Regional Lagging Analysis in Indonesia Using Binary Logistic Regression

Titi Purwandari and Yuyun Hidayat
Department of Statistics, Faculty of Mathematics and Natural Sciences,
Universitas Padjadjaran, Indonesia
titipurwandari@yahoo.com; yuyunhidayat@unpad.ac.id

Sukono
Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Universitas Padjadjaran, Indonesia
sukono@unpad.ac.id; napitupuluherlina@gmail.com

Subiyanto
Department of Marine Science, Faculty of Fishery and Marine Science,
Universitas Padjadjaran, Indonesia.
subiyanto@unpad.ac.id

Abdul Talib Bon
Department of Production and Operations,
University Tun Hussein Onn Malaysia, Malaysia
talibon@gmail.com

Abstract

National development is a process of changing from a particular national situation to a better national condition. Progress in regional development and people's welfare in Indonesia is not always the same and evenly distributed, this has resulted in disparities between regions. This condition is caused by differences in geographical conditions, natural resources, infrastructure, social culture, and human resource capacity. Based on this, a regional development program is needed that is focused on accelerating development in areas where social, cultural, economic, regional finance, accessibility, and infrastructure availability are lagging behind other regions. This study aims to determine the effect of a number of observation variables on the determination of underdeveloped regions and regions not lagging behind in Indonesia. The usefulness of this research is to provide recommendations to relevant agencies in making policies. This study uses secondary data collected by the Central Statistics Agency and the Ministry of Finance of the Republic of Indonesia. The method used is Binary Logistic Regression. Based on the results of the analysis it can be concluded that the variables of the percentage of poor people, per capita consumption, life expectancy, average length of school, percentage of household users of electricity, average distance from the village office to the supervising district office, the percentage of villages with critical land influence to the classification of underdeveloped areas and not left behind.

Keywords
Disadvantaged areas, Binary logistic regression.

1. Introduction

National development is a process of changing from a particular national situation to a better national condition (SKBI, 2015). The progress of regional development and people's welfare in Indonesia is not always the same and evenly distributed, this results in a gap between regions. The condition is caused by differences in geographical conditions, natural resources, infrastructure, socio-cultural, and the capacity of human resources. Based on the foregoing, regional development programs are needed that are focused on accelerating development in areas where social, cultural, economic, regional finance, accessibility, and infrastructure availability are lagging behind other regions (Syafria *et al.,* 2014; Naibaho and elijoi, 2016).

The number of disadvantaged areas in Indonesia in 2014 was 183 districts, while in 2015 it was reduced to 122 districts. Determination of underdeveloped or not lagging regions is based on 27 variables that have been determined by the State Ministry of Development of Disadvantaged Regions and Transmigration (KNPDT). The development of underdeveloped areas was initially focused on eastern Indonesia, but after being evaluated by the government it turned out that underdeveloped areas were also found in parts of other islands in Indonesia such as Java and Sumatra. The government pays attention to disadvantaged areas in Indonesia, so that people in the area have a quality of life not far behind the community in general in other regions. This study aims to determine the effect of a number of observation variables on the determination of underdeveloped regions and regions that are not left behind in Indonesia and the usefulness of this research is to provide recommendations to relevant agencies in making policies.

2. Support Theory

2.1 Underdeveloped regions

Underdeveloped areas are regions whose regions and communities are less developed compared to other regions on a national scale (KNPDT, 2014). Based on Government Regulation Number 78 of 2014 concerning the Acceleration of Development of Disadvantaged Areas, an area designated as a lagging area is based on 6 criteria, namely the community economy, human resources, facilities and infrastructure (infrastructure), regional financial capacity, accessibility, regional characteristics. The government stipulates 122 regions in Indonesia to be in the category of underdeveloped regions, this is regulated in the Presidential Regulation Nonor 131 of 2015 concerning Determination of Disadvantaged Regions in 2015-2019. The government establishes underdeveloped areas once every five years nationally (SKBI, 2015)

3. Literature Review

3.1 Logistic Regression Model

The logistic regression model is used to describe the relationship between the (dependent) response variable with one or several predictor variables (independent) (Bursac et al. 2008; Sarlija *et. al*., 2017). The dependent variable (Y) in logistic regression is generally in the form of dichotomous, with the value of variable $Y=1$ stating an observed event (eg success) and $Y=0$ declaring another event (eg failed). The logistic regression model is a logit transformation of $\pi\left(x\right)$ that is

$$\pi(x) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)} \tag{1}$$

With α : constant, β : regression coefficient, p : many independent variables.

Logit transforms applied to the logistic regression model are:

$$\text{logit}\,[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{2}$$

Odds are the ratio between the chance of a particular event (success) and the chance of another event (failed), formulated as follows:

$$\theta = \frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})} \tag{3}$$

To calculate the association of variables X and Y, the ratio of two odds, called the odds ratio, was formulated as follows (Hosmer, 2000):

$$\psi = \frac{\theta_1}{\theta_2} \tag{4}$$

The logistic regression model does not assume a linear relationship between the independent variable and the dependent variable, does not assume the variable is normally distributed, does not assume homocastricity (Agresti And Allan, 2002; Johnson and Wichern, 2007; Antonogeorgos et al., 2009).

3.2 Parameter Estimation of Logistic Regression Model

To estimate parameters in the logistic regression model the maximum likelihood method is used through iteration (Osibanjo, 2015; Ahmed, 2017). Each observation for a logistic regression model is a Bernoulli distribution variable, the likelihood function of Bernoulli's distribution for *n* independent samples is (Hosmer, 2000):

$$l(\beta) = \prod \pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1 - y_i} \tag{5}$$

Log-likelihood or natural logarithms of joint probability functions are:

$$L(L(\beta)) = \ln l(\beta) = \prod \pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1 - y_i} \tag{6}$$

The estimated parameter β_k is obtained by differentiating the log-likelihood function against β_k with $i = 0,1,\ k = 1,2,...,p$.

3.3 Significance Test of the Parameters of the Logistic Regression Model

Before testing the significance of the parameters individually, the overall significance of the parameters is tested first. Testing is overall referred to as the model significance test using the Likelihood Ratio Test, with the hypothesis:

$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ which states the regression model does not mean, against the alternative hypothesis $H_1 :$ at least two coefficient values β are not the same, meaning that the regression model is significant. The test statistics used are:

$$-2\log\left[\frac{l_0}{l_1}\right] = -2\left[\log[(l_0) - \log(l_1)\right] = -2(L_0 - L_1) \tag{7}$$

$l_0 :$ the maximum value of the likelihood function for the model under H_0.

$l_1 :$ the maximum value of the likelihood function for the model under the alternative hypothesis.

$L_0 :$ the maximum log-likelihood function value for the model under H_0.

$L_1 :$ the maximum log-likelihood function value for the model under the alternative hypothesis.

If $-2(L_0 - L_1) \geq \chi_p^2$ then H_0 is rejected, meaning the model is significant.

The Wald Test is used to significance test of each regression coefficient with $H_0 : \beta_k = 0$ which means that the independent variable to *k* is not significant. Test statistics are:

$$W_k = \left\{\frac{\beta_k}{SE(\beta_k)}\right\}, k = 1,2,...p \tag{8}$$

The hypothesis H_0 is rejected if $W_k \geq \chi^2_{(\alpha,1)}$ means the independent variable to *k* is significant

4. Research Methods

4.1 Research and Variables Objects

The object of observation in this study are districts and cities in Indonesia as many as 491 districts and cities. The variables in this study amounted to 27 variables used by the Ministry of Development of Disadvantaged Regions and Transmigration (KNPDT). The data used in this study are secondary data obtained from the Central Statistics Agency (BPS) and the Ministry of Finance of the Republic of Indonesia, in the form of Village

Potential Data Collection (PODES) 2014, 2014 National Socio-Economic Survey (SUSENAS) and 2014 Regional Financial Capability data for districts and cities in Indonesia.

4.2 Steps to Analyze Binary Logistic Regression

The step of using logistic regression analysis is to form a binary logistic regression model, significance test of the parameters in the binary logistic regression model, compatibility test of the binary logistic regression model, calculation the odds ratio, interpretation of the logistic regression model.

5. Results And Discussion

Based on the results of an analysis of 491 districts and cities in Indonesia and 27 research variables, the following results were obtained:

1. The results of the feasibility testing of logistic regression models can be seen in Table 1, it can be concluded that the binary regression model is feasible to be used for further analysis, because the sig value $=1,000 > 0,05$, meaning that there is no significant difference between the predicted classifications and observed classifications .

Tabel 1. Feasibility Test Results Binary Logistic Regression model

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig
1	.601	8	1.000

2. The results of the significance test of the binary logistic regression model can be seen in Table 2 and Table 3, the value of -2log likelihood in Table 2 is 107,871, while the value of -2log likelihood in Table 3 obtained through the previous iteration is 529,738. This decrease shows that the logistic regression model used is good.

Tabel 2. -2log likelihood value

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Negelkerke R Square
1	107.871[a]	.576	.873

Tabel 3. -2log likelihood value

Iteration History[a,b,c]		
Iteration	-2 Log likelihood	Coefficients
		Constant
Step 0 1	531.198	-1.079
2	529.740	-1.203
3	529.738	-1.208
4	529.738	-1.208

3. From the 27 variables analyzed, 7 independent variables were found that significantly affected the classification of districts and cities as underdeveloped and not underdeveloped regions, namely variable X_1: Percentage of poor people, X_2: consumption per capita, X_3: life expectancy, X_4: long average school, X_5: percentage of household users of electricity, X_6: average distance from village offices to supervising district offices, X_7: percentage of villages with critical land. This can be seen in Table 4 with a sig value < 0.05.

Tabel 4. Variables That Significantly Influence On Regional Classification

	B	S.E	Wald	df	Sig.
Step 1* x1	.212	.048	19.868	1	.000
x2	-.105	.024	19.666	1	.000
x3	-1.119	.176	40.367	1	.000

x4	-1.432	.258	16.033	1	.000
x10	-.136	.0.35	15.397	1	.000
x18	.045	.009	25.908	1	.000
x26	.033	.014	5.251	1	.022
Constant	156.854	24.436	41.202	1	.000

Binary Logistics Regression Model is:

$$\text{logit}\,[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

$$= 156.854 + 0.212\,X1 - 0.105\,X2 - 1.119\,X3 - 1.432\,X4 - 0.136\,X10 + 0.045\,X18 + 0.033\,X26$$

4. The odds ratio is used to interpret the logistic regression coefficient, this can be seen in Table 5 as follows:

Tabel 5 Odds Ratio value for each variable

Variable	Logistic Regression Coefficient	Odds Ratio
X1	.212	1.236
X2	-.105	.900
X3	-1.119	.327
X4	-1.432	.239
X10	-.136	.873
X18	+.045	1.046
X26	+.033	1.034
Constant	156.854	1.320E68

The variable Odds Ratio X_1 value: the percentage of poor people is 1,236, this shows that the percentage of poor people increases by one percent, then the opportunity for an area including underdeveloped areas increases 1,236 times compared to regions that are not left behind, Variable Odds Ratio X_2: consumption per capita is 0,900, this shows that the increase in consumption per capita, the chance of an area including the lagging region is reduced by 0.900 times compared to the area not lagged, the value of the variable X_3 Odds Ratio: life expectancy is 0.327, this indicates that the increase in life expectancy, then the chance of an area including underdeveloped areas is reduced by 0.327 times compared to areas not left behind, X_4 variable Odds Ratio value: the average length of school is 0.239, this shows that every increase in the average length of school is one year, then the probability of a region including underdeveloped regions is 0.239 times dib and if the area is not left behind, the variable X_{10} Odds Ratio value: the percentage of household users of electricity is 0.873, this indicates that each increase. the percentage of household users of electricity, then the opportunity of an area including underdeveloped area is reduced by 0.873 times compared to non-lagging regions, X_{18} variable Odds Ratio value: the average distance from the village office to the supervising district office is 1,046, indicating that each the increase in the average distance from the village office to the district office which is equal to one unit, then the opportunity for an area including underdeveloped area increases by 1,046 times compared to the area that is not left behind, X_{26} Odds Ratio value: the percentage of villages with critical land is 1,034, this shows that each increase in the percentage of villages with critical land is one percent, then the chance of an area including underdeveloped areas increases by 1,034 times compared to regions that are not left behind.

6. Conclusion

Based on the results of the analysis using the steps outlined, it can be concluded that:

1. Variables that significantly influence the classification of underdeveloped or not underdeveloped regions, namely the percentage variable of the poor (X_1), the consumption variable per capita (X_2), these two variables are included in the criteria of the community economy; life expectancy variable (X_3), average length of school variable (X_4), both of these variables include criteria for human resources; variable percentage of electricity user households (X_{10}), these variables include infrastructure criteria; variable average distance from the village office to the supervising district office (X_{18}), these variables include accessibility criteria; the percentage variable of the village has a critical land (X_{26}), this variable includes the characteristics of the region, with the Binary Logistic Regression Model:

$$\operatorname{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right]$$
$$= 156.854 + 0.212\,X_1 - 0.105\,X2 - 1.119\,X3 - 1.432\,X4 - 0.136\,X10 + 0.045\,X18 + 0.033\,X26$$

2. Interpretation through the Odds Ratio coefficient generated from the binary logistic regression model, is able to explain the opportunity that a region is classified as a lagging area or a region not lagging behind.

References

Sekretariat Kabinet Republik Indonesia (SKBI), 8 Desember 2015, *122 Daerah Ini Ditetapkan Pemerintah Sebagai Daerah Tertinggal*,http://Setkab.go.id .

Fadhilah S., Buono A., and Silalahi B. P., A Comparison of Backpropagation and LVQ: A Case Study of Lung Sound Recognition. *Proceedings - ICACSIS 2014: 2014 International Conference on Advanced Computer Science and Information Systems*, 402–7, 2014.

Naibaho and Elijoi, *Perbandingan Backpropagation Neural Network dan Learning Vector Quantization*, Tesis,Universitas Padjadjaran,Bandung. 2016

Kementerian Negara Pembangunan Daerah Tertinggal (KNPDT), 2014, *http://kemendesa.go.id/hal/300027/183-kab-daerah-tertinggal (20/08/2015)*

Bursac, Zoran, Clinton H. G., David K. W., and David W. H., Purposeful Selection of Variables in Logistic Regression, *Source Code for Biology and Medicine* 3: 1886–91, 2008 .

Sarlija N., Bilandžić A and Stanić M., Logistic Regression Modelling: Procedures and Pitfalls in Developing and Interpreting Prediction Models, *Croatian Operational Research Review* 8 (2): 631–52,2017.

Hosmer, D.W and S, Lemeshow, Applied Logistic Regression, Second Edition, John Wiley and Sons,Inc, New York, 2000.

Agresti, Allan, Categorical Data Analysis, 2nd ed,New Jersey, John Wiley and Sons, USA. , 2002.

Johnson, Wichern. *Applied Multivariate Statistical Analysis,* Sixth Edition. Prentice Hall International, Inc., Upper Saddle River, New Jersey, 2007.

George A., Panagiotakos D. B., Priftis K. N., and Tzonou A. , Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence among 10- to 12-Years-Old Children: Divergence and Similarity of the Two Statistical Methods, *International Journal of Pediatrics* 2009: 1–6, 2009.

Osibanjo F. S., Olalude G. A., Akintunde M. O. and Ajala A. G, Application Of Logistic Regression Model To Admission Decision Of Foundation Programme At University Of Lagos. 1 Osibanjo F. S., 2 Olalude G. A. 2 Akintunde M. O.and 2 Ajala A. G." 3 (4): 27–41, 2015.

Ahmed L. A.,Using Logistic Regression in Determining the Effective Variables in Traffic Accidents, 11 (42): 2047–58, 2017.

Biographies

Titi Purwandari is a lecturer at the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, the field of Statistics, with a field of Quality Control Statistics.
Yuyun Hidayat is a lecturer at the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. He currently serves as Deputy of the Head of Quality Assurance Unit, a field of Statistics, with a field of Quality Control Statistics.

Sukono is a lecturer in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. Currently serves as Head of Master's Program in Mathematics, the field of applied mathematics, with a field of concentration of financial mathematics and actuarial sciences.

Subiyanto is a lecturer in the Department of Marine Science, Faculty of Fishery and Marine Science, Universitas Padjadjaran. He received his Ph.D in School of Ocean Engineering from Universiti Malaysia Terengganu (UMT), Malaysia in 2017. His research focuses on applied mathematics, numerical analysis and computational science.

Abdul Talib Bon is a professor of Production and Operations Management in the Faculty of Technology Management and Business at the Universiti Tun Hussein Onn Malaysia since 1999. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France in the year 2008. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business

Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA in the year 1998. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom in 1997. He had published more 150 International Proceedings and International Journals and 8 books. He is a member of MSORSM, IIF, IEOM, IIE, INFORMS, TAM and MIM.