EKSAKTA
Journal of Sciences and Data Analysis
E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol4.iss1.art5

Research Article

Implementation of Panel Data Regression in the Analysis of Factors Affecting Poverty Levels in Bengkulu Province in 2017-2020

Aprilia Dewi Anggraeni Chairunnisa[1], Achmad Fauzan[1,*]

[1] Statistics Department, Faculty of Mathematics and Natural Science Universitas Islam Indonesia, Jl.Kaliurang KM 14.5, Sleman-Yogyakarta, Indonesia

**Corresponding author: achmadfauzan@uii.ac.id*

Received: 1 September 2022; Accepted: 20 January 2023; Published: 27 January 2023

Abstract: Economic resilience is certainly an important target in every country or region. One of the main concerns in the economy of a country is poverty. This study aims to explore data with panel data regression that was formed and find factors that affect poverty in Bengkulu province from 2017 to 2020. The secondary data utilized were obtained from the Central Bureau of Statistics (BPS) of the province of Bengkulu. The independent variables used are Gross Regional Domestic Product (GRDP), Human Development Index (HDI), Life Expectancy (LE), and Average Years of Schooling (AYS), while the dependent variable is the percentage of poverty in the form of per region. The best panel data model obtained is the Fixed Effect Model (FEM) model with a cross-section. Based on the results obtained, the significant variable in this model is the GRDP variable. From the prediction results, the values obtained from Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) respectively are 6.59% for MAPE, 5.48 for MSE, and 2.4 for RMSE indicating that panel data analysis is very very good in terms of predicting poverty in Bengkulu province.

Keywords: Fixed Effect Model, Poverty, Panel Data Regression.

Introduction

From the poverty recorded from 2017 to 2020 according to the Regional Planning and Development Agency (BPPD) of Bengkulu Province, it was noted that poverty cases had decreased. However, based on data from the Indonesian Central Statistics Agency (BPS), Bengkulu Province is still in a row of 10 regions that have the highest poverty rate. Based on these conditions, the study seeks to identify the variables that affect both low and high levels of poverty,which can later be used as observation material for various government agencies to fix and overcome poverty in Bengkulu Province. Furthermore, it is expected to be able to reduce the poverty rate that appears.

One of the methods to analyze related to poverty that contains variations in time and location is panel data regression. Regression analysis that combines cross-sectional data with time series data is called panel data regression [1]. Typically, panel data regression involves making observations on constantly observed data across a number of periods. In other words, panel data is information derived from cross-section data collected throughout time on the same specific unit or item [2]. The advantage of panel data regression analysis is that it considers the diversity that occurs in cross-sectional units and is more informative than simple time series. In addition to panel data regression, there are other factor analysis methods, such as multivariable linear regression. However, multivariable linear regression does not involve elements of cross-section and time series in the analysis.

Panel data regression combines time series data and cross-section data, thus providing several advantages, such as: (1) panel data regression provides more informative data, more variety, less collinearity between variables, more degrees of freedom, and better efficiency, (2) panel data regression is more suitable for studying the dynamics of change because it studies repeated cross-sections (3) panel

data regression can better detect and measure effects that cannot only be observed using time series or cross-section data alone, (4) panel data regression allows us to study more complex behavioral models.

Various studies related to panel data analysis include panel data regression analysis on poverty cases in Indonesia. Indrasetianingsih & Wasik [3] uses a fixed effect model with a cross-section that distinguishes it from the additional dummy variable in its analysis, the analysis of factors that affect the poverty level of provinces in Indonesia.

Aulina & Mirtawati [4] used a regression model with OLS estimation with a fixed effect model approach using a dummy variable with the same research on Poverty, using panel data regression to estimate the variables influencing poverty in East Java regions and cities. Poverty certainly occurs due to many factors, such as Education being too low, laziness to work, and limited natural resources. Limited employment opportunities, limited capital, and family burdens [5].

Based on these references, this research will analyze the factors influencing poverty in Bengkulu province using panel data regression because the data is a combination of time series and cross-section data. From the results of the analysis, it is hoped that it can help policymakers focus more on overcoming the factors that influence poverty to recover the economy, especially in Bengkulu province. cross-section data in this case study is data from 10 districts in Bengkulu province. While the time series data is inter-time data, namely data that is studied in the 2017-2020 period.

Materials and Methods

Research data sourced from BPS Bengkulu province. The data used include four (4) independent variables, namely Gross Regional Domestic Product (GRDP)/X1, Average Years of Schooling (AYS)/X2, Life Expectancy (LE)/X3, Human Development Index (HDI)/X4. While the dependent variable is the percentage of poverty/ Y. The flow chart of the research is presented in Figure 1.

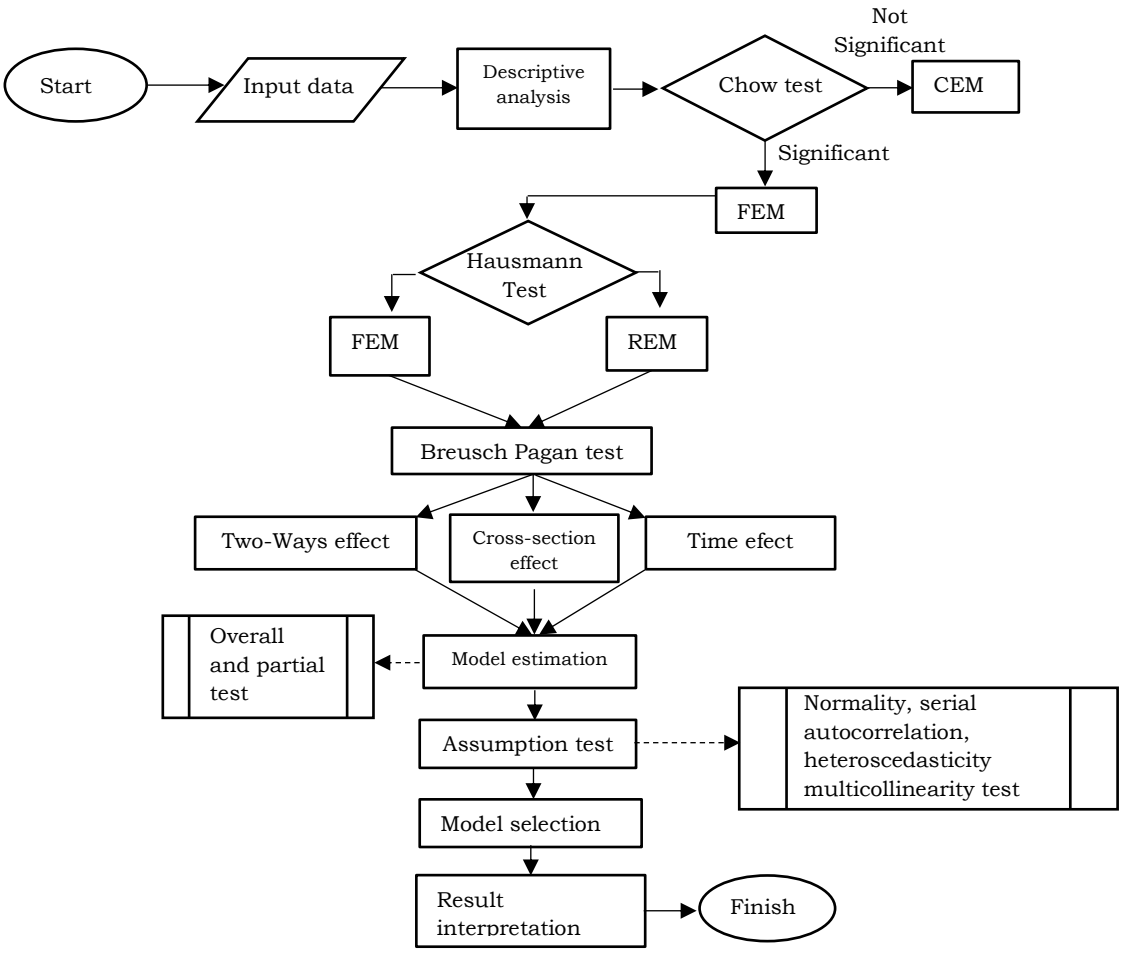

Figure 1. Flow chart of the research.

The first step is characteristic analysis. At this stage, the design of the variables contained in the core analysis that will be carried out, both independent and dependent variables taken must represent the formulation of the problem raised, as is the case with Poverty which is closer to GRDP, AYS, LE, and HDI to see the relationship of influence created.

The second step is parameter estimation. Three models are used in parameter estimation, namely: (1) Common Effect Model (CEM), (2) Fixed Effect Model (FEM), and (3) Random Effect Model (REM) [6]. The approach to the CEM model is an method that combines all existing data, both cross-section and time series data. The FEM model explains that time, and individual characteristics have differences that are symbolized by intercepts in different estimates. While the REM model explains the same thing as the Fixed effect model, the difference between the two is in the errors contained in the model. CEM. model equations [7], FEM [8], and REM [2] presented in Equations (1) to (3) sequentially.

$$Y_{it} = \beta_0 + \beta_1 X_{1/t} + \beta_2 X_{2/t} + \cdots + \beta_j X_{jit} + \varepsilon_{it} \quad (1)$$

$$Y_{it} = \beta_{0it} + \beta_1 X_{1/t} + \beta_2 X_{2/t} + \cdots + \beta_j X_{jit} + \varepsilon_{it} \quad (2)$$

$$Y_{it} = \beta_0 + \beta_1 X_{1/t} + \beta_2 X_{2/t} + \cdots + \beta_j X_{jit} + (\mu_i + \varepsilon_{it}) \quad (3)$$

where β_0 is the intercept of the model , β_j is the slope of the j-th regression, X_{jit} is the independent variable for the i-th cross section and the t-th time series, j is the number of independent variables, t is the unit area, and t is the time.

After the parameter estimation is accomplished. The third step is model selection. In selecting the model, three main tests were used, specifically: the Chow test, Hausman test, and Breusch Pagan test [9]. The Chow test is a test conducted to select between FEM and CEM, which is also called the restricted F-Test. The null hypothesis in the Chow test is that the best model chosen is the CEM model, while the alternative hypothesis is the FEM model [10]. The null hypothesis is rejected if the value of $F_{cal} > F_{table}$. The test statistics used are presented in Equation 4.

$$F_{cal} = \frac{(R_{UR}^2 - R_R^2) / (N-1)}{(1 - R_{UR}^2) / (NT-k)} \quad (4)$$

where N is the number of individualls, T represents the number of series, and k represents the number of parameters, R_{UR}^2 is the determinant coefficient of R^2 (for *FEM*), and R_R^2 is the determinant coefficient R^2 (for *CEM*).

Hausman test is performed to choose between FEM and REM. The REM model is the ideal model, which is the null hypothesis for the Hausman test, while the alternative hypothesis is the FEM model. The null hypothesis is rejected if the value of $X^2{}_{hit} > X^2{}_{k,a}$, e test statistic used is presented in Equation 5.

$$X_{hit}^2 = (b - \beta)' \mathrm{Var}(b - \beta)^{-1} (b - \beta) \quad (5)$$

where b is the coefficient for REM and is the coefficient for FEM, the Breusch Pagan test was carried out to determine the effects contained in the best-selected model, which consisted of (1) two-way effect, (2) cross-section effect, and (3) time effect [11]. The null hypothesis is that there is no two-way effect in two-way effect. In the cross-section effect, the null hypothesis is that there is no cross-section effect, while for the time effect, the null hypothesis is that there is no time effect.

Then, the fourth step is testing the panel data regression model. The tests carried out include selecting the best model, namely the overall test and the partial test. The overall test is used to find out how the independent variables affect the dependent variable with the null hypothesis that there is no independent variable that has a significant effect on the dependent variable. While the partial test is used when knowing which variables have a significant effect on the model from the Overall test results. The null hypothesis in the partial test is that the variable has no significant effect on the dependent variable. Furthermore, testing assumptions, namely the assumption of serial autocorrelation and data heteroscedasticity test. But sometimes, in panel data, it is possible that the assumptions are not satisfied, such as the assumption of homogeneity or the assumption of spatial dependencies. Alternatively, we can

use a Geographically Weighted Regression Model (GWPR) if the assumption of homogeneity is not met or the Spatial Data Panel if the assumption of spatial dependencies is not met. After testing the assumptions, proceed with determining the best model. Three measures are used in determining the best model, namely: (1) Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) [12].

The MSE value that is better or follows the actual data and can be used as a forecasting calculation in the future period is when the MSE value is low or close to 0 (zero). RMSE is a way of evaluating regression models by showing the average model prediction error. The measurement is done by measuring the accuracy of the model's forecast results. In the RMSE model, the prediction results are more accurate when the value is small or close to 0 (zero) [13].

Meanwhile, MAPE is the percentage value generated from the average absolute error [14]. The Accurate MAPE forecasting results are indicated by the smaller the error value in MAPE, indicating the more accurate the forecasting results obtained. MAPE value is said to be very good if it is worth less than 10%, while it is said to be good if it is less than 20% [15]. The equations of MSE, RMSE, and MAPE are presented in Equation 6 to Equation 8.

$$\mathrm{MSE} = \sum_{i=1}^{n} \frac{(\mathrm{Y}_i - \mathrm{A}_\mathrm{i})^2}{\mathrm{n}} \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\mathrm{Y}_i - \mathrm{A}_\mathrm{i})^2}{\mathrm{n}}} \tag{7}$$

$$MAPE = \frac{\sum_{i=1}^{n} \left| \left(\frac{A_i - Y_i}{A_i} \right) 100 \right|}{n} \tag{8}$$

where Y_i is the i-th predicted value, A_i is the i-th actual value, and n is the number of data.

Results and Discussions

Descriptive Analysis

The general picture obtained from existing data is that the poverty rate in Bengkulu Province for 2017-2020 has decreased in several districts, which is visually depicted in Figure 2.

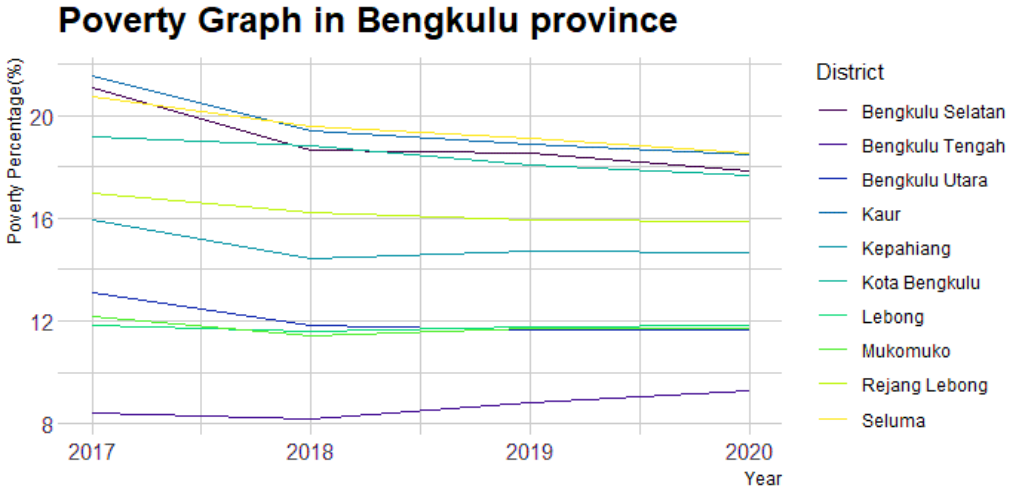

Figure 2. Visualization of the Poverty Level of Bengkulu Province.

Based on Figure 2, it was obtained that the district with the lowest poverty rate was Bengkulu Tengah district, which experienced a minor downward trend in the range of 8.20% - 9.30%. In contrast, the highest poverty rate is in the Kaur district, with the highest upward trend in the range of 18.47% -

21.54%. Based on information from the Regional Personnel Agency (BAPPEDA) of Kaur district, in 2022, Kaur district is currently designated as a priority locus for reducing extreme poverty with a national target of 0% in 2024.

Model Selection from Panel Data Regression

By the method in the previous chapter, based on the results of the chow test, the value of F_{cal}= 35.32 > 2.1 = F_{table} so reject H_0 or choose a fixed effect model. Then proceed to the second test, namely the Hausman test. The value was calculated using the Hausmann test findings of $X^2{}_{cal}$ = 9.92 > 7.8 = $X^2{}_{k,a}$ and it is concluded that H_0 is rejected, which refers to selecting the best estimation model, the fixed effect model. Furthermore, from the test, the researcher continues the next stage, the Breusch pagan test, to see the effects contained in the model. From the results of the Breusch Pagan test for the test hypothesis between the three effects, there are differences, where the results of the study can be explained as in Table 1.

Table 1. Pagan Breusch Test Results

Model	Effect	$X^2{}_{hit}$	Mark	$X^2{}_{k,a}$	decision
Fixed Effect	*two ways*	34.26	>	5.99	Reject H_0
	cross section	33.68	>	3.84	Reject H_0
	time	0.58	<	3.84	Fail to reject H_0

From Table 1. The researcher obtained the results from the test where the model failed to reject H_0 was the time effect, while for the two-way and cross-sectional effect, H_0 was rejected, which means that according to the Breusch Pagan test, there is a two-way and cross-section effect.

The steps taken after selecting the best model from the three tests are then continued with testing the model that has been selected and selecting the effect of the best model. From the results of the Overall test, it is obtained that there is at least one constant (coefficient of the independent variable) that has a significant effect on the model. At the same time, the partial test results showed only the GRDP variable, which is significant to poverty. In the test conducted to compare the coefficients of determination of the two effects, namely two ways and cross-section, it is found that the cross-section effect is better than the two-way effect. Furthermore, on checking the assumptions, it is found that the serial autocorrelation assumption and the heteroscedasticity assumption are fulfilled.

Then, the analysis is carried out by explaining the interpretation of the best model after conducting several previous tests, such as the overall test, partial test, and assumption test. Based on the analysis, from the four factors, a significant GRDP variable was obtained in the panel data analysis of the four factors. Table 2 presents the constant values for each district.

Table 2. Coefficient of each District

Number	Variable (*i*)	Coefficient (β_{0i})
1.	Bengkulu Selatan(BS)	27.39
2.	Rejang Lebong(RL)	24.55
3.	Bengkulu Utara(BU)	18.25
4.	Kaur(K)	26.31
5.	Seluma(S)	24.99
6.	Mukomuko(M)	17.93
7.	Lebong(L)	18.43
8.	Kepahiang(K)	22.23
9.	Bengkulu Tengah(BT)	17.98
10.	Kota Bengkulu(KB)	32.35

From the results of the regression coefficients for each Regency/ City in Bengkulu Province, the panel data regression model is presented in Equation 9.

$$\widehat{y_{i,t}} = -0.24 + GRDP_{i,t} + \beta_{0i} + \varepsilon_{i,t} \quad (9)$$

Based on Table 2 and Equation 9, the following are examples of implementation in predicting Bengkulu Selatan district in year t ($\widehat{\mathrm{BS}_t}$) and Rejang Lebong district in year t ($\widehat{RL_t}$).

$$\widehat{\mathrm{BS}_t} = -0.24 + GRDP_{BS,t} + 27.39$$

$$\widehat{RL_t} = -0.24 + GRDP_{RL,t} + 24.55$$

The MAPE, MSE, and RMSE, respectively, are 6.59% for MAPE, 5.48 for MSE, and 2.4 for RMSE, indicating that panel data analysis is very good in terms of predicting poverty in Bengkulu province.

Conclusion

Based on the results of panel data analysis, the factor that affects the poverty level in Bengkulu Province is GRDP. The model that fits the analysis is a fixed effect model with a cross-section effect or district variable units. The equation of the model is $\widehat{y_{i,t}} = -0.2430 + PDRB_{i,t} + \beta_{0i}$ with β_{0i} regression coefficient per district. From the prediction results, the MAPE value is 6.5%, MSE is 5.5%, and RMSE is 2.3%, where the error obtained is small, which explains that the difference between the actual and predicted values is less.

Acknowledgement

The author would like to thank all those who have helped so that this research can be completed.

References

[1] N. Liao and Y. He, Exploring the effects of influencing factors on energy efficiency in industrial sector using cluster analysis and panel regression model, Energy, 158 (2018) 782–795.
[2] D. N. Gujarati and D. C. Porter, Basic Econometric. New York: McGraw-Hill Education, 2008.
[3] A. Indrasetianingsih and T. K. Wasik, Model regresi data panel untuk mengetahui faktor yang mempengaruhi tingkat kemiskinan di pulau Madura, J. Gaussian, 9(3) (2020) 355–363.
[4] N. Aulina and Mirtawati, Analisis regresi data panel pada faktor-faktor yang mempengaruhi kemiskinan di Indonesia tahun 2015 – 2019, KINERJA J. Ekon. dan Bisnis, 4(1) (2021) 78–90.
[5] Hartomo and A. Aziz, Ilmu sosial dasar. Jakarta: Bumi Aksara, 2004.
[6] A. M. Astuti, Fixed effect model pada regresi data panel, Beta, 3(2) (2010) 134–145.
[7] D. Hanum, Studi tentang SUR untuk Data Panel dengan Model Gravitasi, Insititut Teknologi Sepuluh November, 2014.
[8] W. H. Greene, Econometric Analysis: Global Edition. 2020.
[9] P. R. Sihombing, Analisis regresi data panel brganda, in Statistik Multivariat dalam riset, Widina, (2021) 95–112.
[10] B. H. Baltagi, Econometrics, 6th ed. New York: Springer International Publishing, 2021.
[11] D. Rosadi, Analisis ekonometrika & runtun waktu terapan dengan R: aplikasi untuk bidang ekonomi, bisnis, dan keuangan. Yogyakarta: Gadjah Mada University Press, 2011.
[12] D. S. K. Karunasingha, Root mean square error or mean absolute error? Use their ratio as well, Inf. Sci. (Ny)., 585 (2022) 609–629.
[13] D. Chicco, M. J. Warrens, and G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci., 7 (2021) 1–24.
[14] S. Kim and H. Kim, A new metric of absolute percentage error for intermittent demand forecasts, Int. J. Forecast., 32(3) (2016) 669–679.
[15] C. Chatfield, Time series-Forecasting, 1st ed., no. 1. United Kindom, 2000.