M – 08

Parameter Estimation and Statistical Test in Modeling Geographically Weighted Poisson Inverse Gaussian Regression

Ima Purnamasari[1], I Nyoman Latra[2], Purhadi[3]
[1]Ima Purnamasari (Department of Statistics, Institute of Technology Sepuluh Nopember)
[2]I Nyoman Latra (Department of Statistics, Institute of Technology Sepuluh Nopember)
[3]Purhadi (Department of Statistics, Institute of Technology Sepuluh Nopember)
imapurnama89@gmail.com

***Abstract*—**Poisson regression is a member of Generalized Linear Models (GLMs) family which is derived from a Poisson distribution. Poisson distribution is a discrete distribution with the value of positive integer random variable so that it becomes a good choice for count data modeling. Poisson distribution is only determined by one parameter that defines both the mean and variance of the distribution. In Poisson regression there is an assumption that must be complete, that are mean and variance of the response variable should be the same (equidispersion). While, some of the count data potentially violates these assumptions because due to overdispersion (variance is greater than the mean). Therefore, modeling the count data is not sufficient with a simple Poisson regression. Poisson Inverse Gaussian Regression (PIGR) is a regression which is derived from mixed Poisson distribution that is designed for count data modeling with overdispersi case. PIGR will produce global model that is assumed to be valid in all areas in which the data was taken. But of course every region has different geographical conditions, social, cultural and economic. Thus, the development of a regression model that considers spatial effect, which is Geographically Weighted Regression (GWR) needs to be employed. By providing a weighting based on the location of the region, the GWR models will generate different local models for each region. Furthermore, the response variable must follow the PIG distribution so development will be Geographically Weighted Poisson Inverse Gaussian Regression (GWPIGR). Parameter estimation is done using Maximum Likelihood Estimation (MLE) and hypothesis testing conducted by Maximum Likelihood Ratio Test (MLRT).

Keywords: *Geographically Weighted Poisson Inverse Gaussian Regression, Maximum Likelihood Estimation, Maximum Likelihood Ratio Test.*

I. INTRODUCTION

Poisson regression is one of the family members of Generalized Linear Models (GLMs) derived from a Poisson distribution. Poisson distribution is a discrete distribution with the value of the random variable a positive integer so it becomes a good choice for discrete data modeling. Poisson distribution is determined solely by a single parameter that defines both the mean and variance of the distribution. Thus, in a Poisson regression there is an assumption that must be fulfilled which is the mean and variance of the response variable should be the same (equidispersion). But most of the discrete data found in the case suffer overdispersion [1].

But in reality these assumptions violations often occur when variance is smaller than the mean (underdispersion) or variance is greater than the mean (overdispersion). To overcome overdispersion case, some form of statistical model is employed by mixing Poisson distribution with the other distribution both discrete and continuous (mixed Poisson distribution). Mixed Poisson distribution is an alternative solution to overcome overdispersion case, but only a few distributions that are often used in research due to complicated calculations. One of them is the Poisson Inverse Gaussian distribution (PIG) which is the mix of Poisson distribution with random effects that follow Inverse Gaussian distribution. This distribution was first introduced by Holla in 1966 [2].

The development of regression models that considers the spatial heterogeneity, is called Geographically Weighted Regression (GWR). By providing a weighting based on the position or distance of an observation area with the observation area other then GWR models will produce local models that vary in each region. So this research using modeling Poisson Inverse Gaussian Geographically Weighted Regression (GWPIGR).

II. METHODS

Regression analysis is a statistical method used to model the relationship between the response variable and one or more predictor variables. Not all regressions have a response variable that follows a normal distribution. If the response variable follows the Poisson Inverse Gaussian distribution (PIG), then the regression is called regression PIG. Regression PIG includes global regression, while the local form of PIG called Geographically Weighted Regression Poisson Inverse Gaussian (GWPIGR).

A. Poisson Distribution

Poisson distribution is a probability distribution for the events that happen rarely, where the observation depends on the specific time intervals or in a particular area with a discrete response and one or more independent predictor. The time interval can be measured in minute, day, week, month, even year [3]

Probability density function follows:

$$p(y;\ \mu) = \frac{e^{-\mu}\mu^{y}}{y!} \text{ for } y = 0, 1, 2, \dots \text{ and } \mu > 0 \quad (1)$$

With mean and variance of the same value is determined by:

$$E(Y) = Var(Y) = \mu \quad (2)$$

B. Inverse Gaussian Distribution

Inverse Gaussian has two parameters and probability density function that can be written as follows:

$$f(y) = (2\pi y^{3}\sigma)^{-0,5} e^{-(y-\mu)^{2}/2y\mu^{2}\sigma^{2}}, y > 0 \quad (3)$$

With mean and variance are written as:

$$E(Y) = \mu \text{ and } Var(Y) = \sigma^{2}\mu^{3} \quad (4)$$

Where σ^{2} is the parameter of dispersion. Inverse Gaussian is used in cases with extreme skewness. The name itself comes from the inverse Gaussian cumulant function which has inverse relationship with kumulant function (the natural logarithm of the function of MGF) normal distribution / Gaussian distribution [4].

C. Poisson Inverse Gaussian Distribution

PIG probability density distribution can be calculated as follows:

$$P(Y = y|\mu) = \frac{\mu^{y} e^{\frac{1}{\tau}}}{y!}\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}(2\mu\tau + 1)^{-\left(\frac{y-\frac{1}{2}}{2}\right)} K_{y-\frac{1}{2}}\left(\frac{1}{\tau}\sqrt{2\mu\tau + 1}\right) \quad (5)$$

While Bessel functions [5] is:

$$K_{\frac{1}{2}}(a) = K_{-\frac{1}{2}}(a) = \left(\frac{\pi}{2a}\right)^{\frac{1}{2}} e^{-a},$$
$$K_{\frac{3}{2}}(a) = \left(1 + \frac{1}{a}\right) K_{\frac{1}{2}}(a). \quad (6)$$

D. Poisson Inverse Gaussian Regression

Poisson Inverse Gaussian Regression model can be written as follows:

$$y_i \sim PIG\ [\mu_i]$$
$$\mu_i = e^{x_i^{T}\beta} \text{ or } ln(\mu_i) = \mathbf{x}_i^{T}\boldsymbol{\beta} \quad (7)$$

With

$$\mathbf{x}_i{}^{T} = [1 \quad x_{1i} \quad x_{2i} \quad \dots \quad x_{ki}]$$
$$\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \dots \quad \beta_k]^{T}$$

Where $i = 1, 2, \dots, n$ is the number of observations.

Probability density function follows::

$$P(Y=y_i|\mathbf{x}_i;\boldsymbol{\beta};\tau)=\left\{\frac{e^{x_i^T\beta}y_{ie}^{\frac{1}{\tau}}}{y_i!}\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}\left(2e^{x_i^T\beta}\tau+1\right)^{-\frac{\left(y_i-\frac{1}{2}\right)}{2}}K_{si}(z_i)\right\} \quad (8)$$

E. Procedures

Step 1. Determine the density function of the opportunities GWPIGR models. Step 2. Determine GWPIGR likelihood function on the model. Step 3. Determine the natural logarithm of the likelihood function. Step 4. Calculate first partial derivative of the natural logarithm function. Step 5. Calculate the second partial derivative of the natural logarithm function. Step 6. Calculate estimates $\boldsymbol{\beta}$ and τ. If the previous steps resulting equation that is not close the form, Fisher Scoring Algorithm is employed. Step 7. Do hypothesis testing simultaneously using MLRT and partialy using Z-test.

III. RESULT AND DISCUSSION

Spatial data is data collected from different spatial locations and indicates the existence of dependence between the measurement data by location. Consequently, if data model using ordinary linear regression, it will generate autocorrelation and heterogeneity in the data. There are several methods to overcome the problem, one of them is the Geographically Weighted Regression Poisson Inverse Gaussian method using Maximum Likelihood Estimation (MLE).

A. Parameter Estimation Geographically Weighted Poisson Inverse Gaussian Regression Model

Geographically Weighted Poisson Inverse Gaussian regression is a method that can be used to analyze the data count suffering overdispersion by considering the spatial aspect. This research will be carried out with the parameter estimation using Maximum Likelihood Estimation (MLE).

Geographically Weighted Poisson Inverse Gaussian Model:

$$\begin{aligned}\mu(u_i,v_i)&=exp\left(\beta_0(u_i,v_i)+\beta_1X_{1i}(u_i,v_i)+\beta_2X_{2i}(u_i,v_i)+\cdots+\beta_kX_{ki}(u_i,v_i)\right)\\&=e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}\end{aligned} \quad (9)$$

Where y_i is the value of the response variable observation i[th] location, x_{ik} is the value of predictor variable k observation at (u_i,v_i) locations, dan $\beta_k(u_i,v_i)$ is the regression coefficient for each location (u_i,v_i).

The first step in determining the parameter estimation is to determine the density function opportunities:

$$P(Y=y_i|\boldsymbol{x}_i;\boldsymbol{\beta};\tau)=\frac{e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}y_{ie}^{\frac{1}{\tau}}}{y_i!}\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}\left(2e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}\tau+1\right)^{-\frac{\left(y_i-\frac{1}{2}\right)}{2}}K_{si}(z_i) \quad (10)$$

The second step determines the likelihood function of the density function chances

$$L(\boldsymbol{\beta},\tau)=\prod_{i=1}^{n}\frac{e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}e^{\frac{1}{\tau}}}{y_i!}\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}\left(2e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}\tau+1\right)^{-\frac{\left(y_i-\frac{1}{2}\right)}{2}}K_{si}(z_i) \quad (11)$$

The third step is to transform the likelihood function ln:

$$\begin{aligned}l(\boldsymbol{\beta}(u_i,v_i),\tau)=&\sum_{i=1}^{n}y_i\ln\left(e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}\right)+\frac{n}{\tau}-\ln\left(\sum_{i=1}^{n}y_i!\right)+\frac{n}{2}\ln\left(\frac{2}{n}\right)-\frac{n}{2}\ln\tau\\&-\sum_{i=1}^{n}\left(\frac{2y_i-1}{4}\right)\ln\left(2e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}\tau+1\right)+\sum_{i=1}^{n}\ln K_{si}(Z_i)\end{aligned} \quad (12)$$

The fourth step is to find the first partial derivatives of the parameters $\beta(u_i,v_i)$ of the natural logarithm function by adding a weighting:

$$\frac{\partial l^*}{\partial\boldsymbol{\beta}(u_i,v_i)}=\sum_{i=1}^{n}\left\{\left(y_i-M_{(y_i)}(w_{ij})e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}\right)x_i^T\right\} \quad (13)$$

Next is determine first partial derivatives of the parameters τ:

$$\frac{\partial l^*}{\partial\tau}=\sum_{i=1}^{n}\left\{-\frac{1}{\tau^2}+\frac{M_{(y_i)}(w_{ij})\left(e^{x_i^T\boldsymbol{\beta}(u_i,v_i)}\tau+1\right)}{\tau^2}-\frac{y_i}{\tau}\right\} \quad (14)$$

Where

$$M_{(y_i)} = \frac{1}{\sqrt{2e^{x_i^T\beta(u_i,v_i)}\tau+1}} \frac{K_{y_i+\frac{1}{2}}(z)}{K_{y_i-\frac{1}{2}}(z)} \quad (15)$$

The fifth step is to determine a second partial derivatives against $\boldsymbol{\beta}(u_i, v_i)$ and τ. The second derivatives are written as follows:

$$\frac{\partial^2 l^*}{\partial\beta\partial\beta^T} = \frac{\sum_{i=1}^n \left\{\left(y_i - M_{(y_i)}(w_{ij})e^{x_i^T\beta(u_i,v_i)}\right)x_i^T\right\}}{\partial\beta^T} \quad (16)$$

For the next is the first derivative of the β second descent to τ

$$\frac{\partial^2 l}{\partial\beta\partial\tau} = \frac{\sum_{i=1}^n \left\{\left(y_i - M_{(y_i)}(w_{ij})e^{x_i^T\beta(u_i,v_i)}\right)x_i^T\right\}}{\partial\tau} \quad (17)$$

The next second derivative is the second derivative of the likelihood function of the first derivative τ, is:

$$\frac{\partial^2 l}{\partial^2\tau} = \frac{\sum_{i=1}^n \left\{-\frac{1}{\tau^2} + \frac{M_{(y_i)}(w_{ij})\left(e^{x_i^T\beta(u_i,v_i)}\tau+1\right)}{\tau^2} - \frac{y_i}{\tau}\right\}}{\partial\tau} \quad (18)$$

From the results of the above derivative obtained explicit equation then to solve these equations RS algorithm and CG algorithm are employed. Likelihood function is maximized by using Fisher Scoring Algorithm. If the above equation and non-linear implicit equations within the parameters $\boldsymbol{\beta}$ and τ so as to obtain estimates of the parameters $\boldsymbol{\theta} = [\boldsymbol{\beta}^T \quad \tau]$, function is maximized by using Fisher Scoring Algorithm, by the following equation:

$$\hat{\boldsymbol{\theta}}_{(r+1)} = \hat{\boldsymbol{\theta}}_{(r)} + \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{(m)})\mathbf{D}(\hat{\boldsymbol{\theta}}_{(m)}), \quad (19)$$

Where

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\tau}) \quad (20)$$

$$\boldsymbol{D}(\hat{\boldsymbol{\theta}}) = \left(\frac{\partial l}{\partial\hat{\tau}}, \frac{\partial l}{\partial\hat{\boldsymbol{\beta}}^T}\right)^T \quad (21)$$

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_{(m)}) = -E\begin{bmatrix} \frac{\partial^2 l}{\partial\tau^2} & \frac{\partial^2 l}{\partial\tau\partial\beta} \\ \frac{\partial^2 l}{\partial\beta\partial\tau} & \frac{\partial^2 l}{\partial\beta^T\partial\beta} \end{bmatrix} \quad (22)$$

$$\mathbf{H}(\hat{\boldsymbol{\theta}}_{(m)})_{(k+1)(k+1)} = \begin{bmatrix} \frac{\partial^2 l}{\partial\tau^2} & \frac{\partial^2 l}{\partial\tau\partial\beta} \\ \frac{\partial^2 l}{\partial\beta\partial\tau} & \frac{\partial^2 l}{\partial\beta^T\partial\beta} \end{bmatrix}_{\boldsymbol{\theta}=\boldsymbol{\theta}_{(m)}} \quad (23)$$

Hessian matrix is a matrix that contains the second derivative of the likelihood function of the parameter $\boldsymbol{\beta}$ and τ. The steps as follows Fisher Scoring Algorithm:

1) Determining the initial vector parameter $\hat{\boldsymbol{\theta}}_0$ with assuming the data meet the multiple linear regression model.
2) Forming gradient vector $\mathbf{D}(\hat{\boldsymbol{\theta}})$ by substituting equation (13) and (14) into the equation $\mathbf{D}(\hat{\boldsymbol{\theta}})$
3) Hessian matrix forming (23) by substituting equation (16), (17), and (18) into the equation (22)
4) Fisher information matrix formed $\mathbf{I}(\hat{\boldsymbol{\theta}}_{(0)})$
5) inserting values $\hat{\boldsymbol{\theta}}_{(0)}$ thus obtained gradient vector $\mathbf{D}(\hat{\boldsymbol{\theta}}_{(0)})$ and hessian matrix $\mathbf{H}(\hat{\boldsymbol{\theta}}_{(0)})$
6) Starting from $m = 0$ iterating the equation $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{(m)})$, value $\hat{\boldsymbol{\theta}}_{(m)}$ is a set of parameter estimator convergent iteration to- m.

7) If you have not obtained when the parameter estimation convergent iteration to- m, then proceed back to step 6 to iteration to- $m+1$. Iteration will stop when the value of $\| \hat{\boldsymbol{\theta}}_{(m+1)} - \hat{\boldsymbol{\theta}}_{(m)} \| \leq \varepsilon$ and $\varepsilon > 0$ is a very small number [6].

B. Statistical Test

Parameter and hypothesis testing in the model are done simultaneously using MLRT and partially using Z-test. The test statistic used in the simultaneous test likelihood ratio is a statistical measure that was formed by determining the parameters set under the population and under the null hypothesis.
Hypotheses to test the significance of the parameters $\boldsymbol{\beta}$.
H_0: $\beta_1(u_1, v_1) = \beta_2(u_2, v_2) = \cdots = \beta_k(u_k, v_k) = 0$
H_1: at least one $\beta_l(u_l, v_l) \neq 0$ with $l = 1, 2, \ldots, k$
Here is a statistical test used:

$$G = -2\, ln\left(\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right) \tag{24}$$

in which the value of $L(\hat{\omega})$ and $L(\hat{\Omega})$ are the maximum likelihood value for each model in which $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}$ is the result of parameter estimation. Determine the rejection region H_0 if $G_{hit} > \chi^2_{(\alpha,v)}$.

For the partial test parameter β using the hypotheses:
H_0: $\beta_l(u_i, v_i) = 0$
H_1: $\beta_l(u_i, v_i) \neq 0$, with $l = 1, 2, \ldots, k$
Here is a statistical test used

$$Z = \frac{\hat{\beta}_l}{SE(\hat{\beta}_l)} \tag{25}$$

To reject H_0 if $|Z_{hit}| > Z_{\alpha/2}$ where α is the significance level used.

While the partial test parameter τ using the hypotheses:
H_0: $\tau = 0$
H_1: $\tau \neq 0$
Here is a statistical test used:

$$Z = \frac{\hat{\tau}}{SE(\hat{\tau})} \tag{26}$$

To reject H_0 if $|Z_{hit}| > Z_{\alpha/2}$ where α is the significance level used.

IV. CONCLUSION

Parameter estimation of Geographically Weighted Poisson regression model using the Inverse Gaussian Maximum Likelihood Estimation (MLE). In the process of parameter estimation, equation obtained is not close the form, so it requires iteration method employing Fisher Scoring Algorithm. Hypothesis testing is done simultaneously using Maximum Likelihood Ratio Test and partially using Z-test .

REFERENCES

[1] Consul, P.C. and Famoye, "Smoothing Reference Centile Curves : The LMS Method and Penalized Likelihood," *Statistics in Medicine*, Vol. 11, pp. 1305-1319, 1992.

[2] Karlis, D. and Nikoloupoulos, E. "Mixed Poisson Distribution", International Statistical Review, Vol. 73, No.1, pp 35-58, 2005.

[3] Walpole, Ronald E. “Pengantar Statistika”, Gramedia Pustaka Utama. Jakarta, 2005.

[4] De Jong, P. dan Heller, G.Z,),”Generalized Linear Models for Insurance Data”, 1st edition, Cambridge University, Press., New York, 2008.

[5] Shoukri, M.M., Asyali, M.H., vandorp, R. and Kelton, R, “The Poisson Inverse Gaussian Regression Model in the Analysis of Clustered Counts Data”, Journal of Data Science, Vol. 2, No. 2, hal 17-32, 2004.

[6] Ummah, Z., Suliyanto dan Sediono, “Estimasi Model Linier Tergeneralisasi Gaussian Berdasarkan Maksimum Likelihood estimator dengan menggunakan Algoritma Fisher Scoring”, Jurnal Matematika, Vol. 1, No. 1, pp. 110-120, 2013.