Panel Data Analysis of Two Level Mixed Linear Models for Factors Affecting The Health Index in West Java

Asep Solih Awalluddin[1,a)], Mia Siti Khumaeroh[1,b)], H. Amalia[1], Inge Wahyuni[2,c)]

[1]*Department of Mathematics UIN Sunan Gunung Djati Bandung*
[2]*Regional Research and Development Agency, West Java Province*

[a)]email: aasolih@uinsgd.ac.id
[b)]email: miasitihumairoh@uinsgd.ac.id
[c)]email: inge.awal@gmail.com

Abstract

The purpose of this study is to construct a multilevel mixed linear model for panel data by estimating parameters and testing the hypothesis of fit of the model with case studies in determining the prediction of the health index for the marginal and conditional models on the factors that influence the prediction of the health index in West Java for 2016 data. -2021, with time (year) and region (district and city) variables as factors involved in the model. Multilevel mixed linear model is the development of a mixed linear model that can be used to analyze correlated panel data. Parameter estimation uses the Maximum Likelihood (ML) method to estimate fixed effect parameters and Restricted Maximum Likelihood (REML) to estimate covariance parameters. The results obtained by the health index prediction model in West Java, both for the marginal and conditional prediction models and goodness of fit model.

Keywords : *Panel Data, Mixed Linear Models, Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), Health Data*

Introduction

The use of statistical analysis in various studies and scientific disciplines is used as a basis for decision making, especially by policy makers, namely regional governments. Policy making is no longer based on subjective or political comparisons, but is carried out by taking into account data, design and analysis. One very basic policy is in the health sector. The health index is one of three components that build the human development index (HDI). The HDI value is used as a parameter to measure the development of a region, whether it is a further progress of a country's development or vice versa. Thus, it is important to acknowledge the factors significantly affecting the health index of a region.

Comprehensive statistical analysis taking into account various variables and the completeness of the information will greatly affect whether or not the results of the analysis are good. Data analysis that has been done so far is data analysis with cross section data structure or data with time series. The weakness of each data analysis tends to ignore variations in data changes. For cross-sectional data, it only pays attention to variations in the object of research, while for time series data, variations in the time of observation observed is a fixed object. A more comprehensive alternative is panel data analysis with a combined data structure between cross section and time series data, taking into account variations in data changes for both the research object and the time of study [17],[18],[19].

The aim of the research is to formulate a mathematical model on a two-level mixed linear model on panel data for factors that influence the health index in West Java by estimating model parameters and determining the best model fit test that can be used as a basis for making development priority policies, especially in the health sector. This research will look at the time factor (year) and regional factor (district/city) in the health index panel data in West Java to predict the health index in West Java with the marginal prediction model and the conditional prediction model from West Java health index data in 2016- 2021. Multilevel mixed linear model is the development of a mixed linear model that can be used to analyze correlated panel data.

Method

1. Linear Regression Models

The linear regression model is an approach to modeling the relationship between dependent observations $\boldsymbol{Y} = (y_1, \dots, y_n)'$ and one or more independent variables denoted by $\boldsymbol{X} = \left(x_{ij}\right)$ the matrix model $n \times k$, where x_{ij} is the value of the explanatory variable j for the observation i. The linear regression model with n observations and p the dependent variable for each data studied is as follows[1]:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ij}\beta_j + \varepsilon_i, i = 1,2, \dots, n \; and \; j = 1,2, \dots, p \quad (1)$$

In matrix notation the linear regression model can be stated as follows[2]:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where Y : the vector $n \times 1$ of the dependent or response variable

$\boldsymbol{X}$: the matrix $n \times p$ of the independent or predictor variables

$\boldsymbol{\beta}$: the vector $p \times 1$ of the unknown parameters

$\boldsymbol{\varepsilon}$: represents the error

$E(\boldsymbol{\varepsilon}) : 0$

$Var(\boldsymbol{\varepsilon}) : \sigma^2\boldsymbol{I}.$

2. Mixed Linear Model

The linear regression model for longitudinal data becomes less suitable because longitudinal data is a correlated data. Correlation found in longitudinal data can be overcome by including random effects in the form of parameters specific to the study unit into the model [2]. Mixed linear model or Linear Mixed Model (LMM) is a model consisting of a mixture of fixed effects and random effects. Fixed effects are parameters whose values are unknown and are related to categorical or classification variables [3]. Of all the conditions of fixed effect is an interesting thing that is the main goal of research. Random effect is a random value related to the level of random factor i.e. a classification variable that can be considered as a random sample of the population being studied [3]. In contrast to fixed effects which are represented by parameters in LMM, random effects are represented by random variables which are usually assumed to follow a normal distribution [2]. While random effects usually represent random deviations from the relationship explained by fixed effects, random effects in mixed linear models can be either random intercepts or random slopes [3].

Mixed linear regression models can be used to model the correlation among observations of longitudinal data by assuming that each study unit is associated with a random effect whose value cannot be observed [5]. There are research units, all research units are collected at the same time, . The mixed linear model for the research unit is stated as follows $t = 1, \dots, T_i$ [3]:

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad (3)$$

Where $\boldsymbol{Y}_i$: vector of response for *i*th research unit

$\boldsymbol{X}_i$: matrix T_i x p of as many p covariates as the first column is a constant 1

$\boldsymbol{\beta}$: vector of fixed effect parameters associated with the p covariates in the matrix $\boldsymbol{X}_i$

$\boldsymbol{Z}_i$: matrix representing as many q covariate values as the first column is constant 1

$\boldsymbol{\alpha}_i$: the vector of random effects associated with the q covariates in the matrix $\boldsymbol{Z}_i$.

Random effects specific $\boldsymbol{\alpha}_i$ are assumed to be normally distributed with a mean of zero and the variance covariance matrix $\boldsymbol{D}$ which is a positive definite matrix of size qxq. The random error component is also assumed to follow a normal distribution with a mean of zero and a variance covariance matrix $\boldsymbol{R}_i$ of size T_i x T_i. Random effects $\boldsymbol{\alpha}_i$ with random error components $\boldsymbol{\varepsilon}_i$ are assumed to be mutually independent.

3. Two Level Mixed Linear Model

To perform an analysis using a mixed linear model, the data used must have at least two data levels. The first level of longitudinal data is repeated observations of the research unit. For the unit-*i*th to time study *t*, the linear relationship between the explanatory z_{it} and response variables y_{it} can be written as follow [2]:

$$y_{it} = \beta_{0i} + \beta_{1i} z_{it} + \varepsilon_{it} \quad (4)$$

The response variable at level one is modeled as a function of time and the within subject residual component.[4]. The residual component in the first level model is assumed to be normally distributed with zero mean and variance $\boldsymbol{R}_i = var(\varepsilon_{it})$. Each research unit on longitudinal data has a unique intercept and slope. Therefore the intercept and slope of the first level model are allowed to vary according to the research unit through the notation β_{0i} and β_{1i} . The second level of the model represents the next level of data, namely the research unit. By adding the predictor factor x_i , a second-level model is created with the specification of random effects for both the intercept and for the slope:

$$\beta_{0i} = \beta_{00} + \beta_{01} x_i + \alpha_{0i} \quad (5)$$

$$\beta_{1i} = \beta_{10} + \beta_{11} x_i + \alpha_{1i} \quad (6)$$

Added predictor factors were defined as between-subject variables. Variations between research units that occur in the intercept and slope are assumed to be related to predictor factors x_i. The subject-specific intercept (β_{0i}) for the study unit is related to the overall intercept or average intercept (β_{00}) and the random effect (α_{0i}) is related to the intercept. While the subject-specific slope (β_{1i}) is related to the overall slope (β_{10}) and the random effect (α_{1i}) is related to the slope. The random terms added to the second level model are α_{0i} and α_{1i} which are the random effects for the intercept and slope. This random effect is assumed to be $\boldsymbol{\alpha}_i \sim \boldsymbol{N}(0, \boldsymbol{D})$

Combining the second-order model in equations (5) and (6) to the first-order model in equation (4) will produce a two-level mixed linear model:

$$\begin{aligned} y_{it} &= (\beta_{00} + \beta_{01} x_i + \alpha_{0i}) + (\beta_{10} + \beta_{11} x_i + \alpha_{1i}) z_{it} + \varepsilon_{it} \\ &= \beta_{00} + \beta_{01} x_i + \alpha_{0i} + \beta_{10} z_{it} + \beta_{11} x_i z_{it} + \alpha_{1i} z_{it} + \varepsilon_{it} \end{aligned} \quad (7)$$

Define $\boldsymbol{\alpha}_i = (\alpha_{0i} \quad \alpha_{1i})'$, $\boldsymbol{z}_{it} = (1 \quad z_{it})'$, $\boldsymbol{\beta} = (\beta_{00} \quad \beta_{10} \quad \beta_{01} \quad \beta_{11})'$ and $\boldsymbol{x}_{it} = (1 \quad x_i \quad z_{it} \quad x_i \times z_{it})'$, then a two-level mixed linear model can be written in general form

$$\boldsymbol{y}_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_{it}\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{it} \quad (8)$$

or in matrix form it can be written as in equation (3).

4. Covariance Matrix Structure

One of the advantages of LMM is modeling the structure of the covariance matrix of data which is denoted by $\boldsymbol{R}$, because the covariance matrix determines the pattern of autocorrelation between residual components [4]. Observations at any time have a unique variance and covariance. The

covariance structure is the pattern in the covariance matrix. Some of these patterns appear frequently in several statistical procedures and so these patterns have names. The three most simple and commonly used covariance structures in conducting longitudinal data analysis using a mixed model are: first, the diagonal matrix structure or also known as the independent matrix structure assumes the error term associated with observations in the same research unit is uncorrelated and has the same variance, second, the Compound Symmetry (CS) covariance structure assumes that all variances over time are constant, with correlations between observations also constant, and third, unstructured covariance (UN). The structure of covariance matrix is in Table 1.

Table 1. Type of the covariance matrix

Type	Matrix $\boldsymbol{R}$	Number of Parameter
Diagonal	$\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ & \sigma^2 & & \vdots \\ & & \ddots & 0 \\ & & & \sigma^2 \end{pmatrix}$	1
Compound Symmetry (CS)	$\sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ & 1 & & \vdots \\ & & \ddots & \rho \\ & & & 1 \end{pmatrix}$	2
Unstructured (UN)	$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1T_i} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2T_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T_i1} & \sigma_{T_i2} & \cdots & \sigma_{T_i}^2 \end{pmatrix}$	$T_i(T_i+1)/2$

5. Parameter Estimation

The mixed linear model contained in equation (3) can be seen as a marginal linear model as follows [6]:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^* \tag{10}$$

With $\boldsymbol{\varepsilon}_i^* = \boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i$. The residuals $\boldsymbol{\varepsilon}_i^*$ are assumed to be $\boldsymbol{\varepsilon}_i^* \sim N(0, \boldsymbol{V}_i)$. The covariance matrix $\boldsymbol{V}_i$ is defined as $\boldsymbol{V}_i = Var(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i$[2]. The estimation of the fixed effect parameters in the mixed linear model is carried out using the marginal model approach for convenience. Longitudinal data with the marginal model approach defines the marginal distribution of the vector $\boldsymbol{Y}_i$ as follows

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i) \tag{11}$$

MLE is a method used to obtain an estimator by optimizing the natural logarithm of its likelihood function. The log likelihood function for is $\boldsymbol{y}_i \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}_i)$

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}) == -\frac{1}{2}n\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{12}$$

Looking for the optimal solution to the fixed effect parameters, the estimation results are:$\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}y \tag{13}$$

The estimation results $\widehat{\boldsymbol{\beta}}$ still depend on the covariance parameter $\boldsymbol{\theta}$ that needs to be estimated. Covariance parameter estimation is carried out using the REML method which will be done with R software.

Analysis Step

The steps that must be taken to estimate all parameters contained in a two-level mixed linear model:

1. To simplify the estimation process, modify the two-level mixed linear model into a single mixed linear model.
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{\varepsilon}$$
2. Assume so $\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{D}), \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{R})$ so that $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V} = \boldsymbol{ZDZ'} + \boldsymbol{R})$
3. Estimating the fixed effect parameters $\boldsymbol{\beta}$ using the maximum likelihood method with the following steps:
   1) The form of the likelihood function derived from the normal distribution density function, $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$
   $$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{V}(\boldsymbol{\theta})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})' \boldsymbol{V}(\boldsymbol{\theta})^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})\right)$$
   2) The form of the log likelihood function
   $$\ln(L) = -\frac{1}{2} n \ln(2\pi) - \frac{1}{2} \ln|\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})' \boldsymbol{V}(\boldsymbol{\theta})^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})$$
   3) Assume $\boldsymbol{\theta}$ it is known
   4) Optimize the log likelihood function against $\boldsymbol{\beta}$
   $$\frac{\partial l}{\partial \boldsymbol{\beta}} = \boldsymbol{X'V}(\boldsymbol{\theta})^{-1} \boldsymbol{y} - \boldsymbol{X'V}(\boldsymbol{\theta})^{-1} \boldsymbol{X\beta}$$
   5) Adjust $\frac{\partial l}{\partial \boldsymbol{\beta}} = 0$ so that a closed solution is obtained for $\widehat{\boldsymbol{\beta}}$
   $$\boldsymbol{X'V}(\boldsymbol{\theta})^{-1} \boldsymbol{y} - \boldsymbol{X'V}(\boldsymbol{\theta})^{-1} \boldsymbol{X\beta} = 0$$
   $$\boldsymbol{X'V}(\boldsymbol{\theta})^{-1} \boldsymbol{X\beta} = \boldsymbol{X'V}(\boldsymbol{\theta})^{-1} \boldsymbol{y}$$
   $$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'V}(\boldsymbol{\theta})^{-1} \boldsymbol{X})^{-1} \boldsymbol{X'V}(\boldsymbol{\theta})^{-1} \boldsymbol{y}$$
4. Estimating the covariance parameter using the REML and Newton Raphson methods with the following steps:
   1) Determine the structure of the matrix $\boldsymbol{D}$ and $\boldsymbol{R}$ the matrix that will be used so that the right assumptions are obtained.
   2) For matrix $\boldsymbol{D}$ and $\boldsymbol{R}$ standard structures, the following assumptions are obtained:
   $$\boldsymbol{\alpha} \sim N\left(0, \sigma_i^2 \boldsymbol{I}_{q_i}\right)$$
   $$\boldsymbol{\varepsilon} \sim N\left(0, \sigma_\varepsilon^2 \boldsymbol{I}_{T_i}\right)$$
   So obtained $\boldsymbol{V} = \boldsymbol{Z_i D Z_i'} + \boldsymbol{R} = \sum_{i=1}^{q} \boldsymbol{Z}_i \boldsymbol{Z}_i' \sigma_i^2 + \sigma_\varepsilon^2 \boldsymbol{I}_{T_i}$
   3) Suppose $\boldsymbol{\alpha}_0 = \boldsymbol{\varepsilon}, q_0 = T_i,$ dan $\boldsymbol{Z}_0 = \boldsymbol{I}_{T_i}$, then
   $$\boldsymbol{V} = \sum_{i=0}^{q} \boldsymbol{Z}_i \boldsymbol{Z}_i' \sigma_i^2$$
   4) To overcome biased estimation of the covariance parameter using ML, error contrast $\boldsymbol{K'X} = \boldsymbol{0}$ will be used to $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V}$ obtain it
   $$\boldsymbol{K'y} \sim N(\boldsymbol{0}, \boldsymbol{K'VK})$$
   5) The form of the REML function which is a derivative of the likelihood function by replacing the variables with their error contrasts, namely: replace $\boldsymbol{y}$ with $\boldsymbol{K'y}$, $\boldsymbol{X}$ with $\boldsymbol{K'X} = \boldsymbol{0}$, $\boldsymbol{Z}$ with , $\boldsymbol{K'Z}$and $\boldsymbol{V}$ with$\boldsymbol{K'VK}$
   $$l_{REML} = -\frac{1}{2}(N - r)\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{K'VK}| - \frac{1}{2}\boldsymbol{y'K}(\boldsymbol{K'VK})^{-1}\boldsymbol{K'y}$$
   6) Optimize against $l_{REML}\sigma_i^2$
   $$\frac{\partial l_{REML}}{\partial \sigma_i^2} = -\frac{1}{2} tr[(\boldsymbol{K'VK})^{-1} \boldsymbol{K'Z}_i \boldsymbol{Z}_i' \boldsymbol{K}] - \frac{1}{2} \boldsymbol{y'}(-1) \boldsymbol{PZ}_i \boldsymbol{Z}_i' \boldsymbol{Py}$$
   $$= -\frac{1}{2} tr(\boldsymbol{PZ}_i \boldsymbol{Z}_i') + \frac{1}{2} \boldsymbol{y'PZ}_i \boldsymbol{Z}_i' \boldsymbol{Py}$$
   with $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X'V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'V}^{-1}$

7) The optimal solution can be found using iterative methods. The iterative method that will be used is the Newton Raphson method which has the following iterations:σ_i^2

$$\widehat{\boldsymbol{\theta}}^{(m+1)} = \widehat{\boldsymbol{\theta}}^{(m)} - \left(\boldsymbol{H}(\widehat{\boldsymbol{\theta}})^{(m)}\right)^{-1}\boldsymbol{\nabla} f(\widehat{\boldsymbol{\theta}})^{(m)}$$

with $m = 0,1,2, \ldots$

8) Find the second derivative l_{REML} from point 5) to form the Hessian matrix to be used in the Newton Raphson algorithm

$$\frac{\sigma^2 l_{REML}}{\partial \sigma_i^2 \sigma_j^2} = \frac{1}{2} tr\left(\boldsymbol{P}\boldsymbol{Z}_j\boldsymbol{Z}_j'\boldsymbol{P}\boldsymbol{Z}_i\boldsymbol{Z}_i'\right) - \boldsymbol{y}'\boldsymbol{P}\boldsymbol{Z}_j\boldsymbol{Z}_j'\boldsymbol{P}\boldsymbol{Z}_i\boldsymbol{Z}_i'\boldsymbol{P}\boldsymbol{y}$$

Where$\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$

9) Iteration in point 7) will stop when $\left|\widehat{\boldsymbol{\theta}}^{(m)} - \widehat{\boldsymbol{\theta}}^{(m+1)}\right|$ it converges.

5. Estimating random effect parameters $\boldsymbol{\alpha}$ using the ML method with the following steps:$\boldsymbol{\alpha}$

1) The form of the likelihood function derived from the combined probability density function $\boldsymbol{y}$ and $\boldsymbol{\alpha}$

$$L(\boldsymbol{y}, \boldsymbol{\alpha}) = |\boldsymbol{R}|^{-\frac{1}{2}}|\boldsymbol{D}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha}) + \boldsymbol{\alpha}'\boldsymbol{D}^{-1}\boldsymbol{\alpha}]\}$$

2) The form of the log likelihood function

$$\begin{aligned} l(\boldsymbol{y}, \boldsymbol{\alpha}) &= \ln[L(\boldsymbol{y}, \boldsymbol{\alpha})] \\ &= |\boldsymbol{R}| + |\boldsymbol{D}| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha}) + \boldsymbol{\alpha}'\boldsymbol{D}^{-1}\boldsymbol{\alpha} \\ &= |\boldsymbol{R}| + |\boldsymbol{D}| + \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{\alpha} \\ &\quad + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\alpha}'\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{\alpha} \\ &\quad + \boldsymbol{\alpha}'\boldsymbol{D}^{-1}\boldsymbol{\alpha} \end{aligned}$$

3) Find the partial derivative of the log likelihood function with respect to the parameters $\boldsymbol{\alpha}$

$$\frac{\partial l(\boldsymbol{y}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{\alpha} - \boldsymbol{D}^{-1}\boldsymbol{\alpha}$$

4) Make the equation in point 3) equal to zero by using to show the solution$\widehat{\boldsymbol{\alpha}}$

$$\begin{aligned} \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{\alpha}} + \boldsymbol{D}^{-1}\widehat{\boldsymbol{\alpha}} &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1})\widehat{\boldsymbol{\alpha}} &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1})\widehat{\boldsymbol{\alpha}} &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta} \\ (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1})\widehat{\boldsymbol{\alpha}} &= \boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \widehat{\boldsymbol{\alpha}} &= (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{D}\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

Where $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{R}$

5) Replace $\boldsymbol{\beta}$ with the estimation results, namely $\widehat{\boldsymbol{\beta}}$ those obtained from (13) so that the estimation results (predictions) of the random effect are

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{D}\boldsymbol{Z}'^{\boldsymbol{V}^{-1}}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

Case Study and Results Analysis

The case study was conducted using Health Index data by district and city in West Java Province from 2016 to 2021. The data used is secondary data obtained from the West Java Health Service and the West Java Central Bureau of Statistics. The Health Index is one of the basic dimensions that builds the Human Development Index (IPM). HDI is an important indicator to measure success in efforts to build the quality of human life in a region and country. HDI is strategic data because a part used measure government performance, the level of the Health Index in an area can be seen from indicators of longevity and healthy life, which are calculated by looking at the AHH variable (life expectancy), with the formula:

$$Health\ Index = \frac{AHH - AHH_{min}}{AHH_{max} - AHH_{min}}$$

Data obtained from West Java data in figures for 2016 and 2021 are used as part of a longitudinal study of 27 regions in West Java. Regions are divided into two groups, namely City and Regency areas. The aim of the study was to assess the effect of regional and time category factors on the trajectory of the development of the health index in each region, and whether there were any differences between the two regions and the development of the index values over time in the 2016-2021 period.

Exploration was carried out on the subject profile and mean profile. The subject profile shows the observed subject values as a function of time for each subject at the regional level. While the mean profile shows the average by regional group. The subject profile results in the appearance of a variety of changes in response and the effect of changes in time on the response in each observed data and the mean profile results in the appearance of trends in the observed data. From these two explorations, considerations were obtained to determine whether the data tended to be linear or not, whether intercepts and slopes were treated as random effects or fixed effects, and whether there was an effect of changes in time on the response for construct model at a later stage.

Changes in the value of the Health index for each district and city in West Java can be seen in Figure 1, which shows that in general the index numbers have increased from time to time, both for quite significant increases and gradual increases. It can also be seen that the health index for the city area 3 is the highest among the 24 other urban district areas, as for the lowest health index in the 3 district areas.

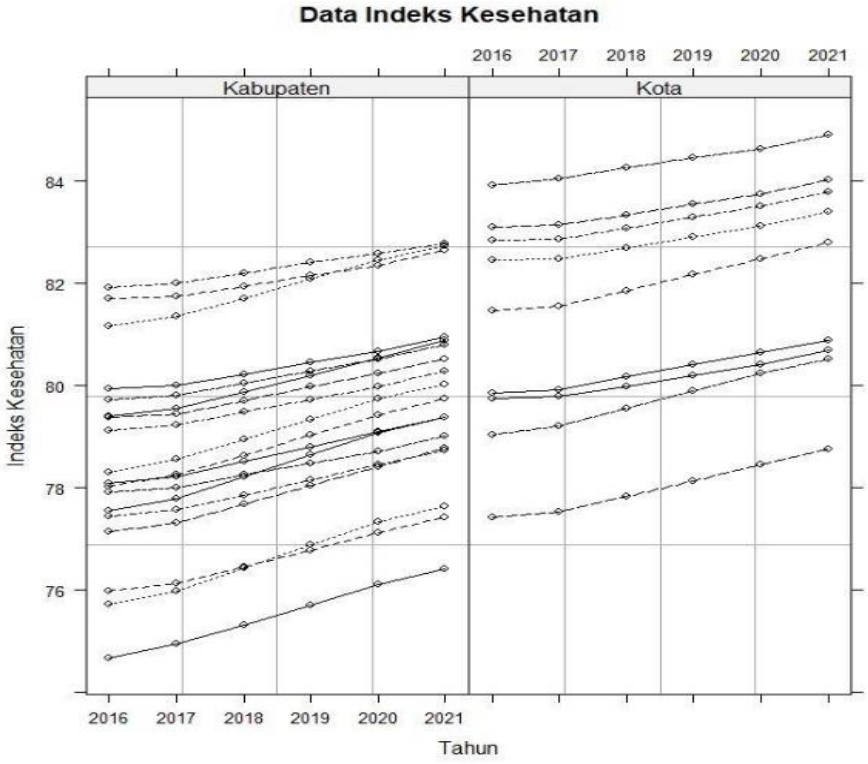

Figure 1. Trend of City District Health Index in West Java

The trend shown in Figure 1 shows that each study unit has a unique intercept and slope, therefore the intercept and slope or time slope will be treated as random effects.

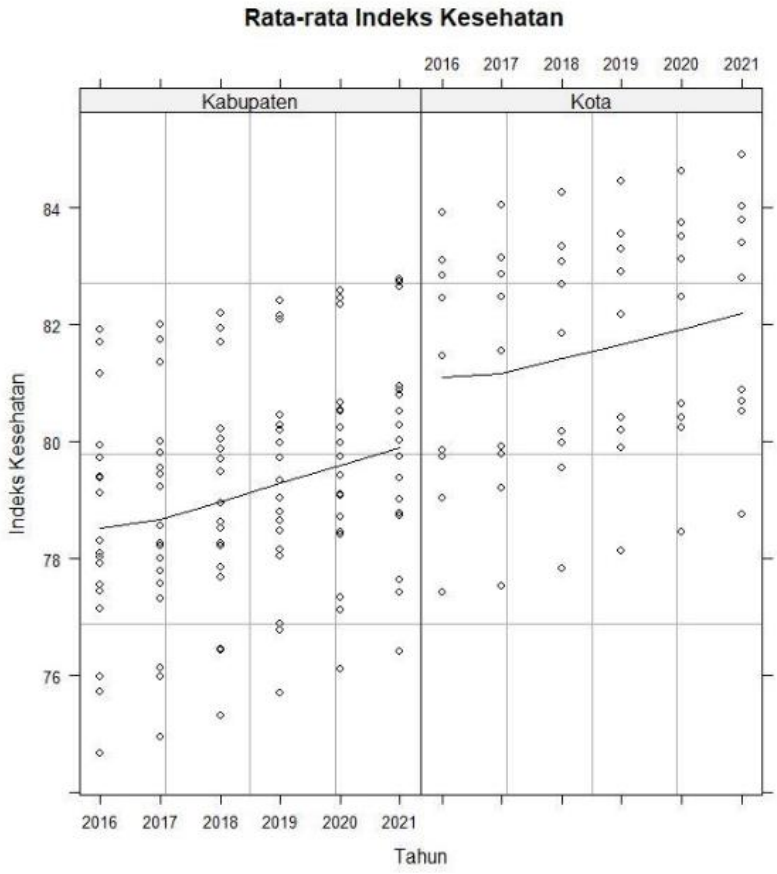

Figure 2. Average City Health Index in West Java

Mean profile shown in Figure 2 shows that the average value of the index generally increases with time, with a trend that tends to be linear for all urban districts. This suggests that the model for predicting the health index includes a linear fixed effect of the year variable and a possible interaction between the linear effect of time and region groups.

The model to be formed is a hierarchical model that has two components that reflect the contributions of two levels, namely time at the first level and cities at the second level. The first-level model denotes a subject-specific linear regression model for the health index based on time (Year) as the first factor. The intercept (β_{0i}) and the time slope (β_{1i}) vary between cities given index i. first-level model can be written as follows:

$$IK_{it} = \beta_{0i} + \beta_{1i} YEAR_{it} + \varepsilon_{it} \tag{4.1}$$

with $\varepsilon_{it} \sim N(0, \boldsymbol{R})$

In the second-level model, the unobserved subject-specific coefficients for the intercept and time slope in the first-level model varied. The intercept and slope of time at the level depend on the fixed effect associated with an additional predictor factor at level two namely REGION as the second factor and random subject effect.

$$\begin{aligned} \beta_{0i} &= \beta_{00} + \beta_{01} REGION_i + \alpha_{0i} \\ \beta_{1i} &= \beta_{10} + \beta_{11} REGION_i + \alpha_{1i} \end{aligned} \tag{4.2}$$

where $\boldsymbol{\alpha}_i = \begin{pmatrix} \alpha_{0i} \\ \alpha_{1i} \end{pmatrix} \sim N(0, \boldsymbol{D})$. The two-level model shows that the intercept(β_{0i}) for a i city depends on the overall intercept (β_{00}), the fixed effect (β_{01}) of the area predictor factors, and the random effect (α_{0i}) associated with the i city. The time slope(β_{1i}) depends on the overall time slope(β_{10}), the region fixed effect (β_{11}) , and the random effect (α_{1i}) associated with the *i* city. The random effect in

the second-order model allows for city-specific intercepts and the time slope effect to vary randomly between cities.

The overall model is by substituting equation (4.2) into equation (4.1). From the results of substitution of the second-level model to the first-level model, we will get the general form of a two-level mixed linear model for the i-th subject health index at and t-th time (t = 0, 1, 2, 3, 4, 5) at period 2016 - 2021.

$$IK_{it} = \beta_{00} + \beta_{01}REGION_i + \beta_{10}YEAR_{it} + \beta_{11}REGION_iYEAR_{it} + \alpha_{0i} + \alpha_{1i}REGION_{it} + \varepsilon_{it} \quad (4.3)$$

The parameters $\beta_{00}, \beta_{01}, \beta_{10}$ and β_{11} represent the fixed effects associated with the intercept, area variable, year variable, and interaction term in the model, respectively. The parameter β_{10} represents the fixed effect β_{11} of time for the reference category of the region group (region = 1). The fixed effect .β_{01} represents the difference in intercept for the first level of the region group by reference category. The fixed effect represents the difference in the year linear effect between the first level of region groups and the linear effect year within the area group reference category. The terms α_{0i} and α_{1i} in the first-level model represent the random effects associated with the intercept and the time linear effects for subjects-*i*. While the terms ε_{it} in the first-level model are residues related to observations on the subject-i at time-t

To complete the estimation of fixed effect parameters $\boldsymbol{\beta}$ in the model from the case study, the calculations were carried out in accordance to the previous steps, with calculations using the Rstudio software, the following results are obtained:

Using the help of software R, the estimated fixed effect parameters for $\boldsymbol{\beta} = (\beta_{00} \quad \beta_{10} \quad \beta_{01} \quad \beta_{11})'$ each residual covariance matrix structure $\boldsymbol{R}$ are obtained in Table 2 below:

Table 2. Fixed effect parameter estimation result

Residual Covariance Matrix Structure (R)	Parameter Type	Estimation Results
Diagonal	Fixed Effect Parameters	
	β00 (Intercept)	78,435
	β10 (YEAR)	0.287
	β01 (REGION)	2,566
	β11 (REGIONAL × YEAR)	-0.058

The random effect parameter $\boldsymbol{\alpha}_i = (\alpha_{0i} \quad \alpha_{1i})'$was also estimated or predicted from the random effect $\boldsymbol{\alpha}_i$ of the two-level mixed linear model as shown in Table 3.

Table 3. Random effect parameter estimation results

Region	ID	$\hat{\alpha}_{0i}$	$\hat{\alpha}_{1i}$	Region	ID	$\hat{\alpha}_{0i}$	$\hat{\alpha}_{1i}$
1	0	-0.616	-0.055	16	0	3,417	-0.104
2	0	-1,378	0.049	17	0	1,194	-0.061
3	0	-2,539	0.014	18	0	-0.942	0.092
4	0	3.155	-0.089	19	1	0.364	0.046
5	0	-0.425	-0.019	20	1	-1,239	-0.008
6	0	-3,804	0.072	21	1	1730	-0.032
7	0	-0.157	0.063	22	1	-1,368	-0.030

Region	ID	$\hat{\alpha}_{0i}$	$\hat{\alpha}_{1i}$	Region	ID	$\hat{\alpha}_{0i}$	$\hat{\alpha}_{1i}$
8	0	2,677	0.033	23	1	2,879	-0.034
9	0	0.595	-0.046	24	1	2,000	-0.037
10	0	-2,763	0.109	25	1	1,346	-0.032
11	0	1,404	-0.075	26	1	-2,034	0.079
12	0	-0.464	0.065	27	1	-3,678	0.049
13	0	0.882	0.016				
14	0	-1,077	-0.016				
15	0	0.842	-0.048				

A two-level mixed linear model is used to predict the value of the Health Index in West Java Province based on the district/city area factors and the time trend of index achievement. From the calculation results, a predictive model for the Health Index (IK) value for cities at time that depends on the random linear time effect is obtained as follows:it

1. Marginal Prediction Model

$$IK_{it} = 78.44 + (2.57 * REGION_i) + (0.29 * YEAR_{it}) \\ -(0.06 * REGION_i * YEAR_{it})$$

The marginal prediction model shows that in general the intercept will be the same for all cities at a certain REGION level, but the trajectory of each city will be different because of the random linear time effects that are included in the model. For districts and cities, a marginal prediction model can be obtained by making a dummy variable for the district to have a value of 0, and the city to have a value of 1, with each model as follows:

a. Marginal prediction model for district (area = 0)

$$\mathrm{IK_{it}} = 78.44 + (0.29{*}\mathrm{YEAR_{it}})$$

b. Marginal prediction model for city (area = 1)

$$\mathrm{IK_{it}} = 81.01 + (0.23{*}\mathrm{YEAR_{it}})$$

Prediction of the marginal health index values for districts and cities can be seen in Fig 3. below:

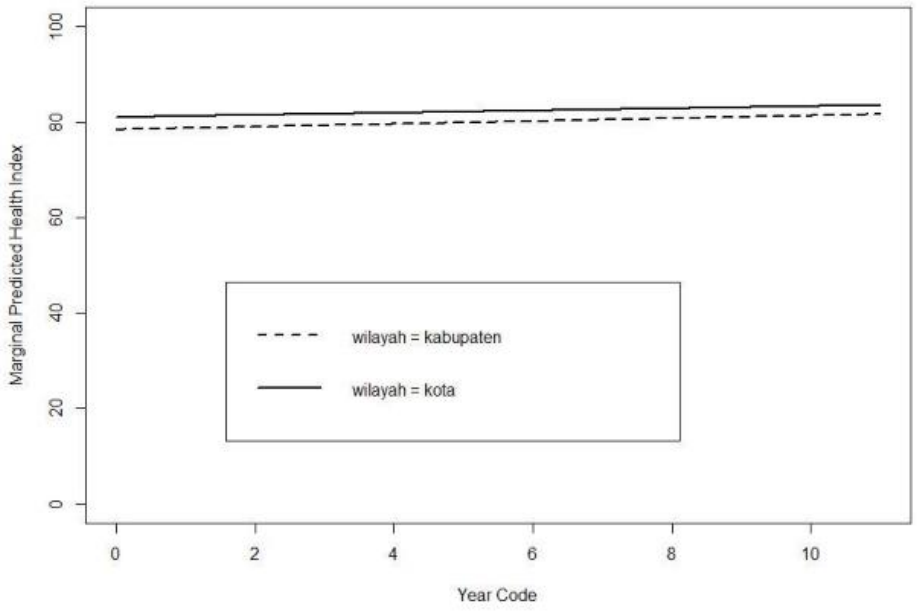

Figure 3. Prediction of Marginal Health Index for West Java Regencies and Cities

2. Conditional Prediction Model

$$IK_{it} = 78.44 + (2.57 * REGION_i) + (0.29 * YEAR_{it}) \\ -(0.06 * REGION_i * YEAR_{it}) + \hat{\alpha}_{0i} + (\hat{\alpha}_{1i} \times YEAR_{it})$$

In the average model or marginal model, the intercept and slope of each region are the same. But in the random effects model, the intercept and slope are different for each subject. Random effect α_{0i} and α_{1i}describes the difference between the intercept and slope values of the fixed values for the average intercept and slope. For districts and cities, a marginal prediction model can be obtained by making a dummy variable for the district to have a value of 0, and the city to have a value of 1, with each model as follows:

a. Marginal prediction model for district (area = 0)

$$IK_{it} = 78.44 + (0.29 + \alpha_{1i})YEAR_{it} + \alpha_{0i}$$

b. Marginal prediction model for city (area = 1)

$$IK_{it} = 81.01 + (0.23 + \alpha_{1i})YEAR_{it} + \alpha_{0i}$$

For the i-th city, the predicted $\hat{\alpha}_{0i}$ and $\hat{\alpha}_{1i}$ appropriate values can be seen in Table 2.

Figure 4 shows a graph of subject-specific conditional predictive values where there are additional changes originating from each subject for the Health Index values of 27 districts and cities in West Java.

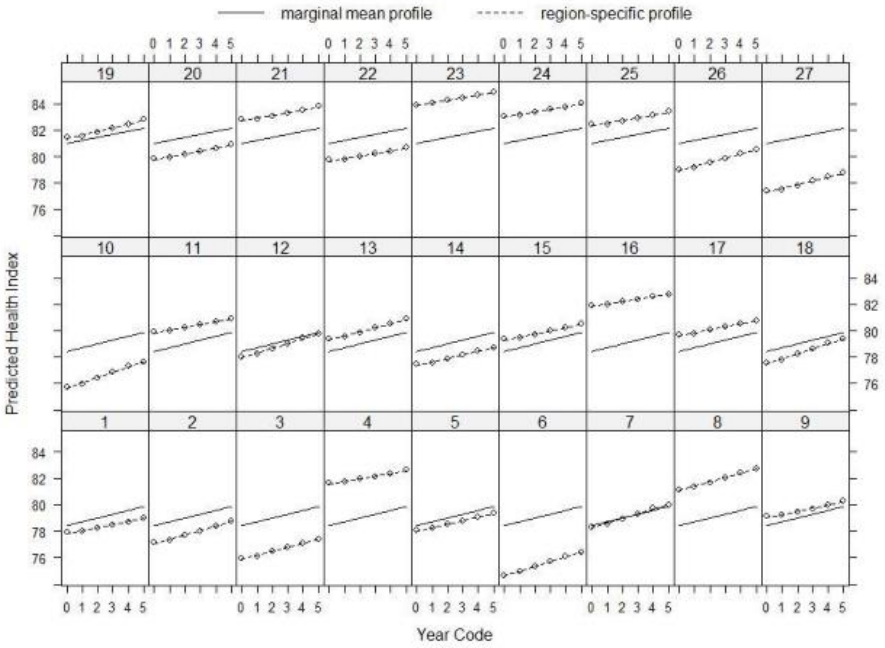

Figure 4. Specific Conditional Prediction of Health Index Regency and City of West Java

To test the fit of the model used the coefficient of determination (R^2) used to test the suitability of the model. The model fit test was carried out to determine the effect of the variables in the model on the response variable. The results of the calculation of the coefficient of determination (R^2) for each model are presented in Table 4.

Table 4. Match test value R^2

Covariance Matrix Structure (R)	Fixed Effects	Random Effects	Fixed Effects R2 (R_F^2)	Random Effects R2 (R_T^2)
Diagonal	Region+Year	Intercepts+Slopes	0.739	0.906

The coefficient of determination R_F^2 for models with fixed effects or marginal models provides information on the variance explained only by fixed effects, whereas the coefficient of determination R_T^2 for models with fixed effects and random effects or conditional models provides information on the variance explained by the entire model, that is, fixed effects and random effects. From the value of the coefficient of determination, the result shows that the value of the health index is explained by each of the fixed variables, namely region and year, which is 73.97%. While the fit test for the model with the overall effect $R_T^2 = 0.906$ means that the variation in the value of the Health Index in West Java is explained by the area variable and the year variable as a fixed effect and the area taking variable is considered as a random effect of 90.6%.

Conclusion

The two-level mixed linear model provides information with a level one describing the trajectory of each unit of research as a function of time and random error which is assumed to be normally distributed N(0,**R**), whereas a level two represents the group category parameters as functions of fixed effects and random errors also assuming to be distributed normally N (0**,D**). This study modeled random effects on level one using diagonal or independent covariance matrix structures.

This two-stage mixed linear model was implemented in the analysis of how regional and time differences influenced the value of the Health Index in West Java, using health index panel data from 27 regions in Western Java during the period 2016-2021, which is divided into two groups, namely city and district regions. The methods used to perform the estimation process are the Maximum Likelihood method and the Restricted Maximum Lieklihood method. The estimate of a model with a diagonal covariance matrix structure was obtained:

$$IK_{it} = 78.44 + (2.57 * REGION_i) + (0.29 * YEAR_{it}) - (0.06 * REGION_i * YEAR_{it}) + \hat{\alpha}_{0i} + (\hat{\alpha}_{1i} \times YEAR_{it})$$

Research can be developed with other covarince matrix structures with longer panel data and more other variables.

Reference

[1] E. W. Frees, "Longitudinal and Panel Data Analysis and Applications" *In The Social Sciences*. New York: Cambridge University Press, 2004.

[2] C. Angelini, "Regression analysis", Vol. 1–3. Elsevier Ltd., 2018.

[3] N. Pandis, "Cross-sectional studies,"*Am. J. Orthod. Dentofac. Orthops.*, vol. 146, no. 1, pp. 127–129, 2014.

[4] S. Searle, G. Casella, and C. McCulloh,"Variance Components", New Jersey: A John Wiley & Sons, Inc., 2006.

[5] N. H. Pusponegoro, R. N. Rachmawati, K. A. Notodiputro and B. Sartono, "Linear Mixed Model for Analyzing Longitudinal Data: A Simulation Study of Children Growth Differences", *Procedia Comput. sci.*, Vol. 116, pp. 284–291, 2017.

[6] S. Ross, "Introduction to Probability and Statistics for Engineers and Scientists", Sixth Edit. London: Katey Birtcher, 2021.
[7] H. Liu, Y. Zheng, and J. Shen, "Goodness-of-fit measures of R2 for repeated measures mixed effect models,"*J. Appl. Stats.*, Vol. 35, No. 10, pp. 1081–1092, 2008.
[8] A. C. Rencher and G. B. Schaalje, "Linear Models in Statistics", Second Edition, Vol. 84, No. 500, New Jersey: John Wiley & Sons, Inc., 2000.
[9] F. N. Gumedze and T. T. Dunne, "Parameter estimation and inference in the linear mixed model", *Linear Algebra Appl.*, Vol. 435, No. 8, pp. 1920–1944, 2011.
[10] X. Liu, "Methods and applications of longitudinal data analysis", Elsevier Inc., 2015.
[11] J. Hilbe and A. Robinson, "Methods of Statistical Model Estimation", Boca Raton: CRC Press, 2013.
[12] B. West, K. Welch, and A. Gałecki, "Linear Mixed Models A Practical Guide Using Statistical Software", Second Edition, vol. 53, no. 9, Boca Raton: CRC Press, 2015.
[13] B. West, K. Welch, and A. Gałecki, "Linear Mixed Models A Practical Guide Using Statistical Software", Second. Boca Raton: CRC Press, 2015.
[14] J. W. Wu, "The Quasi-Likelihood Estimation In Regression", *Ann. Inst. Statist. Math*., Vol. 48, No. 2, pp. 283-294, 1996.
[15] P. Widyaningsih, DRS. Saputro and A.N. Putri, "Fisher Scoring Method for Parameter Estimation of Geographically Weighted Ordinal Logistic Regression (GWOLR) Mode", *Journal of Physics: Conf. Series*, Vol. 855, No.1, p. 012060 2017.
[16] B. P. Carlin and T. A. Louis, "Bayes and Empirical Bayes Methods for Data Analysis, Second Edition", Boca Raton: CRC Press, 2000.
[17] N. Saputri, B. N. Ruchjana and E. S. Hasbullah, "Penerapan Model Regresi Data Panel pada Faktor Fundamental dan Teknikal Harga Saham Sektor Industri Real Estate", *Kubik: Jurnal Publikasi Ilmiah Matematika*, Vol. 5, No. 1, pp. 10-19, 2020.
[18] R. Rahmadeni, and N. Nurjannah, "Model Tingkat Kemiskinan di Kabupaten/Kota Provinsi Riau: Menggunakan Regresi Data Panel", *KUBIK: Jurnal Publikasi Ilmiah Matematika*, Vol.6, No. 2, pp. 98-109, 2021.
[19] D. F. Durrah, R. Cahyandari and A. S. Awalluddin, "Model regresi data panel terbaik untuk faktor penentu laba neto perusahaan asuransi umum Syariah di Indonesia", *KUBIK: Jurnal Publikasi Ilmiah Matematika*, Vol.5, No. 1, p. 27-34, 2020.