Jurnal Ekonomi dan Statistik Indonesia
2022, Vol. 2, No. 1, 19 – 23
http://dx.doi.org/10.11594/jesi.02.01.03
E ISSN : 2777-0028

Research Article

Comparison Of Normal-Based and Beta-Based Regression Models on Ratio/ Proportion Data

(Case Study: Gini Ratio Modeling In 34 Provinces in Indonesia in 2021)

Pardomuan Robinson Sihombing*

BPS-Statistics Indonesia

Article history:
Submission March 2022
Revised March 2022
Accepted April 2022

**Corresponding author:*
E-mail:
robinson@bps.go.id

ABSTRACT

This study compares the regression using the assumption of a normal distribution with a beta distribution on ratio/proportion data. The data used is the Gini ratio data as the dependent variable and the percentage of the poor, economic growth and unemployment as independent variables in 2021. The data used is sourced from the Central Statistics Agency. The criteria for selecting the best model are based on the smallest AIC and BIC criteria. The results obtained by the beta regression model are better than the model based on the normal distribution. This result is reflected by the probability value of the model suitability test and the error value which the smaller AIC and BIC reflect. The poverty variable has a significant effect on the Gini ratio. On the other hand, there is not enough evidence that the variables of economic growth and open unemployment affect the Gini ratio. From the results obtained, it is hoped that the government will be able to implement appropriate policies in overcoming inequality so that every level of society can feel welfare without exception.

Keywords: *Beta, distribution, gini, normal, regression*

Introduction

Classical linear regression models generally assume that the data used are normally distributed(Gujarati, 2004). In addition to assuming the normality of the data distribution, classical linear regression also has non-heteroscedastic assumptions on data variance and non-autocorrelation on inter-time errors. Classical linear regression models are sometimes not fully applicable in various fields. For example, in some cases where the data is in the form of categorical data, count data or data with a certain value interval where the data follows an exponential family distribution. In the case of count data, categories and intervals, the classical assumptions are often not met. General linear modelling can be used for modelling with an exponential family distribution (Agresti, 2002). One of the modellings in GLM is beta regression. Beta regression is used if the data

How to cite:
Sihombing, P R. (2022). Comparison of Normal-Based and Beta-Based Regression Models on Ratio/ Proportion Data (Case Study: Gini Ratio Modeling In 34 Provinces in Indonesia in 2021). *Jurnal Ekonomi dan Statistik Indonesia*. *2* (1), 19 – 23. doi: 10.11594/jesi.02.01.03

used is the ratio or proportion data where the value is in the interval 0 to 1. Beta regression uses a beta distribution approach, where this distribution is very flexible in various uncertainty phenomena (Johnson & Kotz, 1995). Modelling with beta regression will provide an accurate and efficient parameter estimator compared to the ordinary least squares method when the observed response variables are not symmetrical in the distribution or a heteroscedasticity problem (Swearingen, 2010).

One problem still experienced by some developing countries is the problem of inequality in income distribution (Farrah & Yuliadi, 2020). The inequality of income distribution is measured by the Gini ratio, where the Gini ratio value ranges from 0 to 1. The higher the Gini ratio value, the greater the inequality in the region. Many socioeconomic factors affect inequality, including economic growth, poverty and unemployment, human development index, etc. (Farrah & Yuliadi, 2020) and (Hindu. et al., 2019).

Based on the problems above, the authors are interested in comparing the regression method based on the normal distribution and beta distribution in a case study of factors affecting Indonesia's Gini ratio. The criteria for selecting the best model are based on the smallest AIC and BIC values.

Data and Methodology

This study uses data published data from the Central Statistics Agency (BPS). The research time reference is 2021. The research variables used can be seen in Table 1.

Table 1. Research variable

Variable	unit	Data scale
Gini	points	ratio
poverty	percent	ratio
economic growth	percent	ratio
unemployment	percent	ratio

Beta Regression Model

Model beta regression is used if the data follows the beta distribution (the value is 0 to 1). The beta distribution function can be written as follows:(Walpole, 2012):

$$f(y;a,b) = \frac{\Gamma(a+b)}{\Gamma(b)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$

with 0 < y < 1;a > 0, b > 0, and (.) is a gamma function. The equations in beta regression are:

$$g(\mu) = \text{logit}(\mu) = \ln\left[\frac{\mu}{1-\mu}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

With $\mu = \frac{e^{(x^T\beta)}}{1+e^{(x^T\beta)}}$

Criteria for selecting the best model

In this study, the model selection was based on the AIC. criteria(Akaike, 1974) and BIC(Schwarz, 1978). The formulas used are:

$$AIC = -2\,L(\hat{\theta}) + 2p$$
$$BIC = -2\,L(\hat{\theta}) + p\ln(n)$$

1

where is the likelihood value, and p is the number of parameters to be estimated, including constants.The best model is the model that has the smallest AIC and BIC values $L(\hat{\theta})$ (Widarjono, 2007).

2

It is, furthermore, testing the goodness of the model. The testing of the model's goodness can be seen in Table 2(Gujarati, 2004). After all the tests of the goodness of the model have been met, the interpretation of the formed regression equation is carried out.

Table 2. Model goodness test

The goodness of Fit Test	Null Hypothesis	Alternative Hypothesis	Reject Ho
Coefficient of Determination Test/ R square			R square, the bigger the s, the better
Simultaneous Test (Test F or X2)	Incorrect Model/ All variables have no effect	The model fits / at least 1 variable has a significant effect	Prob. Value < 0.05
Partial Test/ T Uji Test	The i-th independent variable has no effect	The i-th independent variable has an effect	Prob. Value < 0.05

Result and Discussion

Before further discussing the modelling in regression analysis, Table 3 presents descriptive statistics for each research variable. On average, the Gini ratio in 34 provinces in Indonesia in 2021 is 0.35, with the highest value of 0.436 in Yogyakarta Province and the lowest 0.247 in Bangka Belitung Province. The higher the Gini ratio, the higher the income ratio in the province. On average, the percentage of poverty in 34 provinces in Indonesia in 2021 is 10.43 percent, with the highest value of 27.38 percent in Papua Province and the lowest 4.56 percent in South Kalimantan Province. On average, the percentage of economic growth in 34 provinces in Indonesia in 2021 is 4.23 percent, with the highest value of 16.4 percent in North Maluku and the lowest of -2.47 percent in Bali Province. The higher the economic growth, the higher the value of economic output in the region. On average, the percentage of open unemployment in 34 provinces in Indonesia in 2021 is 5.49 percent, with the highest value of 9.91 percent in the Riau Islands and the lowest of 3.01 percent in West Nusa Tenggara Province. The higher the value of open unemployment will be the burden of development for the region.

Table 3. Descriptive analysis of research variables

Variable	mean	std,	Min	Max
like this	0.35	0.04	0.247	0.436
poverty	10.43	5.41	4.56	27.38
economic growth	4.23	3.57	-2.47	16.4
unemployment	5.49	1.82	3.01	9.91

Furthermore, in the regression modelling, it is hoped that there will be no high correlation between the independent variables. This result is indicated by the Variant Inflation Factor (VIF) value and the tolerance value (1/VIF). A good model does not contain a high correlation value between the independent variables with a reference value of VIF 10 and a tolerance of 0.1. In Table 4, all the independent variable VIF values are less than ten, and the tolerance value is > 0.1 so that the model used does not experience multicollinearity.

Table 4. Independent variable VIF value

Variable	VIF	1/VIF
poverty	1.22	0.820622
economic growth	1.06	0.947343
unemployment	1.24	0.807553

In **Error! Reference source not found.**, it can be seen in the comparison between the regression models using a regression model based on the OLS normal distribution using a beta distribution based. When viewed from the number of significant independent variables, it can be seen that both methods produce the same method. Only the poverty variable has a significant effect on the Gini ratio. In the beta model, the Gini ratio variable is significant for 1 percent, while it is significant for 5 percent in the OLS model.

The coefficient of determination (R2) value of both normal and beta models produce almost the same value. The coefficient of determination is 0.168, meaning that the variation in the Gini ratio can be explained by the variables of poverty, economic growth and unemployment of 16.8 percent, the rest by other variables outside the model.

Table 5. Comparison of normal and beta regression models

Model	Normal	Beta
constant	0.2932***	-.86944***
poverty	0.0033*	0.0147**
economic growth	-0.0019	-0.0084
unemployment	0.0046	0.0206
F /X2	2.3682	8.0072*
R2	0.168	0.168
AIC	-117.9282	-116,1065
BIC	-111.8228	-108,4747

Legend: * p<0.05; ** p<0.01

From the model fit test results, in the OLS model, the statistical F value is 2.3682, which is smaller than the F table with df (3.30) = 2.9222 and the probability value = 0.0905 is greater than alpha 0.05. This result means that it does not reject Ho, and it is concluded that the model does not fit. For the beta model, the statistical X2 value is 8.0072, greater than the X2 table with df (3) = 7.8147 and the probability value = 0.0459 is smaller than alpha 0.05. This result means that Ho rejects and concludes that the model is fit.

If we look at the AIC and BIC values, it can be seen that the AIC and BIC values of the beta model are smaller than the normal model. This result indicates that the beta model is better at modelling the Gini than the normal model.

Discussion

The poverty variable has a significant positive effect on the Gini ratio; this means that an increase in poverty will increase the Gini ratio in the area. This result is in line with the research of Hindun et al. (2019)and Hindus. et al. (2015) states that the higher the poverty, the higher the income inequality, or vice versa.

The variable of economic growth has not had a significant effect on the Gini ratio; this means that the increase in economic growth has not been able to reduce the Gini ratio in the area. This result is in line with the research of Aprisa & Miyasto (2013) stated that there is not enough evidence that economic growth affects inequality. This result means that economic growth has not been felt equally by every level of society.

The open unemployment rate variable has not significantly affected the Gini ratio; this means that the increase in open unemployment has not affected the Gini Ratio level in the area. This result is in line with the research of Hindun et al. (2019), and Farrah & Yuliadi (2020)stated that there is not enough evidence that unemployment affects inequality.

Conclusions and Suggestions

From the discussion above, it can be concluded that the beta regression model is better than the model with a normal distribution in modelling the dependent variable in the form of proportions or ratios. This result is reflected by the probability value of the model suitability test and the error value reflected by the smaller AIC and BIC. The poverty variable has a significant effect on the Gini ratio. On the other hand, there is not enough evidence that the variables of economic growth and open unemployment affect the Gini ratio. From the results obtained, it is hoped that the government will be able to implement appropriate policies in overcoming inequality so that every level of society can feel welfare without exception.

For further research, it is possible to add other variables that have the potential to affect the Gini ratio, such as Education, Health and other variables. On the other hand, further research can use other regression model applications, such as panel data regression models.

References

Agresti, A. (2002). *Categorical Data Analysis. New York*. Inc. John Wiley and Sons.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Apriesa, L. F., & Miyasto. (2013). Pengaruh Desentralisasi Fiskal terhadap Pertumbuhan Ekonomi Daerah dan Ketimpangan Pendapatan (Studi Kasus: Kabupaten/Kota di Jawa Tengah). *Diponegoro Journal of Economics*, *2*(1), 1–12. https://ejournal3.undip.ac.id/index.php/jme/article/view/1916/1914

C J Swearingen, M. S. M. C. (2010). C J Swearingen, M S M Castro. *Macro. SAS Global Forum 2011: Statistics and Data Analysis*, 335–2011.

Farrah, N., & Yuliadi, I. (2020). Determinan Ketimpangan Distribusi Pendapatan di Indonesia. *Proceedings The 1st UMYGrace 2020*, 129–140.

Gideon Schwarz. (1978). Estimating The Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Gujarati, D. (2004). *Basic Econometrics BY Gujarati* (pp. 1–1002). McGraw-Hill Inc.

Hassan, S. A., Zaman, K., & Gul, S. (2015). The Relationship between Growth-Inequality-Poverty Triangle and Environmental Degradation: Unveiling the Reality. *Arab Economic and Business Journal*, *10*(1), 57–71. https://doi.org/10.1016/j.aebj.2014.05.007

Hindun., Soejoto., A., & Hariyati. (2019). *Pengaruh Pendidikan , Pengangguran , dan Kemiskinan terhadap Ketimpangan Pendapatan di Indonesia: Universitas, Pascasarjana Surabaya, Negeri Soejoto, Ady Universitas, Pascasarjana Surabaya, Negeri Universitas, Pascasarjana Surabaya, Negeri*. *8*(3), 250–265.

Johnson, N., Kotz, S., & Balakrishnan, N. (1995). *Continuous Univariate Distributions*. Wiley.

Walpole, R. E. (2012). *Probability & Statistics for Engineers & Scientists*. Pearson.

Widarjono, A. (2007). *Ekonometrika: Teori dan Aplikasi untuk Ekonomi dan Bisnis*. Ekonosia Fakultas Ekonomi Universitas Islam Indonesia.

Agresti, A. (2002). *Categorical Data Analysis. New York*. Inc. John Wiley and Sons.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Apriesa, L. F., & Miyasto. (2013). Pengaruh Desentralisasi Fiskal terhadap Pertumbuhan Ekonomi Daerah dan Ketimpangan Pendapatan (Studi Kasus: Kabupaten/Kota di Jawa Tengah). *Diponegoro Journal of Economics*, *2*(1), 1–12. https://ejournal3.undip.ac.id/index.php/jme/article/view/1916/1914

C J Swearingen, M. S. M. C. (2010). C J Swearingen, M S M Castro. *Macro. SAS Global Forum 2011: Statistics and Data Analysis*, 335–2011.

Farrah, N., & Yuliadi, I. (2020). Determinan Ketimpangan Distribusi Pendapatan di Indonesia. *Proceedings The 1st UMYGrace 2020*, 129–140.

Gideon Schwarz. (1978). Estimating The Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Gujarati, D. (2004). *Basic Econometrics BY Gujarati* (pp. 1–1002). McGraw-Hill Inc.

Hassan, S. A., Zaman, K., & Gul, S. (2015). The Relationship between Growth-Inequality-Poverty Triangle and Environmental Degradation: Unveiling the Reality. *Arab Economic and Business Journal*, *10*(1), 57–71. https://doi.org/10.1016/j.aebj.2014.05.007

Hindun., Soejoto., A., & Hariyati. (2019). *Pengaruh Pendidikan , Pengangguran , dan Kemiskinan terhadap Ketimpangan Pendapatan di Indonesia: Universitas, Pascasarjana Surabaya, Negeri Soejoto, Ady Universitas, Pascasarjana Surabaya, Negeri Universitas, Pascasarjana Surabaya, Negeri*. *8*(3), 250–265.

Johnson, N., Kotz, S., & Balakrishnan, N. (1995). *Continuous Univariate Distributions*. Wiley.

Walpole, R. E. (2012). *Probability & Statistics for Engineers & Scientists*. Pearson.

Widarjono, A. (2007). *Ekonometrika: Teori dan Aplikasi untuk Ekonomi dan Bisnis*. Ekonosia Fakultas Ekonomi Universitas Islam Indonesia