Analysis of Longitudinal Regression Model Using the Generalized Estimating Equation (GEE) for the Child Welfare Composite Index (CWCI) in West Java

Asep S. Awalluddin[1,*] Inge Wahyuni[2] Hilda Nurmuslimah[3]

[1,3] *Department of Mathematics, UIN Sunan Gunung Djati*
[2] *BP2D West Java Provincial Government*
[*] *Email: aasolih@uinsgd.ac.id*

ABSTRACT

The purpose of this study was to determine the Generalized Estimating Equation (GEE) analysis as one of the longitudinal data regression methods for composite index data on children's welfare (CWCI) in West Java Province. GEE analysis uses four different correlation structures to determine the best model, namely: independent, exchangeable, AR (1), and unstructured. The determination of the best model of the four structures uses the Quasi-likelihood Information Criterion (QIC). The West Java CWCI longitudinal data used consists of 5 independent variables, each of which is a dimension of survival, protection, growth and development, participation, and identity in 27 districts and cities in West Java for the period 2015 – 2017. This research is important as an alternative approach of longitudinal data regression analysis when we found a correlation problem in the condition of observed data.

Keywords: *Longitudinal data, GEE, Child welfare composite index, Estimation.*

1. INTRODUCTION

Longitudinal data is data obtained from observations made repeatedly on each subject of the same observation [1]. Longitudinal data analysis has been initiated in the research conducted by Balestra and Nerlove [2] and Hoch [3] who reconcile it as a combination of cross-sectional and time series data and some other researchers call it panel data. Several studies of longitudinal data analysis have been carried out including in the field of econometrics [4] and the social field [1], besides research results in the fields of education [5], health [6], social [7], industry [8] [9], epidemiology [10], tourism [11], and other fields

Longitudinal data regression analysis is used to see the effect of the independent variable / covariate on the response variable by paying attention to changes in subject data and time of observation. Measurement of data subjects repeatedly will cause a correlation between the outcome of the observation, so that the classical regression method cannot be done. The most commonly used longitudinal regression analysis methods include general effects models, fixed effects models, and random effects models. Another method that can be used for data that is not normally distributed is the Generalized Linear Model (GLM). Another method developed is Generalized Estimating Equation (GEE). GEE is an extension of the GLM model to solve the problem of correlation, as a statistical approach that is suitable for marginal models on longitudinal data. GEE analysis is used in various fields including biomedical studies [12] [13], social education [14], economics [15], and other fields.

This paper will explain the longitudinal data regression analysis using the GEE model to see a model of the variables that affect the composite index of child welfare (CWCI) in West Java. CWCI is a tool to measure the level of welfare of Indonesian children, both nationally and province, by using a reference to the fulfillment of children's rights. The composite index has an interval of 0 to 100 which shows the value of the level of child welfare. The greater the index value, the higher the level of child welfare. With the Indonesian Child Rights Convention (KHA) used in the CWCI measurement, the fulfillment of children's rights is formulated in five variable dimensions which are the basis for the formulation of the CWCI measurement framework, including the right to obtain: (1) survival, (2) protection, (3) development, (4) participation, and (5) identity [16]. The longitudinal data used are CWCI data for the period 2015 - 2017.

2. GENERALIZED ESTIMATING EQUATION (GEE)

2.1. Assumptions and Components of the GEE

The Generalized Estimating Equation (GEE) was introduced by Liang and Zeger in several writings [17] [18], this method is an extension of the Generalized Linear Model (GLM) and can analyze correlated data both discrete and continuous [19]. The GEE model in this study was used to analyze longitudinal data whose responses were correlated. The parameter estimation method that is often used is the maximum likelihood method. However, not all exponential family distributions have a reference to the likelihood function, when there is none, it can be done using the quasi-score method by determining the mean form as the first moment and the variance-covariance matrix as the second moment. The quasi-score is nothing but a model of the GEE estimation equation. So that the parameter estimation in the GEE is not by using the maximum likelihood method but by using the quasi-likelihood.

The GEE model models a known function of the marginal expectation of a response variable which is a linear function of one or more independent variables. The marginal expectation $E(Y_{ij})$ is modeled as a function of the independent variable. The marginal expectation is the average response of a subpopulation that has the same explanatory variables with the assumptions:

- $E(Y_{ij}|X_{ij}) = \mu_{ij}$depending on the vector of the independent variable x_{ij} through a relationship $g(\mu_{ij}) = x_{ij}\beta$. $g(.)$ is the known link function according to its distribution.
- The marginal variance of each Y_{ij}depends on the average, so the equation can be written as $Var(Y_{ij}) = \phi v(\mu_{ij})$, where $v(.)$ Is the known variance function and ϕ is the scale of parameters you may need to estimate.
- The correlation between $y_{ij}$and y_{ik} is a function of the marginal mean and possibly an additional parameter of $corr(y_{ij}, y_{ik}) = \psi(\mu_{ij}; \mu_{ik}; \alpha$ where ψ (.) Is a known function.

With the marginal model, longitudinal data only need to determine the form of the mean (first moment) and the variance-covariance matrix (second moment), namely $E(Y_{ij}) = \mu_{ij}$ and $Var(Y_{ij}) = \sigma_{ij}^2$, respectively. For a normal distribution, these two moments are sufficient to form the likelihood function, but they cannot be ascertained for the distribution. Some of the assumptions in the GEE are: correlated response variables, do not have to meet the assumption of homogeneity on the variance, multicollinearity does not occur in the independent variables, the correlation structure needs to be determined, parameter estimation uses quasi-likelihood.

The components in the GEE model which are an extension of GLM are as follows:

- Random component: the response variable (y_{ij}) satisfies the exponential family distribution where *i* denotes the order of the *i*th subjects, *i* = 1,2,…, N and j denotes the number of repetitions of the observations, *j* = 1,2,…, n_t.
- Fixed component or linear predictor
$$\eta_{ij} = x_{ij}^T\beta$$
- The link function is a transformation or liaison function on the dependent variable that connects the average response variable μ_ij with its linear predictor.
$$g(\mu_{ij}) = \eta_{ij}$$

GEE has advantages over other methods of analyzing longitudinal data, including GEE using quasi-likelihood estimation so that the complete likelihood function of the data does not have to be defined. In addition, the GEE model is formed in a marginal model so that the average response only depends on the first and second moments, namely the mean and variance. Another thing, the data analyzed using GEE can be in discrete or continuous form. In the traditional longitudinal regression method, correlations to the data are ignored whereas in the GEE model, the correlation can be calculated. The basis of the GEE model is that the common distribution of the response variable $y_i$for each subject does not have to be determined, only the marginal distribution of y_{ij} must be determined. For example, there are two repeated observations with a continuous normal response. Assume that the y_{i1} and y_{i2} distributions are two univariate normal data so that GEE can avoid the use of a more complicated multivariate distribution [19].

2.2. Estimation of the GEE Model Equation

Suppose y_{ij} is the response variable for *n* the number of objects of observation with the time that the observation is made *T* times for each object *i*th, *i* = 1,2, ... *n*. For each y_{ij} there is a covariate $\boldsymbol{x}_{it}$ of size *p*. Data can be expressed as a response vector $\boldsymbol{y}_i$ and a matrix $\boldsymbol{X}_i$ with the assumption that the data pair $(\boldsymbol{y}_i , \boldsymbol{X}_i)$ iid with $E(y_{ij}| \boldsymbol{X}_i) = E(y_{ij}| \boldsymbol{x}_{it}) = \boldsymbol{x}^T_{it} \boldsymbol{\beta}$

The GEE U (β) equation can be as follows [20]:

$$U(\beta) = \left[\left\{\left(\sum_{i=1}^{N} x_{mi}^{T}\, D\left(\frac{\partial\mu}{\partial\eta}\right)[V(\mu_i)]^{-1}\left(\frac{y_i-\mu_i}{a(\phi)}\right)\right)\right\}_{m=1,\dots,p}\right]_{(px1)} = [0]_{(px1)}$$

D (.) Is the matrix derived from the mean of the parameter, a (ϕ) a function of ϕ for overdispersed data and V ((μ_i) is a diagonal matrix derived from the form:

$$V(\mu_i) = \left[D(V(\mu_{ij}))^{1/2} I_{(n_t x n_t)} D(V(\mu_{ij}))^{1/2}\right]_{n_t x n_t}$$

on longitudinal data with an independent correlation structure. In general, for other forms of correlation structure the matrix $V(\mu_i)$ can be written in the following form:

$$V(\mu_i) = \left[D\left(V(\mu_{ij})\right)^{\frac{1}{2}} R(\alpha)_{(n_t x n_t)} D\left(V(\mu_{ij})\right)^{\frac{1}{2}}\right]_{n_t x n_t}$$

In simple terms it can be written as follows:

$$V(\hat{\alpha}) = A_i^{1/2} R_i(\alpha) A_i^{1/2} \text{ or } V(\hat{\alpha}) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

for certain distributions that are overdispersed with $A_i = D(V(\mu_{ij}))$

R (α) is the correlation structure matrix which will be estimated using the parameter vector α. other than that,

$$x_{mi}{}^{T} D\left(\frac{\partial\mu}{\partial\eta}\right) = x_{mi}{}^{T} D\left(\frac{\partial\mu}{\partial x_{mi}{}^{T}\beta}\right) = x_{mi}{}^{T}\frac{1}{x_{mi}{}^{T}} D\left(\frac{\partial\mu}{\partial\beta}\right)$$

$$= D\left(\frac{\partial\mu}{\partial\beta}\right)$$

In simple terms, the GEE model estimation equation can be written as follows:

$$U(\hat{\beta}) = \sum_{i=1}^{N} D_i{}^{T}[V(\hat{\alpha})]^{-1}(y_i - \mu_i) = 0 \qquad (1)$$

2.3 Variance-Covariance Matrix in the GEE Model

The variance-covariance estimator matrix in the GEE model, namely Naïve estimator and robust estimator, is used to test which variables are significant in the model and will also be used to test the correlation structure. The form of naïve and robust estimators is shown in equations (2) and (3):

$$V(\hat{\beta}) = \left[\sum_{i=1}^{N} D_i{}^{T}(\hat{V}_i^{-1}) D_i\right]^{-1} \qquad (2)$$

for data with normal distribution, $D_i = X_i$, so

$$V(\hat{\beta}) = \left[\sum_{i=1}^{N} X_i{}^{T}(\hat{V}_i^{-1}) X_i\right]^{-1}$$

Naïve estimator is used when research with small sample data (N <20).

$$V(\hat{\beta}) = \sum_{i=1}^{N} M_0^{-1} M_1 M_0^{-1} \qquad (3)$$

with

$$M_o = \left[\sum_{i=1}^{N} D_i{}^{T}(\hat{V}_i^{-1}) D_i\right]$$

$$M_1 = \left[\sum_{i=1}^{N} D_i{}^{T}(\hat{V}_i^{-1})(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^{T} D_i\right]$$

robust estimators are suitable for research with large amounts of sample data.

2.4 Correlation Structure of the GEE Model

The correlation structure is denoted by ***R_i(α)***, where α is the average parameter of dependence between repeated observations on the subject. Whereas αˆ is the estimated correlation of the α parameter which will be calculated for each correlation structure. Before testing the correlation structure, the residual error calculation is determined first with equation (4):

$$\hat{e}_{ij} = \frac{(y_{ij} - \mu_{ij})}{\sqrt{\frac{v(\mu_{ij})}{\phi}}} \qquad (4)$$

The value of $\hat{e}_{ij}$ will be used to obtain estimates of $\hat{\alpha}$ and ϕ needed to calculate the value of the correlation structure. If the data is overdispersed, then the dispersion parameter ϕ must be calculated by equation (5):

$$\phi = \frac{1}{K-p}\sum_{i=1}^{N}\sum_{j=1}^{n_t} e_{ij}^{2} \qquad (5)$$

where $K = \sum_{i=1}^{N} n_i$ and p is the number of covariates. Several types of correlation structures that will be used in this study are:

2.4.1. Independent with $R_i(\alpha) = I$

This structure assumes that the response data over time is uncorrelated. The form of the independent correlation matrix is:

$$Corr\left(y_{ij}, y_{i,jN}\right) = \begin{cases} 1, j = N \\ 0, j \neq N \end{cases}, \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

The α value does not need to be calculated for this structure.

2.4.2. Exchangeable with $R_i(\alpha) = \rho$

This structure assumes that the correlation of each subject's repetitions is always the same. The form of the correlation matrix in the exchangeable structure is:

$$Corr\left(y_{ij}, y_{iN}\right) = \begin{cases} 1, j = N \\ \rho, j \neq N \end{cases}, \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

The estimators for this structure are:

$\hat{\alpha} = \frac{1}{(K_1 - p)\phi} \sum_{i=1}^{N} \sum_{j \neq N} e_{ij} e_{ik}$, with

$$K_1 = \sum_{i=1}^{N} n_t (n_t - 1)$$

2.4.3. Autoregressive (AR(1)) with $R_i(\alpha) = \rho^{|j-j^i|}$

In this structure, the magnitude of the (positive) correlation decreases rapidly with each iteration time for each response. The form of the correlation matrix in this structure is:

$Corr\left(y_{ij}, y_{i+N}\right) = \rho^N$, dengan $N = 0,1,2 \dots, n_i - j$

$$Corr\left(y_{ij}, y_{i+N}\right) = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \dots & 1 \end{bmatrix}$$

The estimators for this structure are:

$\hat{\alpha} = \frac{1}{(K_2 - p)\phi} \sum_{i=1}^{N} \sum_{j \leq n_t - 1} e_{ij} \, e_{i(j+1)}$, with

$K_2 = \sum_{i=1}^{N} (n_t - 1)$.

2.4.4. Unstructured

This structure does not have a special formula because it is assumed that each time of observation, the correlation of each subject will vary arbitrarily regardless of whether the difference is getting smaller or bigger. The form of the correlation matrix in this structure is:

$$Corr\left(y_{ij}, y_{iN}\right) = \begin{cases} 1, \; j = N \\ \rho_{jN}, j \neq N \end{cases},$$

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \rho_{N1} & \rho_{N2} & \rho_{N3} & \cdots & 1 \end{bmatrix}$$

The estimators for this structure are:

$$\hat{\alpha}_{jN} = \frac{1}{(K - p)\phi} \sum_{i=1}^{N} e_{ij} e_{iN}$$

The correlation structure that is more than one means that tests must be carried out to find out which correlation structure is most suitable for use in research. The number of coefficients estimated is K (K-1) / 2 where K is the amount of time observed. Methods for testing the correlation structure are AIC and QIC. AIC stands for Akaike's Information Criterion is a correlation structure test when the parameter estimation method uses the maximum likelihood method, while for the correlation structure test when the parameter estimation method uses quasi-likelihood is the Quasi-likelihood Information Criaterion (QIC).

2.5. GEE Model Parameter Estimation

Estimation of parameters in the GEE model is to solve equation (1), by first knowing the shape of the longitudinal data distribution to be analyzed. Furthermore, the data distribution model is formed into the exponential family distribution function and looks for its mean and variance. For example, longitudinal data which is assumed to be normally distributed, has the form of an exponential distribution function as follows:

$$f(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right]$$

$$= \exp\left\{ln\left((2\pi\sigma^2)^{-\frac{1}{2}}\right) - \left[\frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right]\right\}$$

$$= \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + ln\left((2\pi\sigma^2)^{-\frac{1}{2}}\right) - \frac{y^2}{2\sigma^2}\right\}$$

It is obtained that $\theta = \mu$, $b(\theta) = \frac{1}{2}\mu^2$ and $a(\phi) = \sigma^2$. For a normally distributed response variable with mean = 0 and variance = 1 (no overdispersion occurs), the resulting variance is:

$$V(\mu_{it}) = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & 1 & 0 \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix}$$

The variance in the GEE model for the normal distribution is produced as follows:

$$V(\hat{\alpha}) = \phi A_i^{1/2} R_i(\hat{\alpha}) A_i^{1/2}$$

$$= \phi \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & 1 & 0 \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix} R_i(\alpha) \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & 1 & 0 \\ 0 & 0 & \dots & \dots & 1 \end{bmatrix}$$

$$= \phi \, R_i(\hat{\alpha})$$

So that the GEE model equation for longitudinal data that is normally distributed is:

$$\sum_{i=1}^{N} X_i^T [\phi \, R_i(\hat{\alpha})]^{-1} (y_i - X_i\beta) = 0 \qquad (6)$$

can then be broken down into:

$$\sum_{i=1}^{N} X_i^T [\phi \, R_i(\hat{\alpha})]^{-1} y_i - \sum_{i=1}^{N} X_i^T [\phi \, R_i(\hat{\alpha})]^{-1} X_i\beta = 0$$

$$\sum_{i=1}^{N} X_i^T [\phi \, R_i(\hat{\alpha})]^{-1} X_i\beta = \sum_{i=1}^{N} X_i^T [\phi \, R_i(\hat{\alpha})]^{-1} y_i$$

$$\hat{\beta} = \left[\sum_{i=1}^{N} X_i^T [\phi \, R_i(\hat{\alpha})]^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i^T [\phi \, R_i(\hat{\alpha})]^{-1} y_i\right]$$

From the example of the normal distribution model above, it can be concluded that in general the parameter β can be estimated by calculating the value:

$$\hat{\beta} = \left[\sum_{i=1}^{N} D_i^T [V(\hat{\alpha})]^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{N} D_i^T [V(\hat{\alpha})]^{-1} y_i\right]$$

Estimation of the $\hat{\beta}$ parameter in the GEE model is carried out numerically to produce a convergent estimate.

3. GOODNESS OF FIT MODELS

The suitability test that will be carried out is the correlation structure test and hypothesis testing. The correlation structure test will be calculated using the Quasi-likelihood Information Criterion (QIC), while the hypothesis test will be calculated using the Wald test. QIC is used to test the correlation structure in the GEE models. Some of the correlation structures used in this study are independent, exchangeable, and AR (1). the QIC equation, namely:

$$QIC = -2Q\left(\hat{\beta}, R(\hat{\alpha})\right) + 2trace\left[\left(V^{-1}V(\hat{\alpha})\ \right)\right]$$

with:

$Q\left(\hat{\beta}, R(\hat{\alpha})\right)$: quasi-likelihood function with β^estimator with selected correlation structure.

V^{-1} : inverse of the diagonal matrix of size *NxN* of the variance function of the dependent variable

$V(\hat{\alpha}))$: variance estimation function of the selected correlation structure matrix

The model with the smallest QIC value is the model with the best correlation structure hypothesis testing on longitudinal data can be done simultaneously or partially using the Wald test. This test is conducted to see the significance of the independent variable on the dependent variable. The Wald test equation is:

$Z_i = \left(\frac{\hat{\beta}_{ip}}{SE(\hat{\beta}_{ip})}\right)$ with $SE\left(\hat{\beta}_{ip}\right) = \sqrt{V(\hat{\beta}_{ip})}$ with statistics test $Z_i = \left(\frac{\hat{\beta}_{ip}}{SE(\hat{\beta}_{ip})}\right)^2 \sim Z_{tabel}$ and reject test criteria H_0 if Z_i is outside the interval Z_{tabel} or P-value < α

4. COMPOSITE INDEX OF CHILD WELFARE IN WEST JAVA

CWCI is a measure of the achievement of the level of child welfare that can show the level of achievement of fulfilling children's rights. Based on the Convention on the Rights of the Child (KHA) based on Presidential Decree No. 36 of 1990, contained in Article 5 of Law No. 35 of 2014 concerning child protection. Children's rights that must be fulfilled include: the right to survival, the right to protection, the right to development (development), the right to participate (participation), and the right to identity. Furthermore, the five rights are used as a dimension which consists of indicator variables used in determining the CWCI [16].

The exploration of CWCI data in West Java for 27 districts and cities for the 2015 - 2017 period was carried out as an implementation of the model that was built in the first phase [16] [21] [22]. Data mining in the form of primary data and secondary data. Primary data that will be explored to each district and city is needed to complete data that is not yet available at the West Java Central Statistics Agency (BPS). There are 5 dimensions of measurement with 11 indicators consisting of: under-five mortality rate, prevalence of basic and complete immunization, APS PAUD,% traveling,% having a birth certificate (for groups of 0-4 years old). Morbidity rate, prevalence of having been married, prevalence of child labor, APS 5-17 years,% traveling,% having a certificate (for the group of children aged 5-17 years). Table 1 shows indicators for the formation of CWCI at the district and city levels, including the method for determining the index [22].

Tabel 1. Indicators for the Formation of CWCI at the District and City Level [22]

Dimensions	Need and Risk Stages	
	Toddler (0-4 years)	Children (5-17 years)
Survival Rate	IMR	Morbidity
Protection	Prevalence of basic and complete immunization	Prevalence of ever married Prevalence of child labor
Growth and development	APS PAUD	APS 5 – 17 years
Participation	% traveling	% traveling

Identity	% have a birth certificate	% have a birth certificate

5. CASE STUDIES AND DATA ANALYSIS

The case study used is data on CWCI 27 districts and cities in West Java in 2015 - 2017. The independent variables as covariates are 5 dimensions, namely: survival (X1), protection (X2), growth and development (X3), participation (X4) and Identity (X5) with the response variable is CWCI (Y) [16] [21] [22].

5.1 Test assumptions

The assumption tests carried out are the correlation test, multicollinearity test, normality test, and heteroscedasticity test. The results obtained are as follows:

5.1.1. Correlation Test

The GEE model is effective if the response variable in the analyzed data is correlated for each observation time. Therefore, it is necessary to prove that the CWCI data for 2015-2017 are correlated. Using EViews the output is as shown in Table 2 below:

Table 2. Response Variable Correlation Value

	2015	2016	2017
2015	1	0.989662	0.985012
2016	0.989662	1	0.998991
2017	0.985012	0.998991	1

The results in Table 2 show that the CWCI response variable data correlates at each observation time (2015-2017) with a correlation value almost close to 1.

5.1.2. Multicollinearity Test

The multicollinearity test that may appear on the free variable covariates from the longitudinal data needs to be checked whether there is multicollinearity between the independent variables. The test results are shown in Table 3.

Table 3. Multikolinearitas Test Results

	X_1	X_2	X_3	X_4	X_5
X_1	1	-0,247885	0,16217	0,00404	0,00866
X_2		1	0,46675	0,36165	0,54127
X_3			1	0,32207	0,62335
X_4				1	0,37478
X_5					1

from the test results, it can be seen that multicollinearity does not occur between the independent variables, this is indicated that all values are less than 0.8.

5.1.3. Normality test

The normality test is carried out on the CWCI data, with the test results shown in Table 4 which is the output of Eviews. It can be seen that the Jarque-Bera value is 0.577410 with a p-value of 0.749233, meaning that the hypothesis that the data is normally distributed can be accepted.

Table 4. Normality Test Results

	value
jarque-bera	0,577410
probabiliy	0,749233

so it can be assumed that the response variable is included in the distribution of the Gaussian family and its link function is identity with (μ_i) = η_ (i) = μ_i. So that the general form of the GEE equation model for CWCI data is as follows:

$$\mu_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \varepsilon_{ij}$$

5.1.4. Heteroscedasticity Test

Furthermore, it will be tested whether the CWCI data has heteroscedasticity. By using the Glejser method on EViews, the output is as shown in Table 5. Based on the results in the table above, it can be seen that all variables have Prob <0.05, except X_5, so it can be concluded that the data has heteroscedasticity problems. This can be done by using the GEE model.

Table 5. Heteroscedasticity test results using the Glejser method

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
Intersep	0,13941	0,140208	0,994331	0,3233
X_1	-0,00085	0,001401	-0,60907	0,5443
X_2	0,00203	0,001081	1,88115	0,0638
X_3	0,00374	0,002050	1,82539	0,0719
X_4	0,00028	0,000367	0,77165	0,4427
X_5	-0,00501	0,000782	-6,40938	0,0000

The rule of the Glejser method is that if Prob <α then the data experiences heteroscedasticity. The next step is to estimate the parameters on the CWCI data and the independent variables using the GEE model using four different types of correlation structures. Estimating parameters is assisted by using R software.

5.2 Estimation of model parameters

The correlation structures used in the GEE model are: independent, exchangeable, AR (1) and unstructured. The response variable is CWCI normally distributed, so the equation for estimating the GEE model parameters that are normally distributed is in equation (2.6). Calculating the parameter values for each model was done using R software. Table 6 shows the estimation results of the four correlation structures used. It can be seen that the estimation results are not significantly different for each of the regression parameters for the independent variables. The difference is seen in the estimation results of the intercept parameters. The standard error (SE) values for the Naive estimator and robust are all significant.

Table 6. Estimation Parameter Results

coef	Independent			Exchangeable		
	Estimation	naive SE	robust SE	Estimation	naive SE	robust SE
Intercept	-0,0068	0,0395	0,0132	-0,0106	0,0360	0,0190
X_1	0,2000	0,0003	0,0002	0,2001	0,0003	0,0002
X_2	0,1995	0,0003	0,0002	0,1995	0,0002	0,0002
X_3	0,1994	0,0005	0,0003	0,1994	0,0004	0,0003
X_4	0,1999	0,0001	0,0000	0,1999	0,0000	0,0000
X_5	0,2008	0,0002	0,0003	0,2008	0,0001	0,0003

coef	AR (1)			Unstructured		
	Estimation	naive SE	robust SE	Estimation	naive SE	robust SE
Intercept	-0,0258	0,0310	0,0022	-0,0300	0,0306	0,0221
X_1	0,2002	0,0003	0,0002	0,2002	0,0003	0,0002
X_2	0,1994	0,0002	0,0002	0,1994	0,0002	0,0002
X_3	0,1992	0,0003	0,0004	0,1992	0,0004	0,0004
X_4	0,1998	0,0006	0,0000	0,1998	0,0000	0,0000
X_5	0,2010	0,0001	0,0004	0,2011	0,2011	0,0004

The rule of the Glejser method is that if Prob <α then the data experiences heteroscedasticity. The next step is to estimate the parameters on the CWCI data and the independent variables using the GEE model using four different types of correlation structures. Estimating parameters is assisted by using R software.

5.3 Goodness of Fit Models

The suitability test that will be carried out is the correlation structure test and hypothesis testing. correlation structure test using QIC in software R is generated respectively QIC_ind = 9.0344, QIC_exc = 9.2024, QIC_ (AR (1)) = 14.5439. From the resulting QIC value, it can be concluded that the best correlation structure is a model with an independent correlation structure. Meanwhile, to test the suitability hypothesis, namely by using the Wald test, with the R software produced in Table 7.

By looking at the Wald value with the p-value, it can be seen that the p-value is <0.005 which indicates that H0 is rejected, which means that all independent variables (X1, X2, X3, X4, X5) are significant to the model. Only the intercept is not significant because the p-value is> 0.005, so it must be removed from the model, so that the estimation of model parameters used in general does not differ significantly for all correlation structures.

Table 7. Wald Test Result

	Wald	*p-value*
Intersep	0.27	0.61
X_1	8.73×10^5	< 2e-16
X_2	9.49×10^5	< 2e-16
X_3	3.87×10^5	< 2e-16
X_4	1.1×10^7	< 2e-16
X_5	2.72×10^5	< 2e-16

6. CONCLUSION

In this study, a research was carried out on the estimation steps of the longitudinal regression model parameters with four types of correlation structures using the Generalized Estimating Equation (GEE) method. Determination of the best model using the QIC and Wald test. The application of the model was carried out on the Composite Child Welfare Index (CWCI) data as a response variable and dimensions of survival, protection, growth and development, participation and identity as independent variables for 2015-2017 in 27 districts / cities in West Java Province.

The results of the case study analysis show that CWCI data is good for analysis using the GEE model because it takes into account the correlation that appears in the response variable due to repetition in data collection. The GEE model can also be used for data experiencing heteroscedasticity such as CWCI data. The best GEE models in general do not differ significantly for the four types of correlation structures, with the intercept parameter fit test being insignificant.

ACKNOWLEDGMENTS

This research was funded by the Ministry of Religion's BOPTN Litapdimas fund in 2019. We would like to thank the Ministry of Religion for funding support through the Research Center for LP2M UIN Sunan Gunung Djati Bandung.

REFERENCES

[1] E.W. Frees, Longitudinal and Panel Data Analysis and Applications in the Social Science, Cambridge University Press, USA, 2004.

[2] M. Nerlove, P. Balestra Pooling Cross Section and Time Series Data in the Estimation of Dynamic Model : The Demand for Natural Gas, vol 34, Ecometrica, 1966, pp. 585-612. DOI : https://doi.org/0012-682(196607)34:3<585:PCSATS>2.0.CO;2-F

[3] I. Hoch, Estimation of Production Function Parameters Combining Time Series and Cross Section Data, vol 30, Econometrica, 1962, pp. 34-53. DOI : https://doi.org/0012-9682(196201)30:1<34:EOPFPC>2.0.CO;2-8

[4] B.H. Baltagi, Econometric Analysis of Panel Data, West Sussex, John Wiley & Sons Inc, 2005.

[5] G. T. Sav, Panel Data Estimates of Public Higher Education Scale and Scope Economies, vol.39, Atl Econ J, 2011, pp. 143-153. DOI:10.1007/S11293-011-9273-3

[6] J. Novignon, S.A. Olakojo, J. Nonvignon, The effects of public and private health care expenditure oh health status in sub-Saharan Africa : new evidence from panel data analysis, vol 2, Health Economics Review Springer**,** 2012, pp. 2-8. DOI:10.1186/2191-1991-2-22

[7] M.S. Ucal, Panel Data Analysis of Foregin Direct Investment and Povert from the Perspective of Developing Countries, vol 109 Procedia-Social and Behavioral Sciences, 2014, pp. 1101-1105. DOI : https://doi.org/10.1016/j.sbspro.2013.12.594

[8] J.I. Park and S.Lee, Examining the spatial patterns of green industries and the role of green industries and the role of goverment policies in south korea : Application of a panel regression model (2006-2012), vol 78, Renewable and Suistainable Energy Reviews, 2017, pp. 614 - 623. DOI : https://doi.org/10.1016/j.rser.2017.04.061

[9] A. Fitrianto, N.F. K. Musakkal, Panel Data Analysis for Sabah Construction Industries : Choosing the Best Model, vol 35, Procedia Economics and Finance, 2016, pp. 241 – 248. DOI : https://doi.org/10.1016/S2212-5671(16)00030-7

[10] J.W.R. Twisk, Applied Longitudinal Data Analysis for Epidemiology, Cambridge University Press, 2003.

11] P.K. Narayan, S. Narayan, A. Prasad, Tourism and economic growth : a panel data analysis for pasific island countries tour, vol 16, Econ, 2010, pp. 169-189. DOI : https://doi.org/10.5367/000000010790872006

[12] G.M. Fitzmaurice, N.M. Larid, J.H. Ware, Applied Longitudinal Data, John Wiley & Sons, 2004.

[13] J.W. Hardin, J.M. Hilbe, Generalized Estimating Equations, Chapman and Hall / CRC Press, Boca Raton, Fla, USA, 2003.

[14] P. Ghistela, D. Spini, An Introduction to Generalized Estimating Equations and an Application to Assess Selectivity Effects in a Longitudinal Study on Very Old Individuals, vol 29, Journal of Educational and Behavioral Statistics, 2004, pp. 421-437. DOI : https://doi.org/10.3102/10769986029004421

[15] S. Hudecova, M. Pesta, Modeling dependencies in claim reserving with GEE, vol 53, Insurance : Mathematics and Economics, 201, pp. 786 - 794. DOI : https://doi.org/10.1016/j.insmatheco.2013.09.018

[16] A. Ahnaf, W. Imawan, A. Hernawa, H.B. Surbekti, Indeks Komposit Kesejahteraan Anak (IKKA) 2016., Kementrian Pemberdayaan Pemberdayaan Perempuan dan Perlindungan Anak Indonesia, 2016.

[17] S.L. Zeger, K.Y. Liang, Longitudinal Data Analysis for Discrete and Continuous Outcomes, vol 42, Biometric, 1986, pp. 121-130. DOI : https://doi.org/10.2307/2531248

[18] S.L. Zeger, K.Y. Liang, P.S. Albert, Models for Longitudinal Data : A Generalized Estimating Equation Approach, vol 4, Biometrics, 1988, pp. 1049-1060. DOI : https://doi.org/10.2307/2531734

[19] D. Hedeker, R.D. Gibbons, Longitudinal Data Analysis, John Wiley & Sons, Canada, 2006.

[20] J.W. Hardin, J. M. Hilbe, Generalized Estimating Equations, Chapman & Hall, USA, 2003

[21] W. Imawan , Indeks Komposit Kesejahteraan Anak (IKKA) 2015 Kementrian Pemberdayaan Pemberdayaan Perempuan dan Perlindungan Anak Indonesia, 2017.

[22] W. Imawan, Indeks Komposit Kesejahteraan Anak (IKKA) 2017 Kementrian Pemberdayaan Pemberdayaan Perempuan dan Perlindungan Anak Indonesia, 2018.