**Eksakta** : Berkala Ilmiah Bidang MIPA
VOLUME 24 NO 03 2023, pp 409-415
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol24-iss03/297

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

*Article*

# Multilevel Modeling on Underdispersion Data

**Article Info**

***Article history :***

Received December 22, 2021
Revised February 17, 2022
Accepted March 07,2022
Published September 30, 2023

***Keywords :***
Multilevel modeling, underdispersion, binomial negative regression

**Corry Sormin[1*], Niken Rarasati[1], Gusmanely Z[1], Hamidreza Kashefi[2]**

[1]Department of Mathematics, Faculty of Science and Technology (FST), Universitas Jambi, Jambi, Indonesia
[2]Department of Mathematics Education, Farhangian University, Tehran, Iran

**Abstract.** Binomial negative regression is able to handle poisson regression problem with underdispersion assumption. When the data has hierarchy and level that need to be calculated, regression is no longer appropriate to solve this problem, therefore binomial negative regression is used. To solve multilevel binomial negative regression modeling, several steps need to be fulfill: poisson assumption test and underdispersion assumption test, parameter estimation with expectation-maximization algorithm, variance components estimation, feasibility test with likelihood ratio test, significance parameter test with wald test and determining the best model. This research modeled the numbers of neonatal death in district as cluster 1 and small public health center as cluster 2, in the correlation with the number of visit on trimester 1 and 3, number of pregnant mother who have Tetanus Diphtheria vaccination, assumed number of neonatal babies with complication disease, numbers of babies who got Hepatitis B vaccination less than 24 hour, numbers of babies who got BCG vaccination and also number of visit neonatal 1 and 3. The result shows that number of neonatal death is only affected by number of babies who had Hepatitis B vaccination less than 24 hour.



***Corresponding Author :***
Corry Sormin
Department of Mathematics, Faculty of Science and Technology (FST),
Universitas Jambi, Jambi, Indonesia
Email : corry.sormin@unja.ac.id

## 1. Introduction

Regression is one of statistical analysis that aim to understand how much the relationship and the effect of independent variable against dependent variable. Normally in regression analysis, the scale of dependent variable data is continuous variable, but not all event are continuous scale data, they could be also in the form of discrete data [1-2]. For discrete scale data, one of regression analysis that we can use is Poisson regression. Poisson regression is one of regression that didn't required normal distribution data, it only required that the data is poisson distributed. When the data is not able to fulfil the assumption variant equal to mean, that mean that this regression is not appropriate on modeling the problem [3-5].

One of the appropriate regression to model this problem is binomial negative regression. This regression can solve problem with variance is smaller than mean data, or what is known as underdispersion [6-13]. And also, when the data has hierarchy and has clusters that need to be considered during analyzing, the regression need to be reconsidered therefore the best result for the model could be achieved. Multilevel model is the best solution for this kind of problem [14].

A study used zero inflated negative binomial regression to see the effect of metal exposure on Charlson's comorbidity. It was found that cadmium and manganese affect the increase in death, while selenium and lead are good for health [15]. Then, other research regarding neonatal death resulting a binomial negative regression model, which is not good enough to use. One of the parameter that assumed to affect that result is cluster, which is need to be considered [16].

On that research, there are data from district and small public health center that is not considered on the research. In order to try to improve the research, in this research we try to reconsider the effect of both cluster using multilevel modeling. This research is important in order to minimize or even dismissed neonatal death by determining the factors that cause the death.

## 2. Method

Multilevel modelling for two level on binomial negative regression is defined as followed [17]

$$P(Y_{ij}) = \exp(\beta_{0j} + \beta_{1j}x_{1ij} + \cdots + \beta_{pj}x_{pij}) + \varepsilon_{ij}$$

In binomial negative distribution, correlation function that being used is log link in the form of $\log(\mu_i) = \boldsymbol{X_i^T\beta}$. To estimate the parameter, penalized log-likelihood approach was done, which is noted by $l = l_1 + l_2$ with $l_1$ is log-likelihood fixed effect function and $l_2$ is log-likelihood from random effect function. Negative value of $l_2$ could be seen from penalty function for random effect if random effect is assumed as parameters:

$$\begin{aligned} l_1 &= \textstyle\sum_{y_{ij}=0} \log\left(\frac{\exp(\xi_{ij})+\exp(-t_{ij})}{1+\exp(\xi_{ij})}\right) + \sum_{y_{ij}>0} y_{ij}\log(t_{ij}) \\ &\quad + \textstyle\sum_{y_{ij}>0} y_{ij}\log(1+ry_{ij}) - \sum_{y_{ij}>0}\log(1+ry_{ij}) - \sum_{y_{ij}>0}\log(y_{ij}!) \\ &\quad - \textstyle\sum_{y_{ij}>0} t_{ij}(1+ry_{ij}) - \sum_{y_{ij}>0}\log(1+\exp(\xi_{ij})) \\ l_2 &= -\tfrac{1}{2}[m\log(2\pi\sigma_u^2) + \sigma_u^{-2}u^Tu + n\log(2\pi\sigma_v^2) + +\sigma_v^{-2}v^Tv] \\ &\quad -\tfrac{1}{2}[m\log(2\pi\sigma_w^2) + \sigma_w^{-2}w^Tw + n\log(2\pi\sigma_s^2) + \sigma_s^{-2}s^Ts] \end{aligned}$$

With

$$t_{ijk} = \frac{\exp(\eta_{ij})}{1 + rexp(\eta_{ij})}$$

Estimation by maximizing $l_1$, using variant component fixed in current value, later on, updating value is achieved by using restricted maximum likelihood estimation through inclusion from $l_2$.

Algorithm expectation-maximization (EM) is used to make sure conversion and stabilization on parameter and random effect $l_1$ estimation. Algorithm EM is done by processing:

a. Changing $z_{ijk}$ with conditional expectation $z_{ijk}^{(g)}$, where $g$ notating iteration-$g$ under *current value* from estimation $\hat{\alpha}^{(g)}, \hat{w}^{(g)}, \hat{s}^{(g)}, \hat{\beta}^{(g)}, \hat{u}^{(g)}$ dan $\hat{v}^{(g)}$ (as **step *E***):

$$z_{ij}^{(g)} = \begin{cases} \dfrac{1}{1+\exp\left(-\left(\boldsymbol{a}_{ij}^T\widehat{\boldsymbol{\alpha}}^{(\boldsymbol{g})} + \widehat{\boldsymbol{w}}_i^{(\boldsymbol{g})} + \hat{\boldsymbol{s}}_{ij}^{(\boldsymbol{g})}\right) - \hat{t}_{ij}^{(g)}\right)}, & \text{if } y_{ij} = 0 \\ 0, & \text{if } y_{ij} \geq 1 \end{cases}$$

were

$$\hat{t}_{ij}^{(g)} = \frac{\exp\left(\boldsymbol{x}_{ij}^T\widehat{\boldsymbol{\beta}}^{(\boldsymbol{g})} + \widehat{\boldsymbol{u}}_i^{(\boldsymbol{g})} + \hat{v}_{ij}^{(g)}\right)}{1 + rexp\left(\boldsymbol{x}_{ij}^T\widehat{\boldsymbol{\beta}}^{(\boldsymbol{g})} + \widehat{\boldsymbol{u}}_i^{(\boldsymbol{g})} + \hat{v}_{ij}^{(g)}\right)}$$

b. With $z_{ijk}$ fix on $z_{ijk}^{(g)}$ and maximizing $l_\xi$ and $l_\eta$ separately (as **step M**) to $\left\{\widehat{\boldsymbol{\alpha}}^{(\boldsymbol{g+1})}, \widehat{\boldsymbol{w}}^{(\boldsymbol{g+1})}, \hat{\boldsymbol{s}}^{(\boldsymbol{g+1})}\right\}$ and $\left\{\widehat{\boldsymbol{\beta}}^{(\boldsymbol{g+1})}, \widehat{\boldsymbol{u}}^{(\boldsymbol{g+1})}, \widehat{\boldsymbol{v}}^{(\boldsymbol{g+1})}\right\}$, using two set of recursive equation where partition orthogonal $l_c = l_\xi + l_\eta$.

Next estimation of the component is done by:

1. Estimation of variant from random effect need calculation from information matrix. Information matrix achieved by:

$$\Im_{\alpha,w,s,\beta,u,v} = H + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_w^{-2}I_m & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_s^{-2}I_n & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_u^{-2}I_m & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_v^{-2}I_n \end{bmatrix}$$

with

$$H = \begin{bmatrix} A^T & 0 \\ R_w^T & 0 \\ R_s^T & 0 \\ 0 & X^T \\ 0 & R_u^T \\ 0 & R_v^T \end{bmatrix} \begin{bmatrix} E\left(-\frac{\partial^2 l}{\partial\xi\partial\xi^T}\right) & E\left(-\frac{\partial^2 l}{\partial\xi\partial\eta^T}\right) \\ E\left(-\frac{\partial^2 l}{\partial\xi\partial\eta^T}\right) & E\left(\frac{\partial^2 l}{\partial\eta\partial\eta^T}\right) \end{bmatrix} \begin{bmatrix} A & R_w & R_s & 0 & 0 & 0 \\ 0 & 0 & 0 & X & R_u & R_v \end{bmatrix}$$

2. Square root of diagonal element $V_{11}$ and $V_{44}$ is a standard error from regression coefficient $\alpha$ and $\beta$.
3. Variant asymptotic from estimator in variant component is achieved by inversion of information matrix restricted maximum likelihood as follow

$$V = var\begin{bmatrix} \sigma_w^2 \\ \sigma_s^2 \\ \sigma_u^2 \\ \sigma_v^2 \end{bmatrix} = 2\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

The following are the stages of the research presented in the flowchart in **Figure 1**:

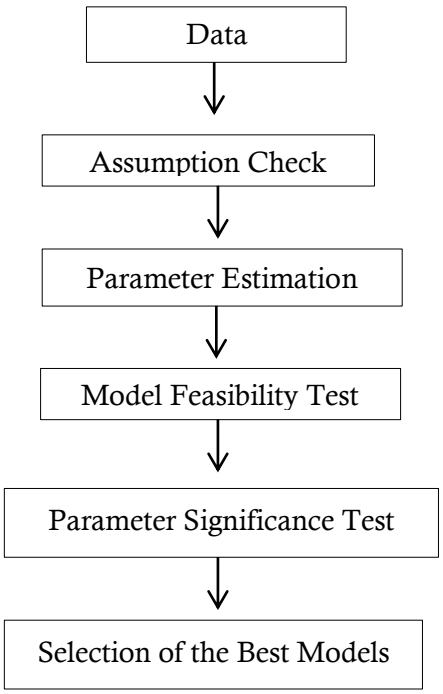


**Figure 1**. Flowchart of research

## 3. Results and Discussion

This study used secondary data obtained from the Jambi City Health Office. The research variables used were the numbers of neonatal death as a response variable ($Y$) and as a predictor variable, namely the number of visit on trimester 1 ($X_1$) and 3 ($X_2$), number of pregnant mother who have Tetanus Diphtheria vaccination ($X_3$), assumed number of neonatal babies with complication disease ($X_4$), numbers of babies who got Hepatitis B vaccination less than 24 hour ($X_5$), numbers of babies who got BCG vaccination ($X_6$) and also number of visit neonatal 1 ($X_7$) and 3 ($X_8$).

As cluster one is the district and cluster two is the public health center. Then it is tested whether the assumptions fulfill, namely the assumptions of a Poisson distribution. This research shows that variable dependent is Poisson distributed can be proved from **Asymp.Sig. (2-tailed)** result which is 1.000 which indicate that the data is Poisson distributed and variant is 0.197 and mean is 0.25. This is indicated that variant is smaller than mean or in other word the data is underdispersion [18-24]. Next, by using parameter estimation, multilevel model is achieved as:

$$P(Y_{ij}) = \exp(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{5ij} + \beta_7 x_{5ij} + \beta_8 x_{5ij})$$

With estimation result as follow:

**Table 1.** Parameter estimation for early model

| Parameters | Estimations | P-Value |
|---|---|---|
| $\boldsymbol{\beta_0}$ | -3.798646 | 0.046 |
| $\boldsymbol{\beta_1}$ | 0.005321 | 0.808 |
| $\boldsymbol{\beta_2}$ | 0.001793 | 0.945 |
| $\boldsymbol{\beta_3}$ | 0.002735 | 0.778 |
| $\boldsymbol{\beta_4}$ | -1.635887 | 0.418 |
| $\boldsymbol{\beta_5}$ | 0.007468 | 0.667 |
| $\boldsymbol{\beta_6}$ | -0.000362 | 0.982 |
| $\boldsymbol{\beta_7}$ | 0.229368 | 0.420 |
| $\boldsymbol{\beta_8}$ | 0.004048 | 0.641 |
| $\boldsymbol{\sigma^2}$ (Public health center) | 1.69e-07 | |
| $\boldsymbol{\sigma^2}$ (District) | 8.991e-07 | |

It can be seen that there is no significant variable. This is proved by likelihood ratio test resulting log-likelihood is -9.16631 which mean model is well fit to be used. Next significant parameter test shows that only parameter $\beta_0$ is significant. It is needed a better model by extracting unsignificant variable one by one. Parameter estimation for final model is shown as follow:

**Table 2.** Parameter Estimation for Final Model

| Parameters | Estimations | P-Value |
|---|---|---|
| $\boldsymbol{\beta_0}$ | -4.06859 | 0.018 |
| $\boldsymbol{\beta_5}$ | 0.00386 | 0.064 |
| $\boldsymbol{\sigma^2}$ (Public health center) | 3.645e-06 | |
| $\boldsymbol{\sigma^2}$ (district) | 1.135e-07 | |

Resulting multilevel model as follow:

$$P(Y_{ij}) = \exp(-4.06859 + 0.00386x_{5ij})$$

It can be seen that there are significant variable using 10% error. It also can be proved by likelihood ratio test resulting log-likelihood is 9.98621 which mean that the model can be used. Next by using parameter significant test it can be seen that only parameter $\beta_0$ and parameter $\beta_5$ are significant. Multilevel model resulting the probability of neonatal death in Jambi is 0.0171 and also if the baby got Hepatitis B vaccination less than 24 hours, then the probability of neonatal death is equal to 0.00172 or less than 1% by considering public health center and district as level.

## 4. Conclusion

Although the chance of neonatal death is small, namely 1%, it is necessary to take into account the feasibility of the public health center and existing facilities in each district. Also, to parents of babies to always supervise the administration of vaccinations, especially newborns who need several types of vaccinations that are less than 24 hours, in this study especially the administration of the Hepatitis B vaccination.

## References

[1] Sormin, C. (2013). *Aplikasi Regresi Poisson pada Faktor-Faktor yang Mempengaruhi Banyaknya Pasien Diabetes Mellitus*. Skripsi. Tidak diterbitkan. Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Negeri Yogyakarta: Yogyakarta.

[2] Dobson, A. J., & Barnett, A. (2008). An Introduction to Generalized Linear Models. *CRC press*. Boca Raton.
[3] Inan G & Das K. (2017). A Score Test for Testing a Marginalized Zero-Inflated Poisson Regression Model Against a Marginalized Zero-Inflated Negative Binomial Regression. *Journal of Agricultural Biological and Environmental Statistics*, 23, 113-128.
[4] Almasi, A, dkk. (2015). Multilevel zero-inflated Generalized Poisson regression modelling for dispersed correlated count data. *Statistical Methodology*, 30, 1-14.
[5] Ismail, N & Jemain, A. A. (2007). Handling overdispersion with Negative Binomial and Generalized Poisson Regression Models. *Casualty Actuarial Society Forum*, 103-158.
[6] Febritasari P, Wardhani NWS, & Sa'adah U. (2019). Generalized Linier Autoregressive Moving Average (GLARMA) Negative Binomial Regression Models with Metropolis Hasting Algorithm. *IOP Conf. Series: Materials Science and Engineering*, 546, 1-6.
[7] Hafemeister C & Satija R. (2019). Normalization and Variance Stabilization of Single-cell RNA-seq data using regularized Negative Binomial Regression. *Genome Biology*, 20, 296-311.
[8] Dadaneh S.Z, Zhou M, & Qian X. (2018). Bayesian Negative Binomial Regression for Differential Expression with Confounding Factors. *Bioinformatics*, 35, 2346-2359.
[9] Gomes MJTL, Cunto F, & Dilva AR. (2017). Geographically Weighted Negative Binomial Regression Applied to Zonal level safety performance models. *Accident Analysis & Prevention*, 106, 254-261.
[10] Najaf P, Duddu VR & Pulugurtha SS. (2017). Predictability and interpretability of hybrid link-level crash frequency models for urban arterial compared to cluster-based and general negative binomial regression models. *International Journal of Injury Control and Safety Promotion*, 25, 3-13.
[11] Hilbe, J. M. (2011). Negative Binomial Regression. *Cambridge University Press*. New York.
[12] Su Z, Hu H, Tigabu M, Wang G, Zeng A, & Guo F. (2019). Geographically Weighted Negative Binomial Regression Model Predicts Wildfire Occurrence in the Great Xing'an Mountains Better Than Negative Binomial Model. *Forests*, 10, 377-393.
[13] Zou Y, Ash JE, Park B & Lord D. (2017). Empirical Bayes Estimates of Finite Mixture of Negative Binomial Regression Model and its Application to Highway Safety. *Journal of Applied Statistics*, 45, 1652-1669.
[14] Sormin, C. (2017). *Model Multilevel Regresi Poisson Tergeneralisasi Zero-Inflated*. Tesis. Tidak diterbitkan. Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Gadjah Mada: Yogyakarta.
[15] Zhao H, Pan Y, Wang C, Guo Y, Yao N, Wang H & Li B. (2021). The Effects of Metal Exposures on Charlson Comorbidity Index Using Zero-Inflated Negative Binomial Regression Model. *NHANES 2011-2016 Biological trace element research*, 199, 2104-2111.
[16] Sormin, C. & Gusmanely Z (2020). Generalized Poisson Regression Type-II at Jambi City Health Office. *Eksakta Berkala Ilmiah Bidang MIPA*, 21, 54-58.
[17] Moghimbeigi, A, dkk. (2008). Multilevel zero-inflated negative binomial regression modelling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, 35, 1193-1202.
[18] Oztig LI & Askin OE. (2020). Human Mobility and coronavirus disease 2019 (COVID-19): A Negative Binomial Regression Analysis. *Public Health (London)*, 185, 364-367.
[19] Yee, Thomas W. (2020). The VGAM Package for Negative Binomial Regression. *Australian & New Zealand journal of statistics*, 62, 116-131.
[20] Tohari A, Chamidah N & Fatmawati. (2019). Modeling of HIV and AIDS in Indonesia Using Bivariate Negative Binomial Regression, *IOP conference series. Materials Science and Engineering*; 546, 52079.
[21] Kim J & Lee W. (2019). On testing the hidden heterogeneity in negative binomial regression models. *Metrika*, 82, 457-470.

[22] Weng J, Yang D, Qian T & Huang Z. (2018). Combining Zero-inflated Negative Binomial Regression with MLRT Techniques: An Approach to Evaluating Shipping Accident Casualties. *Ocean Engineering*, 166, 135-144.
[23] Ardiles, L.G. et. Al (2018). Negative Binomial Regression Model for Analysis of the Relationship Between Hospitalization and Air Pollution. *Atmospheric Pollution Research*, 9, 333-341.
[24] Liu C, Zhao M & Sharma A. (2018). Multivariate Random Parameters Zero-Inflated Negative Binomial Regression for Analyzing Urban Midblock Crashes. *Analytic Methods in Accident Research*, 17, 32-46.