BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2293–2306
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2293-2306

ROBUST LEAST MEDIAN OF SQUARE MODELLING USING SEEMINGLY UNRELATED REGRESSION WITH GENERALIZED LEAST SQUARE ON PANEL DATA FOR TUBERCULOSIS CASES

Amanda Adityaningrum [1*], Resmawan[2], Annisa Maharani Brahim[3], Dewi Rahmawaty Isa[4], La Ode Nashar[5], Asriadi[6]

*[1,2,3,4,5,6]Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Gorontalo
Jln. Prof. Dr. Ing. B.J. Habibie, Bone Bolango, Gorontalo, 96554, Indonesia*

*Corresponding author's e-mail: * **amanda@ung.ac.id***

ABSTRACT

Article History:

Received: 23rd February 2024
Revised: 20th May 2024
Accepted: 16th July 2024
Published:14th October 2024

Keywords:

Robust Least Median of Square;
Seemingly Unrelated Regression;
Generalized Least Square;
Panel Data;
Tuberculosis.

Tuberculosis, primarily affecting the lungs and other organs, was the leading cause of death worldwide before the COVID-19 pandemic and continues to be a significant health concern. This research examined tuberculosis (TB) using a panel dataset. As a consequence, the datasets may contain outliers and contemporaneous correlations. A Robust Least Median of Square (LMS) model was developed in this research by combining Seemingly Unrelated Regression (SUR) with Generalized Least Square (GLS) on panel data to provide an analysis overview to overcome outliers and contemporaneous correlations. Based on secondary data obtained from the Central Bureau of Statistics of the Gorontalo Province and the Ministry of Health of the Gorontalo Province, this research examines TB cases between 2017 and 2021. The Chow test result suggests that CEM is the most appropriate model for analyzing panel data for TB cases in Gorontalo Province between 2017 and 2021. Due to the presence of outliers and influential observations in the data, robust LMS is employed. Furthermore, there is a problem of contemporaneous correlation in this research. Each regency or city can mitigate this problem by implementing robust LMS using SUR with GLS.

How to cite this article:

A. Adityaningrum, Resmawan, A. M. Brahim, D. R. Isa, L. O. Nashar and Asriadi., " ROBUST LEAST MEDIAN OF SQUARE MODELLING USING SEEMINGLY UNRELATED REGRESSION WITH GENERALIZED LEAST SQUARE ON PANEL DATA FOR TUBERCULOSIS CASES," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2293-2306, December, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Tuberculosis (TB) is an infection caused by the bacteria Mycobacterium tuberculosis, primarily attacking the lungs and other body organs, with approximately 80% of cases worldwide being pulmonary TB [1][2]. According to [3], until the COVID-19 pandemic, TB was the foremost cause of death in the world, surpassing HIV/AIDS. TB accounted for 354 cases per 100,000 in Indonesia in 2021, and in Gorontalo province, the number fluctuated between 2019 and 2021 [3][4].

Generally, the analysis uses three types of data: time-series, cross-sectional, and panel data [5]. Currently, most TB research utilizes cross-sectional data, in which one or more objects are collected simultaneously. Time-series data, on the other hand, are measurements taken over a period of time [6]. Panel data are a combination of these two datasets, encompassing samples made up of several individuals over a period of time [7][8]. Panel data offers several advantages, including studying more complex behavioral models that cannot be detected using time-series or cross-sectional data alone [9]. The panel data regression method can be used to model this behavior, namely an analysis that utilizes panel data to analyze the influence of several independent variables on a single dependent variable [10].

A panel dataset may, however, contain observations that are inconsistent with other datasets and can have a significant impact on the regression model, commonly referred to as outliers [11]. There is also a term for outliers in statistics, which describes observations whose values differ from those exhibited by the data group [12]. Outliers can result in significant and non-homogeneous residual variances, making obtaining the Best Linear Unbiased Estimators (BLUE) for regression models impossible [13]. Outliers can arise from various factors, including data entry errors, measurement system inaccuracies, or unforeseen events, such as crises or disasters. These factors can make an outlier an influential observation and can change the meaning of a regression model if the observation is discarded or rejected before analysis [12][14][15]. An approach that overcomes this issue is to use an analysis unlikely to be affected by outliers, namely robust regression analysis [13][16]. There are several robust regression estimates, including Maximum Likelihood (M), Method of Moment (MM), Scale (S), Least Trimmed Square (LTS), and Least Median of Squares (LMS) [17][18].

In previous research, [15] researched simulation data and student's online test scores that compared linear regression using OLS (Ordinary Least Square), robust regression using LMS, and using M, finding that the robust LMS had a significantly higher R-square value than the two other regressions, demonstrating that robust LMS was an effective regression coefficient estimation model for data with outliers. Further, [19] compared robust regression using LMS and S (case: Budget Revenues and Expenditures data), where it was discovered that robust LMS would yield smaller AIC and SIC values than robust S when modeling data with outliers. In addition, prior research conducted by [12], comparing robust regression using LMS and linear regression using OLS in the context of stack loss data, found that the MAPE value for robust LMS is much lower than for linear OLS. Also, a comparison between robust regression using LMS and MM was conducted by [20] about rice production data, concluding that robust LMS is most suitable for modeling data with outliers because the RMSE value is lower than and the R-square is higher than the robust MM. Therefore, this research used robust regression using LMS to overcome the outlier problem on panel data.

In this research, the case study is TB cases in Gorontalo Province. Previous research on TB cases on cross-sectional data with outliers has been conducted by [13], [16] using quantile regression, robust M and LTS regression. Additionally, TB cases have also been analyzed using panel data by [7], [8], [28] using panel data regression.

Aside from outliers, contemporaneous correlation can also pose a problem in panel data regression models. The contemporaneous correlation refers to the correlation between the panel data regression model and the individual residuals [21]–[23]. It is also referred to as cross-sectional dependence in residuals and occurs when residuals are correlated between individuals or units, resulting in deviations from the model [21]. Seemingly Unrelated Regression (SUR) is used to address this problem [21][24]. The SUR method is a development of linear OLS regression and is used when residuals of several regression models (in which each model has a different dependent variable and/or an independent variable) are correlated between one residual and the residuals of another equation [25][26]. A Generalized Least Square (GLS) estimator is used in SUR, a modification of OLS estimation, and does not disregard the correlation between blocks when estimating this technique. Moreover, parameter estimation in SUR is carried out simultaneously through contemporaneous correlation. As a result, SUR can handle any correlation, including peer correlation, and produce a BLUE estimator rather than linear regression, which was carried out on each block separately [21]–

[25], [27]. Hence, SUR with GLS was employed in this research to tackle contemporaneous correlation on panel data.

Previously, there has been research about SUR using GLS, including research on panel data by [21], panel data and outliers by [22] with robust LTS, and by [24] with robust MM and S. Until now, no research has been conducted about robust LMS using SUR with GLS on panel data and outliers. Thus, this research used robust LMS using SUR with GLS on panel data with outliers.

However, no studies on TB cases using panel data with outlier and contemporaneous correlation have been conducted. Hence, this research aims to conduct robust LMS modeling using SUR with GLS on panel data to provide an analysis overview to overcome outlier and contemporaneous correlation. This research studied a TB case from the Province of Gorontalo.

2. RESEARCH METHODS

2.1 Data

This study relies on secondary data from the Central Bureau of Statistics of Gorontalo Province and the Ministry of Health of Gorontalo Province between 2017 and 2021. In this research, all five regencies and one city of Gorontalo Province, namely Boalemo (*BR*), Gorontalo (*GR*), Pohuwato (*PR*), Bone Bolango (*NR*), Gorontalo Utara (*UR*), and Gorontalo (*GC*), were analyzed as samples. Based on the data collected, the dependent variable used is the number of TB cases (Y), while the independent variables are the number of doctors (X_1), population density (X_2), and the number of poor individuals (X_3).

2.2 Methods

2.2.1. Robust Regression using LMS

The algorithm of robust regression using LMS minimizes the median of the squared residuals of the linear regression using OLS, which has been sorted [12]. Robust regression using LMS (Robust LMS) is a robust regression with a high breakdown point, defined as the proportion of observations that can be resolved before influencing the regression model [12][29]. As a result of this regression, a robust and outlier-resistant model is obtained [12][19][30][31].

The median of the squared residuals of the linear regression using OLS is given in **Equation (1)**:

$$M_j = \text{median}\, \varepsilon_i^2 \tag{1}$$

thus, the $M_1, M_2, \dots, M_S$ value will be obtained, which is the median ε_i^2 of every iteration [12][30][31]. A robust LMS consists of the following steps [12], [29]–[31]:

a. Calculate the β parameter of linear regression using OLS based on **Equation (2)**;

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2}$$

b. Calculate the residual (ε_i) value of linear regression using OLS, which $\varepsilon_i = y - X\beta$;
c. Square the ε_i value (ε_i^2);
d. Sort the ε_i^2 value;
e. Calculate the M_j value based on the **Equation (1)**;
f. Calculate the h_k value, which $h_k = \left(\frac{n+p+1}{2}\right)$ and an integer, h_k is rounded up if the result obtained is not an integer;
g. Create new samples of h_k samples (consisting of dependent and independent variables data) based on the sorted ε_i^2 value ;
h. Using the new sample obtained from (f), repeat step (a) and stop at the k^{th} iteration, where $h_k = h_{k+1}$;
i. Determine the minimum value of all M_j ($\min M_j$, $j = 1,2, \dots, s$);

j. Calculate the initial weight value (w_{0i}) for each observation based on **Equation (3)**:

$$w_{0i} = \begin{cases} 1 & ; \left|\dfrac{\varepsilon_i}{S_0}\right| \leq 2.5 \\ 0 & ; \left|\dfrac{\varepsilon_i}{S_0}\right| > 2.5 \end{cases} \tag{3}$$

which S_0 is the initial value, where $S_0 = 1.4826\left(1 + \frac{5}{(n-p)}\right)\sqrt{\min M_j}$ **[12][32]**;

k. Calculate the final weight (w_i) for each observation based on **Equation (4)**:

$$w_i = \begin{cases} 1 & ; \left|\dfrac{\varepsilon_i}{\hat{\sigma}}\right| \leq 2.5 \\ 0 & ; \left|\dfrac{\varepsilon_i}{\hat{\sigma}}\right| > 2.5 \end{cases} \tag{4}$$

which $\hat{\sigma}$ is a robust parameter, where

$$\hat{\sigma} = \sqrt{\frac{\left(\sum_{i=1}^{n} w_i \varepsilon_i^2\right)}{\left(\sum_{i=1}^{n} w_i - p\right)}}$$

and p is the number of parameters in the model (including intercept);

l. Calculate the β parameter of robust LMS based on **Equation (5)**:

$$\hat{\beta} = (X'WX)^{-1}X'Wy \tag{5}$$

which

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix};\ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}; W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

2.2.2. Seemingly Unrelated Regression

A SUR model consists of q equations interconnected by residuals from different equations, referred to as blocks **[23]**, **[24]**. The **Equation (6)** represents the SUR model as a linear regression equation **[24][33]**.

$$\begin{gathered} y_1 = X_1\beta_1 + \varepsilon_1 \\ y_2 = X_2\beta_2 + \varepsilon_2 \\ \vdots \\ y_q = X_q\beta_q + \varepsilon_q \end{gathered} \tag{6}$$

In matrix notation, **Equation (6)** can be expressed by **Equation (7) [24][27]**.

$$y = X\beta + \varepsilon \tag{7}$$

The **Equation (7)** is represented by the matrix design shown below:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_q \end{bmatrix}; X = \begin{bmatrix} X_1 & 0 & 0 & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_q \end{bmatrix}; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_q \end{bmatrix}$$

with y represents the dependent variable as a $(n \times 1)$ column vector; X represents the independent variable as a $(n \times p_i)$ matrix, and each block may have a different number of independent variables; βrepresents the unknown parameter as a $(p_i \times 1)$ column vector; ε represents the residual as a $(n \times 1)$ column vector; and suppose $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma_{ii}I_n$, for each question $(i = 1, \ldots, q)$ **[24][34]**. A particular characteristic of the SUR model is that $cov(\varepsilon_i, \varepsilon_j) = \sigma_{ij}I_n(i,j = 1, \ldots, q)$ **[33]**. **Equation (8)** depicts the estimation of parameters β for the SUR model using GLS:

$$\hat{\beta} = (X^{\prime V^{-1}}X)^{-1}X^{\prime V^{-1}}y \tag{8}$$

which $V = \boldsymbol{\Sigma} \otimes I_n = var(\varepsilon)$; thus, the **Equation (9)** can be expressed by **Equation (9) [24][26]**.

$$\hat{\beta} = \left(X'^{(\boldsymbol{\Sigma}\otimes I_n)^{-1}}X\right)^{-1} X'^{(\boldsymbol{\Sigma}\otimes I_n)^{-1}} y$$
$$\hat{\beta} = \left(X'^{(\boldsymbol{\Sigma}^{-1}\otimes I_n)}X\right)^{-1} X'^{(\boldsymbol{\Sigma}^{-1}\otimes I_n)^{-1}} y \tag{9}$$

The design matrix of $\boldsymbol{\Sigma}$ is represented by **Equation (10) [22][33]**.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \cdots & \sigma_{qq} \end{bmatrix} \tag{10}$$

The **Equation (9)** is used when the V value is known, and as V is seldom known, an estimator that can be calculated is presented in the **Equation (11)**:

$$\hat{\beta}_c = \left(X'^{(\hat{\boldsymbol{\Sigma}}^{-1}\otimes I_n)}X\right)^{-1} X'^{(\hat{\boldsymbol{\Sigma}}^{-1}\otimes I_n)^{-1}} y \tag{11}$$

which $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$ **[33], [34]**. A matrix of residuals is used to compute $\hat{\boldsymbol{\Sigma}}$ as shown in **Equation (12) [33][34]**.

$$\hat{\boldsymbol{\Sigma}} = \left[\sigma_{ij}\right]_{i,j=1,\dots,q} = \frac{\varepsilon_i{}'\varepsilon_j}{n} \tag{12}$$

In the SUR model, Kronecker ($\otimes$) multiplication is used (for each matrix $A_{k\times l}, B_{m\times n}, A \otimes B = C_{km\,\mathrm{x}\,ln}$) **[21]**.

2.3 Analysis Steps

R software is used for all steps, which the research involves the following steps:

(a) Identify multicollinearity;

A VIF (Variance Inflation Factor) test statistic is used to identify this, which is illustrated by **Equation (13)**:

$$\frac{1}{VIF_j} = \left(1 - R_j^2\right) = TOL_j \tag{13}$$

which R_j^2 is a coefficient of determination ($R^2 = \frac{ESS}{TSS}$, where ESS is "Explained Sum of Squares" and TSS is "Total Sum of Squares") in the regression of X_j on the remaining regression (there are $[k-1]$ regressors in the k-variable regression model) **[35]**. In this test, those whose VIF exceeds 10, which occurs when R_j^2 exceeds 0.90, reject the null hypothesis **[21][35]**.

(b) Identify outliers;

Boxplot graphs are used to identify outliers. An observation is considered an outlier if it is outside the boxplot **[36]**.

(c) Conduct panel data regression;

Panel data regression utilizes the Chow test to determine whether CEM or FEM provides the best model **[28]**. In **Equation (14)**, the statistics of the Chow test are presented:

$$F = \frac{\dfrac{(RRSS - URSS)}{(N-1)}}{\dfrac{URSS}{(NT - N - p)}} \tag{14}$$

which $RRSS$ is CEM's Restricted Residual Sum of Squares; $URSS$ is FEM's Unrestricted Residual Sum of Squares; N is the total unit of cross-section; T is the total unit of time-series; and p is the total estimated parameter including intercept [37]. It is rejected the null hypothesis if either the Fvalue is greater than $F_{N-1;N(T-1)-p}$ or the $p - value < \alpha$, implying that the best model is FEM [37][38].

Meanwhile, to determine whether REM of FEM provides the best model, the Hausman test is used [8]. In Equation (15)**,** the statistics of the Hausman test are presented [37].

$$\begin{gathered} m_1 = \hat{q}_1^{\iota}[\mathrm{var}(\hat{q}_1)]^{-1}\hat{q}_1 \\ \hat{q}_1 = \hat{\beta}_{REM} - \hat{\beta}_{FEM} \end{gathered} \tag{15}$$

Equation (15) is asymptotically distributed as chi-square distribution (χ^2_K), which K indicates the dimension of the slope vector β [37]. Either m_1 value is greater than $\chi^2_{K;\alpha}$ or $p - value < \alpha$ will reject H_0, suggesting that FEM is the most suitable model [10][28][37].

(d) Identify influential observations in the regression model;

As illustrated in Equation (16), DFFITS can be computed as follows:

$$DFFITS_i = t_i\left(\frac{h_{ij}}{1-h_{ij}}\right)^{\frac{1}{2}}; t_i = \frac{\varepsilon_i}{\sqrt{s^2_{(i)}\left(1-h_{ij}\right)}}; s^2_{(i)} = \frac{(n-p)s^2 - \dfrac{\varepsilon_i^2}{\left(1-h_{ij}\right)}}{n-p_i-1} \tag{16}$$

which $i = 1,2,\ldots,n$; h_{ij} are the diagonal elements of the matrix $H = (X'X)^{-1}X'$ ($n \times n$ matrix); $t_i$is the R-student or studentized deleted residual; ε_i is the i^{th} residual; and $s^2 = \sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{n-p}$ [29][39]. When $|DFFITS|$ exceeds $2\sqrt{\frac{p_i}{n}}$, the i^{th} observation is considered influential, which p_i denotes the number of regression parameters, including the intercept [11][31][39].

(e) Identify assumptions of residuals in the regression model (normal distribution, homoscedasticity, and autocorrelation);

In order to determine whether residuals are from a normal population, the Kolmogorov-Smirnov test is employed [39]–[41]. Equation (17) presents the statistics of the Kolmogorov-Smirnov test:

$$D = \max_X|F^*(X) - S_N(X)| \tag{17}$$

which $S_N(X)$ is the sample cumulative distribution function, $F^*(X)$ is the cumulative normal distribution function with the sample mean $\mu = \overline{X}$, and the sample variance $\sigma^2 = s^2$, defined with a denominator $n - 1$ [40]. The null hypothesis (observations are from a normal population) is rejected if the value D exceeds the critical value in the table or if $p - value < \alpha$ [40][41].

Identifying whether the residuals are independent of one another is the next assumption; otherwise, an autocorrelation condition will occur; thus, the Breusch-Godfrey test is used [5][35][42]. The Breusch-Godfrey is illustrated in Equation (18) [35].

$$BG = (n-p)R^2 \tag{18}$$

The $(n-p)R^2$ value follows the chi-square distribution with p degrees of freedom (χ^2_p) [5][35]. *H*$_0$ (no serial correlation of any order) is rejected if the chi-square computed is greater than the critical chi-square value at the chosen significance level, or if $p - value < \alpha$ [5][35].

One method for detecting homoscedasticity (which assumes an equal variance of the residual) is the Breusch-Pagan-Godfrey test, which can be seen in Equation (19) [35].

$$\varTheta = \frac{1}{2}(ESS) \tag{19}$$

The $\varTheta$ value follows the chi-square distribution with $(p-1)$ degrees of freedom (χ^2_{p-1}) [9][35]. One can reject the null hypothesis (which states that homoscedasticity occurs in the model) if the computed $\varTheta$ value exceeds the critical value (χ^2_{p-1}) or when $p - value < \alpha$ [9][35][43].

(f) Conduct robust LMS;

The algorithm of robust regression is explained in section 2.2.1. Furthermore, the β parameter of robust LMS is based on Equation (5).

(g) Identify contemporaneous correlation;

A Lagrange Multiplier test is conducted to detect contemporaneous correlation in the regression analysis of panel data [21][23][24]. As demonstrated in Equation (20) Lagrange Multiplier can be calculated in the following way:

$$LM = n\sum_{i=2}^{p}\sum_{j=1}^{i-1} r_{ij}^2\,;\; r_{ij}^2 = \frac{\varepsilon_i{}'\varepsilon_j}{\sqrt{(\varepsilon_i{}'\varepsilon_i)(\varepsilon_j{}'\varepsilon_j)}} \tag{20}$$

which r_{ij}^2 denotes the sample correlation matrix elements derived from the residual vectors [24][34]. Equation (20) follows a $\chi^2_{\left(p\frac{(p-1)}{2}\right)}$ distribution asymptotically under the null hypothesis, which is that there is no contemporaneous correlation [21][22][24][34]. If the computed LM value is greater than the critical value ($\chi^2_{\left(p\frac{(p-1)}{2}\right)}$) or when $p - value < \alpha$, then the null hypothesis is rejected; otherwise, it failed to reject [21][24][34].

(h) Conduct robust LMS using SUR with GLS;

Robust LMS using SUR with GLS is explained in section 2.2.2. The estimation of parameters $\beta$is illustrated in Equation (8).

(i) Conclude.

3. RESULTS AND DISCUSSION

3.1 Identify Multicollinearity

An analysis of multicollinearity identification using Equation (13) is summarized in Table 1.

Table 1. Identify Multicollinearity

Variable	VIF
X_1 and X_2	2.37751
X_1 and X_3	1.00206
X_2 and X_3	1.25194

Multicollinearity occurs when there is a linearly strong relationship between explanatory variables, and if multicollinearity exists, the standard errors of each coefficient are increased, which changes the outcome of the analysis [21][44]. Table 1 shows that the value of VIF is less than 10 ($\text{VIF} < 10$) for all independent variables. This result indicates that the independent variables do not exhibit highly collinear relationships.

3.1 Identify Outliers

Figure 1 illustrates the boxplot graphs for four variables, including the dependent variable.

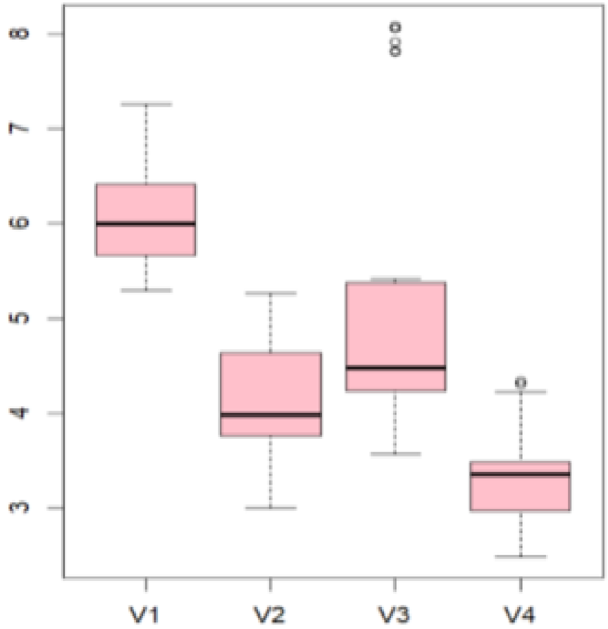

Figure 1. **Boxplot Graph for Four Variables**

Figure 1 shows that there are observations that are outside of the boxplot. Therefore, those observations indicate outliers.

3.2 Panel Data Regression

According to the result of the Chow test using Equation (14), the $p-value$ is 0.05312 ($p-value > \alpha$), which indicates that it fails to reject the null hypothesis. Therefore, CEM is this research's most appropriate model for panel data regression. The model for the panel data regression of TB cases from 2017 to 2021 in the Province of Gorontalo is shown in Equation (21).

$$y = 2.133 + 0.474X_{1it} + 0.099X_{2it} + 0.445X_{3it} + \varepsilon_{it} \tag{21}$$

The Hausman test in Equation (15) is no longer needed as CEM is the most suitable model in this research.

3.3 Identify Influential Observations

As a result of the *DFFITS* test using Equation (16), two observations (8th and 28th) were identified as influential since their $|DFFITS|$ values (0.76885 and 0.87050, respectively) exceeded the $2\sqrt{\frac{p_i}{n}} = 2\sqrt{\frac{4}{30}} =$ 0.73030. Thus, a further analysis of the outlier and influential observations is needed.

3.4 Identify Assumptions of Residuals

Equation (17), Equation (18), and Equation (19) are used to identify assumptions of residuals. The results are presented in Table 2.

Table 2. **Residual Assumptions**

Test	*p-value*
Kolmogrov-Smirnov	0.15500
Breusch-Godfrey	0.07450
Breusch-Pagan-Godfrey	0.58050

As it is illustrated in Table 2, all $p - values$ are greater than α, suggesting that it failed to reject the null hypothesis. Hence, the residuals are from a normally distributed population, have no autocorrelation, and have constant variance (no heteroscedasticity).

3.5 Robust LMS

The first step in a robust regression using LMS is calculating the linear regression parameter β using OLS based on the Equation (2). The results are presented in Table 3.

Table 3. $\hat{\boldsymbol{\beta}}$ of Linear Regression using OLS

City / Regency		$\hat{\beta}$			
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Boalemo	(Regency)	-1.77180	0.21980	1.63570	-0.24120
⋮	⋮	⋮	⋮	⋮	⋮
Gorontalo	(City)	31.55570	0.87180	-0.20430	-11.11950

Table 3 shows the $\hat{\beta}$ value for each regency and city using linear regression. The value is then applied to the next step.

Calculating the residual value of linear regression (ε_i) is the next step, followed by squaring the ε_i value (ε_i^2) and sorting it to determine its median. The median of the ε_i^2 value is referred to as the M_j value, as expressed in Equation (1). The h_k value should then be calculated $h_k = \left(\frac{n+p+1}{2}\right)$. Table 4 shows the results.

Table 4. Linear Regression's Residual and Squared Residual, Median of the Squared Residual, and $\boldsymbol{h_k}$

City/Regency		ε_i	ε_i^2	M_j	h_k
Boalemo	(Regency)	-0.00578	0.00003	0.00033	$h_{BR} = \left(\frac{5+4+1}{2}\right) = 5$
		0.02440	0.00060		
		0.00438	0.00002		
		0.01830	0.00033		
		-0.04130	0.00170		
⋮	⋮	⋮	⋮	⋮	⋮
Gorontalo	(City)	-0.23800	0.05680	0.05680	$h_{GC} = \left(\frac{5+4+1}{2}\right) = 5$
		0.11000	0.01210		
		0.28300	0.07990		
		-0.30900	0.09560		
		0.15400	0.02400		

Table 4 shows the h_k value for each regency and city, which is 5. Since the value $h_2 = h_{1+1} = h_1 = 5$, the iteration stops (at h_2). The value is then applied to the next step. Therefore, the M_j value in Table 4 is the minimum value of all median residuals for each regency and city.

Afterward, each observation's initial weight value (w_{0i}) and final weight (w_i) value will be calculated based on Equation (3) and Equation (4). The results are presented in Table 5.

Table 5. The Minimum Value, Initial Weight Value, and Final Weight Value

City / Regency		M_j	S_0	w_{0i}	$\hat{\sigma}$	w_i
Boalemo	(Regency)	0.00033	0.16293	1 ⋮ 1	0.05190	1 ⋮ 1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Gorontalo	(City)	0.05680	2.11939	1 ⋮ 1	0.51795	1 ⋮ 1

The last step in robust LMS is calculating the β parameter based on **Equation (5)**. Results are shown in **Table 6**.

Table 6. $\hat{\beta}$ of Robust LMS

City / Regency		$\hat{\boldsymbol{\beta}}$			
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Boalemo	(Regency)	-1.77175	0.21977	1.63570	-0.24121
⋮	⋮	⋮	⋮	⋮	⋮
Gorontalo	(City)	31.55600	0.87200	-0.20400	-11.11900

Using robust LMS, the value $\hat{\beta}$ for each regency and city is presented in **Table 6**. Afterward, the value is applied to the next step.

3.6 Identify Contemporaneus Correlation

Accordingly, the Lagrange Multiplier test using **Equation (20)** result is 0.00472 ($p - value < \alpha$), so it rejects the null hypothesis, which provides evidence that residuals are correlated between units. SUR is used when contemporaneous correlation happens; hence, Robust LMS using SUR is performed.

3.7 Robust LMS using SUR with GLS

Since panel data exhibit a contemporaneous correlation, robust LMS using SUR with GLS is conducted in this analysis. Firstly, calculate the $\hat{\boldsymbol{\Sigma}}$ value in **Equation (12)** using the result of Robust LMS in **Table 6**. The result is illustrated in **Equation (22)**.

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1q} \\ \vdots & \ddots & \vdots \\ \sigma_{q1} & \cdots & \sigma_{qq} \end{bmatrix} = \begin{bmatrix} \frac{\varepsilon_1{}'\varepsilon_1}{5} & \cdots & \frac{\varepsilon_1{}'\varepsilon_6}{5} \\ \vdots & \ddots & \vdots \\ \frac{\varepsilon_6{}'\varepsilon_1}{5} & \cdots & \frac{\varepsilon_6{}'\varepsilon_6}{5} \end{bmatrix} = \begin{bmatrix} 0.00054 & \dots & -0.00135 \\ \vdots & \ddots & \vdots \\ -0.00135 & \dots & 0.05360 \end{bmatrix} \tag{22}$$

Secondly, calculate the $\hat{\boldsymbol{\Sigma}}^{-1} \otimes I_n$ value. **Equation (23)** presented the $\hat{\boldsymbol{\Sigma}}^{-1}$value and the $\hat{\boldsymbol{\Sigma}}^{-1} \otimes I_n$ value, which is a (30x30) matrix referred to as a V matrix.

$$\begin{gathered} \hat{\boldsymbol{\Sigma}}^{-1} = \begin{bmatrix} 30.60602 & \dots & -11.04690 \\ \vdots & \ddots & \vdots \\ -11.04690 & \dots & 22.33947 \end{bmatrix} \\ \hat{\boldsymbol{\Sigma}}^{-1} \otimes I_n = \begin{bmatrix} 30.60602 & \dots & -11.04690 \\ \vdots & \ddots & \vdots \\ -11.04690 & \dots & 22.33947 \end{bmatrix} \otimes \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \\ \hat{\boldsymbol{\Sigma}}^{-1}_{6\times 6} \otimes I_{5\times 5} = V_{30\times 30} = \begin{bmatrix} 30.60602 & \dots & 0.00000 \\ \vdots & \ddots & \vdots \\ 0.00000 & \dots & 22.33947 \end{bmatrix} \end{gathered} \tag{23}$$

Finally, based on the **Equation (11)**, calculate the β parameter of robust LMS using SUR with GLS. **Table 7** summarizes the results.

Table 7. $\hat{\beta}$ of Robust LMS using SUR with GLS

City / Regency		$\hat{\boldsymbol{\beta}}$			
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Boalemo	(Regency)	0.45018	0.10004	11.29399	-2.83803
Gorontalo	(Regency)	0.55373	1.35182	-0.78715	1.04621
Pohuwato	(Regency)	1.61068	0.00343	1.88613	-3.98042
Bone Bolango	(Regency)	1.10680	-0.09795	1.72387	1.57233
Gorontalo Utara	(Regency)	1.33749	0.26049	-0.10992	-2.19965
Gorontalo	(City)	-1.11663	0.15239	-0.72172	2.64162

Based on Robust LMS using SUR with GLS, Table 6 illustrates the $\hat{\beta}$ value for each regency and city. For example, the model of robust LMS using SUR with GLS for TB cases in Boalemo Regency, Gorontalo Province, from 2017 to 2021 can be expressed in Equation (24).

$$\hat{\boldsymbol{y}}_{\boldsymbol{BR}} = \mathbf{0.450} + \mathbf{0.100}\boldsymbol{X}_{\mathbf{1}} + \mathbf{11.294}\boldsymbol{X}_{\mathbf{2}} - \mathbf{2.838}\boldsymbol{X}_{\mathbf{3}} \tag{24}$$

Models of robust LMS using SUR with GLS are developed for each regency or city to overcome contemporaneous correlation in panel data. Since outliers and influential observations are presented in the data, robust LMS is used. Moreover, there is a problem of contemporaneous correlation in the panel data used in this research. Each regency or city can mitigate this problem by implementing robust LMS using SUR with GLS. The result aligns with previous research that said that when contemporaneous correlation happens, SUR can be used as an alternative [21][24].

Nevertheless, the model is better suited when outliers exist in the panel data within each regency and city. A similar finding was made by [22], who found that robust regression alone is impractical when each panel data consists of outliers and in the data and contemporaneous correlation between different equations individuals/units.

4. CONCLUSIONS

According to the findings of this research, it appears that CEM is the most suitable model to use for the regression analysis of panel data for TB cases in Gorontalo Province from 2017 to 2021. However, outliers and influential observations are present in this research, necessitating robust LMS. Additionally, contemporaneous correlation poses a problem in TB cases panel data, which can be compensated using robust LMS using SUR with GLS.

REFERENCES

[1] H. Jusuf, M. Sakti, I. Husein, M. Elveny, R. Syah, and S. Tuba, "Modelling Optimally to the Treatment of TB Patients for Increase Medical Knowledge," *Syst. Rev. Pharm.*, vol. 11, no. 4, pp. 742–748, 2020, [Online]. Available: https://www.sysrevpharm.org/articles/modelling-optimally-to-the-treatment-of-tb-patients-for-increase-medical-knowledge.pdf

[2] S. S. Rajak, S. Ismail, and R. Resmawan, "Metode Conditional Autoregressive dalam Analisis Penyebaran Kasus Penyakit Tuberculosis," *Jambura J. Probab. Stat.*, vol. 2, no. 1, pp. 28–34, Apr. 2021, doi: 10.34312/jjps.v2i1.9771.

[3] WHO, *Global Tuberculosis Report 2022*. WHO, 2022.

[4] Dinas Kesehatan Provinsi Gorontalo, "Profil Kesehatan 2021." Dinas Kesehatan Provinsi Gorontalo, 2021.

[5] D. N. Gujarati, *Econometrics by Example*. Macmillan Education Palgrave, 2015. [Online]. Available: https://books.google.co.id/books?id=ONpdyQEACAAJ

[6] M. Irwansyah, R. Ruliana, and M. K. Aidid, "Analisis Regresi Balanced Panel dengan Komponen Galat Dua Arah pada Kasus Melek Huruf Masyarakat di Provinsi NTB," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 3, no. 1, p. 10, Mar. 2021, doi: 10.35580/variansiunm14644.

[7] A. Kustanto, "The role of socioeconomic and environmental factors on the number of tuberculosis cases in Indonesia," *J. Ekon. Pembang.*, vol. 18, no. 2, pp. 129–146, Dec. 2020, doi: 10.29259/jep.v18i2.12553.

[8] E. D. Sihaloho, I. L. Alfarizy, and E. B. Sagala, "Indikator Ekonomi dan Angka Tuberkulosis di Kabupaten Kota di Jawa Barat," *J. Ilmu Ekon. dan Pembang.*, vol. 19, no. 2, pp. 128–138, 2019, [Online]. Available: https://jurnal.uns.ac.id/jiep/article/view/33698

[9] D. Suliyanto, *Ekonometrika Terapan: Teori dan Aplikasi dengan SPSS*, 1st ed. Yogyakarta: Andi, 2011. [Online]. Available: https://scholar.google.com/scholar?cluster=3601403273910001182&hl=en&oi=scholarr

[10] I. Alamsyah, R. Esra, S. Awalia, and D. Nohe, "Analisis Regresi Data Panel untuk Mengetahui Faktor yang Memengaruhi Jumlah Penduduk Miskin di Kalimantan Timur," *Pros. Semin. Nas. Mat. dan Stat.*, vol. 2, 2022, [Online]. Available: http://jurnal.fmipa.unmul.ac.id/index.php/SNMSA/article/view/861

[11] C. O. Arimie, E. O. Biu, and M. A. Ijomah, "Outlier Detection and Effects on Modeling," *OALib*, vol. 07, no. 09, pp. 1–30, 2020, doi: 10.4236/oalib.1106619.

[12] F. Daniel, "Mengatasi Pencilan pada Pemodelan Regresi Linear Berganda dengan Metode Regresi Robust Penaksir LMS," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 13, no. 3, pp. 145–156, Oct. 2019, doi: 10.30598/barekengvol13iss3pp145-156ar884.

[13] H. Khotimah and Suwanda, "Pemodelan Quantile Regression untuk Menentukan Faktor-Faktor Penyebab Penyakit Tuberkulosis Paru di Kabupaten Tasikmalaya," *Bandung Conf. Ser. Stat.*, vol. 2, no. 2, pp. 61–70, Jul. 2022, doi: 10.29313/bcss.v2i2.3068.

[14] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decis. Anal. J.*, vol. 6, p. 100164, Mar. 2023, doi: 10.1016/j.dajour.2023.100164.

[15] H. Sugiarti and A. Megawarni, “Tingkat efisiensi penaksir M terhadap penaksir LMS dalam menaksir koefisien garis regresi,” *J. Mat. Sains dan Teknol.*, vol. 11, no. 2, pp. 90–98, 2010.

[16] D. Rohmah, Y. Susanti, and E. Zukhronah, “Perbandingan Model Regresi Robust Estimasi M dan Estimasi Least Trimmed Squares (LTS) pada Jumlah Kasus Tuberkulosis di Indonesia,” *Kontinu J. Penelit. Didakt. Mat.*, vol. 4, no. 2, p. 136, Nov. 2020, doi: 10.30659/kontinu.4.2.136-146.

[17] S. Heritier, E. Cantoni, S. Copt, and M.-P. Victoria-Feser, *Robust Methods in Biostatistics*. Germany: Wiley, 2009. [Online]. Available: https://scholar.google.com/scholar?cluster=4506501391474728644&hl=en&as_sdt=0,5

[18] H. Riazoshams, H. Midi, and G. Ghilagaber, *Robust Nonlinear Regression: with Applications using R*. United Kingdom: Wiley, 2018. [Online]. Available: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=18421333600214340159

[19] L. S. Febrianto, P. Hendikawati, and N. K. Dwidayati, “Perbandingan Metode Robust Least Median of Square (LMS) dan Penduga S Untuk Menangani Outlier Pada Regresi Linier Berganda,” *Unnes J. Math.*, vol. 7, no. 1, pp. 83–95, 2018.

[20] K. Khotimah, K. Sadik, and A. Rizki, “Study of Robust Regression Modeling Using MM-Estimator and Least Median Squares,” 2019. doi: 10.4108/eai.2-8-2019.2290533.

[21] W. Jannah, I. D. Sulvianti, Indahwati, P. Silvianti, and A. Kurnia, “Analysis of Credit Bank Distribution with Seemingly Unrelated Regression Method on Panel Data,” *J. Phys. Conf. Ser.*, vol. 1863, no. 1, p. 012057, Mar. 2021, doi: 10.1088/1742-6596/1863/1/012057.

[22] D. A. Yulianto, S. Sugiman, and A. Agoestanto, “Estimasi Parameter Regresi Robust Model Seemingly Unrelated Regrssion (SUR) dengan Metode Generalized Least Square (GLS),” *Unnes J. Math.*, vol. 7, no. 2, pp. 216–227, 2018, doi: https://doi.org/10.15294/ujm.v7i2.21463.

[23] A. Widyaningsih, M. Susilawati, and I. W. Sumarjaya, “Estimasi Model Seemingly Unrelated Regression (SUR) dengan Metode Generalized Least Square (GLS),” *J. Mat.*, vol. 4, no. 2, pp. 102–110, 2014, [Online]. Available: https://ojs.unud.ac.id/index.php/jmat/article/view/12554

[24] K. Peremans and S. Van Aelst, “Robust Inference for Seemingly Unrelated Regression Models,” *J. Multivar. Anal.*, vol. 167, pp. 212–224, Sep. 2018, doi: 10.1016/j.jmva.2018.05.002.

[25] A. Fitriana, W. Pramesti, and E. M. P. Hermanto, “Pemodelan Seemingly Unrelated Regression (SUR) pada Faktor Pertumbuhan Perekonomian di Provinsi Jawa Timur Tahun 2018,” *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 13, no. 2, pp. 12–19, Dec. 2020, doi: 10.36456/jstat.vol13.no2.a2950.

[26] A. Anisa, “Seemingly Unrelated Regression (SUR) Penderita Penyakit DBD RS. Wahidin Sudirohusodo Dan RS. Stella Maris Makassar,” *J. Mat. Stat. dan Komputasi*, vol. 13, no. 1, pp. 20–25, 2016, [Online]. Available: https://journal.unhas.ac.id/index.php/jmsk/article/view/3475

[27] A. G. Ghazal and S. A. Hegazy, “The two feasible seemingly unrelated regression estimator,” *Int. J. Sci. Technol. Res.*, vol. 4, no. 4, pp. 247–253, 2015, [Online]. Available: https://www.ijstr.org/final-print/apr2015/The-Two-Feasible-Seemingly-Unrelated-Regression-Estimator.pdf

[28] A. Y. Kartini, N. Cahyani, and N. Himawati, “Regresi Data Panel untuk Pemodelan Jumlah Penderita Tuberculosis di Kabupaten Bojonegoro,” *J. Stat. dan Apl.*, vol. 6, no. 2, pp. 264–275, Dec. 2022, doi: 10.21009/JSA.06212.

[29] A. R. Widyaningrum, Y. Susanti, and I. Slamet, “Pemodelan Penyakit Diare Balita di Jawa Timur Menggunakan Regresi Robust,” in *SINASIS (Seminar Nasional Sains)*, 2021, vol. 2, no. 1. [Online]. Available: https://proceeding.unindra.ac.id/index.php/sinasis/article/view/5393

[30] M. Kurniati, Y. Yundari, and S. W. Rizki, “Metode Least Median Square (LMS) dalam Analisis Regresi Robust Ketika Terdapat Outlier,” *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 8, no. 4, Oct. 2019, doi: 10.26418/bbimst.v8i4.36553.

[31] T. Tusilowati, L. Handayani, and R. Rais, “Simulasi Penanganan Pencilan pada Analisis Regresi Menggunakan Metode Least Median Square (LMS),” *J. Ilm. Mat. DAN Terap.*, vol. 15, no. 2, pp. 238–247, 2018.

[32] P. J. Rousseeuw, “Least Median of Squares Regression,” *J. Am. Stat. Assoc.*, vol. 79, no. 388, p. 871, Dec. 1984, doi: 10.2307/2288718.

[33] M. Bilodeau and P. Duchesne, “Robust estimation of the SUR model,” *Can. J. Stat.*, vol. 28, no. 2, pp. 277–288, Jun. 2000, doi: 10.2307/3315978.

[34] J.-M. Dufour and L. Khalaf, “Exact tests for contemporaneous correlation of disturbances in seemingly unrelated regressions,” *J. Econom.*, vol. 106, no. 1, pp. 143–170, Jan. 2002, doi: 10.1016/S0304-4076(01)00093-8.

[35] D. N. Gujarati and D. C. Porter, *Basic Econometrics*, 5th ed. McGraw-Hill Education, 2008. [Online]. Available: https://books.google.co.id/books?id=zJlDPgAACAAJ

[36] I. R. Akolo and A. Nadjamuddin, “Analisis Regresi Robust Estimasi Least Trimmed Square dan Estimasi Maximum Likelihood pada Pemodelan IPM di Pulau Sulawesi,” *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 10, no. 2, pp. 211–221, Oct. 2022, doi: 10.34312/euler.v10i2.16708.

[37] B. H. Baltagi, *Econometric Analysis of Panel Data*. United Kingdom: Wiley, 2005. [Online]. Available: https://books.google.co.id/books?id=31ruAAAAMAAJ

[38] M. J. Hidayat, A. F. Hadi, and D. Anggraeni, “Analisis Regresi Data Panel Terhadap Indeks Pembangunan Manusia (IPM) Jawa Timur Tahun 2006-2015,” *Maj. Ilm. Mat. dan Stat.*, vol. 18, no. 2, p. 69, Sep. 2018, doi: 10.19184/mims.v18i2.17250.

[39] D. A. Wulandari, D. Kusnandar, and N. Imro’ah, “Estimasi-S Model Regresi Robust menggunakan Pembobot Welsch pada Data Indeks Pembangunan Manusia di Indonesia,” *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 11, no. 4, 2022, [Online]. Available: https://jurnal.untan.ac.id/index.php/jbmstr/article/view/57009

[40] H. W. Lilliefors, “On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown,” *J. Am. Stat. Assoc.*, vol. 62, no. 318, p. 399, Jun. 1967, doi: 10.2307/2283970.

[41] G.- MARDIATMOKO, “PENTINGNYA UJI ASUMSI KLASIK PADA ANALISIS REGRESI LINIER BERGANDA,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 14, no. 3, pp. 333–342, Oct. 2020, doi: 10.30598/barekengvol14iss3pp333-342.

[42] A. Adityaningrum, N. Arsad, and H. Jusuf, “Faktor Penyebab Stunting di Indonesia: Analisis Data Sekunder Data SSGI Tahun 2021,” *Jambura J. Epidemiol.*, vol. 2, no. 1, pp. 1–10, 2023, [Online]. Available: https://ejurnal.ung.ac.id/index.php/jje/article/view/21542

[43] T. Tuwanakotta, M. W. Talakua, Y. W. A. Nanlohy, and L. J. Sinay, “PENERAPAN REGRESI DATA PANEL UNTUK MEMODELKAN APBD DI PROVINSI MALUKU,” *Var. J. Stat. Its Appl.*, vol. 2, no. 1, pp. 15–26, Jul. 2020, doi:

10.30598/variancevol2iss1page15-26.

[44] N. Shrestha, “Detecting Multicollinearity in Regression Analysis,” *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, Jun. 2020, doi: 10.12691/ajams-8-2-1.