IOP Conf. Series: Journal of Physics: Conf. Series **855** (2017) 012064 doi:10.1088/1742-6596/855/1/012064

Parameters Estimation of Geographically Weighted Ordinal Logistic Regression (GWOLR) Model

Shaifudin Zuhdi, Dewi Retno Sari Saputro, Purnami Widyaningsih
Mathematics Department, Mathematics and Natural Sciences Faculty, Universitas Sebelas Maret, Indonesia

E-mail: shaifudinzuhdi@gmail.com, dewi.rss@gmail.com

Abstract. A regression model is the representation of relationship between independent variable and dependent variable. The dependent variable has categories used in the logistic regression model to calculate odds on. The logistic regression model for dependent variable has levels in the logistics regression model is ordinal. GWOLR model is an ordinal logistic regression model influenced the geographical location of the observation site. Parameters estimation in the model needed to determine the value of a population based on sample. The purpose of this research is to parameters estimation of GWOLR model using R software. Parameter estimation uses the data amount of dengue fever patients in Semarang City. Observation units used are 144 villages in Semarang City. The results of research get GWOLR model locally for each village and to know probability of number dengue fever patient categories.

1. Introduction

A regression model is the representation of relationship between independent variable and dependent variable. The dependent variable has categories used in the logistic regression model to calculate odds on.

Logistic regression is extended to handle outcome variables that have more than two ordered categories, there are polytomous logistic regression and ordinal logistic regression. Polytomous logistic regression is used when the categories of the outcome variable are nominal, that is, they do not have any natural order. When the categories of the outcome variable have a natural order, ordinal logistic regression may be appropriate [5].

Statistical methods have been developed to model the relationship between dependent variable and independent variables that depend on the geographic location where the data is observed. For categorical dependent variable have been developed geographically weighted logistic regression (GWLR) model [1]. GWLR model can also be developed for ordinal scale dependent variable, namely geographically weighted ordinal logistic regression (GWOLR) model.

Generally to estimate the parameter of GWOLR model, the likelihood maximum method needs to be employed. This method results in a system of nonlinear equation which is hard to solve. This method includes iteration process which is a repetition process. In the field of programming, iteration is a process consisted of more than one algorithm and is conducted in a loop program which often refers to a repetition program. One of the programming languages which allows the repetition of instruction is the R language (R software).

R software is an integrated software which has the facility to manipulate the data, calculation, and graphical display. R software is supported by a community whose members are actively interacting to each other via internet and guided by a manual or R-help. Parameter estimation on R software can be done by creating an algorithm-based program.

2. Ordinal logistic regression model

The OLR model is a model representing the relationship between ordinal-scaled response variable and independent variable which has the characteristic of category and/or continual [3]. Logistic regression model can be called as logit model. If the response variable which has G category employs ordinal scale and x_i refers to the vector of independent variable in i-observation, thus the OLR model can be formulated as

$$logit[P(Y_a \leq g|x_a)] = \alpha_g + \boldsymbol{x_i}^T\boldsymbol{\beta_i}, \tag{1}$$

with $g = 1,2,\dots,G-1$, $a = 1,2,\dots,n$, and $i = 1,2,\dots,p$. $P(Y_a \leq g|x_i)$ refers to the cumulative probability which is less or equal to g-category on x_i. The parameter α_g refers to the intercept which fulfils the condition of $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_{g-1}$ and β_i refers to the coefficient vector of the regression which is in mutual accord with x_i. Thus, the probability of each G response category is formulated as

$$P_g(x_i) = \frac{exp(\alpha_g + \boldsymbol{x_i}^T\boldsymbol{\beta_i})}{1 + exp(\alpha_g + \boldsymbol{x_i}^T\boldsymbol{\beta_i})} - \frac{exp\left(\alpha_{g-1} + \boldsymbol{x_i}^T\boldsymbol{\beta_i}\right)}{1 + exp\left(\alpha_{g-1} + \boldsymbol{x_i}^T\boldsymbol{\beta_i}\right)}, g = 1,2,\dots,G. \tag{2}$$

The probability value for each response category is used as guidance for classification. An observation will be included to the g-response category based on the biggest probability value.

The parameter estimation of OLR model is conducted with maximum likelihood estimation (MLE) method. This method is employed to examine the maximum solution from the likelihood function, but can only derive nonlinear equation system. This nonlinear equation system is conducted using numerical approach, i.e. Newton Raphson method. The Newton Raphson method is employed to conduct the nonlinear equation system until the maximum parameter value is derived. In the OLR model, for instance take n amount of sample on random variable vector which is $Y_1, Y_2, \dots, Y_n$, with $Y_a = [y_{a1} y_{a2} \dots y_{a,G-1}]^T$ and probability of result from the g-category is $p_g(x_i)$, thus the likelihood function is

$$L = \prod_{a=1}^{n}\prod_{g=1}^{G}\left(\frac{exp(\alpha_g + \boldsymbol{x_i}^T\boldsymbol{\beta_i})}{1 + exp(\alpha_g + \boldsymbol{x_i}^T\boldsymbol{\beta_i})} - \frac{exp\left(\alpha_{g-1} + \boldsymbol{x_i}^T\boldsymbol{\beta_i}\right)}{1 + exp\left(\alpha_{g-1} + \boldsymbol{x_i}^T\boldsymbol{\beta_i}\right)}\right)^{y_{ag}}. \tag{3}$$

The MLE principle is employed to estimate the parameter vector $V = [\alpha_1 \alpha_2 \dots \alpha_{G-1} \beta_1 \beta_2 \dots \beta_p]^T$ by maximize the likelihood function. To simplify the calculation, the likelihood function is transformed into *ln-likelihood* which is formulated as

$$\ell = \sum_{a=1}^{n}\sum_{g=1}^{G} y_{ag}\, ln\left(\frac{exp(\alpha_g + \boldsymbol{x_i}^T\boldsymbol{\beta_i})}{1 + exp(\alpha_g + \boldsymbol{x_i}^T\boldsymbol{\beta_i})} - \frac{exp\left(\alpha_{g-1} + \boldsymbol{x_i}^T\boldsymbol{\beta_i}\right)}{1 + exp\left(\alpha_{g-1} + \boldsymbol{x_i}^T\boldsymbol{\beta_i}\right)}\right). \tag{4}$$

3. Geographically weighted ordinal logistic regression model

The function which is used for GWOLR model is the cumulative logit function. In GWOLR model, each regression parameter depends on geographical location of the data. GWOLR model with response variable in G amount of ordinal-scaled categories is formulated as [7]

$$Logit[P(Y_a \leq g|x_a)] = \boldsymbol{\alpha_g}(u_a, v_a) + \boldsymbol{x_i}^T \boldsymbol{\beta_i}(u_a, v_a) \quad (5)$$

with $P(Y_a \leq g|x_a)$ refers to the cumulative probability of g-category toward x_a; $g = 1,2, \dots, G$ is the amount of response variable category; $\alpha_g(u_a, v_a)$ is the intercept parameter which fulfills the condition of $\alpha_1(u_a, v_a) \leq \alpha_2(u_a, v_a) \leq \cdots \leq \alpha_{G-1}(u_a, v_a)$; (u_a, v_a) is the coordinate point (latitude, longitude) for the location of a, and $\beta_i(u_a, v_a) = [\beta_1(u_a, v_a)\beta_2(u_a, v_a) \dots \beta_p(u_a, v_a)$ is the regression parameter vector for the a-location. The probability of each G response category is

$$P_g(x_a) = \frac{exp(\boldsymbol{\alpha_g}(u_a,v_a)+\boldsymbol{x_i}^T\boldsymbol{\beta_i}(u_a,v_a))}{1+exp(\boldsymbol{\alpha_g}(u_a,v_a)+\boldsymbol{x_i}^T\boldsymbol{\beta_i}(u_a,v_a))} - \frac{exp(\boldsymbol{\alpha_{g-1}}(u_a,v_a)+\boldsymbol{x_i}^T\boldsymbol{\beta_i}(u_a,v_a))}{1+exp(\boldsymbol{\alpha_{g-1}}(u_a,v_a)+\boldsymbol{x_i}^T\boldsymbol{\beta_i}(u_a,v_a))}. \quad (6)$$

The parameter estimation of GWOLR model employs the weighted maximum likelihood method. The geographical factor is the weighting factor in GWOLR model. This factor has different value for each location. The weight is given in the form of weighted ln-likelihood for local GWOLR model. For instance if the weight for each location (u_a, v_a) is $w_{ab}(u_a, v_a)$ then the weighted ln-likelihood function is

$$l = \sum_{a=1}^{n}[y_{a1} ln(p_1(x_a)) + y_{a2} ln(p_2(x_a)) + \dots + y_{aG} ln(p_G(x_a))] w_{ab}(u_a, v_a). \quad (7)$$

Further, it uses kernel function as the weight. Based on [6], the weight based on kernel function which is used is adaptive Gaussian and stated as

$$w_{ab}(u_a, v_a) = exp\left(-\frac{1}{2}\left(\frac{d_{ab}}{b}\right)^2\right) \quad (8)$$

with d_{ab} refers to the distance between the location of (u_a, v_a) and the location of (u_b, v_b), $d_{ab} = \sqrt{(u_a - u_b)^2 + (v_a - v_b)^2}$, b refers to the already known nonnegative parameter and called as bandwidth.

The ln-likelihood function and the weighted ln-likelihood will reach maximum point if the first partial derivation toward the estimated parameter equals to zero and the derivation matrix of the both weighted ln-likelihood functions have the characteristic of negative definite. Hence, the derivation of the first and second weighted ln-likelihood functions toward each estimated parameter can be determined.

4. R software

R software is a language programming that one of the object-oriented programming with syntax like S language. R can also computing on the language [2]. Instruction that was done by R software is based on code described.

Within computing on the language, the user can be create the goal program that appropriate for statistical modelling or graphic. Parameters estimation in R is used one of the loop instruction, that is instruction was done by R with any condition, if the condition has been correct then R would be stopped the process. Three instruction in R software is: for, while, and repeat. The understanding about statistical method is needed before did implementation in R software [6].

5. Data

The data used in this study are secondary data obtained from Badan Pusat Statistik (BPS) and public health office of Semarang city. The observation unit is villages in Semarang city, Central Java, consisting of 144 villages.

In this study, dependent variable (Y) is incidence rate (IR) of dengue fever of village. The incidence rate have three categories: IR $< 20/100.000$ (low), IR $20 - 50/100.000$ (medium), and IR $> 50/100.000$ (high) [4]. Independent variable used are the population density (X_1), the ratio of 0-14 years old group (X_2), the ratio of semi-permanent houses (X_3), the ratio of healthy facilities (X_4), and the larva free index (X_5).

6. Main Results

Based on the variables above, parameters were needed is two intercept parameters and five regression parameters. Parameters estimation has done by R software, create the program consisted loop instruction. For example, parameters estimation of GWOLR model for 10 locations shown in Table 1.

Table 1. The Parameters Estimation Result of 10 Villages

Villages	$\widehat{\alpha_1}$	$\widehat{\alpha_2}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$
Pekunden	53.3676	78.1957	-0.0005	0.0084	0.0030	0.4175	-0.9071
Karang Kidul	-0.8665	8.2843	-0.0001	0.0016	0.0011	0.1998	-0.1132
Jagalan	15.5292	19.6942	-4.8703×10^{-5}	-0.0004	5.0770×10^{-5}	0.0687	-0.2192
Brumbungan	101.7097	123.7707	-0.0005	0.0066	0.0206	-0.1492	-1.4097
Miroto	119.6938	129.1980	-0.0009	0.0057	0.0278	-0.4242	-1.3881
Gabahan	84.1002	94.1645	-0.0002	-0.0017	0.0102	-0.1063	-1.0284
Kranggan	61.6866	68.4917	-7.0066×10^{-5}	-0.0025	0.0033	0.0318	-0.7502
Purwodinatan	52.3171	58.8539	-9.8767×10^{-5}	-0.0015	0.0019	0.0168	-0.6436
Kauman	59.7941	65.9429	-8.6251×10^{-5}	-0.0019	0.0021	0.0381	-0.7235
Bangunharjo	25.4967	29.0148	-7.2769×10^{-5}	-0.0008	0.0004	0.0599	-0.3156

The parameters estimator could have used to construct the GWOLR model and calculate probability of IR in each villages. From the models could have calculated probability of dependent variable of each villages. Other villages is also estimated the parameters using R software. For the example, GWOLR model of Pekunden village is :

$$P_1(x_{Pekunden}) = \frac{\exp(53.3676 - 0.0005\mathbf{x}_1 + 0.0084\mathbf{x}_2 + 0.0030\mathbf{x}_3 + 0.4175\mathbf{x}_4 - 0.9071\mathbf{x}_5)}{1 + \exp(53.3676 - 0.0005\mathbf{x}_1 + 0.0084\mathbf{x}_2 + 0.0030\mathbf{x}_3 + 0.4175\mathbf{x}_4 - 0.9071\mathbf{x}_5)}$$

$$P_2(x_{Pekunden}) = \frac{\exp(78.1957 - 0.0005x_1 + 0.0084x_2 + 0.0030x_3 + 0.4175x_4 + 0.9071x_5)}{1 + \exp(78.1957 - 0.0005x_1 + 0.0084x_2 + 0.0030x_3 + 0.4175x_4 + 0.9071x_5)} - \frac{\exp(53.3676 - 0.0005x_1 + 0.0084x_2 + 0.0030x_3 + 0.4175x_4 + 0.9071x_5)}{1 + \exp(53.3676 - 0.0005x_1 + 0.0084x_2 + 0.0030x_3 + 0.4175x_4 + 0.9071x_5)}$$

$$P_3(x_{Pekunden}) = 1 - \frac{\exp(78.1957 - 0.0005x_1 + 0.0084x_2 + 0.0030x_3 + 0.4175x_4 + 0.9071x_5)}{1 + \exp(78.1957 - 0.0005x_1 + 0.0084x_2 + 0.0030x_3 + 0.4175x_4 + 0.9071x_5)}.$$

P_1 is the probability of low IR, P_2 is the probability of medium IR, and P_3 is the probability of high IR. The probability of response category are affected by independent variable value. β_1 is negative so that if the population density increases then low category and medium category have decreased but high category have increased.

7. Conclusion

In this paper we have discussed parameters estimation of GWOLR model using R software. Estimation in R software could be easier than manually estimate, because R have package/library for analizing statistics data. User would need to create the syntax program if the package/library not available in R software. Run the program can get the parameters estimator quickly. From the

parameters estimator could have construct the model and calculate the probability each categories of dependent variable. GWOLR model is the local model, so one village have one GWOLR model, not the global model. The model parameters obtained by modified ordinal logistic regression model with add coordinate locations of research. From the coordinate locations, get the local model for each village.

References

[1] Atkinson P M, S E German, D A Sear and M J Clark 2003 *Exploring The Relations Between Riverbank Erison and Geomorphological Control Using Geographically Weighted Logistic Regression* (Ohio: Ohio State University vol 35) pp 58-82

[2] Becker R A, Chambers J M and Wilks A R 1988 *The New S Language: A Programming Environment for Data Analysis and Graphics* (Chapman & Hall)

[3] Hosmer D W and S Lemeshow 2000 *Applied Logistic Regression* (USA: John Willey and Sons Inc.)

[4] Kemenkes RI 2010 *Buletin Jendela Epidemiologi: Demam Berdarah Dengue, Pusat Data dan Surveilans Epidemiologi* (Jakarta)

[5] Kleinbaum D G and Klein M 2010 *Logistic Regression – A Self-Learning Text* 3rd Ed. (New York: Springer)

[6] Micheaux L P, Remy D and Benoit L 2013 *The R Software: Fundamentals of Programming and Statistical Analysis* (New York: Springer)

[7] Purhadi, Rifada M and Wulandari P 2012 Geographically Weighted Ordinal Logistic Regression Model *International Journal of Mathematics and Computation* 16