J. Ris. & Ap. Mat. Vol. 9 No. 1 (2025) pp. 77-89
Jurnal Riset dan Aplikasi Matematika (JRAM)
e-ISSN: 2581-0154
URL: journal.unesa.ac.id/index.php/jram

ROBUST MINIMUM COVARIANCE DETERMINANT SCALE FOR ADDRESSING OUTLIERS IN FOOD SECURITY INDEX DATA

MOHAMAD ISWANTO RAHMAN[1], SALMUN K. NASIB[2*], AMANDA ADITYANINGRUM[3]

[1,2,3]Program Studi Statistika, Jurusan Matematika, Universitas Negeri Gorontalo, Indonesia

*salmun@ung.ac.id

ABSTRACT

Food Security Index (FSI) data from regions with systemic challenges, such as Papua, Indonesia, often present significant extreme outliers (because of geographic isolation or poverty). These outliers can affect regression analysis. To address this problem, this research conducted the Minimum Covariance Determinant S-estimator (MCD-S), a robust regression method designed to manage datasets with up to 50% outliers. This research aims to address this issue by employing robust regression with the MCD-S estimator on the 2023 FSI data from Papua, Indonesia, as well as identifying key determinants of FSI in Papua. The research used secondary data from the 2023 Food Security and Vulnerability Atlas report, with a sample of 42 regencies and municipalities in Papua. The results showed that according to the model performance, the MCD-S model showed superiority with a higher R^2 compared to the Ordinary Least Square (OLS) model. The result also showed that conducting MCD-S resulted in all independent variables significantly related to FSI. Meanwhile, OLS uncovered some independent variables as key determinants (as in OLS, some variables were considered statistically insignificant in FSI). These results demonstrate the importance of robust methods in FSI data, especially for outlier-prone regions. For future research analyzing similar datasets, MCD-S is recommended.

Keywords: Food Security Index, Outliers, Robust Regression, Minimum Covariance Determinant, S estimator

1 Introduction

Diversity in data can make data analysis challenging, especially when the data is collected from multiple sources, as it can present various characteristics [1]. This diversity can result in issues such as the presence of outliers [2]. An outlier refers to a data point significantly differing from other group values [3]–[5]. The outliers, which may result from errors in data entry, inaccuracies in measurement systems, or unexpected crises, can become influential observations that alter the meaning of a regression model if discarded or rejected, potentially resulting in non-normally distributed data and failing to generalize to broader contexts [4]–[8]. It is, however, not realistic to directly eliminate this aspect or ignore it since this would result in inaccurate parameter estimations in regression models [8]–[10].

In order to address the outlier, robust regression can be used [8]. One of the robust regressions is the Minimum Covariance Determinant (MCD) estimator. The MCD estimator is a highly robust estimator for the dispersion matrix of a multivariate, linearly symmetric distribution and is one of the first robust equivariant estimators of multivariate location and

2020 *Mathematics Subject Classification*: 62F35, 62J05, 62P20
Diterima: 25-02-25; direvisi: 12-03-25; diterima: 29-04-25

scatter [11]–[14]. Referring to [12], [15], MCD is designed to be resistant to outliers, effectively safeguarding against up to 50% of them, making it particularly useful for outlier detection. With a breakdown point equal to α (as n approaches infinity), MCD (α) approximates the arithmetic mean in large datasets while yielding very robust covariance estimators [15]. This allows MCD to remain close to the classical empirical average and covariance matrix, but it does so by excluding outlying points from its computations [15], [16]. As a result, MCD provides a highly reliable measure of dispersion in multivariate data. However, the 50% breakdown MVE and MCD estimators have low asymptotic efficiencies, as cited in [15]. This limitation can be addressed by applying a one-step M-estimator subsequently, thereby increasing their asymptotic efficiency [15]. In addition to M-estimators, there are several other robust estimators available, including Scale (S), Method of Moments (MM), Least Trimmed Squares (LTS), and Least Median Squares (LMS) [8].

Previous research using robust MCD estimators was [6], [14], then MCD with other robust estimators, such as Method of Moment or MCD-MM by [17], Least Trimmed Squares or MCD-LTS by [18]–[20], and Least Median Squares or MCD-LMS [10], [19]. At present, there is a limited amount of research on robust MCD estimators, particularly S-estimators (MCD-S). This gap is particularly important for outlier-prone datasets, such as FSI, where extreme values (for example, regencies with near-zero infrastructure access) can affect conclusions. Thus, the objective of this research is to conduct a review of robust MCD-S aimed at addressing outliers in real data.

Food security is when individuals have access to adequate, safe, diverse, nutritious food that is affordable and fair, while respecting the beliefs and cultures of different communities, thus allowing them to lead healthy, active, and productive lives sustainably [21]. Citing from [21], in Indonesia, food-insecure regencies are identified by several key indicators: high ratios of per capita consumption to net production, high numbers of poor people, high numbers of population per health worker to population density, high stunted children rates, and high numbers of households without access to clean water. Similarly, food-insecure municipalities are linked to high numbers of poor people, high stunted children rates, and high numbers of households without access to clean water, as well as lower life expectancy [21]. The study case focuses on the 2023 Food Security Index (FSI) data, specifically examining the situation in Papua, Indonesia.

Papua consistently reports the lowest FSI values in Indonesia, highlighting significant systemic challenges. For example, Papua Province has the highest percentage of households without electricity at 28%, particularly in Puncak and Dogiyai regencies, at 97% and 96%, respectively. It also had the highest poverty rate at 28% in 2019, followed by Papua Barat at 22% [21]. These factors worsen food insecurity by limiting storage and access to markets. Refer to data from [22], in 2023, Intan Jaya Regency in Papua Province had the lowest FSI at 14.54, down 2.67 from the previous year, while Tambrauw Regency in Papua Barat increased slightly to 32.88 [22]. The above data shows extreme outliers representing real crises (including geographic isolation or poverty). However, these outliers may lead to bias in regression models if they are not addressed.

Furthermore, the FSI in Papua exhibits fluctuations, as indicated in [22]. This variability may lead to outliers and influential observations within the data. Similarly, the previous analysis said that countries represented by outlier data points experienced a change in ranking after these points were winsorized, in which outliers were replaced with the closest available values to mitigate their impact on the 2016 Global Food Security Index (GFSI) [23]. Also, [24] did the same analysis using 2019 Kenya's GFSI. All of the prior research ignores the focus on outliers. Robust MCD-S, resistant to datasets with up to 50% outliers, provides a statistically reliable alternative [15], [25]. Until now, no research has applied MCD-S to FSI data despite its suitability for high-disparity regions like Papua. However, there has been no research on FSI

data using robust regression to address these outliers. So, this research aims to apply the regression of robust MCD-S to Papua's 2023 FSI data to handle outliers, as well as identify key determinants of FSI (for example, poverty or stunting) through hypothesis testing.

2 Literature Review

2.1 Robust with Minimum Covariance Determinant (MCD)

Citing from [12], [13], in an $n \times p$ data matrix $\mathbf{X} = \left(\mathbf{x}_1, ..., \mathbf{x}_n\right)^T$, where each observation $x_i = \left(x_{i1}, ..., x_{ip}\right)^T$, the goal is to identify *h* observations (where $\left\lfloor \left(n + p + 1/2\right) \right\rfloor \le h \le n$) that minimize the determinant of the covariance matrix. The MCD estimate of location or center, denoted as $\hat{\boldsymbol{\mu}}$, is the average of these *h* points (mean), and the estimate of scatter, denoted as $\hat{\boldsymbol{\Sigma}}$, is determined as a multiple of the covariance matrix. When *h* equals to $\left\lfloor \left(n + p + 1/2\right) \right\rfloor$, the MCD reaches its maximum breakdown value, which is 50% [12], [13].

The observation is represented as $\mathbf{x}_i \left(i = 1, ..., n\right)$, while $\mathbf{X}_j \left(j = 1, ..., p\right)$ represented the columns of the data matrix. For the data matrix $\mathbf{X}$, with estimated center, $\hat{\boldsymbol{\mu}}$, and scatter matrix, $\hat{\boldsymbol{\Sigma}}$, the statistical distance of the *i*-th observation $\mathbf{x}_i$ is written as in Equation (1) [12], [13].

$$D\left(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right) = \sqrt{\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}\right)^T \hat{\boldsymbol{\Sigma}}^{-1} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}\right)} \tag{1}$$

Robust MCD uses the FASTMCD algorithm [12], [13]. The first step is to calculate initial estimates for the center, $\hat{\boldsymbol{\mu}}_{old}$, and scatter matrix, $\hat{\boldsymbol{\Sigma}}_{old}$. Perform C-Step (Concentration Step) is the next step, which is written below [12], [13].

a. Calculate the distance for all data points $d_{old} = D\left(\mathbf{x}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right), i = 1, ..., n$ using Equation (1).
b. Sort the distance, resulting in a permutation π for which $d_{old}\left(\pi(1)\right) \le d_{old}\left(\pi(2)\right) \le ... \le d_{old}\left(\pi(n)\right)$, and select the subset *H* of size *h* corresponding to the distance $H = \left[\pi(1), \pi(2), ..., \pi(h)\right]$.
c. Calculate the new estimates for the center $\hat{\boldsymbol{\mu}}_{new}$ and scatter matrix $\hat{\boldsymbol{\Sigma}}_{new}$ using Equation (2) and (3).

$$\hat{\boldsymbol{\mu}}_{new} = \sum_{i \in H} \mathbf{x}_i / h \tag{2}$$

$$\hat{\boldsymbol{\Sigma}}_{new} = \sum_{i \in H} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{new}\right)\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{new}\right)^T / h - 1 \tag{3}$$

The C-Steps should be repeated in iterative steps until convergence (for example, until the determinant of $\hat{\boldsymbol{\Sigma}}_{new}$ is minimized). As there is no guarantee that the final outcome of the iteration step is the global minimum of the MCD objective function, an approximation to the MCD solution is found by first selecting a large number of *h*-subsets (typically 500) $H_1 \subset \left\{1, 2, ..., n\right\}$. C-steps are then applied to each subset, and the solution with the lowest overall determinant in $\hat{\boldsymbol{\Sigma}}_{new}$ is kept [12], [13].

The next step is to construct the initial subset. To create the initial subset H_1, draw a random subset *J* that contains $(p+1)$, then calculate $\hat{\boldsymbol{\mu}}_0 = \sum_{i \in J} \mathbf{x}_i / \left(p + 1\right)$ and $\hat{\boldsymbol{\Sigma}}_0 = \sum_{i \in J} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0\right)\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_0\right)^T / p$. If $\hat{\boldsymbol{\Sigma}}_0$ is singular, add more random points to *J* until it

becomes nonsingular. Next, the refinement process is performed by applying the C-step to the initial estimates ($\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0$) to compute ($\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1$). Repeat the C-step for each initial subset, applying it only twice, and for the ten subsets with the lowest determinant, continue to perform additional C-steps until convergence is achieved. The number of iterations should be limited to avoid excessive computations since all C-steps require the calculation of the covariance matrix and its inverse, as well as the distances associated with these covariance matrices [12], [13].

For final estimates, calculate raw FASTMCD estimates ($\hat{\boldsymbol{\mu}}_{rawMCD}, \hat{\boldsymbol{\Sigma}}_{rawMCD}$). Afterward, to increase statistical efficiency, calculate reweighted estimates ($\hat{\boldsymbol{\mu}}_{FASTMCD}, \hat{\boldsymbol{\Sigma}}_{FASTMCD}$) with weight $w_i = 1$ when $D\left(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{RAWMCD}, \hat{\boldsymbol{\Sigma}}_{RAWMCD}\right) \leq \sqrt{\chi^2_{p,0.975}}$ and 0 in other cases. $\chi^2_{p,\alpha}$ is the α-quantile of the χ^2_p distribution [12], [13].

MCD estimators have low asymptotic efficiencies, and to address this limitation, [15] suggests applying a one-step M-estimator to increase the asymptotic efficiency. Other robust estimators, such as S-estimator, can be used.

2.2 Robust with Minimum Covariance Determinant Scale (MCD-S)

MCD-S refers to a combination of the MCD estimator and the S-estimator approach. S-estimators exhibit similar asymptotic performances as regression M-estimators [25]. Citing from [25], the S-estimators minimize residual dispersion:

$$\underset{\hat{\theta}}{\text{Minimize}}\ s\left(r_1(\theta),...,r_n(\theta)\right) \tag{4}$$

where the final scale estimate $\hat{\sigma} = s(r_1(\hat{\theta}),\ldots,r_n(\hat{\theta}))$. The solution of the Equation (5) is what defines the dispersion $s\left(r_1(\theta),...,r_n(\theta)\right)$ [25].

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i}{s}\right) = K \tag{5}$$

K is frequently equivalent to $E_{\Phi}[\rho]$, with Φ representing the standard normal. The function ρ needs to meet several conditions: (1) it must be continuously differentiable and symmetric, and $\rho(0)$should equal 0; (2) there must be a positive constant c such that ρ is strictly increasing on the interval [0,c] and remains constant thereafter on [c,∞]; (3) extra condition $K/\rho(c) = 1/2$ [25]. As cited in [25], an example of ρ functions, is written in Equation (6). This has a derivative represented by Tukey's biweight function.

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c^2} + \dfrac{x^6}{6c^4}, & \text{for } |x| \leq c \\ \dfrac{c^2}{6}, & \text{for } |x| > c \end{cases}, \tag{6}$$

For Equation (6) with, to achieve a 50% breakdown point, it should be $c = 1,547$ and $K = 0,1995$[25].

Furthermore, the Equation (5) specifies an M-estimator for scale, typically calculated iteratively, beginning with an initial value $S^0 = 1,4826 \underset{i}{\text{med}}|r_i|$. Due to limitations on computation time, it might be suitable to choose a quick and straightforward approximation of the objective function, like the one-step estimate:

$$S^1 = S^0 \sqrt{\frac{1}{K}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i}{S^0}\right)\right\}} \tag{7}$$

with ρ is given in Equation (6) [25].

2.3 Data

This research used secondary data from the 2023 Food Security and Vulnerability Atlas (FSVA) report [22] and its website (https://fsva.badanpangan.go.id/). The FSVA was developed based on three key food security aspects: availability, affordability, and utilization. Based on these aspects, nine indicators are outlined below [21], [22].

1) Availability aspect:
   - The ratio of per capita normative consumption to food availability.
2) Affordability aspects:
   - the percentage of people living below the poverty line;
   - the percentage of households with a proportion of expenditure on food of more than 65 percent;
   - the percentage of households without access to electricity.
3) Utilization aspects:
   - The average length of schooling for women over 15 years;
   - The percentage of households without access to clean water;
   - The ratio of population per health worker to population density;
   - The percentage of stunted toddlers;
   - Life expectancy.

The population for this research includes the FSI from all regencies and cities in Papua as of 2023. This research used a total sampling technique, resulting in a sample consisting of all 42 regencies and municipalities in Papua, to maintain a geographically balanced representation.

This research uses dependent and independent variables to conduct a regression of robust MCD-S. The dependent variable in this research is the FSI (Y). The independent variables are based on nine indicators from the FSVA report [21], [22]. Those indicators are the ratio of normative consumption per capita to food availability (X_1), the percentage of people living below the poverty line (X_2), the percentage of households with a proportion of expenditure on food of more than 65 percent (X_3), the percentage of households without access to electricity (X_4), the average length of schooling for women over 15 years (X_5), the percentage of households without access to clean water (X_6), the ratio of population per health worker to population density (X_7), the percentage of stunted toddlers (X_8) and life expectancy (X_9).

2.4 Step Analysis

The analysis steps in this research include identifying multicollinearity, identifying outlier and influential observations, and estimating robust regression with MCD-S. All the steps were analyzed using the R software.

3 Result and Discussion

3.1 Identification of Multicollinearity

Multicollinearity is a condition in which two or more independent variables in a dataset correlate. This condition can lead to increased variance in the regression; thus, identifying this is important [18]. One approach to identifying multicollinearity is using the Variance Inflation Factor (VIF). Citing from [26], the calculation of VIF is written in Equation (8):

$$(VIF)_k = \left(1 - R_k^2\right)^{-1} \tag{8}$$

where R_k^2 is the determination coefficient. A VIF value that exceeds 10 is commonly seen as a sign that multicollinearity might significantly affect the estimates produced by Ordinary Least Squares (OLS) [26]. The results of the VIF values are displayed in Table 1.

Table 1. VIF Value of Independent Variables

	X_1	X_2	X_3	⋯	X_7	X_8	X_9
X_1	1	1.597	1.181	⋯	1.001	1.324	1.035
X_2	1.597	1	1.708	⋯	1.049	1.188	1.002
X_3	1.181	1.708	1	⋯	1.009	1.155	1.003
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
X_7	1.001	1.049	1.009	⋯	1	1.001	1.060
X_8	1.324	1.188	1.155	⋯	1.001	1	1.052
X_9	1.035	1.002	1.003	⋯	1.060	1.052	1

Table 1 indicates that all VIF values are below ten. These values suggest no multicollinearity in the FSI data in Papua for the year 2023.

3.2 Identification of Outliers and Influential Observations

An outlier is a data point that significantly differs from other values in the group, which may result from errors in data entry, inaccuracies in measurement systems, or unexpected crises. Additionally, if discarded, influential observations, which can alter the meaning of a regression model, are often outliers that may lead to non-normally distributed data and fail to generalize to broader contexts [3]–[10]. For this reason, it is important to identify outliers and influential observations in step analysis.

Citing from [27], a simple method for identifying outliers is creating a box plot, also known as a box-and-whisker diagram, visually representing data variability. The box plot displays the first and third quartiles at the bottom and top, with the median in the middle, and the whiskers indicate the range of data within 1.5 IQR of the quartiles. Any data point outside this range is marked as a potential outlier with a special symbol ('*') [27]. Figure 1 presents the box plot of ten variables (which are one dependent variable and nine independent variables).

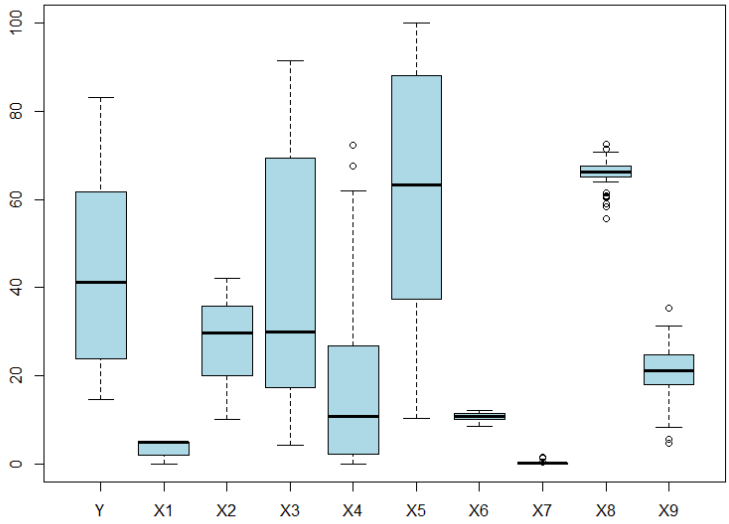

Figure 1. Box Plot of Ten Variables

As shown in Figure 1, there are observations outside the boxplot, specifically in variables X_4, X_7, X_8, and X_9. Outliers are therefore evident in those observations.

To identify if the *i*-th observation influences the regression model, DFFITS (Difference-in-fits) is applied using the predicted value from the regression model [28]. Before calculating the value of DFFITS, it is necessary to estimate the linear regression parameters by applying OLS. As cited [29], the estimation of regression parameters is written in Equation (9).

$$\widehat{\beta} = (X'X)^{-1}X'y \tag{9}$$

By applying the Equation (9), the coefficient of the linear regression parameter for the FSI in Papua for the year 2023 is displayed in Table 2.

The computed determination coefficient (R^2) of OLS is also displayed in Table 2. The value is computed using Equation (10) [29].

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{10}$$

R^2, which ranges between 0 and 1, measures how effectively the independent variable explains the dependent variable, with higher values indicative of a more reliable and accurate model [33].

Table 2. The Coefficient and R^2 Value of Linear OLS

		Full Model	**After Stepwise**
Coefficient	β_0	0.000	0.000
	β_1	-0.481	-0.502
	β_2	-0.169	-0.155
	β_3	-0.192	-0.176
	β_4	-0.117	-0.113
	β_5	-0.208	-0.218
	β_6	-0.029	-
	β_7	0.023	-
	β_8	0.070	0.070
	β_9	-0.028	-
R^2 *Value*		0.978	0.976

The next step is then calculating the value of DFFITS, which is written in Equation (11):

$$DFFITS_i = t_i \left(\frac{h_{ij}}{1-h_{ij}} \right)^{\frac{1}{2}} \quad ; t_i = \frac{\varepsilon_i}{\sqrt{s^2_{(i)}(1-h_{ij})}} \tag{11}$$

which $i = 1,2,...,n$; h_{ij} are matrix diagonal elements $H = (X'X)^{-1}X'$ (a $n \times n$ matrix); ε_i is the *i*-th residual; $s^2_{(i)} = \frac{(n-p)s^2 - \varepsilon_i^2 / (1-h_{ij})}{n - p_i - 1}$; and $s^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-p}$ [28], [30]. A $|DFFITS|$ value of greater than $2\sqrt{p_i/n}$ is considered influential for the *i*-th, which p_i is the number of parameters in the regression, including the intercept [28], [31], [32]. The DFFITS test indicated that the 29th and 42nd observations were influential, as their $|DFFITS|$ values (1.585 and 2.411, respectively) were greater than the value of $2\sqrt{p_i/n} = 2\sqrt{10/42} = 0.976$. Consequently, it is important to acknowledge the presence of outliers in the data and use the right estimation to address the problem.

3.3 Estimation of Robust with MCD-S

A robust regression with MCD-S is obtained by performing the C-Steps. To conduct C-Steps, it is mandatory to calculate the covariance matrix and statistical distance. The C-Steps procedure is thereafter reiterated through a series of iterative steps until convergence is achieved.

The covariance matrix for final estimates is presented in Equation (12).

$$Cov(X) = \begin{bmatrix} 1.398 & -1.124 & -1.138 & \dots & -0.147 & 0.724 & -0.351 \\ -1.124 & 1.123 & 0.761 & \dots & 0.106 & -0.616 & 0.384 \\ -1.138 & 0.761 & 1.344 & \dots & 0.126 & -0.433 & 0.016 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ -0.147 & 0.106 & 0.126 & \dots & 0.057 & -0.004 & -0.394 \\ 0.724 & -0.616 & -0.433 & \dots & -0.004 & 1.035 & 1.053 \\ -0.351 & 0.384 & 0.016 & \dots & -0.394 & 1.053 & 1.053 \end{bmatrix} \quad (12)$$

Meanwhile, the result of the distance values for final estimates using Equation (1) is displayed in Table 3.

Table 3. The Value of Statistical Distance

Observation	Distance Value
1	7.138
2	7.141
3	6.106
⋮	⋮
40	1024.016
41	50.965
42	259.169

To increase efficiency while maintaining high robustness, a weighting step can be applied. In this research, the Equation (6) is applied to the weight to find the estimator in the Equation (5), which is an S-estimator (or M-estimator for scale). The coefficient of the regression model using robust with MCD-S for the FSI in Papua for the year 2023 is then displayed in Table 4. The R^2 value, computed using Equation (10), is also shown in Table 4.

Table 4. The Coefficient and R^2 Value of Robust MCD-S

		Robust MCD-S
Coefficient	β_0	0.119
	β_1	-0.494
	β_2	-0.105
	β_3	-0.226
	β_4	-0.125
	β_5	-0.297
	β_6	-0.026
	β_7	0.167
	β_8	0.044
	β_9	-0.064
R^2 *Value*		0.999

The R^2 value, according to the result in Table 4, is 0.999. The MCD-S model demonstrates a higher value than the OLS model, which has an R^2 of 0.976 (shown in Table 2). This higher R^2 value indicates that the MCD-S model increases the overall accuracy and reliability of the analysis by effectively addressing outliers in the FSI in Papua for the year 2023. The value also suggests that the regression model accounts for 99.9% of the variability in strength, leaving 0.01% influenced by factors that this research did not consider.

The next step involves conducting significance tests for the regression parameters in Table 4, including both the simultaneous and partial tests. In a simultaneous regression analysis, the independent and dependent variables are evaluated simultaneously to determine whether they significantly affect one another [34]. An *F*-test statistic is used for this test, which can be computed using Equation (13) [35].

$$F = \frac{R^2/k-1}{\left(1-R^2\right)/n-k} \tag{13}$$

The null hypothesis, suggesting that no regression coefficients for the independent variables differ from zero $\left(\beta_1 = \beta_2 = ... = \beta_k = 0\right)$, is defined as rejected when the computed *F* value surpasses the critical value of $F_{(k-1;n-k;\alpha)}$ [35]. The computed *F* value is 14304.79, leading to the rejection of the null hypothesis, which is 2.494 ($F_{(9;17;0.05)}$). Based on this result, it can be concluded that at least one independent variable has a significant impact on the FSI (*Y*) in Papua for the year 2023.

Following that, the partial regression test, commonly referred to as the t-test, is conducted next. This test determines the significance of the effect of each independent variable on the dependent variable individually [34]. The *t*-test statistic is used for this test, which can be computed using Equation (14).

$$t = \frac{\hat{\beta}_i}{se\left(\hat{\beta}_i\right)} \tag{14}$$

The null hypothesis, suggesting that the parameter $\beta_i = 0$ (which i = 1,2,…,k), is defined as rejected when the computed $|t|$ value surpasses the critical value of $t_{\left(n-k;\alpha/2\right)}$ [29]. The results of the $|t|$ computation using both $\hat{\beta}$ from MCD-S and OLS are displayed in Table 5.

Table 5. The Value of *t*-test for Robust MCD-S and Linear OLS

Variable	$\|t\|$	
	Robust MCD-S	***Linear OLS***
X_1	-41.386 *	-10.559 *
X_2	-13.757 *	-3.290 *
X_3	-21.293 *	-4.197 *
X_4	-8.238 *	-3.382 *
X_5	-22.656 *	-2.843 *
X_6	-4.380 *	-0.929
X_7	5.540 *	0.750
X_8	7.007 *	2.108 *
X_9	-8.905 *	-0.898

According to the result shown in Table 5, when using MCD-S, it was found that all nine independent variables rejected the null hypothesis, as $t_{(17;0.025)} = 2.110$. It can, therefore, be concluded that these nine variables have a significant impact on the FSI (Y) in Papua for the year 2023.

In contrast, in Table 5, when using OLS, only six variables rejected the null hypothesis, excluding the percentage of households without access to clean water (X_6), the ratio of population per health worker to population density (X_7), and life expectancy (X_9). The result indicated that OLS missed some important variables.

These results show that MCD-S can provide a more reliable estimation in datasets containing outliers, such as Papua's FSI. It is therefore recommended that MCD-S be used for future research analyzing similar datasets.

In addition, the result of the hypothesis testing (partial test) using MCD-S in Table 5 aligns with the theory about nine indicators explained in the FSVA report [21], [22]. Therefore, it shows that the nine indicators used in the research significantly influence the FSI in Papua and are key determinants of FSI.

4 Conclusion

The analysis results suggest that there are outliers and influential observations within the dataset. To address these issues, robust regression with MCD-S is employed. This research result illustrated that MCD-S effectively addressed outliers within the FSI in Papua for the year 2023. It showed that unlike OLS, which is significantly affected by extreme values, MCD-S down-weighted outliers (with higher $R^2 = 99.9\%$), thus maintaining the consistency of the data. The MCD-S model also identified all nine independent variables as statistically significant. In contrast, the OLS model failed to identify important variables, such as the percentage of households without access to clean water (X_6), the ratio of population per health worker to population density (X_7), and life expectancy (X_9). The result supported the claim that MCD-S offers more reliable estimates in datasets characterized by extreme values, as seen in Papua's FSI. Therefore, for future research analyzing similar datasets, MCD-S is recommended.

References

[1] S. D. Kurniawan *et al.*, *Big Data: Mengenal Big Data & Implementasinya di Berbagai Bidang*. PT. Sonpedia Publishing Indonesia, 2024.

[2] S. Abdullah and T. E. Sutanto, *Statistika tanpa stres*. TransMedia, 2015.

[3] A. P. A. Pangesti, S. Sugito, and H. Yasin, "PEMODELAN REGRESI RIDGE ROBUST S, M, MM-ESTIMATOR DALAM PENANGANAN MULTIKOLINIERITAS DAN PENCILAN (Studi Kasus: Faktor-Faktor yang Mempengaruhi Kemiskinan di Jawa Tengah Tahun 2020)," *J. Gaussian*, vol. 10, no. 3, pp. 402–412, 2021, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/article/view/32799

[4] F. M. Barus and Sutarman, "Mendeteksi Outlier pada Data Multivariat dengan Metode Jarak Mahalanobis-Minimum Covariance Determinant (MMCD)," *IJM Indones. J. Multidiscip.*, vol. 1, no. 3, 2023, [Online]. Available: https://journal.csspublishing.com/index.php/ijm/article/view/287/200

[5] A. Husain and S. R. W. Jamaluddin, "Pemodelan Data Angka Kematian Bayi Menggunakan Regresi Robust," *J. Sains, Teknol. Komput.*, vol. 1, no. 1, pp. 1–7, 2024,

[Online]. Available: https://doi.org/10.56495/saintek.v1i1.326
[6] A. Pandu and K. Nisa, “PERBANDINGAN MVE-BOOTSTRAP DAN MCD-BOOTSTRAP DALAMANALISIS REGRESI LINEAR BERGANDA PADA DATA BERUKURAN KECIL YANG MENGANDUNG PENCILAN,” 2018. [Online]. Available: http://repository.lppm.unila.ac.id/11782/
[7] P. R. Sihombing, S. Suryadiningrat, D. A. Sunarjo, and Y. P. A. C. Yuda, “Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya,” *J. Ekon. Dan Stat. Indones.*, vol. 2, no. 3, pp. 307–316, Jan. 2023, doi: 10.11594/jesi.02.03.07.
[8] A. Adityaningrum, R. Resmawan, A. M. Brahim, D. R. Isa, L. O. Nashar, and A. Asriadi, "Robust Least Median of Square Modelling using Seemingly Unrelated Regression with Generalized Least Square on Panel Data for Tuberculosis Cases," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 18, no. 4, pp. 2293–2306, Oct. 2024, doi: 10.30598/barekengvol18iss4pp2293-2306.
[9] N. Nurcahyani, B. Pratikno, and S. Supriyanto, “FUNGSI PEMBOBOT TUKEY BISQUARE DAN WELSCH PADA REGRESI ROBUST ESTIMASI-S,” *J. Ilm. Mat. dan Pendidik. Mat.*, vol. 14, no. 2, p. 133, Dec. 2022, doi: 10.20884/1.jmp.2022.14.2.6668.
[10] M. Kurniadi, M. Aritonang, and M. N. Mara, “Mendeteksi Outlier Dengan Metode Minimum Covariance Determinant,” *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 1, no. 01, 2012, [Online]. Available: https://jurnal.untan.ac.id/index.php/jbmstr/article/view/651
[11] C. Croux and G. Haesbroeck, "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *J. Multivar. Anal.*, vol. 71, no. 2, pp. 161–190, Nov. 1999, doi: 10.1006/jmva.1999.1839.
[12] M. Hubert, M. Debruyne, and P. J. Rousseeuw, "Minimum covariance determinant and extensions," *WIREs Comput. Stat.*, vol. 10, no. 3, p. e1421, May 2018, doi: 10.1002/wics.1421.
[13] M. Hubert, P. J. Rousseeuw, and T. Verdonck, "A Deterministic Algorithm for Robust Location and Scatter," *J. Comput. Graph. Stat.*, vol. 21, no. 3, pp. 618–637, Jul. 2012, doi: 10.1080/10618600.2012.672100.
[14] D. Rosadi, E. P. Setiawan, M. Templ, and P. Filzmoser, "Robust covariance estimators for mean-variance portfolio optimization with transaction lots," *Oper. Res. Perspect.*, vol. 7, p. 100154, 2020, doi: 10.1016/j.orp.2020.100154.
[15] P. Rousseeuw, "Multivariate Estimation with High Breakdown Point," in *Mathematical Statistics and Applications*, Dordrecht: Springer Netherlands, 1985, pp. 283–297. doi: 10.1007/978-94-009-5438-0_20.
[16] C. Fauconnier and G. Haesbroeck, "Outliers detection with the minimum covariance determinant estimator in practice," *Stat. Methodol.*, vol. 6, no. 4, pp. 363–379, Jul. 2009, doi: 10.1016/j.stamet.2008.12.005.
[17] N. A. Balqis, “Analisis Regresi Komponen Utama Robust Menggunakan Metode Minimum Covariance Determinant (MCD) Dengan Pendekatan MM-Estimator Untuk Mengatasi Multikolinieritas Dan Beragam Proporsi Tingkat Pencilan,” Universitas Brawijaya, 2021. [Online]. Available: https://repository.ub.ac.id/id/eprint/184847/
[18] S. D. A. Larasati, K. Nisa, and E. Setiawan, “Analisis Regresi Komponen Utama Robust dengan Metode Minimum Covariance Determinant – Least Trimmed Square (MCD-LTS),” *J. Siger Mat.*, vol. 1, no. 1, Mar. 2020, doi: 10.23960/jsm.v1i1.2472.
[19] S. Wulandari, N. Salam, and D. Anggraini, “PERBANDINGAN METODE ROBUST MCD-LMS, MCD-LTS, MVE-LMS, DAN MVE-LTS DALAM ANALISIS REGRESI KOMPONEN UTAMA,” *Epsil. J. Mat. MURNI DAN Terap.*, vol. 4, no. 1, pp. 57–64,

2010.
[20] A. Criszardin, “ROBUST PRINCIPAL COMPONENT REGRESSION DENGAN METODE MINIMUM COVARIANCE DETERMINANT (MCD) DAN MINIMUM VOLUME ELLIPSOID (MVE) MENGGUNAKAN ESTIMATOR LEAST TRIMMED SQUARE (LTS) (STUDI KASUS: DATA KEMISKINAN INDONESIA MENURUT PROVINSI PADA TAHUN 2022),” UIN SUNAN KALIJAGA YOGYAKARTA, 2024. [Online]. Available: https://digilib.uin-suka.ac.id/id/eprint/63758/
[21] Badan Ketahanan Pangan Kementerian Kesehatan, “FOOD SECURITY AND VULNERABILITY ATLAS 2020: Data Indikator Tahun 2019,” 2020. [Online]. Available: https://badanpangan.go.id/storage/app/media/2021/fsva-2020-202101261020fix.pdf
[22] Deputi Bidang Kerawanan Pangan dan Gizi Badan Pangan Nasional, “FOOD SECURITY AND VULNERABILITY ATLAS 2023: Data Indikator Tahun 2022,” 2023. [Online]. Available: https://drive.google.com/file/d/1tWwYUXmE2fKEkCr5ogzPzXTINJs4hFsj/view
[23] A. Thomas, B. D'Hombres, C. Casubolo, F. Kayitakire, and M. Saisana, *The use of the Global Food Security Index to inform the situation in food insecure countries*. European Commission: Joint Research Centr, 2017. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/29b5de92-f103-11e7-9749-01aa75ed71a1/language-en
[24] P. Atieno and S. L. Hendriks, "The effects of outdated data and outliers on Kenya's 2019 Global Food Security Index score and rank," *CABI Agric. Biosci.*, vol. 4, no. 1, p. 6, Mar. 2023, doi: 10.1186/s43170-023-00140-y.
[25] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Wiley, 1987. doi: 10.1002/0471725382.
[26] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied linear statistical models*. McGraw-hill, 2005.
[27] N. N. R. R. Suri, N. M. M, and G. Athithan, *Outlier Detection: Techniques and Applications: A Data Mining Perspective*. Springer, 2019. [Online]. Available: https://books.google.co.id/books?id=wTSDDwAAQBAJ&printsec=copyright&redir_esc=y#v=onepage&q&f=false
[28] 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
[29] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. United Kingdom: Wiley, 2012.
[30] 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
[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] 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.
[33] M. K. Najib *et al.*, “Prediksi Angka Harapan Hidup Menggunakan Regresi Linear Berganda, Lasso, Ridge, Elastic Net, dan Kuantil Lasso,” *J. Sains Mat. dan Stat.*, vol. 10, no. 2, 2024, doi: 10.24014/jsms.v10i2.27916.

[34] M. S. Priyono, *Analisis Regresi dan Korelasi untuk Penelitian Survei (Panduan Praktis Olah Data dan Interpretasi: D*. GUEPEDIA, 2021.
[35] D. N. Gujarati, *Econometrics by Example*. Macmillan Education Palgrave, 2015. [Online]. Available: https://books.google.co.id/books?id=ONpdyQEACAAJ