Parameter Estimation and Hypothesis Testing
Geographically Weighted Bivariate Zero-Inflated Poisson

Joice Pangulimang¹, Purhadi², Sutikno³
¹Student of Magister Statistics, Institute of Technology SepuluhNopember, Surabaya, Indonesia
²,³Department of Statistics, Institute of Technology SepuluhNopember, Surabaya, Indonesia
pangulimang_joice@yahoo.com

Abstract—Statistical methods which often used to analyze count data is Poisson regression. However, Poisson regression is not appropriate to be used in analyzing a Zero-inflated count data so that the method used is the Zero-inflated Poisson (ZIP). To model a pair of count data with Poisson distribution and has correlation with some Zero-inflated predictor variable, Bivariate Zero-inflated Poisson Regression (BZIPR) can be used. Therefore, this study was developed in Geographically Weighted Bivariate Zero-inflated Poisson Regression (GWBZIPR). GWBZIP regression parameter estimation was conducted using Maximum Likelihood Estimation (MLE), whereas hypothesis testing was conducted using Maximum Likelihood Ratio Test (MLRT).

Keywords: Bivariate Poisson Regression, MLE, MLRT, BZIPR, GWBZIPR

I. INTRODUCTION

Regression analysis is a statistical method that used in various fields, as it provides a simple concept to investigate the functional relationship between the response variable and the predictor variable. If the response consists of positive integers or non-negative valued and stated the number of observations, the response variable is called as a discrete count data [2]. The statistical method used to analyze count data is the Poisson regression [7]. Count data exaggerated with zero value is referred to as Zero-inflated. Zero-inflated can lead to over-dispersion or the mean and variance are not the same [1]. It is therefore not appropriate to use poisson regression in analysing the Zero-inflated count data. Zero-inflated Poisson (ZIP) regression considered as more proper method in dealing with Zero-inflated count data[3]. ZIP regression can be applied to the case of univariate, bivariate and multivariate. Bzipped regression model can produce estimated values of parameters that are global or equal to the entire location. This study developed a Geographically Weighted Bivariate Zero-inflated Poisson Regression (GWBZIPR) model to perform estimation and hypothesis testing using Maximum Likelihood Estimation (MLE) and the determination of the test statistics is using Maximum Likelihood Ratio Test (MLRT).

II. LITERATURE

A. Bivariate Zero-Inflated Poisson Regression

Let $Y_1$ and $Y_2$ is a random variable that jointly bivariate Poisson distribution with probability function as follow:

$$f(y_1, y_2) = \begin{cases} e^{-(\mu_1 + \mu_2 + j_0)} \sum_{k=0}^{\min(y_1, y_2)} \frac{\mu_1^{y_1-k} \mu_2^{y_2-k} \mu_0^k}{(y_1-k)! (y_2-k)! k!} & (y_1, y_2) = 0, 1, 2, \ldots, (y_1, y_2) \text{ others} \\ 0 \end{cases}$$

(1)

Where the regression equation as follows:

M - 73
\[(Y_{ij}, Y_{2j}) - PB(\mu_{i1}, \mu_{21}, \mu_{0})\]
\[\mu_{ji} + \mu_{0} = e^{x_{ji}^{\beta_{j}}}; j = 1, 2\]

Estimation method used in Bivariate Poisson Regression is Maximum Likelihood Estimation (MLE). The method for calculating test statistic on parameter test is Maximum Likelihood Ratio Test (MLRT):

\[D(\hat{\beta}) = -2\ln \left( \frac{L(\hat{\phi})}{L(\hat{\Theta})} \right) = 2 \left( \ln L(\hat{\Theta}) - \ln L(\hat{\phi}) \right)\]

where

- \(L(\hat{\phi})\): Maximum Likelihood function for complete model with predictor variable
- \(L(\hat{\Theta})\): Maximum Likelihood function for simple model without predictor variable

B. Generalized Poisson Regression

A pair of count data poisson distribution which has zero value deal on response variables can be analyzed by Bivariate Zero-inflated Poisson Regression. Distribution of Bivariate Zero-inflated Poisson is as follows [6]:

\[f(Y_{1}, Y_{2}) = \begin{cases} 
(1 - \pi) + \pi e^{-(\mu_{1} + \mu_{2})} \left( 1 + \alpha \left( 1 - e^{-\mu_{r}} \right) \right), & (y_{1}, y_{2}) = (0,0) \\
\pi e^{-(\mu_{1} + \mu_{2})} \mu_{1}^{y_{1}} \mu_{2}^{y_{2}} \left( 1 + \alpha (e^{\mu_{1}} - e^{-\mu_{r}}) (e^{\mu_{2}} - e^{-\mu_{r}}) \right), & (y_{1}, y_{2}) \neq (0,0) 
\end{cases}\]

The regression equation is:

\[\mu_{1} = e^{x_{1}^{\beta_{1}}} \text{ dan } \mu_{2} = e^{x_{2}^{\beta_{2}}} \]
\[\pi = e^{x_{1}^{\gamma}} \text{ dan } (1 - \pi) = \frac{1}{1 + e^{x_{1}^{\gamma}}} \]

where \(c = 1 - \frac{1}{e}\)

C. Correlation Test

Correlation analysis is usually used to measure the linear relationship between the response variable and the predictor variables through a number called the coefficient of correlation. The value of the correlation coefficient ranges between -1 and 1, which shows the relationship of positive and negative. If the correlation value is positive or negative approach of 1 means the two variables have a close relationship. The hypothesis test of correlation between response variable following:

- \(H_{o}\): there is no relationship between \(Y_{1}\) and \(Y_{2}\)
- \(H_{i}\): there is relationship between \(Y_{1}\) dan \(Y_{2}\)

The test statistic

\[t = \frac{r_{yy}}{\sqrt{1 - (r_{yy})^2}} \text{ where } r_{yy} = \frac{\sum_{i=1}^{n} (y_{i1} - \bar{Y}_{1}) (y_{i2} - \bar{Y}_{2})}{\sqrt{\sum_{i=1}^{n} (y_{i1} - \bar{Y}_{1})^2 \sum_{i=1}^{n} (y_{i2} - \bar{Y}_{2})^2}}\]

\(H_{o}\) rejected if \(T > t_{\nu/2}\)

D. Multicollinearity Test

Multicollinearity among predictor variables may result inaccurate parameter estimation. Detection of multicollinearity in Poisson regression modeling is very important as correlation between predictor variables with other predictor variables indicate that those two variables have a comparable value. According to [5], the detection of multicollinearity can be done using Variance Inflation Factor (VIF).

VIF calculation using the following formula:
value \( VIF = \frac{1}{1 - R_i^2} \)

E. Spatial Effects

Spatial data on each observation has the characteristics that identify a pair of geographic coordinates or the location of each data covering areas such as agriculture, geology, environmental science and economics [8]. Modeling spatial data can be grouped by two types of spatial and spatial point of the area.

In modeling the spatial data, spatial weighting matrix is required. Weighting matrix used to represent the scope of information and spatial effects from a location in the system such as geography or the coordinates of latitude and longitude. Spatial heterogeneity between one and other locations indicated by the weighting matrix \( W(u_i, v_i) \) whose elements are a function of the Euclidean distance between locations.

Form of weighting function of the Euclidean distance use the kernel functions follow,

\[
W_d = \begin{cases} 
\left(1 - \left(\frac{d_d}{h}\right)^2\right)^2, & \text{if } d_d \leq h \\
0, & \text{if } d_d > h
\end{cases}
\]

The identification of spatial heterogeneity is tested using Koenker-Basset.

Hypothesis:

\( H_0 \) : no heterogeneity

\( H_1 \) : heterogeneity

Statistics test :

\[
Z = \frac{\hat{\gamma}_1}{SE(\hat{\gamma}_1)}
\]

\( H_0 \) rejected if \( |Z| > Z_{0.025} \).

F. Geographically Weighted Bivariate Zero-Inflated Poisson Regression (GWBZIPR)

Model GWBZIPR is a local form of the Zero-inflated Poisson regression model estimator that will generate local model parameters to each location. GWBZIP distribution is:

\[
f(Y_{ij}, Y_{ij}^*) = \begin{cases} 
(1 - \pi_j) + \pi_j e^{-(\mu_j + \mu_j^*)} \left(1 + \alpha(1 - e^{-\mu_j^*})\right), & (y_{ij}, y_{ij}^*) = (0,0) \\
\pi_j e^{-(\mu_j + \mu_j^*)} \mu_j^{y_{ij}} \mu_j^{y_{ij}^*} \left(1 + \alpha(e^{-\mu_j^*} - e^{-\mu_j})e^{-\mu_j^*} - e^{-\mu_j}\right), & (y_{ij}, y_{ij}^*) \neq (0,0)
\end{cases}
\]

Where

\[
\pi_j = e^{s_j^1}, \mu_j = e^{s_j^2}
\]

III. Method

Steps to get parameter estimation of GWBZIPR is to determine the likelihood function based on GWBZIP distribution, form and function of the natural logarithm likelihood. And the find the first derivative function of the natural logarithm likelihood under population and define the equation to be equal to zero. If there is no close form solution, Newton-Raphson iteration method is performed to find out the estimator. To test the hypothesis a maximum likelihood ratio test is performed.
IV. RESULT AND DISCUSSION

GWBZIPR model was developed from BZIPR model which use MLE as a method in estimating parameter and resulting a global parameter estimation applied to all locations. Therefore, the estimation of model parameters GWBZIPR will use the MLE method with the likelihood function as follows:

\[ L(\alpha, \gamma(u, v), \beta_1(u, v), \beta_2(u, v), j = 1, 2, \ldots, n) = \prod_{j=1}^{n} g_j + \prod_{j=1}^{n} h_j \]  

(7)

Where

\[ g_j = \frac{1}{1 + e^{s_j y_j (u, v)}} + e^{s_j y_j (u, v)} \exp\left(-e^{s_j y_j (u, v)} - e^{s_j y_j (u, v)}\right) \left[ 1 + \alpha \left(1 - \exp\left(-e^{s_j y_j (u, v)}c\right)\right) \right] \]

\[ h_j = (r(s)) \]

\[ r = \frac{e^{s_j y_j (u, v)} \exp\left(-e^{s_j y_j (u, v)} - e^{s_j y_j (u, v)}\right) (e^{s_j y_j (u, v)})}{1 + e^{s_j y_j (u, v)}} \]

\[ s = 1 + \alpha (e^{-\gamma_j} - \exp(-e^{s_j y_j (u, v)}c))(e^{-\gamma_j} - \exp(-e^{s_j y_j (u, v)}c)) \]

To estimate the parameters \((\alpha, \gamma(u, v), \beta_1(u, v), \beta_2(u, v))\) is

\[ \ell = \ln L(\alpha, \gamma(u, v), \beta_1(u, v), \beta_2(u, v)) = \ln \prod_{j=1}^{n} (g_j)^{a_j} (h_j)^{n_j} w_j \]

(8)

\[ \ell = \sum_{j=1}^{n} (1 - a_j) \ln(g_j) w_j + \sum_{j=1}^{n} (a_j) \ln(h_j) w_j \]

Where

\[ \mu_j = e^{s_j y_j (u, v)} \] and \[ \mu_j = e^{s_j y_j (u, v)} \]

\[ \pi_j = \frac{e^{s_j y_j (u, v)}}{1 + e^{s_j y_j (u, v)}} \]

\[ a_j = 1 \text{ if } (y_{ij}, y_{ij}) \neq (0, 0) \text{ and } a_j = 0 \text{ otherwise. With } w_j \text{ is the geographical weighting. Ln likelihood of existing functions in equation (8) downgraded to each parameter } \beta_1^T(u, v), \beta_2^T(u, v), \gamma^T(u, v), \alpha \]

\[ \frac{\partial \ell}{\partial \beta_1^T(u, v)} = \sum_{j=1}^{n} (1 - a_j) P(S)(w_j) + \sum_{j=1}^{n} (a_j) R w_j \]

(9)

With

\[ P = \frac{e^{s_j y_j (u, v)} (\exp(-e^{s_j y_j (u, v)} - e^{s_j y_j (u, v)}))^{x_j^T}}{1 + e^{s_j y_j (u, v)} (\exp(-e^{s_j y_j (u, v)} - e^{s_j y_j (u, v)}))(1 + \alpha (1 - \exp(-e^{s_j y_j (u, v)}c))(1 - \exp(-e^{s_j y_j (u, v)}c))} \]

\[ S = 1 + \alpha (1 - \exp(-e^{s_j y_j (u, v)}c))(1 - \exp(-e^{s_j y_j (u, v)}c)) + \alpha c(1 - \exp(-e^{s_j y_j (u, v)}c))(1 - \exp(-e^{s_j y_j (u, v)}c)) \]

\[ R = \sum_{j=1}^{n} a_j (-e^{s_j y_j (u, v)}) x_j^T + y_j x_j^T + \sum_{j=1}^{n} a_j \alpha c x_j^T (e^{s_j y_j (u, v)})(1 - \exp(-e^{s_j y_j (u, v)}c))(1 - \exp(-e^{s_j y_j (u, v)}c)) \]

\[ \frac{\partial \ell}{\partial \beta_2^T(u, v)} = \sum_{j=1}^{n} (1 - a_j) P/S w_j + \sum_{j=1}^{n} (a_j) R w_j \]

(10)

Where
\[ P = \left[ \left( e^{\gamma u v} \left( \exp(-e^{\gamma u v}) - e^{\gamma u v} \right) \right) \left( e^{\gamma u v} \right) \right] \]

\[ S = 1 + e^{\gamma u v} \left( \exp(-e^{\gamma u v}) - e^{\gamma u v} \right) \left[ 1 + \alpha \left( 1 - \exp(-e^{\gamma u v}) \right) \right] \]

\[ R = \sum_{j=1}^{n} a_j \left( e^{\gamma u v} \right) x_j^T + y_j x_j^T + \sum_{j=1}^{n} a_j \frac{\partial \ell}{\partial a_j} \left( e^{\gamma u v} \right) w_j^T + \]

\[ \frac{\partial \ell}{\partial \alpha} = \sum_{j=1}^{n} \left( 1 - a_j \right) e^{\gamma u v} \left( \exp(-e^{\gamma u v}) - e^{\gamma u v} \right) \left[ 1 + \alpha \left( 1 - \exp(-e^{\gamma u v}) \right) \right] w_j^T \]

The results of the first derivative of each parameter is explicit or not close the form that will be solved by Newton-Raphson iteration method as follows:

\[ \theta_{m+1} = \theta_m - H^{-1}(\theta_m) g(\theta_m) \]

Where

\[ \theta = (\alpha, \gamma, u, v) \]

\[ g(\theta) = \left( \frac{\partial \ell}{\partial \alpha}, \frac{\partial \ell}{\partial \gamma}, \frac{\partial \ell}{\partial u}, \frac{\partial \ell}{\partial v} \right)^T \]

\[ H(\theta_m) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \alpha^2} & \frac{\partial^2 \ell}{\partial \alpha \gamma} & \frac{\partial^2 \ell}{\partial \alpha u} & \frac{\partial^2 \ell}{\partial \alpha v} \\ \frac{\partial^2 \ell}{\partial \gamma \alpha} & \frac{\partial^2 \ell}{\partial \gamma^2} & \frac{\partial^2 \ell}{\partial \gamma u} & \frac{\partial^2 \ell}{\partial \gamma v} \\ \frac{\partial^2 \ell}{\partial u \alpha} & \frac{\partial^2 \ell}{\partial u \gamma} & \frac{\partial^2 \ell}{\partial u^2} & \frac{\partial^2 \ell}{\partial u v} \\ \frac{\partial^2 \ell}{\partial v \alpha} & \frac{\partial^2 \ell}{\partial v \gamma} & \frac{\partial^2 \ell}{\partial v u} & \frac{\partial^2 \ell}{\partial v^2} \end{bmatrix} \]

Hessian matrix is a matrix containing the second derivative of each parameter \( \theta = (\alpha, \gamma, u, v) \). Steps in parameter estimation using Newton Raphson iteration is as follows:

1. Determining the initial value of parameter.
2. Forming vector \( g(\theta_m) \) by substituting equation (9), (10), (11),and (12) into the equation (15)
3. Form the Hessian matrix \( H(\theta_m) \) by substituting the second derivative of equation (9), (10), (11) and (12) to the equation (16).
4. Including the value of $\mathbf{\theta}_{(m)}(u, v)$ into vector elements $g(\mathbf{\theta}(u, v))$ and matrix $\mathbf{H}(\mathbf{\theta}_{(m)}(u, v))$ thus obtained gradient vector $g(\mathbf{\theta}_{(m)}(u, v))$ and Hessian matrix $\mathbf{H}(\mathbf{\theta}_{(m)}(u, v))$.

5. Starting from $m = 0$ to iterate the equation (13). $\mathbf{\theta}_{(m)}(u, v)$ is a set of parameter estimator which converges when iteration to $m$.

6. If you have not obtained when the parameter estimation convergent iteration to $m$, then proceed back to step 5 to $m+1$ iteration. Iteration will stop when the value of $\| \mathbf{\theta}_{(m+1)}(u, v) - \mathbf{\theta}_{(m)}(u, v) \|$ is less than $\varepsilon$, $\varepsilon$ is a number that is very small.

The test statistic in testing parameters for the GWR ZIP model determined by the likelihood function of the model is $L(\Omega)$ which is a maximum likelihood of complete models involving the predictor variables and $L(\alpha)$ value is the maximum likelihood for a simple model without involving predictors. To determine the test statistic in testing parameters in this study, Maximum Likelihood Ratio Test (MLRT) is used.

$$G = -2 \ln \left( \frac{L(\alpha)}{L(\Omega)} \right)$$

$$= 2 \ln \left( L(\Omega) - L(\alpha) \right)$$

RESULT

Estimation of model parameters of Geographically Weighted Bivariate Zero-inflated Poisson Regression using Maximum Likelihood Estimation by Newton-Raphson iteration, produces parameter estimates which do not close the form. The hypothesis is tested using Maximum Likelihood Ratio Test conducted simultaneously and partially by comparing $H_0$ and the following possibilities population.

REFERENCES