Geographically Weighted Logistic Regression (GWLR) with Adaptive Gaussian Weighting Function in Human Development Index (HDI) in The Province of Central Java

Indah Manfaati Nur 1), M. Al Haris 2)
1,2) Department of Statistics, Universitas Muhammadiyah Semarang, Semarang

Corresponding email: indahmnur@unimus.ac.id

Abstract. The Human Development Index (HDI) is an indicator to measure the success of quality human life. Efforts to calculate HDI to the district/city level are very important, so a better understanding of local conditions is needed with more adequate data support for all districts/cities in Indonesia. Modelling of HDI in Central Java with the factors suspected to influence it is done by Geographically Weighted Logistic Regression (GWLR) with the Adaptive Gaussian Kernel weighting function. The analysis result shows that the deviance value is 32.3992, where the value is greater than the value of χ² (0.1.4) table, which is 7.77943. So, there is at least one independent variable that significantly affects of HDI in Central Java. GWLR Model Parameter Testing Results with Adaptive Gaussian Kernel weighting function obtained factors that influence HDI in Central Java Province is health facilities.

Keywords : HDI, GWLR, Adaptive Gaussian Kernel

1. Introduction

The Human Development Index (HDI) is one of the most important measures to determine the quality of human development that has been successfully achieved. HDI is very suitable to be used as a measurement of development performance, especially human development carried out in an area at a certain time because HDI is considered more able to reflect the results of development that focuses on human development. HDI is a value that indicates the level of community welfare measured from 3 (three) main components, namely: health as measured by Life Expectancy (AHH), education is measured by Literacy Rate (AMH) and Average Length of School (RLS) and the economy is measured by a decent standard of living with the Gross Domestic Product per capita approach at the level of real per capita consumption or the purchasing power of the people [5].

HDI values in Central Java is different in each district/city. This is because the development focus is different in each district/city and other supporting factors are different. According to HDI calculations, HDI figures in 2011 were higher than in previous years. There are several reasons for not comparing the HDI values for consecutive years, including: HDI is an independent variable are state, namely a variable whose changes take place very slowly and will increase or decrease gradually in response to changes in various physical, social conditions, economy and environment.

The HDI data per district/city in Central Java is very diverse because each district/city has different location characteristics. The analysis used to determine the relationship between HDI and its
constituent components that pay attention to location factors is used Geographically Weighted Logistic Regression (GWLR). GWLR is a method for obtaining regression parameters by considering location factors with binomial distribution response variable data.

This research applies the Geographically Weighted Logistic Regression (GWLR) method which is a development of logistic regression that takes into location factors to determine the factors that influence HDI in Central Java.

2. Research Methods

The method used in this research consisted of data collection methods and data analysis method

2.1 Method of collecting data

The data used in this study are secondary data obtained from Badan Pusat Statistik (BPS) in the form of Human Development data. This data includes HDI p-values and HDI components in Central Java which cover 35 districts/cities.

The variables used in this study were 5 variables consisting of 1 response variable and 4 independent variables. Variable responses (Y) is form of Human Development Index (HDI) with category 0 = low HDI (<70) and category 1 = high HDI (≥70). The independent variables used are Literacy Numbers (X1), Number of Health Facilities (X2), Open Unemployment Rate (X3) and Pain Rate (X4).

2.2 Logistic Regression

Binary logistic regression is one method that can be used to analyze the relationship of binary variables responses with one or more free variables that are continuous, categorical or a combination of both [1]. The logistic regression model with k independent variables can be written in the form of an equation as follows:

\[
\pi(x) = \frac{\exp(\beta^T x)}{1 + \exp(\beta^T x)}
\]

If a logit transformation is performed on equation (1), the logistic regression model is obtained as follows:

\[
g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta^T x
\]

With \( x = [1 \ x_1 \ \cdots \ x_k]^T \) represents the vector of the independent variable and \( \beta^T = [\beta_0 \ \beta_1 \ \cdots \ \beta_k] \) is the parameter vector of the coefficient. The logistic regression model parameters can be obtained by the Maximum Likelihood Estimation (MLE) method and Newton Raphson iteration [3].

2.3 Multicollinearity Test

Multicollinearity checks are performed to see if there is a strong correlation between independent variables in the regression model. Multicollinearity testing between explanatory variables is done by looking at the Pearson correlation value and VIF (Variance Inflation Factor). Pearson correlation values are obtained by the following formula:

\[
r_{xy} = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}
\]

while the VIF value is formulated as follows:

\[
VIF = \frac{1}{(1 - R_k^2)}
\]

\( R_k^2 \) is the coefficient of determination for the k-explanatory variable which is considered as a response to the remaining explanatory variables [4].
2.4 Weight Selection
Weighting is used to provide different parameter estimation results at one location to other. One of the weighting that can be used is the Adaptive Gaussian Kernel function which is formulated as follows:

\[ w_j = \exp \left[ -\frac{1}{2} \left( \frac{d_{ij}}{h_i} \right)^2 \right] \] (2.3)

\( d_{ij} \) is the euclidean distance between location \((u_i, v_i)\) and location \((u_j, v_j)\) formulated with

\[ d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \]

and \( h_i \) is a refining parameter (bandwidth). According to Fotheringham, et all. (2002) one of the methods used to determine optimal bandwidth is the Cross Validation (CV) method formulated with

\[ CV(h) = \sum_{i=1}^{n} \left( y_i - \hat{y}_{\neq i}(h) \right)^2 \]

where observations at location \( i \) are removed from the assessment process.

2.5 Geographically Weighted Logistic Regression (GWLR) Method
The Geographically Weighted Logistic Regression (GWLR) method is a regression model developed by Fotheringham, et all. (2002) for variables responses that are binomial distributions that consider location aspects. GWLR model is a local linear regression model (locally linear regression) which results in the parameter model estimator that has a local characteristic at each point or location. In the GWLR model, the parameter estimator \( p \)-values are different at each geographical location, because each parameter \( p \)-value is calculated at each geographical location.

The GWLR model with \( k \) independent variables can be written in the form of an equation as follows:

\[ \pi(x) = \frac{\exp(\beta^T(u_i,v_i)x)}{1 + \exp(\beta^T(u_i,v_i)x)} \] (2.4)

The GWLR model is a nonlinear model so transformation is needed to become a linear function. The transformation used is the logit transformation of \( \pi(x) \) and the following equation is obtained:

\[ g(x) = \ln \left( \frac{\pi(x)}{1-\pi(x)} \right) = \beta^T(u_i, v_i)x \] (2.5)

As with the logistic regression model, GWLR model parameters can be obtained by estimating using the Maximum Likelihood Estimation (MLE) method and Newton Raphson iteration.

3. Research Result and Discussion

3.1 Multicollinearity Test
Multicollinearity testing between explanatory variables is done by looking at the VIF (Variance Inflation Factor) value. VIF values that indicate less than 10 indicate there is no strong correlation between independent variables.

| Independent Variable | VIF   |
|----------------------|-------|
| X_1                  | 1.2392|
| X_2                  | 1.2042|
| X_3                  | 1.4191|
| X_4                  | 1.4312|

Table 1 above shows the VIF \( p \)-values of the four independent variables that are less than 10, so it can be concluded that there is no strong correlation between independent variables.
3.2 Modeling HDI with GWLR

The stages of HDI modeling in Central Java are done by determining the geographical location of districts/cities in Central Java based on latitude (longitude) and longitude (longitude) coordinates and then euclidean distance is calculated between the location \((u_i, v_i)\) and location \((u_j, v_j)\). The results of euclidean distance calculations between locations are presented in table 2.

| Region                        | Cilacap        | Banyumas       |
|-------------------------------|----------------|----------------|
| Cilacap Regency               | 0              | 0.190262976    |
| Banyumas Regency              | 0.190262976    | 0              |
| PurworejoRegency              | 0.574891294    | 0.388973007    |
| BanjarnegaraRegency           | 0.945304184    | 0.755380699    |
| KebumenRegency                | 0.927631392    | 0.744647568    |
| WonosoboRegency               | 0.752728371    | 0.581377674    |
| MagelangRegency               | 0.652763357    | 0.462709412    |
| BoyolaliRegency               | 0.991261822    | 0.802246845    |
| Purbalingga Regency           | 1.351813597    | 1.61550688     |
| Banjarnegara Regency          | 1.792344833    | 1.605147968    |
| Purbalingga Regency           | 1.863035158    | 1.675857989    |
| PurworejoRegency              | 2.392822601    | 2.209660607    |
| KaranganyarRegency            | 1.697557068    | 1.509602597    |
| SragenRegency                 | 1.701440566    | 1.511191583    |
| GroboganRegency               | 1.129291813    | 0.945780101    |
| BloraRegency                  | 2.138831457    | 1.949871791    |
| RembangRegency                | 2.305666932    | 2.121909517    |
| PatiRegency                   | 1.935458602    | 1.758010239    |
| KudusRegency                  | 2.001924074    | 1.820027472    |
| JeparaRegency                 | 1.41509717     | 1.315180596    |
| DemakRegency                  | 1.491207564    | 1.311830782    |
| SemarangRegency               | 1.211981848    | 1.022594739    |
| TemanggungRegency             | 0.89738509     | 0.711688134    |
| KendalRegency                 | 0.98386991     | 0.813019065    |
| BatangRegency                 | 0.738173421    | 0.586941224    |
| PekalonganRegency             | 0.56639209     | 0.451884941    |
| PemalangRegency               | 0.468401537    | 0.444072066    |
| TegalRegency                  | 0.653069675    | 0.586941224    |
| BrebesRegency                 | 0.528015151    | 0.654369926    |
| Magelang City                 | 1.480033783    | 1.290155029    |
| Surakarta City                | 1.982523644    | 1.793376703    |
| Salatiga City                 | 1.734877517    | 1.54468249     |
| Semarang City                 | 1.640121947    | 1.454991409    |
| Pekalongan City               | 1.06           | 0.898109125    |
| Tegal City                    | 0.653069675    | 0.586941224    |

The results of the euclidean distance calculation are then used to determine the optimal bandwidth with the Cross Validation (CV) method. The results of optimal bandwidth calculation are presented in Table 3.
Table 3. Optimal Bandwidth Value

| Weighting          | Bandwidth Optimum | CV Score |
|--------------------|-------------------|----------|
| Adaptive gaussian kernel | 20                | 8.6612   |

Furthermore, the *Adaptive Gaussian kernel* weighting function is used to estimate the parameters of the GWLR model. The summary statistical results of the GWLR model parameter estimator are presented in Table 4.

Table 4. Summary of Parameter Estimator Statistics

| Variable | Min     | Quantil 1 | Media | Quantil 3 | Max     |
|----------|---------|-----------|-------|-----------|---------|
| Intercep | -22.936 | -13.386   | -5.878| -3.865    | 0.607   |
| X₁       | 0.0589  | 0.1065    | 0.1376| 0.2205    | 0.3220  |
| X₂       | -0.00122| -0.00118  | -0.00114| -0.00084 | -0.0007 |
| X₃       | -0.5216 | -0.4663   | -0.3648| -0.1705   | -0.0862 |
| X₄       | -0.8517 | -0.6522   | -0.5476| -0.4730   | -0.3305 |

After the GWLR model parameter estimator is obtained, the next step is to test the hypothesis in the GWLR model which includes simultaneous testing of the GWLR model and partial testing. Simultaneous testing is used to test the significance of the parameters of the GWLR model together. The hypothesis used is:

\[ H₀ : β₁(uᵢ, vᵢ) = β₂(uᵢ, vᵢ) = β₃(uᵢ, vᵢ) = β₄(uᵢ, vᵢ) = 0, \]
with \( i = 1, 2, ..., 35 \)

Table 5. Deviance Value

| Statistics | P-Value |
|------------|---------|
| Deviance   | 32.3992 |

Based on Table 5, the deviance value is 32.3992, where the value is greater than the value of \( \chi² (0.1.4) \) table, which is 7.77943. So it can be concluded that there is at least one independent variable that has a significant effect on HDI in Central Java.

For example, a partial test is performed at the Cilacap District location, the hypothesis used is as follows:

\[ H₀ : β_j(uᵢ, vᵢ) = 0, \text{ with } i = 1, 2, ..., 35 \]
\[ H₁ : β_j(uᵢ, vᵢ) ≠ 0 \]

The results of partial testing of the GWLR model in Cilacap district are shown in Table 6.

Table 6. Testing Results of GWLR Model Parameters

| Parameter | Estimation | Error Standard | Z_hit |
|-----------|------------|----------------|-------|
| \( β₀ \)  | -17.15692338| 20.56129564    | -0.834428125 |
| \( β₁ \)  | 0.251833007 | 0.254319068    | 0.990224635  |
| \( β₂ \)  | -0.001151787| 0.000724619    | -1.589507732 |
| \( β₃ \)  | -0.499029999| 0.369989234    | -1.348750055 |
| \( β₄ \)  | -0.390362988| 0.845597597    | -0.461641553 |

Based on Table 6 above, \( Z_{hit} \)p-value for all parameters is less than \( Z_{table} \) with \( α = 10% \) of 1.64. This shows that there are no factors that influence HDI in Cilacap district. The GWLR model with the *Adaptive Gaussian kernel* weighting in the HDI data in Central Java is:

\[
\hat{r} = \frac{\exp(-17.157 + 0.2518X₁ - 0.0012X₂ - 0.4990X₃ - 0.3904X₄)}{1 + -17.157 + 0.2518X₁ - 0.0012X₂ - 0.4990X₃ - 0.3904X₄}
\]
Then the logit function is as follows

\[ g(x) = -17.157 + 0.2518X_1 - 0.0012X_2 - 0.4990X_3 - 0.3904X_4 \]

A summary of the significant variables in each district/city in Central Java is shown in Table 7.

| Region                  | Significant Variable |
|-------------------------|----------------------|
| CilacapRegency          | -                    |
| BanyumasRegency         | -                    |
| PurbalinggaRegency      | -                    |
| BanjarnegaraRegency     | X_2                  |
| KebumenRegency          | X_2                  |
| PurworejoRegency        | X_2                  |
| WonosoboRegency         | -                    |
| MagelangRegency         | X_2                  |
| BoyolaliRegency         | -                    |
| KlatenRegency           | -                    |
| SukoharjoRegency        | -                    |
| WonogiriRegency         | -                    |
| KaranganyarRegency      | -                    |
| SragenRegency           | -                    |
| GroboganRegency         | -                    |
| BloraRegency            | -                    |
| RembangRegency          | -                    |
| PatiRegency             | -                    |
| KudusRegency            | -                    |
| JeparaRegency           | -                    |
| DemakRegency            | -                    |
| SemarangRegency         | -                    |
| TemanggungRegency       | X_2                  |
| KendalRegency           | -                    |
| BatangRegency           | -                    |
| PekalonganRegency       | -                    |
| PemalangRegency         | -                    |
| TegalRegency            | -                    |
| BrebesRegency           | -                    |
| Magelang City           | -                    |
| Surakarta City          | -                    |
| Salatiga City           | -                    |
| Semarang City           | -                    |
| Pekalongan City         | -                    |
| Tegal City              | -                    |

Based on table 7 above, the factors that influence the value of HDI in Central Java are number of health facilities (X_2)

4. Conclusion

Based on the results of the analysis above, GWLR modeling with the *Adaptive Gaussian kernel* weighting function on HDI in Central Java obtained 35 different models for each district/city. Factors affecting HDI in Central Java are health facilities. Research on HDI modeling in Central Java still be developed by using different weights and using HDI data with a more than two category approach.
5. References
[1] Agresti, A. (2002). *Categorical Data Analysis, Second Edition*. John Wiley & Sons. Inc.
[2] Fotheringham, A.S., Brunsdon, C. & Charlton, M. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley & Sons Ltd.
[3] Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression, Second Edition*. John Wiley & Sons. Inc.
[4] Myers R. H. (1990). *Classical and Modern Regression with Applications Second Edition*. Boston: PWS-KENT.
[5] Nur. & Fatimah, C. (2010). *Pemodelan IPM Provinsi Jawa Timur, Jawa Tengah, Jawa Barat dan Sumatera Utara dengan Metode Regresi Logistik Ordinal*. Thesis. Jurusan Statistika. FMIPA Institut Teknologi Sepuluh Nopember. Surabaya