BAREKENG: Journal of Mathematics and Its Application
September 2022 Volume 16 Issue 3 Page 861–868
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss3pp861-868

MULTILEVEL NON-LINIER REGRESSION FOR REPEATED MEASURMENT DATA AS STUDY OF PEANUT GROWTH

Arie Purwano [1]*, Umul Aiman[2]

[1]Mathematic Study Program, Faculty of Teacher Training and Education,
Mercu Buana Yogyakarta University
[2]Agrotechnology, Faculty of Agroindustry, Mercu Buana Yogyakarta University
Wates St. Km 10, Yogyakarta, 55753, Indonesia

Corresponding author's e-mail: [1] arie@mercubuana-yogya.ac.id*

Abstract. *Peanut is one of the most important legume commodities in Indonesia. In its implementation, a lot of research has been done related to this plant. However, in studies conducted by growth models, it is very rarely studied. Therefore, researchers are interested in modeling the growth of peanuts. One of the models that can be used is a multilevel regression model for the case of repeated measurement data. Multilevel regression was chosen because it is considered to provide more information than other regression models. On the other hand, the nonlinear model was chosen based on the tendency of the initial plot of the data obtained. The research method used is a case study in the study of peanut growth. This study aims to build the best model based on the tested model. The Restricted Estimator Maximum Likelihood (REML) parameter estimation method was chosen because it is considered to have unbiased parameter estimates. The best model is based on the lowest Akaike Information Criterion (AIC) generated from a predetermined model. The results obtained indicate that the multilevel parabolic regression model is the model with the best AIC size. In addition, it was found that there was an Interclass Correlation (ICC) of 81.19% which indicated a difference in variability between levels.*

Keywords: *peanut,regression, multilevel*

Article info:

Submitted: 9th May 2022 *Accepted: 9th August 2022*

How to cite this article:
A. Purwano and U. Aiman, "MULTILEVEL NON-LINIER REGRESSION FOR REPEATED MEASURMENT DATA AS STUDY OF PEANUT GROWTH", *BAREKENG: J. Math. & App.*, vol. 16, iss. 3, pp. 861-868, September, 2022.

1. INTRODUCTION

The rapid development of science in the last few decades is no doubt the result of human curiosity about the universe. Along with scientific developments, the need for analytical data also plays a role in analyzing every existing scientific study. Data analysis seems to play a fundamental role in research in various sciences, both scientific and social. One way to create superior research is to conduct research collaborations. This is very important to produce innovative researches by giving each other different scientific contributions in research. It is understandable that the development of statistics itself could not have been born without the problems that arise in relation to data analysis. Research collaboration seems to provide the key in developing problem-based statistics in other scientific fields. One of the scientific fields in question is in the field of agriculture. In agriculture, especially mathematics and statistics, it has important functions and roles, including as a communication tool for data producers and data lovers, as a tool or method for describing agricultural data, both with regression, correlation, and comparison methods [1]. The role of statistics in agriculture is very basic in order to produce compatible data analysis. In agriculture there are many things that can be studied, one of the interesting study materials is plant growth modeling.

Peanuts or in Latin Arachis hypogaea L., is one of the second most important legume commodities after soybeans in Indonesia [2]. Peanut is a staple commodity that is very valuable in Indonesia. Peanuts are a commodity that is rich in nutritional content of protein, fat, iron, vitamin E, vitamin B, vitamin A, vitamin K, phosphorus, lecithin, choline, and calcium. In his study, it was stated that peanut seeds contain 40–48% oil elements, 25% protein, and 18% carbohydrates and B complex vitamins [3]. In the field of agriculture itself has been found related to the analysis of production or agricultural products of peanuts. However, studies on growth or better known as growth models are still rarely carried out. It could be because of the economic factors from the results of growth modeling that are less attractive to researchers or in the implementation of models in the field that are rarely used. However, every problem about data becomes interesting to be discussed in statistical modeling. In some literacy, the growth model is generally carried out using a linear regression model approach, either simple or embellished by taking into account the factors that influence growth. However, the development of the complexity of the data structure directly becomes one of the interesting studies in the selection of the model to be used. The complexity in question is the discovery of a hierarchical or multilevel data structure, where each object under study basically has a growth model of each. Therefore, one of the analyzes that can be used is multilevel regression model analysis. Multilevel regression is characterized by a nested data structure. The data are characterized by nested membership relationships among observation units [4]. Thus the growth data for each collection of individual objects observed in the growth study can be modeled in a multilevel regression model.

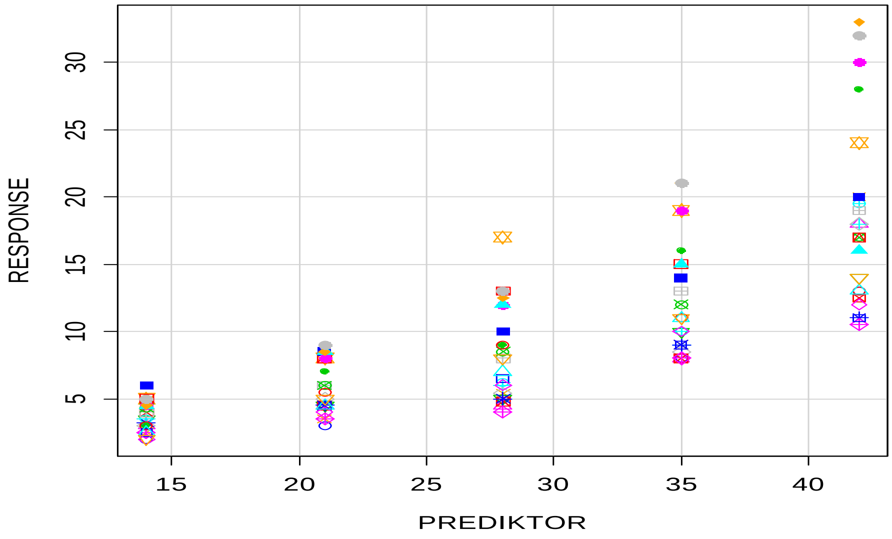
Multilevel regression analysis is considered very full power. Multilevel regression methodological approach, researchers can analyze the relationships between variables on or at least two different levels of analysis [5]. In general, the equation in a multilevel regression model can be partitioned in two parts. The partition in question is better known as the fix effect (fixed effect) and random effect (random effect). The fixed effects section in the multilevel regression model includes multilevel regression coefficients and predictor variables, while the random effects section includes random parameters that include errors at each level formed. The two parts that make up the regression model equation are known as mix models. In line with this [6]. In an ordinary one-level regression model, the assumption is that all individuals, even if from different centres, belong to one common population. In a multilevel model, we consider that there may be genuine differences between center populations which are themselves a sample from a superpopulation [7]. In relation to mixed models related to regression models, an analytical model known as Generalized Linear Mixed Models (GLMMs) was also developed. The development of GLMMs contributed to the analysis carried out. The development of GLMMs gave birth to several estimation methods for further analysis, namely Gauss Hermite quadratur, Laplace Approximation, and Penalized Quasi-Likelihood (PQL).

The Penalized Quasi-Likelihood estimation method is very helpful in estimating the parameters in the multilevel regression equation model. The PQL method has developed theoretically until now it can be used. Penalized Quasi-Likelihood aims to obtain values that are useful for approaching parameter inference and the realization of random effects in multilevel models [8]. Along with scientific developments, the PQL method that is widely known today is a numerical procedure. In line with this, the procedures applied in the PQL method [9]. The first procedure is Iterative Generalized Least Square (IGLS) which in its development is considered to be biased towards the estimation of the variance value. To obtain an unbiased estimate of the variance value, the Resticted (residual) Maximum Likelihood (REML) estimation method is used. The REML method can at least greatly reduce the bias and even completely eliminate it in some situations [10]. Therefore, Goldstein offers a new iterative procedure known as the Restricted (residual) Iterative Generalized Least Square (RIGLS) which is unbiased in the estimation of the variance value.

Based on the things that have been written, peanut growth is an event that can be approached using a multilevel regression model for repeated measurement data. This can be seen from the data structure where if each plant is measured repeatedly, the plant measurement is at level-1 and the plant that is measured repeatedly is at level-2. The estimation method used is the REML method because this method is unbiased with respect to the estimated variance parameter. On the other hand, in the initial study of the growth model, it was found that the model to be estimated was non-linear. Therefore, the model approach is both quadratic, cubic, and logarithmic considering that the estimation procedure used is still based on the procedure in the estimation of the linear model. Thus, more and more models will be formed so that it is hoped that a much better model estimate will be obtained. Furthermore, the modeling procedure used using the forward selection method by taking into account the Akaike Information Criteria (AIC) value was chosen as the benchmark for selecting the best model formed. In addition, to see whether there is a correlation between treatment classes or what is commonly called Interclass Correlation (ICC). The intraclass correlation coefficient (ICC) is recommended for the assessment of the reliability of the measurement scale [11]. Correlation between classes refers to the variability within each level that is formed.

2. RESEARCH METHODS

In this research, the method used is a case study with a literature method approach. The method chosen was based on the context used, namely the acquisition of data on the growth of peanuts. The literature approach was chosen considering the context of multilevel regression modeling which is still rarely used, especially in non-linear multilevel regressions, namely polynomials and powers. The main focus in this research is the growth of peanuts. The number of samples used in this study were 125 plants and were taken randomly. The data used is primary data with the initial aim of knowing the impact of weeding and non-weeding from the treatment given. Peanut seeds in the plantation laboratory of Mercu Buana Yogyakarta University with initial observations starting on October 10, 2019 and ending on November 19, 2019. Observations on the growth of peanut plants were carried out every week by taking into account the variables of height, number of branches, and day of planting (DAT). The results of observations of peanut growth were then analyzed using the R program package. The R program was chosen considering the advantages of open-source and free programs. The determination of the non-linear model is obtained by looking at the tendency of the scatterplot which is formed based on the multilevel data held as follows:



Picture 1. Scatter plot of repeated measurement multilevel on peanut

The analytical method used is the forward selection method. The forward selection method was chosen with the aim of testing each model formed and obtaining the best model based on the variables used. Forward selection starts with no predictors and then builds the model by adding predictors one at a time, backward stepwise selection begins with a model with all k predictors and then removes them **[12]**. Furthermore, the selection of the best model is carried out by taking into account the lowest Akaike Information Criterion (AIC) value for each model. According to the theory of generalized linear regression modeling, relative goodness-of-fit of several models may be compared based on a number of criteria, including the Akaike information criterion **[13]**. The application of the AIC value itself is a method used mainly in selecting the best regression model with the aim of forecasting (forcasting), which can explain the suitability of the model with existing data (in-sample forecasting) and values that will occur in the future (out of sample casting) **[14]**. The initial modeling concept used is 1-level linear regression modeling, this aims to see the adequacy of statistics in building level-2.

2.1 Modeling Procedure

Suppose i=(1,2,...,n) with h=(1,2,…,m) each describes the index of measurement or observation at the i-th time nesting in the h-th plant. Suppose r=(1,2,3) denotes the polynomial form constructed in the model with r=1 indicating simple or multiple linear regression, r=2 indicating a parabolic regression model, and r=3 indicating a cubic regression model such that the building blocks of the model are used is formed as follows:

y_{ih} is the response variable in this case is the observation of peanut height at time i-th which is measured on h-th plant

x_{1ih}^{r} is a predictor variable in this case is the observation of the number of peanut branches at the i-th time measured on the h-th plant

x_{2ih}^{r} is a predictor variable in this case is the age of the i-th plant measured on the h-th plant

β_{0h} is random intercept on h-th plant

β_{1h} is random slope number of peanut branches of h-th plant

β_{2h} is random slope age of h-th plant

u_{0h} is the h-th random intercept

u_{1h} is an error in random slope number of peanut branches of h-th plant

u_{2h} is an error in random slope age of h-th plant

e_{ih} is random error of the i-th plant measured on the h-th plant

β_0 is parameter intercept in 2-level regression model

β_1 is parameter slope of number of peanut branches in 2-level regression model

β_2 is parameter slope of number of peanut age in 2-level regression model

σ_{0h}^2 is variance in parameter intercept model 2-level

σ_{1h}^2 is variance in parameter slope of branches model 2-level

σ_{2h}^2 is variance in parameter slope of age model 2-level

The equations of the model to be built are generally limited and classified into four, namely:
1. Equation of linear regression model
2. Equation of multilevel linear regression model, both random intercept and random slope models
3. Equation of multilevel parabolic regression model, both random intercept and random slope models
4. Equation of multilevel cubic regression model, both random intercept and random slope models

The five models in general can be solved in a linear iterative procedure. The general equation used in the context of GLLMs can be formed in a general matrix structure as follows:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \qquad (1)$$

where:
y is vector $n \times 1$
β is vector $(p+1) \times 1$
X is matrix $n \times (p+1)$
Z is explanatory matrix $n \times n$
u is vector of random effect in level-2
e is vector of random effect in level-1

The estimation procedure is carried out using the Restricted Iterative Generalized Least Square (RIGLS) iteration procedure to obtain the parameter estimation results using the REML method which is unbiased on variance. The REML estimator in the iteration procedure in obtaining parameters can be done in the following way:

$$\widehat{\boldsymbol{\beta}}^{(t)} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{y} \quad (2)$$

$$\widehat{\boldsymbol{\theta}}^{(t)} = \left(\mathbf{Z}^{*\mathbf{T}}\mathbf{V}^{*^{-1}}\mathbf{Z}^{*}\right)^{-1}\mathbf{Z}^{*\mathbf{T}}\mathbf{V}^{*^{-1}}\mathbf{y} \quad (3)$$

dengan

t is n-th iteration; n=0,1,2,

$\widehat{\boldsymbol{\beta}}^{(t)}$ is t-th iteration of fix parameter estimation

$\widehat{\boldsymbol{\theta}}^{(t)}$ is t-th iteration of random parameter estimation

$\mathbf{V}$ is diagonal block matrix $n \times n$

$\mathbf{Z}^{*}$ is explanatory matrix $n \times n$

For each model using a forward study design approach, the Akaike Information Criteria (AIC) value will be measured as a reference in determining the best model formed. The model with the lowest AIC value is the model chosen to represent peanut growth. In addition, Interclass Correlation (ICC) is also a concern. In multilevel repeated measurements, where the observations are nested in the corresponding individuals, the intraclass correlation can be interpreted as the dependence of the observations nested in the corresponding individuals. This can be interpreted as the value used to determine how much influence the individual can explain the observations made. Intraclass correlation can explain the proportion of the variable values between observations nested in the specified object [15]. In addition, in this case multilevel regression can be performed if the value of the ICC is above 20%. The relationship between levels in the case of 2-level multilevel regression can be calculated using the following formula:

$$ICC = \frac{var(\mu_h)}{var(\mu_h)+var(e_{ih})} \quad (4)$$

In equation (4) it can be understood that $var(\mu_h)$ is the individual or unit variance at level-2 and $var(e_{ih})$is the variance of each nested observation on the corresponding individual at level-1.

3. RESULTS AND DISCUSSION

The best model selection method used in this research is forward study design. Therefore, the diwali analysis by modeling each predictor variable is used, starting from a simple linear regression model with one variable to a more complex model structure. The parameter estimation method used is the REML method with an iterative procedure, namely RIGLS. Based on the results of the data analysis carried out, as many as 25 possible model equations were selected to be formed. The results obtained are presented as follows:

Table 1. Parameter Estimation of linear regression model

Parameter		Linear model	
	Model 1	Model 2	Model 3
Fix			
$\boldsymbol{\beta_0}$	-4.690	3.291	-2.542
$\boldsymbol{\beta_1}$	0.437		0.690
$\boldsymbol{\beta_2}$		1.268	-1.466
Random			
Level 1			
$\boldsymbol{\sigma^2_{e0}}$	4.057	6.100	3.627
AIC	713.817	808.777	688.097

Model 3 is a model for multiple linear regression equations with two variables. Based on the table, it can be seen that the AIC value is 688,097. The parabolic and cubic regression models are presented in Table 2. As follows:

Table 2. Parameter estimation of parabolic dan cubic regression

Parameter	Parabolic		Cubic	
	Model 4	Model 5	Model 6	Model 7
Fix				
$\boldsymbol{\beta_0}$	1.460	0.032	0.000	0.009
$\boldsymbol{\beta_1}$	0.009		3.543	
$\boldsymbol{\beta_2}$		5.274		8.015
Random				
Level 1				
$\boldsymbol{\sigma^2_{e0}}$	15.361	4.140	15.273	39.801
AIC	713.694	549.949	720.346	828.082

Based on Table 2. it can be seen that the best equation model by taking into account the lowest AIC value is Model 5. Equation 5 is a parabolic or quadratic equation model with an AIC value of 549,694. Based on Table 2. the comparison of the very high spike in model variability occurred in Model 7 with a variance value of 39,801. This is possible because the variable used is the third power of the predictor variable, namely the age of the plant. This results in increased model variability. The estimation of the multilevel linear regression equation in either the random intercept or the random slope model is presented as follows:

Table 3. Random intersep dan slope model of multilevel linear regression

Parameter	Random Intercept			Random Slope		
	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Fix						
$\boldsymbol{\beta_0}$	0.150	-4.690	-2.710	-2.119	-4.670	-3.488
$\boldsymbol{\beta_1}$	1.904		-1.352	2.481		-0.691
$\boldsymbol{\beta_2}$		0.508	0.676		0.509	0.579
Random						
Level-2						
$\boldsymbol{\sigma^2_{0h}}$	12.681	8.152	6.426	0.000	5.006	0.445
$\boldsymbol{\sigma^2_{1h}}$				0.906		0.806
$\boldsymbol{\sigma^2_{2h}}$					0.035	0.053
Level-1						
$\boldsymbol{\sigma^2_{e0}}$	26.142	8.507	6.894	22.472	4.260	3.726
AIC	798.599	676.786	652.344	797.962	619.624	616.260

Based on the parameter estimation results presented in Table 3. It can be seen that the best model equation that can be formed is Model 13 with the lowest AIC value of 616,260. However, it is also necessary to pay attention to another regression model, namely multilevel parabolic regression which is presented as follows:

Table 4. Random intersep dan slope model of multilevel parabolic regression

Parameter	Random Intercept			Random Slope		
	Model 14	Model 15	Model 16	Model 17	Model 18	Model 19
Fix						
$\boldsymbol{\beta_0}$	0.163	1.459	2.820	4.374	1.46	2.992
$\boldsymbol{\beta_1}$	2.003		0.021	0.045		0.027
$\boldsymbol{\beta_2}$		0.009	0.004		0.009	0.004
Random						
Level-2						
$\boldsymbol{\sigma^2_{0h}}$	11.467		1.442	1.048	2.706	3.864
$\boldsymbol{\sigma^2_{1h}}$		8.425		0.000		0.000
$\boldsymbol{\sigma^2_{2h}}$					0.000	0.000
Level-1						
$\boldsymbol{\sigma^2_{e0}}$	4.14	7.142	0.909	1.660	2.133	0.568
AIC	551.949	667.537	433.789	487.938	584.605	428.445

Based on Table 4. it can be seen that the model with the lowest AIC is Model 19 with an AIC value of 428,445. In this case, it was found that there was a significant decrease in the AIC values in other models. However, it is also necessary to pay attention to the estimation of the multilevel cubic regression model as follows:

Table 5. Random intersep dan slope model of multilevel cubic regression

Parameter	Random Intercept			Random Slope		
	Model 20	Model 21	Model 22	Model 23	Model 24	Model 25
Fix						
$\boldsymbol{\beta_0}$	6.966	3.543	4.007	3.615	3.543	3.044
$\boldsymbol{\beta_1}$	0.015		-0.008	0.057		0.007
$\boldsymbol{\beta_2}$		0.000	0.000		0.000	0.000
Random						
Level-2						
$\boldsymbol{\sigma_{0h}^2}$	10.581	8.447	6.923	0.000	6.682	0.622
$\boldsymbol{\sigma_{1h}^2}$				0.003	0.000	0.000
$\boldsymbol{\sigma_{2h}^2}$					0.000	0.000
Level-1						
$\boldsymbol{e_{ih}}$	30.682	7.032	6.507	20.905	1.643	1.230
AIC	821.875	673.674	673.177	828.316	592.406	589.724

Based on Table 5. it can be seen that the model with the lowest AIC value is Model 25, which is 589,724. In general, it is found that the best model that can be formed is a multilevel parabolic model with the lowest AIC value of 428,445, namely in Model 19. This can be a study that peanut growth can be modeled in the form of multilevel parabolic regression. In addition, it can also be determined that the ICC value is 0.8719 or 87.19% which shows very high variability at each level. However, further studies showed that there was a very significant difference in the growth of peanuts. On the other hand, the high ICC value is possible due to differences in treatment in the experiments carried out at the time of the study. Therefore, it is necessary to carry out further analysis regarding this assumption. Either by modeling with a higher level based on clustering the treatment or other statistical analysis. Furthermore, it should be realized that in the context of multilevel regression modeling, the study or use of various models certainly provides many alternatives for selecting the best model. This is in line with the expression of George W. Box, an expert in mathematical modeling who stated "All Models are wrong, but some are usefull".

4. CONCLUSIONS

Based on the results of the analysis that has been done, it can be concluded that one of the best models is obtained based on multilevel regression analysis. Multilevel regression would be an alternative development of regression modeling. Based on the results of the parameter estimates carried out, the best model is obtained, namely the multilevel random slope model, precisely the 15th model with the following equation:

$$\hat{y}_{ij} = -2.992 + 0.027x_{1ij}^2 + 0.003x_{2ij}^2 \qquad (5)$$

The model equation was chosen based on the lowest AIC value formed from the 25 models tested. The AIC value obtained is 428,445. In addition, it is known that the ICC value is 87.18% which can be interpreted as the high variability between the levels used. This is possible due to the experimental treatment given, namely with weeding and without weeding. For this reason, it is necessary to carry out further analysis to explain the impact.

AKNOWLEDGEMENT

We would like to express our deepest gratitude to the entire community of the Institute for Research and Community Service, Mercu Buana University, Yogyakarta for all forms of support until this research was completed. To the big family of the Agrotechnology Study Program, Faculty of Agroindustry, Mercu Buana University, Yogyakarta for all forms of support in research experiments and agricultural laboratories used. Not forgetting the entire academic community of the Faculty of Teacher Training and Education, especially the big family of the Mathematics Education Study Program, Mercu Buana University, Yogyakarta for all the encouragement and media for self-development of researchers so far.

REFERENCES

[1] A. F., “Peranan Matematika Dan Statistika Dalam Pertanian Industrial Untuk Mewujudkan Ketahanan Pangan Nasional,” in *Prosiding Seminar Nasional Matematika*, Jember, 2014.

[2] E. e. a.] Respati, “Kacang tanah. Buletin Konsumsi Pangan,” *Pusdatin,* vol. 4 (1), no. Buletin Konsumsi Pangan, pp. 6-15, 2013.

[3] C. R. R. V. O. &. V. R. Kumar, “Correlation and path coefficient analysis in groundnut (Arachis hypogaea L.),” *International Journal of Applied Biology and Pharmaceutical Technology,* vol. 5 (1), pp. 8-11, 2014.

[4] R. V. D. Leeden, “Multilevel Analysis of Repeated Measure Data,” *Kluwer Academic Publisers,* vol. 32, pp. 15-29, 2010.

[5] J. P.-M. M. D. L.-G. E. M. P.-O. a. J. J. T. José F Molina-Azorín, “Multilevel research: Foundations and opportunities in management,” *Business Research Quarterly ,* vol. 23 (4), p. 319 –333, 2020.

[6] G. e. a. Weinmayr, “Multilevel regression modelling to investigate variation in disease prevalence across locations,” *International Journal of Epidemiology,* vol. 46 (1), p. 336–347 , 2017.

[7] A. J. F. N. e. a. Hubbard AE, “To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health,” *Epidemiology,* vol. 21 (4), pp. 467-474, 210.

[8] N. a. C. Breslow, “Aproximate Inference in GLMM,” *Journal of American statistical Association,* vol. 88, pp. 95-25, 1993.

[9] N. S. e. al, “Analisa Metode Forward dan Backward Untuk Menentukan Persamaan,” *Saintia Matematika,* vol. 2 (4), pp. 235-360, 2014.

[10] M. Elff, J. P. Heisig, M. Schaeffer and S. Shikano, “Multilevel Analysis with Few Clusters: Improving Likelihood-Based Methods to Provide Unbiased Estimates and Accurate Inference,” *British Journal of Political Science,* vol. 51 , p. 412–426, 2021.

[11] e. a. Carly A. Bobak, “Estimation of an inter-rater intra-class correlation coefficient that over comes commonassumption violations in the assessment of health measurement scales,” *MC Medical Research ,* vol. 18 (1), pp. 1-11, 2018.

[12] . F. H. Richard B. Darlington, Regression Analysis and Linear Models, New York: Guildford Press, 2016.

[13] O. Korosteleva, Advanced Regression Models with SAS and R, Florida: CRC Press, 2018.

[14] M. Faturahman, “Pemilihan Model Regresi Terbaik Menggunakan Akaike’s Information Criterion,” *Eksponensial,* vol. 1 (2), pp. 26-33, 2010.

[15] R. Bickel, Multilevel Analysis for Applied Research: It's Just Regression, New York: Guilford Press, 2013.