Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 4 (2015), pp. 2255-2264

http://www.ripublication.com

Spatial Autoregressive Quantile Regression Modelling for Gross Domestic Regional Product Data (Case: 113 Districts/Cities in Java in 2010)

Azzikra Febriyanti[1], Anik Djuraidah[2], Aji Hamim Wigena[3]

Departemen of Statistics

Bogor Agricultural University

Jalan Pajajaran, Kampus IPB Baranangsiang, Bogor 1615, Indonesia

fazzikra.af@gmail.com, 2.anikdjuraidah@gmail.com, 3.ajihamim@yahoo.com

Abstract

Spatial Model Autoregressive (SAR) is one of spatial modeling which indicates that the response variable has a spatial dependence. SAR modeling has one weakness that the heterogeneity of variance is unknown. Spatial Autoregressive Quantile Regression (SARQR) Model is a combination of SAR models and Quantile Regression (QR). This is an alternative model to solve the problems of heterogeneity in SAR models. In addition, not only solve an issue of spatial heterogeneity, SARQR modeling can also be a solution to handling problems in non-normality data which caused by outliers. The purpose of this study is to establish a modeling Gross Domestic Regional Product (GDRP) in 113 districts/ cities in Java in 2010 by using a model SARQR to obtain some homogeneous models for each particular quantile. The GDRP case in Java can be concluded that the heterogeneity that occurs in SAR modeling can be overcome by establishing a SARQ model. This result is shown by significant value in Breusch Pagan test. SARQR Modeling which produces several models separately for each particular quantile interpretation models required for some districts/ cities which have a value of GDRP is far from the average value of the overall GDRP in Java.

Keywords: Heteroscedasticity, Spatial Autoregressive, Quantile Regression, Spatial Autoregressive Quantile Regression.

1 Introduction

Regression analysis is a method used to modeling the relationship between the response variable (y) with the explanatory variables (x). This method has several assumptions that must be completed, one of them is independent between observations. If an observation has spatial effects, the circumstance of an observation at a particular location is influenced by observations that are around, the method used spatial regression analysis.

A spatial effect is divided into two the spatial dependence and spatial heterogeneity. Spatial dependency occurs when one observation in one location depends on other observations. The spatial heterogeneity occurs when there are variations in the relationship between observations. There are several regression models that involve spatial dependency in modeling, 1) Spatial Autoregressive Model (SAR) with spatial dependence on the response variable, 2) Spatial Error Model (SEM) with spatial dependence on the error and 3) General Spatial Model (GSM) with spatial dependence on the response or error. When the data has a spatial heterogeneity, Geographically Weighted Regression model (GWR) is being used.

Spatial Autoregressive Regression quantile (SARQR) is one of alternative to solve the problems of spatial heterogeneity in addition to modeling the GWR. SARQR model is a model that combines modeling SAR with Quantile Regression (QR). Based on first research by Koenker and Basset (1978), quantile regression is a method that aims to minimize the weighted absolute error which is not symmetrical to suspect conditional quantile function on a distribution of the data, so as to eliminate the heterogeneity that occurs in the data. Another advantage possessed quantile regression modeling is to produce a model that robust of data outliers. The combination of these two models produce a fairly good model for addressing problems and heteroskedasticity dependence on spatial data modeling, and also stout against the data outliers. This method has been applied by Kostov (2009) for modeling the price of agricultural land, Liao and Wang (2010) and Ziet et.al (2010) on the modeling of house prices.

The value of Gross Regional Domestic Product (GDRP) in 113 districts / cities on the island has a very diverse in distributing data, and it is dependent of the result of the spatial effects that occur between the district and the city. Fatulloh (2013) has done modeling GWR to the value of GDRP in Java to solve the problems on the spatial heterogeneity. Study will be used to address issues SARQR modeling spatial effects on the data value of GDRP in Java.

2 Materials and Methods

2.1 *Spatial Weight Matrix*

Spatial weighting matrix is a matrix that shows the relationship between regions. The weighting matrix of size $n \times n$, where n is the number of observations. Weighting matrix form is as follows:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & w_{ij} & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

The matrixes used in the present study were weighted based on the nearest neighbor matrix, which is defined as follows:

$$w_{ij} = \begin{cases} 1, \text{for } i \text{ and } j \\ 0, \text{other} \end{cases}$$

Lee and Wong (2001) stated that the provision of value in the weighting can be done in the form of lines that have been standardized. This process is based on the number of neighbors in the same row on the matrix weighting. Standardized formula of the line is as follows:

$$w_{ij}^{*} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$$

with the value of w_{ij}^{*} is a matrix element that has been normalized, then the number of each row is equal to 1.

2.2 *Quantile Regression*

Quantile regression model is a model that provides an overview relationship between a set of independent variables and certain percentile (quantile) of the response variable. This method is a regression method with the approach of separating or dividing the data into a particular quantiles that are likely to have different estimate value. Roger Koenker (1978) analyzes the case of quantile regression to see the problem of regression can be used to solve problems in the average example:

$$min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2 \quad (1)$$

If $\mu = X'\beta$ then the equation (1) becomes:

$$min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2, \quad (2)$$

with:

y_i = variable respond to-i

β = parameter of predictors varible

x_i = variable predictor to-i

Furthermore, the model evolved into the median sample stated:

$$min_{\beta \in \mathbb{R}^P} \sum_{i=1}^{n} \left| y_i - x_i'\beta \right|. \quad (3)$$

Models are generally specified in the conditional quantile function to-τ by considering estimators for $\beta(\tau)$, with the notation $\beta(\tau)$, so it can be stated:

$$min_{\beta \in \mathbb{R}^P} \sum_{i=1}^{n} \rho_\tau(y_i - Q_\tau(Y|X)), \quad (4)$$

with:

τ = indects for each quantile $0 < \tau < 1$,

$\rho_\tau(u) = (\tau - I(u < 0))$, $0 < \tau < 1$ and I(.) is loss function.

$Q_\tau(Y|X) = X'\beta(\tau)$ = Quantile function of $-\tau$ of Y with X terms.

(Koenker 2005)

The first thing that should be a concern in the process of quantile regression analysis is conditional quantile function. If Y is a continuous random variable and x is one of the explanatory variables vector X, then the conditional quantile function to-τ can be defined:

$$Q_\tau(Y|X) = inf\{y: |F_y(y|X) \geq \tau\}, \quad (5)$$

with $F_y(y|X)$ is the distribution function of Y with X terms and the conditional density function th $f_y(y|X)$.

Estimation of the value of β in each specific quantile is obtained by minimizing the unconditional quantile regression to estimate the median quantile defined:

$$\beta_{(\tau)} = \arg\min_{(\tau)} E[\rho_\tau(Y - X'\beta)], \quad (6)$$

so as to estimate β as equation (9) general model of quantile regression equation can be formed as follows:

$$Y = X\beta_\tau + \varepsilon, \quad (7)$$

2.3 *Spatial Autoregressive Quantile Regression (SARQR)*

SAR model equations form (Fotheringham & Rogerson 2009) can be written as follows:

$$y_i = \lambda \sum_{j=1}^{n} w_{ij}^* y_j + x_i \boldsymbol{\beta} + \varepsilon_i, \quad (8)$$

with λ is the coefficient of spatial autoregressive, w_{ij}^* is a spatial weighting matrix that has been standardized in all areas *i* and *j* neighbors, ε_i is random error which stochastic identical.

If the SAR models written in matrix form, it can be formulated as follows:

$$Y = \lambda W^* Y + X\beta + \varepsilon \; ; \varepsilon \sim N(0, \sigma^2 I) \quad (9)$$

with W^* a spatial weighting matrix with size $(n \; x \; n)$, Y is a response variable sized vector$(n \; x \; 1)$, X is a matrix of explanatory variables measuring$(n \; x \; k)$, β is a vector of parameters to be suspected of size$(k \; x \; 1)$, and ε is the vector error of size$(n \; x \; 1)$.

SAR reduced form into the following equation:

$$Y = A^{-1}X\beta + \varepsilon^* \quad (10)$$

with $A = I - \lambda W^*$, A^{-1} is invers matrix from A and $\varepsilon^* = A^{-1}\varepsilon$. A^{-1} can be expressed as $\begin{bmatrix} a_i \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$ with a_i the vector lines in the area to-*i* with size (*1 x n*).

The combination of SAR modeling with quantile regression that was called by SARQR, the specifics can be defined as follows:

$$Y = \lambda_\tau W^* Y + X\beta_\tau + \varepsilon. \quad (11)$$

In other side, the case in equation (9) which is a classical modeling SAR, SARQR models have spatial lag parameter (λ) and the parameter vector regression (β), which relies on quantile value for a particular sequence.

Some previous researchers have done estimating model parameters SARQR by several methods. Generalized Method of Moment of Method (GMM) by Kelejian and purcha (1998) and Lee and Lin (2010), Two Stage quantile method Regression (2SQR) by Liao and Wang, and modeling SARQR with Instrumental Variable Quantile Regression (IVQR).

IVQR method first introduced by Chenozhukov and Hansen (2004) and adapted by Su and Yang (2007) to model SARQR. This method is based on the understanding of the methods Instrumental Variable (IV) estimation method which involves new exogenous variables which are in a regression equation and acts as a variable that is not correlated with the error but correlated with endogenous variables. SARQR modeling using instrumental variable in estimate the parameters, which is variable $\boldsymbol{Z}$. The variables have size ($n \times n$) and consist of a group explanatory $\boldsymbol{WX}$ spatial dependence.

SARQR parameter estimation model has similarities stage with Two Stage Least Square (2SLS), which will set up a parameter alleged to eventually substitute with real parameters. The second difference is the method of parameter estimation method, by 2SLS using the least squares method, whereas at this stage will be used quantile regression.

The following will be carried estimation parameters $\lambda_{0\tau}$ and $\boldsymbol{\beta}_{0\tau}$ the equation SARQR:

$$Y = \lambda_{0\tau} W^* Y + X\beta_{0\tau} + \varepsilon, \quad (12)$$

with the following step:

1) Forming a model for W^*Y with the explanatory variables are *X* dan *Z*. Then substitute the value $\widehat{W^*Y}$ in initialization model, so it can be estimated paramter λ with OLS method.

2) At a certain λ value, quantile regression modeling will be performed *Y* and

W^*Y with explanatory varible X and Z, namely:

$$y_i - \lambda\overline{y_i} = x_i^{'}\beta + z_i^{'}\gamma \quad (13)$$

with: $\overline{y_i} = \sum_{j=1}^{n} w_{ij}^{*} y_j$,

γ = parameters of instruments variable. .

The model will be used to estimate the parameters of $\beta_\tau(\lambda)$ and $\gamma_\tau(\lambda)$:

$$\left(\hat{\beta}_\tau(\lambda), \hat{\gamma}_\tau(\lambda)\right) \equiv \arg\min_{(\beta,\gamma)} \left\{\frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(y_i - \lambda\overline{y_j} - x_i^{'}\beta - z_i^{'}\gamma\right)\right\} \quad (14)$$

3) Minimizing the standard vector estimation of $\hat{\gamma}_\tau(\lambda)$ for λ to calculate the value of IVQR of $\lambda_{0\tau}$.

4) Establish a quantile function:

$$y_i - \lambda_{0\tau}\overline{y_j} = x_i^{'}\beta_\tau, \quad (15)$$

which used to estimate the value of the parameter IVQR explanatory variables ($\boldsymbol{\beta}_{0\tau}$).

This process will be repeated for each quantile (τ), so we get different parameter estimators for each quantile.

3 Results and Discussion

This research used value of GDRP 113 districts / cities in Java in 2010 with eighteen explanatory variables from BPS 2010. Selection of the best models of classical linear regression using stepwise method produces modeling the value of GDRP with seven factors, namely the percentage of poor population (X1), Education (X4), Life Expectancy (X6), Potential sea village (X11), Potential mining village (X12), the ratio of head of the family to use electricity (X14), and number of stores and permanent market (X17).

The fisrt step is modeling by least square method. This method produces a model Y =-75.229 – $0.432X_1$ + $25.616X_4$ + $0.899X_6$ + $29.2792X_{11}$ + $217.063X_{12}$ + $1.410X_{14}$ + $6.7931X_{17}$ with R^2 values of 0.579. The values indicate that the predictor variables explain the changes in the value of GDRP (Y) amounted 57.9%. The next test will be performed classical assumption of the normality assumptions and assumptions homoscedasticity to determine whether the results of these estimates contain heteroskedasticity or not.

The assumption of normality test shows the error did not follow a normal distribution with a P-value of 0.01 which is a smaller than the α of 0.05. In homoscedasticity assumptions test, Breusch Pagan test has value of 0.00473 which is showing that the value less than of α. The occurrence of various violations from the basic assumptions of the regression results in this model shows that the model is not valid to use.

Another thing that must be considered is whether there is a spatial dependence between the observation areas by using moran test. It shows that there is spatial dependence between observation points with p-values for 2.08×10^{-8}, where the value

is smaller than α.

This study will use the SAR modeling due to the spatial Otoregresif LM test shows significant results but not for the LM test on spatial error. SAR models result in the formation of the model as follows:

$Y = -84.117 + 0.484WY - 0.269X_1 + 27.601X_4 + 0.9308X_6 + 33.563X_{11} + 2.299X_{12}$

$+ 0.5286X_{14} + 6.3708X_{17}$

SAR modeling works well to form a new model which no longer contain the effects of spatial dependence, as shown by the results of the LM test conducted after the formation of the model have p-values for 0.571, greater than the α. But modeling is still not able to say a good model because still have a violation of assumptions which is homoscedasticity assumptions. In addition, there are some outliers that the data cannot be defined by the model.

Alternative models that will be used to overcome the problems of heterogeneity and the presence of data outliers are SARQR models. This modeling will form a different model for each quantile is different so as to include the data residing in the tail of the distribution. Here are the results of modeling the value of GDRP using a model SARQR:

Table 1. Estimate value from SARQR model (p-value)

	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
(Intercept)	-6.994	-9.045	-45.269	-47.851	-6.078
	(0.656)	(0.664)	(0.116)	(0.022)	(0.831)
X_1	-0.044	-0.039	-0.218	-0.231	-0.396
	(0.697)	(0.800)	(0.380)	(0.239)	(0.144)
X_4	9.085	7.658	17.953	25.826	28.497
	(0.093)	(0.272)	(0.052)	(0.000)	(0.014)
X_6	0.025	0.054	0.483	0.454	-0.048
	(0.905)	(0.845)	(0.189)	(0.121)	(0.904)
X_{11}	3.865	7.735	27.316	42.484	15.491
	(0.695)	(0.574)	(0.054)	(0.075)	(0.674)
X_{12}	0.563	30.933	-1.648	-0.852	14.322
	(0.556)	(0.814)	(0.361)	(0.863)	(0.000)
X_{14}	-0.517	0.057	0.802	1.447	2.026

	(0.534)	(0.947)	(0.851)	(0.182)	(0.138)
X_{17}	3.387	3.847	5.862	5.933	1.797
	(0.000)	(0.001)	(0.000)	(0.000)	(0.304)
WY	0.050	0.050	0.200	0.350	0.900
	0.608	0.700	0.309	0.102	0.004
BP Test	0.1289	0.5889	0.1411	0.1023	0.1083

SARQR model analysis value of GDRP in Java in 113 districts/ cities shows that there are different factors that influence in each quantile and not at all quantile group is influenced by the value of GDRP spatial effects. Based on Table 1, it appears that a significant effect of spatial interaction occurs in the last quantile. It specifies that in areas with high GDRP values (greater than 15.476 T) are influenced by the area surrounding it. The value of the coefficient λ states that if an area surrounded by other regions of *n*, then the influence of each area surrounding can be measured by an average of 0.9 multiplied by the surrounding region.

Each quantile has different factors influencing changes in the value of GDRP, but did not happen on the value of the life expectancy factor (X_6). Predictor factor has no significant value in each quantile group, so it can be concluded that in the event of changes in the value of GDRP for each quantile group there are no effects of changes in life expectancy. It also occurs in ratio factor family head using electricity (X_{14}), the percentage of poor (X_1) and Potential sea village (X11) do not change the value of GDRP if there is a change of the ratio of household heads that use electricity and changes in the percentage of poor people in each region for each quantile group.

The ratio of population/ education facilities (X_4) factor is a significant factor in the latter two groups of quantile which is quantile 0.75 and quantile 0.9. The exhibit significantly value can be stated that the increase in the ratio of population/ education facilities is one of the important factors that influence on the increase in the value of GDRP in areas with above average GDRP value, but it does not happen in areas with medium value of GDRP.

At the last quantile we can see that the factor of potential mining village (X_{12}) became one of the factors affecting changes in the value of GDRP. This explains the growing ptotential of the mining village will result in the increase in the value of GDRP in areas that are at the quantile groups, and it does not happen in other quantile groups. In addition, this model has Breusch Pagan value greater than α. It shows that the model can solve the heterogeneity problem with divides the data into five quantile.

4 Conclusion and Remarks

The GDRP case in Java can be concluded that the heterogeneity that occurs in SAR modeling can be overcome by establishing a model SARQR. SARQR Modeling

which produces several models separately for each particular quantile. It also interprets models for some districts/ cities which have value that far from the average of the overall GDRP in Java.

Acknowledgment

I would like to thank you to Statistics Indonesia Agency (BPS) of Indonesia for providing GDRP data.

References

[1] Anselin,L (1988), "Spatial Econometrics: Methods and Models," Dordrecht, Kluwer Academic Press

[2] Anselin,L. Kelejian,H.H (1997), "Testing for Spatial Error Autocorrelation in the Presence of Endogenous Regressors," International Regional Science Review, 20 (1-2): 422-448.

[3] [BPS]. Badan Pusat Statistik (2011), "Perkembangan Beberapa Indikator Utama Sosial-Ekonomi Indonesia," Jakarta (ID), BPS.

[4] [BPS]. Badan Pusat Statistik (2012). www.bps.go.id. [Mei 2012]

[5] Chernozhukov, V. Hansen C (2006), "Instrumental Quantile Regression Inference For Structural And Treatment Effect Models," Journal of Econometrics 127, 491-525.

[6] Draper,N.R., Smith,H (1992), " Analisis Regresi Trapan. Sumantri B," penerjemah. Jakarta (ID): Gramedia Pustaka Utama. Translated from: "Applied Regression Analysis".

[7] Fatulloh (2013),"Penerapan Regresi Terboboti Geografis Untuk Data Produk Domestik Regional Bruto dengan Studi Kasus: 113 Kabupaten/kota di Pulau Jawa Tahun 2010 [Undergraduate Thesis]". Bogor, IPB

[8] Kim,T.H., Muller. C (2004), "Two-stage Quantile Regression When the First Stage is Based on Quantile Regression", Econometric Journal 7, 218-231.

[9] Koenker,R (2005), "Quantile Regression," Cambridge, Cambridge University Press.

[10] Kostov,P (2009), "A spatial quantile regression hedonic model of agriculture land prices," Spatial Economic Analysis 4, 53-72.

[11] Liao,W.C., Wang.X (2010), "Hedonic House Prices and Spatial Quantile Regression," IRES Working Paper, Institute of Real Estate Studies, National University of Singapore.

[12] Lin,X., Lee,L.F. (2010), "GMM estimation of spatial autoregressive models with unknown heteroscedasticity," Journal of Econometrics 157, 34-52.

[13] Su,L., Yang, Z (2011), "Instrumental Variable Quantile Estimation of Spatial Autoregressive Models," EABER, 05-2007.

[14] Zietz,J., Zietz,E.N, Sirmans.G.S (2008), "Determinants of House Prices: A Quantile Regression Approach," Journal of Real Estate Finance and Economics 37, 317-333.