among regions. In addition, the fixed effects results showed that except for the toilet facility all the dummy and explanatory variables were significant in both models. Whereas the information criterion, the LRT, and the Wald statistics showed that we have strong evidence to fit the multilevel Poisson regression model to the data. Similarly, Fig 3 result showed that the predicted probability of the proposed method was closer to the probabilities of the true values.Therefore, these results showed that the proposed method is superior to the existing method.

Conclusion

This paper used simulation and application data to illustrate the proposed method, the power of the multilevel Poisson regression model has presented in Table 2. The results revealed that the proposed score test is preferable to the existing model. Application results revealed that there would be a great variability of child mortality among regions, and the predicted probability showed that the proposed model is better than the standard model.

Simulation and application study results showed that when we considered the random-effects parameter in clustered count data, the proposed method gives accurate and valid results whereas the Poisson regression model doesn't handle heterogeneous data. For fixed sample size, when the regression and random-effects parameters are increasing the powers are increasing, whereas, for fixed

regression coefficient and random-effects parameters, when the sample size increasing as the power is increasing.

The simulation results showed that the power is smaller for small values of the random effect parameter while the power is increasing as the random effect is increasing. Hence, the score test is appropriate to model the number of child mortality and the results showed that there would be a significant variation of deaths of children among regions (Table 4). Therefore, the information criteria and the predicted probability results revealed that the proposed model is better than the standard Poisson model.

Abbreviations

GLMM	Generalized linear mixed model
EM	Expectation and maximization
GEE	Generalized estimating equation
LRT	Likelihood ratio test
GLM	Generalized linear model
AIC	Akakie information criteria
BIC	Bayesian information criteria
IC	Information criteria
PM	Poisson model
MPM	Multilevel Poisson model
MP	Multilevel Poisson
EDHS	Ethiopian demographic and health related survey

Declarations

Ethics approval and consent to participate

Not Applicable .But, the necessary Permission to undertake the study was obtained from the Central Statistical Agency of Ethiopian.

Consent for publication

Not Applicable

Availability of Data and Materials

The data that support the findings of this study are available from the Central Statistical Agency of Ethiopia but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly applicable. Data are however available from the authors and online upon reasonable request and with permission of the Central Statistical Agency of Ethiopian, the forms for requesting access to raw data which will be available on the website. www.csa.gov.et

Competing Interest

The authors declare that they have no competing interest.

Funding

No fund was obtained from any source to conduct the study as well as in the processing of the Manuscript for publication.

Authors' contributions

Aragaw Eshetie Aguade has contributed on conceptualizing the research problem,research, concept and designed, collection and or assembly of data, data analysis and interpretation and writing the article. **B. Muniswamy Begari** has made a great role in re-vision of the research design, data analysis, editing the entire manuscript and ready for publication. Finally, all authors have read and approved the final manuscript.

Acknowledgment

The authors would like to thank the Ethiopia central Statistics Agency for making available the data used in this research.

References

[1] K. Pearson, “Life and Latters and Labours of Francis Galton,” *University Press,Cambridge,* 1914.

[2] E. M.A., “Multiple Regression Analysis.,” *In A. Ralston and H.S. Wile (Eds.),* 1960.

[3] F. Anseombe, “Topics in the Investigation of Linear Relations Fitted by the Method of Least Squares,” *J.R.S.S,* vol. 29, pp. 1-32, 1967.

[4] S. Searle, Linear Models., New York: John Wiley & Sons, 1971.

[5] Johnston, Econometric Methads, New York: McGraw-Hill, 1972.

[6] J. a. W. R. Nelder, “Generalized Linear Models,” *J.R. Stat. Soc. Ser. A.,* vol. 153, pp. 370-384., 1972.

[7] S. Weisberg, Applied Linear Regression., New York: John Wiley & Sons, 1980.

[8] J. K. M. N. C. a. W. W. Neter, Applied Linear Statistical Models, Homewood: 4th ed Richard D. Irwin, 1996.

[9] N. a. S. I. Draper, Applied Regression Analysis, New York: Third Edition, John Wileg & Sons, 1988.

[10] a. T. Shi.P., “Regression Model selection-A residual likelihood Approach,” *J.R.S.S., Series B,* vol. 64, pp. 237-252, 2002.

[11] S. a. H. R. Horn, “ Comparison of Estimators of Heteroscedastic Variances in linear Models,” *Journal of the American Statistical Association,* vol. 70, p. 872 – 879, 1975.

[12] R. Hocking, “The Analysis and Selection of Variables in Linear Regression,” *Biometrics,* vol. 32, pp. 1-49, 1976.

[13] F. Graybill, Theory and Application of the Linear Model, North Scituate, Mass: Duxbury, 1976, pp. North Scituate, Mass..

[14] P. E. a. G. V. G. Montgomery. D.C., Introduction to Linear Regression Analysis,, New York: John Wiley and sons., 2003.

[15] J. a. L. N. Kadane, “Methods and Criteria for Model Selection,” *JASA,* vol. 99, pp. 279-290, 2004.

[16] A. A. M. a. K. B. Saleh, Theory of Ridge Regression Estimation with Applications, Wiley: New York, 2019.

[17] P. L. K. a. Z. S. Diggle, The Analysis of Longitudinal Data, Clarendon: Oxford, 1994.

[18] H. J.-G. &. D. Commenges, “ Tests of Homogeneity for Generalized Linear Models,” *Journal of the American Statistical Association,* vol. 90, no. 432, pp. 1237-1246, 1995.

[19] D. E. Prentice, “Correlated Binary Regression With Covariates Specific to Each Binary Observation,” *Biometrics,* vol. 44, pp. 1033-1048., 1988.

[20] B. Rosner, “Multivariate Methods in Ophthalmology With Applications to Other Paired Data Situations,” *Biometrics,* vol. 40, pp. 1025-1035., 1984.

[21] N. a. W. J. H. Laird, “Random-Effects Models for Longitudinal Data,” *Biometrics,* vol. 38, pp. 963-974, 1982.

[22] R. L. N. a. W. J. H. Stiratelli, “Random-Effects Models for Serial Observations with Binary

Response," *Biometrics,* vol. 40, pp. 961- 972., 1984.

[23] T. A. B. &. B. R. J. Snijders, "Multilevel analysis," in *An introduction to basic and advanced multilevel modeling.*, Thousand Oaks, Sage Publications, 1999.

[24] B. a. A. E. A. (2019.), "Proposed Score Test of Homogeneity of Groups in the Multilevel Poisson Regression Model for Clustered Discrete Dat," *International Journal of Scientific Research and Review,* vol. 8, no. 1, pp. 1036-1052, 2019.

[25] K. Liang, "A locally most powerful test for homogeneity with many strata," *Biometrical,* vol. 74, p. 259 – 264., 1987.

[26] Chescher, "Testing for Neglected Heterogeneity.," *Econometrical,* vol. 5, no. 2, p. 865 – 872, 1984.

[27] D. Cook D. and Hinkley, Theoretical statistic, London: Chapman and Hall, 1974.

[28] M. a. S. A. Kendell, "The Advanced theory of statistics.," *Machillan, New York,* vol. 1, p. 168, 1977.

[29] K. Azad, Some Inference Problems in Clustered (Longitudinal) Count Data with Over-dispersion, Windsor : Electronic Theses and Dissertations., 2011.

[30] "Lawless, J.F.; Negative binomial and mixed Poisson regression.," *The Canadian Journal of statistics,* vol. 15, p. 2009 – 225., 1987.

[31] N. a. H. J. Jansaku, " Linear mean variance negative binomial models for analysis orange tissue culture," *Journal of Science and Technology,* vol. 26, no. 5, pp. 683-689, 2004.

[32] N. a. J. A. Ismail, "Handling over dispersion with negative binomial and generalized Poisson regression models," *Causality Society forum,* p. 103 – 158, 2007.

[33] R. Hocking, "The Analysis and Selection of Variables in Linear Regression,," *Biometrics,* vol. 32, pp. 1-49, 1976.

[34] T. A. B. &. B. R. J. nijders, "Multilevel analysis," in *An introduction to basic and advanced multilevel modeling.*, Thousand Oaks, Sage Publications., 1999.

[35] D. Cook D. and Hinkley, Theoretical statistic, London: Chapman and Hall, 1974.

302

Table 1: Characteristics of study participants in EDHS, 2013

Variable	level	Frequency	Percent	P- value
Region of Mother`s	Tigray	1960	7.7	0.000
	Affar	1440	5.7	
	Amhara	3760	14.8	
	Oromiya	4700	18.5	
	Somali	1840	7.2	
	Ben-Gumz	2140	8.4	
	SNNP	3780	14.9	

	Gambela	1620	6.4	
	Harari	1000	3.9	
	Addis	1620	6.4	
	Dire Dawa	1560	6.1	
	Total	25420	100.0	
Residence Area of Mother`s	Urban	4680	18.4	0.000
	Rural	20740	81.6	
	Total	25420	100.0	
Education level of mother`s	Higher	260	1.0	0.000
	No education	19860	78.1	
	Primary	3380	13.3	
	Secondary	1920	7.6	
	Total	25420	100.0	
Toilet Facility	others	40	0.2	0.462
	Flush toilet	200	0.8	
	Improved (ventilated) pit latri	120	0.5	
	No facility, bush, field	18960	74.6	
	Not de-jure resident	320	1.3	
	Traditional pit latrine	5780	22.7	
	Total	25420	100.0	
Religion of Mother`s	Orthodox	10940	43.0	0.000
	Catholic	180	0.7	
	Protestant	3780	14.9	
	Muslim	9740	38.3	
	Traditional	720	2.8	
	Other	60	0.2	
	Total	25420	100.0	
Current Marital Status of Mother`s	Married	25160	99.0	0.018
	Living together	260	1.0	
	Total	25420	100.0	
Source of Drinking water	Piped into dwelling	20	0.1	0.000

	Piped into compound	1900	7.5	
	Piped outside compound	2980	11.7	
	Open well	1840	7.2	
	Open spring	8280	32.6	
	Covered well	1240	4.9	
	Covered spring	1000	3.9	
	River	6860	27.0	
	Pond/lake/dam	860	3.4	
	Rainwater	60	0.2	
	Other	60	0.2	
	Not a de-jure resident	320	1.3	
	Total	25420	100.0	

Descriptive Statistics for continues variables

Variables	N	Minimum	Maximum	Mean	Std. Deviation	P-value
Age of Mother`s for first birth	23040	11	41	18.86	3.778	0.000
Current Age of Mother`s	25420	15	49	30.52	8.328	0.000
Age of Mother`s at first Marriage	25420	5	33	16.26	3.431	0.000

303

304

305

306 **Table 2. Goodness of fit tests via simulated models.**

Group(k)	(n)	Model	-2logl	AIC	BIC
5	10	MPM	140.6	146.6	147.5
		PM	101.4	104.4	109.3
	20	MPM	224.5	228.5	231.5
		PM	226	227.6	238
	50	MPM	647.2	652	658.9
		PM	640.7	644.7	651.8
	100	MPM	1457	1459.7	1467.5

		PM	1284	1287.4	1294.8
10	10	MPM	298.7	304.7	304.6
		PM	235	237.5	247
	20	MPM	525	531	533.9
		PM	507	511	517.6
	50	MPM	1299	1298.9	1304.6
		PM	1247.6	1251.6	1260
	100	MPM	2720.5	2726.5	2734.4
		PM	2413	2417.3	2427.2
20	10	MPM	477.9	483.9	484.9
		PM	526.6	530.6	537.2
	20	MPM	1039	1045	1048
		PM	988.9	999	1000.9
	50	MPM	2466.7	2477	2478.5
		PM	2474.9	2478.9	2488.7
	100	MPM	5244.8	5250.8	5258.6
		PM	4959.4	4964	4974.6
50	10	MPM	1394.7	1400.7	1401.6
		PM	1336	1337.6	1346
	20	MPM	2377.1	2381	2386.1
		PM	2535	2537.5	2547.4
	50	MPM	6434	6439.4	6442
		PM	6197.1	6201.1	6217
	100	MPM	12993	12999	13007
		PM	12179	12183	12196

Table 3. Estimated Empirical Power Test of the MP Regression Model

Data are generated from the multilevel Poisson distribution under the null hypothesis on 1000 replication. In the simulation, the levels (10%; 5% and 1%), the sample size (10, 20, 50 and 100) and the number of groups (5, 10, 20 and 50) are considered.

Cluster (k)	Observation(n)	D=0.05			D=0.10			D=0.15		
		Power			Power			Power		
		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	0.99979	0.98849	0.95125	0.94446	0.89942	0.74614	0.97221	0.94512	0.83729
	20	0.98565	0.96943	0.89554	0.99565	0.98952	0.95474	0.99307	0.98406	0.93701
	50	0.99730	0.99378	0.97017	0.99141	0.98067	0.92674	0.97598	0.95174	0.85224
	100	0.99513	0.98840	0.95095	0.99601	0.99029	0.95744	0.97808	0.95548	0.86097

10	10	0.99991	0.99971	0.99766	0.99913	0.99757	0.98613	0.99926	0.99790	0.98769
	20	0.99753	0.99373	0.96998	0.98599	0.97007	0.89725	0.99136	0.98057	0.92644
	50	0.99580	0.98984	0.95584	0.98903	0.97593	0.91314	0.97843	0.95611	0.86245
	100	0.99535	0.98887	0.95253	0.99433	0.98667	0.94527	0.98096	0.96070	0.87348
20	10	0.99589	0.99003	0.95951	0.98562	0.96938	0.89542	0.94187	0.89536	0.73881
	20	0.99708	0.99269	0.96606	0.99540	0.98898	0.95292	0.94689	0.90326	0.75316
	50	0.98714	0.97227	0.90309	0.97897	0.95708	0.86476	0.98101	0.96080	0.87372
	100	0.99770	0.99411	0.97148	0.99218	0.98222	0.93138	0.98768	0.97332	0.90592
50	10	0.99997	0.99988	0.99888	0.92218	0.86538	0.68756	0.96108	0.92623	0.79741
	20	0.99934	0.99811	0.98869	0.99672	0.99187	0.96303	0.91279	0.85153	0.66537
	50	0.99767	0.99403	0.97117	0.99337	0.98467	0.93891	0.99780	0.99433	0.97234
	100	0.99090	0.97964	0.92372	0.99607	0.99044	0.95793	0.98751	0.97298	0.90501

Table 4. Estimated the multilevel Poisson model via EDHS, 2013

Number of children	Poisson				Multilevel			
	Estimate	S.E	Z-value	p-value	Estimate	S.E.	Z-value	P-value
Residence	0 .422053	0.024755	17.05	0.000**	0.447924	0.025148	17.81	0.000**
Educ. level	-0.649558	0.024491	-26.52	0.000**	-0.649343	0.024561	-26.44	0.000**
Toiletfac	-0.013130	0.017863	-0.74	0.462	0.027909	0.018049	1.55	0.122
Religion	0.439330	0 .012739	34.49	0.000**	0.377617	0.013224	28.56	0.000**
HHSMembers	0 .009177	0.002426	78	0.000**	0 .008618	0 .002420	56	0.004**
Agemother	0.056723	0 .000701	80.93	0.000**	0.056305	0 .000704	80.04	0.000**
CurrentMari	0 .189184	0.080185	36	0.018*	0.213766	0.080234	66	0.008**
Agemarriage	-0.040287	0.001901	-21.19	0.000**	-0.038131	0.001908	-19.98	0.000**
Sourcdrinkwat	0.006536	0.011999	4.45	0.000**	0.031170	0.012184	56	0.011*
Constant	-100907	0.065720	-20.77	0.000**	-156525	0 .066296	-353	0.000**
Region|var(cons)					0 .101253	0.101781		

*: Significant at 0.05 level, **: Significant at .01 level.

Table 5. The observed and predicted probabilities of the fitted models via EDHS, 2013.

No. Death	Observed frequency	Observed probability	Predicted probability	
			PM	MPM
0	12000	0.4721	0.3269	0.3271
1	5680	0.2234	0.2229	0.2202

2	3560	0.1400	0.1956	0.1938
3	1880	0.0740	0.1068	0.1080
4	1140	0.0448	0.0688	0.0683
5	500	0.0197	0.0325	0.0326
6	320	0.0126	0.0224	0.0232
7	80	0.0031	0.0049	0.0052
8	120	0.0047	0.0075	0.0076
9	40	0.0016	0.0025	0.0029
10	40	0.0016	0.0039	0.0051
11	20	0.0008	0.0017	0.0016
13	20	0.0008	0.0006	0.0006

Model	Model selection criteria		
	-2l	AIC	BIC
Poisson model	74230	732699	733559
Multilevel Poisson model	73240	73095	73185.06

Table 6. Estimated model selection criteria for the Poisson and MP models via EDHS, 2013

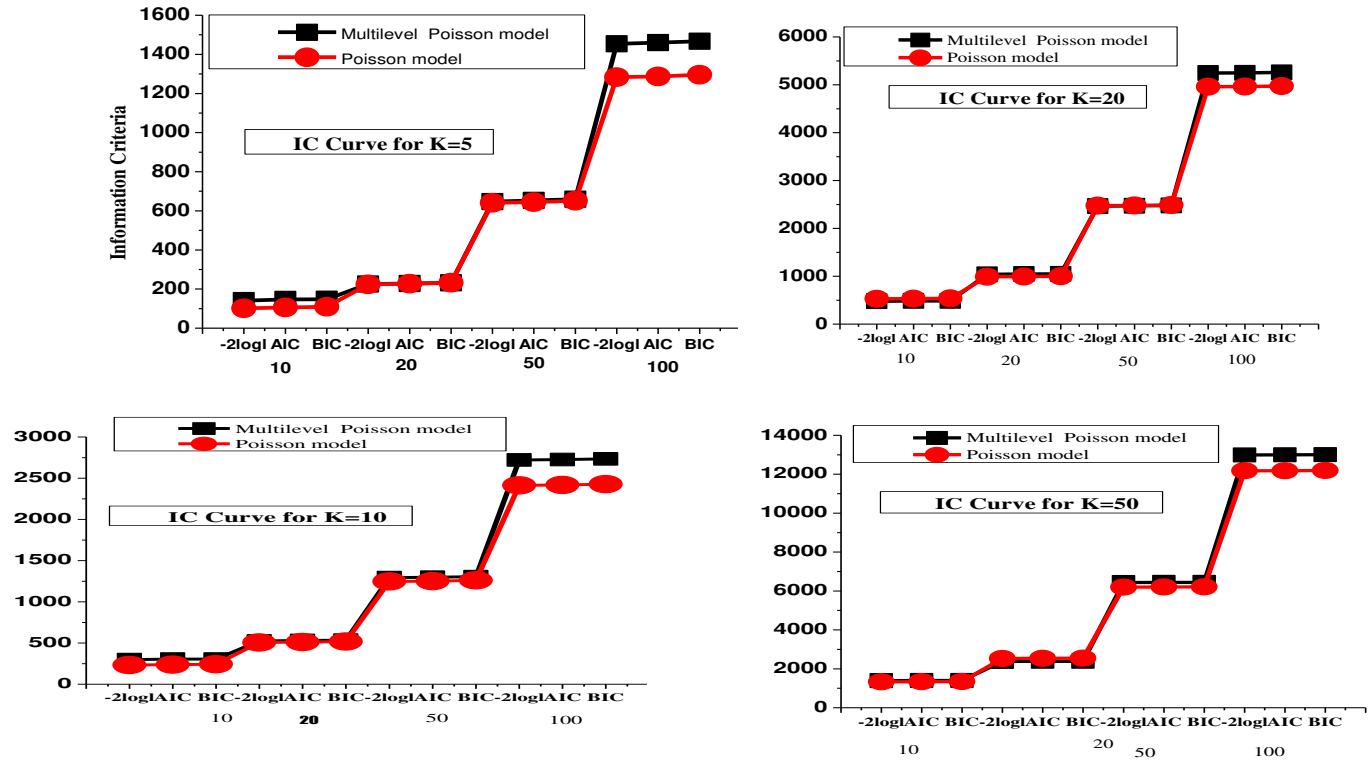

Figure 1. Simulated Data Using LRT and IC

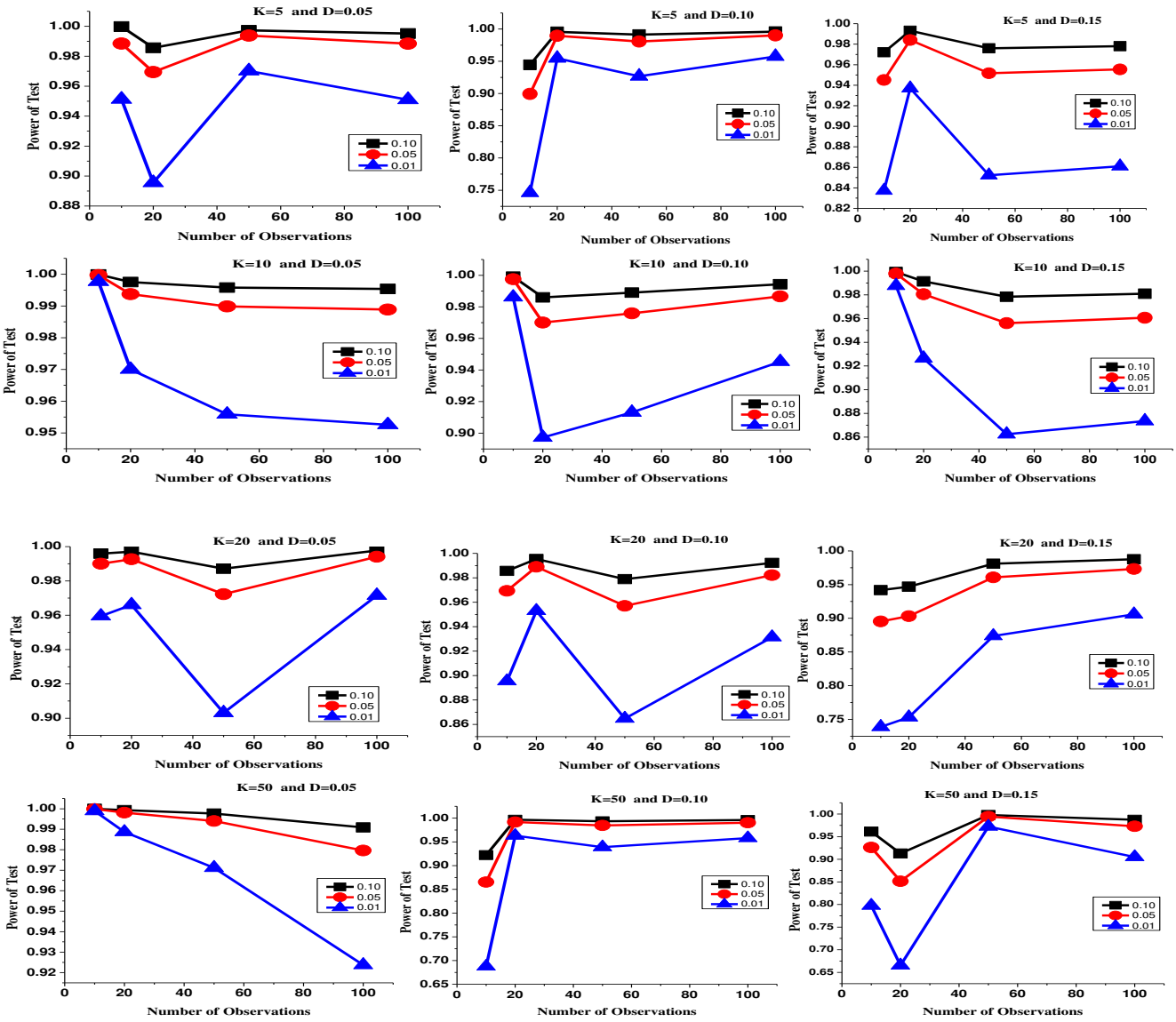

Figure 2. Simulated curve for power and goodness of fit test of the models

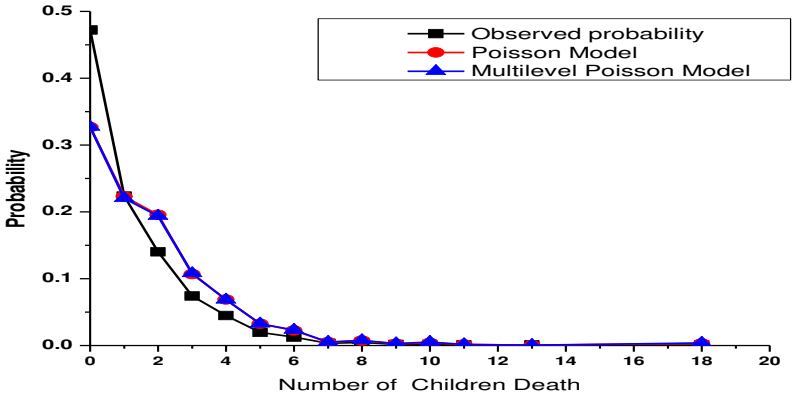

Figure 3. Predicted probabilities of the fitted models with covariates

Figures

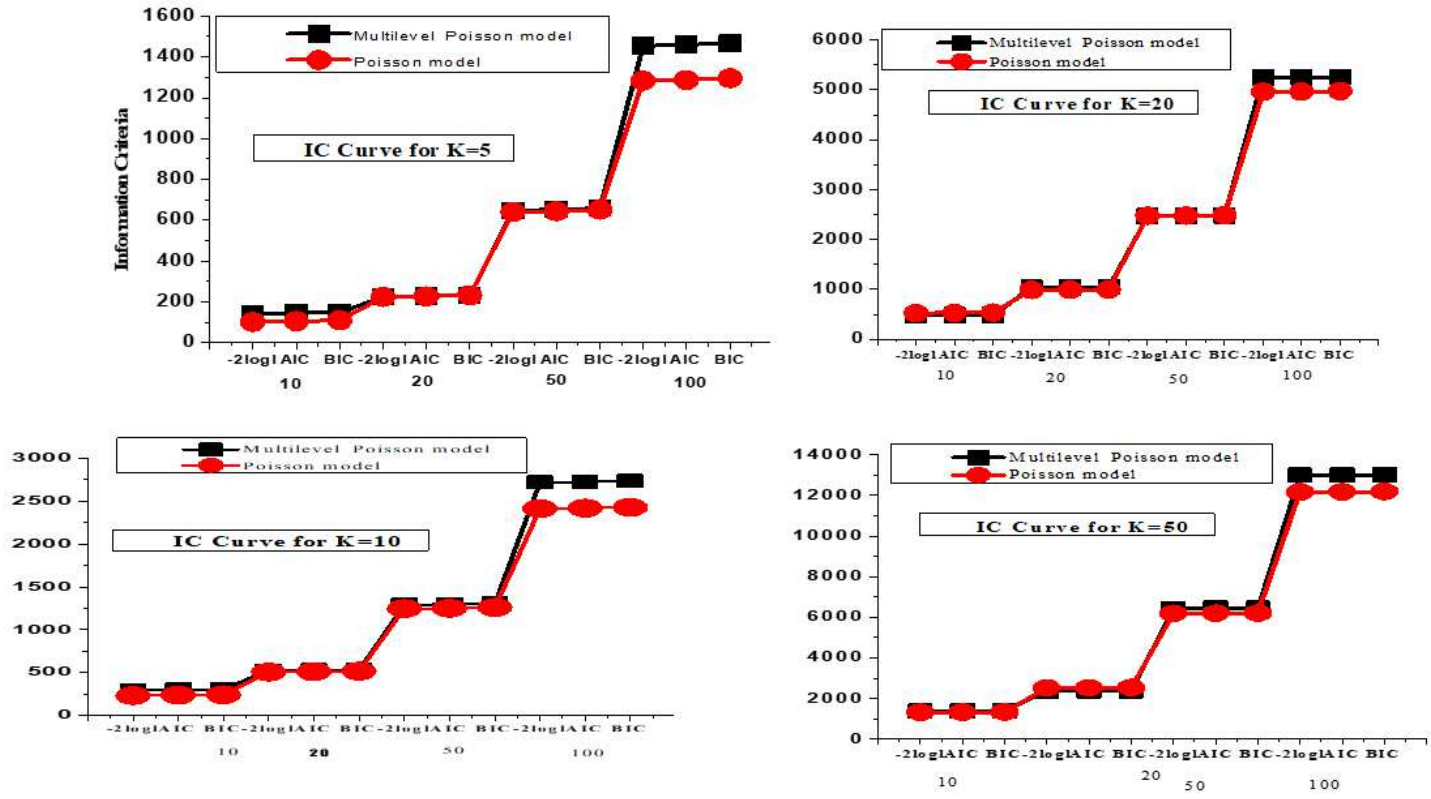

Figure 1

Simulated Data Using LRT and IC

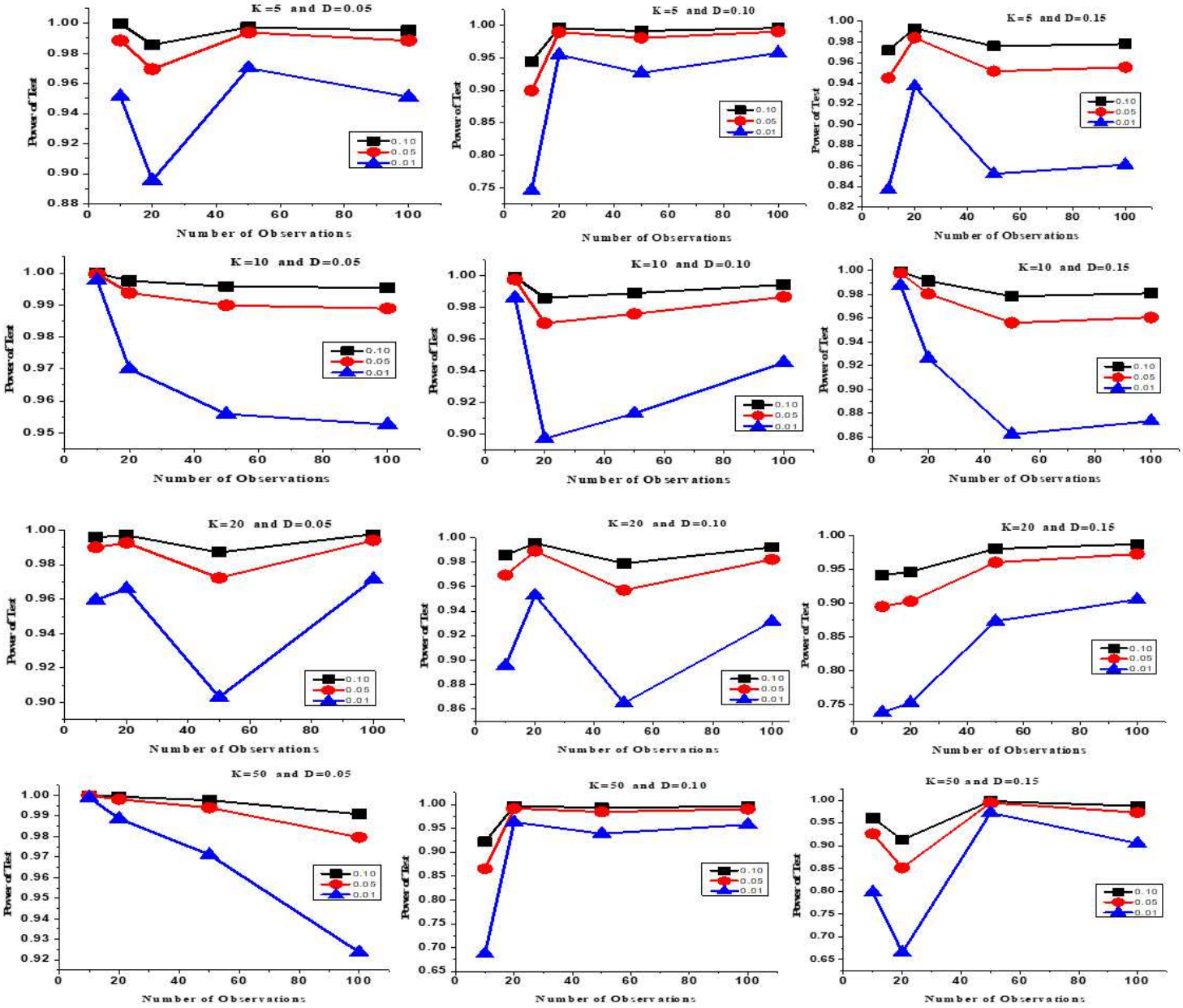

Figure 2

Simulated curve for power and goodness of fit test of the models

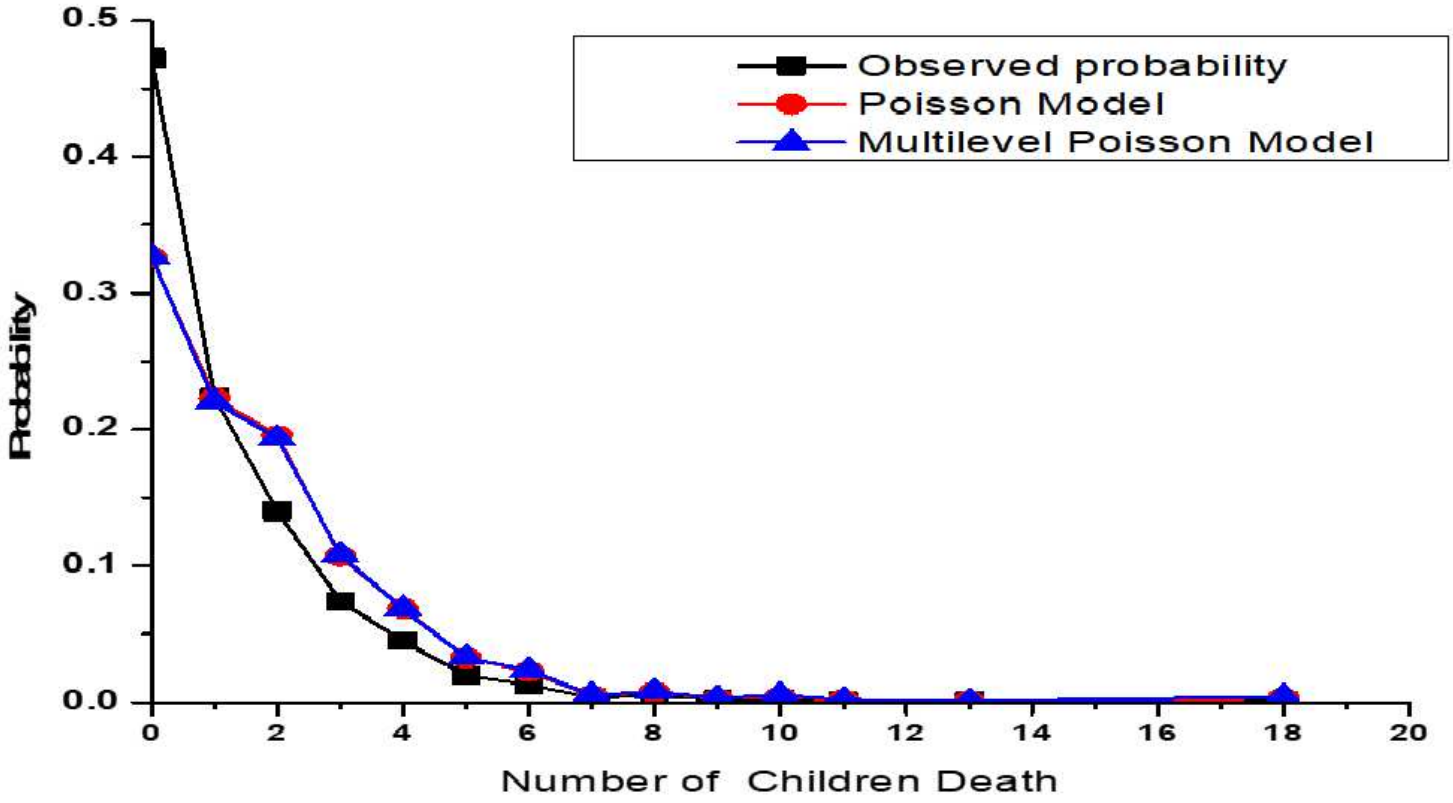

Figure 3

Predicted probabilities of the fitted models with covariates