Robust Geographically Weighted Regression Modeling using Least Absolute Deviation and M-Estimator

Puteri Pekerti Wulandari, Anik Djuraidah*, Aji Hamim Wigena
Department of Statistics, Bogor Agricultural University, Bogor, West Java, Indonesia

ABSTRACT

Geographically weighted regression (GWR) is a development of multiple regression that has spatial varying, so that the estimator of GWR is different for each location. Parameter estimation in GWR uses weighted least square method which is vulnerable to outliers and can cause biased parameter estimation. The robust GWR (RGWR) with LAD and M-estimator is resistance to outliers. This research estimated parameters on RGWR using LAD and M-estimator method and uses data of Java gross domestic product (GRDP) in 2015 containing several outliers. The result showed that RGWR model was better than GWR with M-estimator, and the predictions were closer to the actual values.

Keywords: Geographically Weighted Regression, Least Absolute Deviation, M-Estimator

I. PENDAHULUAN

Geographically weighted regression (GWR) is a model that can be used for data with spatial varying (Fotheringham et al. 2002). The estimator of GWR parameters using weighted least square method is known susceptible to outliers (Zhang dan Wei 2011). In implementation of GWR, outliers are often found. Outliers can be overcome using robust (RGWR) method which is resistance to outliers, such as least absolute deviation (LAD) method and M-estimator.

LAD method introduced by Roger Joseph Boscovich in 1757 used WLS to obtain parameter estimator by minimizing absolute number of residual (Mutan 2009). Application of simplex method uses linier programming as optimum solution and it is considered to be efficient for computing LAD that cannot be solved analytically (Chen 2002).

M-estimator method was introduced by Huber 1973. This method minimizes objective function of residual. Parameter estimation uses weighted least square method iteratively (Chen 2002).

Implementation of LAD method in GWR was studied by Afifah (2015) in the case of Java poverty in 2015, while implementation of M-estimator method was studied by Sari et al. (2014) on mapping potential of agriculture of East Java in 2012, and Azizah (2015) compared to least square (LS), median absolute deviation (MAD) and M-estimator methods. Based on the results of this studies, LAD method which minimized the residuals influence of outliers was more robust than GWR and M-estimator.

Gross regional domestic product (GRDP) is one important indicator to determine economic conditions in an area and certain period (BI 2015). GWR and geographically weighted panel regression (GWPR) to estimate GRDP parameters was studied by Fatulloh
The steps of data analysis

I. METHOD AND MATERIAL

A. Data

Data in this study were Java GRDP in 118 districts/cities in 2015 obtained from website Central Bureau of Statistics (www.bps.go.id). The variables (Table 1) refer to the research by Handayani (2017).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>GRDP districts/cities of Java at Constants Price 2010 in 2015</td>
<td>Thousand Rupiah</td>
</tr>
<tr>
<td>$X_1$</td>
<td>Number of Labor</td>
<td>People</td>
</tr>
<tr>
<td>$X_2$</td>
<td>Regional Income</td>
<td>Million Rupiah</td>
</tr>
<tr>
<td>$X_3$</td>
<td>Wage</td>
<td>Rupiah</td>
</tr>
<tr>
<td>$X_4$</td>
<td>Human Development Index</td>
<td>Percent</td>
</tr>
</tbody>
</table>

B. Procedure Data Analysis

The steps of data analysis are as follows:

1. Data description and exploration.
2. Spatial effects test:
   a. Heterogeneity spatial test
      i. Hypothesis
      $H_0 : \alpha_2^2 = \cdots = \alpha_k^2 = 0$
      (no heterogeneity between location)
      $H_1 : \text{minimal ada satu } \alpha_k^2 \neq 0 \text{ (there is a heterogeneity between location)}$
   ii. Test Statistics
      $BP = \frac{1}{2} f^T Z (Z^T Z)^{-1} Z^T f \sim \chi^2_{(k-1)}$
      with : $f = (f_1, \ldots, f_n)^T$
          $f_i = (\frac{\alpha_i^2}{e_i} - 1)$
          $e_i = y_i - \hat{y}_i$
   iii. Critical region, if $BP > \chi^2_{(k-1)}$ than reject $H_0$
   iv. Conclusion
   b. Spatial autocorrelation
      i. Hypothesis
      $H_0 : I = 0 \text{ (no autocorrelation between location)}$
      $H_1 : I \neq 0 \text{ (there is a autocorrelation between location)}$
      ii. Test Statistics
      $I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (X_i - \bar{X})(X_j - \bar{X})} \sum_{i=1}^n (X_i - \bar{X})^2$
      $S^2_{(I)} = \frac{n^2 \sum_{i,j} W_{ij}^2 + 3(\sum_{i,j} W_{ij}^2)^2 - n \sum_{i,j} (\sum_{k=1}^n W_{ij})^2}{(n^2 - 1)(\sum_{i,j} W_{ij})^2}$
      iii. Critical region, if $Z > Z_{\alpha/2}$ than reject $H_0$
      iv. Conclusion

3. Outlier detection using boxplot
4. Analysis GWR model
   a. Determine longitude-latitude coordinates observation area
   b. Calculates euclidean distance ($d_{ij}$) between observation areas
      $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$
   c. Determining bandwidth using ACV and CV optimum criteria
\[ ACV(h) = \sum_{i=1}^{n} |y_i - \hat{y}_{i|h}(h)| \]
\[ CV(h) = \sum_{i=1}^{n} (y_i - \hat{y}_{i|h}(h))^2 \]

d. Calculates weighting matrix \((w_{ij})\) with kernel function

e. Determine \(\hat{\beta}^{(0)}\) value as first estimate selected spatial regression parameters.

5. Analysis GWR model using LAD method

a. Using step 4 as first step to determine ACV value

b. Determine regression coefficient by minimizing residual using optimum bandwidth

\[ \min \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \beta_j(u_0, v_0)x_{ij} w_j(d_{0i}) - \hat{y}_i| \]

c. Estimating robust GWR LAD parameters

\[ y_i = \sum_{j=1}^{p} \beta_j(u_0, v_0)x_{ij} + \epsilon_i + \epsilon_i^+ - \epsilon_i^- \]

6. Analysis GWR model using M-estimator method

a. Using step 4 as first step to determine CV value, estimate \(\hat{\beta}^{(0)}\) and get \(\epsilon_i^{(0)}\)

b. Calculate robust to get value of influence function

c. Determine objective function and calculate weighting value

\[ w_i^{(0)}(u_i) = \frac{\psi(u_i)^{(0)}}{u_i^{(0)}} \]

d. Determine \(\hat{\beta}\) value with WLS method

\[ \hat{\beta}^m = (X^TW^mX)^{-1}X^TW^my \]

e. Set residual step (d) as residual step (b)

f. Iterating IRLS on new weighting until \(\hat{\beta}^m\) convergent

7. Determine the best model based on the MAPE value

\[ MAPE = \frac{\sum_{i=1}^{n} |e_i| \times 100%}{n} ; |e_i| = y_i - \hat{y} \]

III. RESULT AND DISCUSSION

A. Data Description

Table 2 shows the highest and lowest GRDP percentage in Java. The highest percentage was in DKI Jakarta and the lowest in West Java. Rate of economic growth between districts/cities in Java shown a varying level.

<table>
<thead>
<tr>
<th>Percentage of Province</th>
<th>Kabupaten/Kota</th>
</tr>
</thead>
<tbody>
<tr>
<td>The highest</td>
<td>Jakarta 6.8%, South Jakarta 6.3%, North Jakarta 5.2%, East Jakarta and West Jakarta 4.7%</td>
</tr>
<tr>
<td></td>
<td>East Java Surabaya 6.2%</td>
</tr>
<tr>
<td></td>
<td>West Java Bekasi district 3.9%</td>
</tr>
<tr>
<td>The lowest</td>
<td>West Java Pangandaran district 0.01%, Banjar 0.05%, Indramayu district 0.1%</td>
</tr>
<tr>
<td></td>
<td>East Java Pasuruan 0.9%</td>
</tr>
<tr>
<td></td>
<td>Central Blitar and Java Mojokerto 0.07%, Magelang 0.1%</td>
</tr>
</tbody>
</table>

Multicollinearity test uses variance inflation factor (VIF) on each explanatory variables. If VIF value > 5 then there is an indication of multicollinearity. Table 3 shows that VIF value obtained from all explanatory variables are less than 5, so it can be concluded that there is no multicollinearity.
TABLE 3
VIF VALUE EXPLANATORY VARIABLES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>1.56</td>
</tr>
<tr>
<td>X₂</td>
<td>1.34</td>
</tr>
<tr>
<td>X₃</td>
<td>1.81</td>
</tr>
<tr>
<td>X₄</td>
<td>1.26</td>
</tr>
</tbody>
</table>

B. Spatial Effect Test
Spatial effect to test heterocedasticity uses Breuch-Pagan (BP) test. Breuch-Pagan value is 31.32 with a p-value of 0.00 that is less than 5%, so it can be concluded that there was variance spatial heterocedasticity at 5% significant level.

Spatial effect to test spatial autocorrelation using Moran Index test. The Moran Index value is 0.05 with a p-value of 0.79 which was more than 5%, so it can be concluded spatial autocorrelation at 5 significant level.

C. Outliers Detection
Figure 1 shows standardized residual of GWR model. In the boxplot, there was indicated the outliers. The outliers means that outliers in data cannot be handled by GWR.

Figure 1 : Residual Boxplot GWR

D. Comparison GWR LAD and M-estimator
Bar chart in Figure 2 shows the estimation value of GWR, GWR LAD, and GWR M-estimator models for each districts/cities in Java province. M-estimator model had estimate value that was almost same as GWR model. Overall robust model was closer to actual value than GWR.
Figure 2: Comparison of estimation value of Java districts/cities

Figure 3 shows distribution parameter estimation of GWR, LAD and M estimator methods. Measure of dispersion used was range. Changes coefficient estimated using robust method were seen from median, range, and outliers. LAD model had smaller range and higher median position.
Mean absolute percentage error (MAPE) value also used to compared. Criteria to measure the accuracy of model is smallest MAPE. Smaller value is best model. Table 4 shows the comparison estimation of robust methods. Robust methods had the MAPE value was smaller than GWR, so this model was better applied in outliers data. LAD model had smallest MAPE value and better applied for Java GRDP data in 2015. Coefficient determination of M-estimator model has highest value, it means M-estimator model was be able to explain Java GDP variation Java in 2015 better than LAD. However the M-estimator method in this study had not been as good as LAD method.

**TABLE 4**

<table>
<thead>
<tr>
<th></th>
<th>MAPE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWR</td>
<td>47.4%</td>
<td>84.53%</td>
</tr>
<tr>
<td>RGWR LAD</td>
<td>31.08%</td>
<td>87.21%</td>
</tr>
<tr>
<td>RGWR M</td>
<td>37.61%</td>
<td>88.55%</td>
</tr>
</tbody>
</table>

Figure 4, Figure 5, and Figure 6 shows group of factors having significant effect in each location on GWR, LAD and M-estimator methods. Figure 4 shows that group 1 GWR including all factors. Group 2 had $X_1$, $X_3$, dan $X_4$. Group 3 had $X_1$, $X_3$ and $X_4$. Group 4 had $X_1$ and $X_3$. Group 5 had $X_3$. Group 6 had $X_2$. Group 7 had $X_3$ and $X_4$. Group 8 had $X_1$ dan $X_4$. However, in group 9 there were no significant factor.

Figure 5 shows that group 1 LAD had overall factors effecting significantly. Group 2 had $X_1$, $X_3$, and $X_4$. Group 3 had $X_1$, $X_3$, and $X_4$. Group 4 had $X_1$. Group 5 had $X_3$, $X_4$. Group 6 had $X_2$. Group 7 had $X_3$. Group 8 had overall factor effecting significantly.
IV. CONCLUSION

RGWR LAD and M-estimator methods can be used to handle outliers and result estimators closer to actual value. The best estimator of GRDP data was LAD with smallest MAPE value. Factors in explanatory variables, number of labor (X_1), regional income (X_2), regional minimum wages (X_3), and human development index (X_4) had significant influence on each district/cities. Robust methods is an alternative when there is outlier in data, so there is no need to throw out outlier data.

V. REFERENCES


Cite this article as:


Journal URL : http://ijsrset.com/IJSRSET196123