J. Pijar MIPA, Vol. 18 No. 5, September 2023: 699-706 ISSN 1907-1744 (Print)
DOI: 10.29303/jpm.v18i5.5643 ISSN 2460-1500 (Online)

# MODELING THE PROPORTION OF MEASLES CASES USING SPARSE LEAST TRIMMED SQUARES

**Shelly Kilan Cahaya Pulungan and Rina Filia Sari***
Department of Mathematics, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
*Email: rinafiliasari@uinsu.ac.id

Received: September 8, 2023. Accepted: September 14, 2023. Published: September 25, 2023

**Abstract:** Measles is a highly contagious disease and a health problem in several countries, including Indonesia. In 2022, Indonesia will experience an extraordinary situation (KLB) of measles cases, with the number of measles cases reaching 3,341 across 223 districts/cities. This data shows an increase of 32 times compared to 2021. North Sumatra is one of the provinces included in the list of regions and outbreak status, with 127 measles cases recorded in 2022. This study aims to find the factors that influence the number of measles cases in North Sumatra: one dependent variable, 34 independent variables, and 33 observations made up the study's variables. The data model chosen contains information on the percentage of measles cases linked to health, economic, human resource, and environmental variables. In addition, this study employs high-dimensional (data with many explanatory factors) data and includes outliers. Data with a large number of explanatory factors and outliers can be handled with LTS sparse analysis. The 34 independent variables were successfully chosen and reduced to 14 using the LTS sparse model. In addition, based on the $R^2$ and RMSE values for model evaluation, sparse LTS shows satisfactory results compared with classical LASSO, with $R^2$ and RMSE values for sparse LTS being 93.75% and 0.2933, respectively. Then, the $R^2$ and RMSE values for LASSO are -62.4% and 2.1734. The government can use these elements to guide lowering the number of measles cases in North Sumatra.

**Keywords:** *Measles, Outliers, LASSO, Robust Regression, Sparse Regression*

## INTRODUCTION

Measles is a disease that can be prevented by immunization [1]. Measles is also called Morbilli, a highly contagious disease of the genus Morbillivirus, and belongs to the RNA virus group (Ministry of Health RI, 2018) [2]. Measles is spread worldwide and is included in the ten most infectious diseases in several developing countries, including Indonesia (RI Ministry of Health, 2022) [3]. The Indonesian Child Protection Commission (KPAI, 2023) reported 55 extraordinary incidents in 34 districts/cities in 12 provinces. Based on the release of the Ministry of Health of the Republic of Indonesia (Kemenkes RI), the number of measles cases in 2022 was reported to have reached 3,341 cases spread across 223 districts/cities. This data has increased 32 times compared to 2021 [4].

In 2022, North Sumatra will become one of the provinces that has designated measles as an extraordinary event. According to data from the North Sumatra Provincial Health Service, 127 positive cases of measles were recorded in 2022. The city of Medan had the highest number of positive measles cases, namely 66. Then followed by Deliserdang with 14 cases and Batubara with 8 cases. Next, Serdangbedagai, Langkat and Sibolga each had 6 cases. Furthermore, Tebingtinggi 4 positive cases. Central Tapanuli and Binjai City each had 3 cases. North Labuhanbatu and Simalungun 2 Cases. Then, South Tapanuli, Labuhanbatu, Nias, Samosir, Padang Lawas, and Gunung Sitoli City each had 1 case. It is suspected because the data contains outliers [5]. Therefore, more research is required to determine the variables that affect the percentage of measles cases in North Sumatra [6].

Finding the variables that affect the proportion of measles cases can be done using regression analysis. Ordinary least squares (OLS), a mathematical method, is one way to obtain the regression coefficient. High-dimensional data is not compatible with OLS because the estimations that result will be erroneous. Because LASSO regression may select variables by lowering the regression coefficient to zero, it could be used to solve the problem of high-dimensional data [7]. However, the existence of outliers can have an impact on LASSO. A robust strategy is advised in cases where the data contains outliers. A technique that can address these two issues simultaneously is LASSO robust regression to obtain a simplified model. Therefore, a technique that combines strong regression with LASSO regression is required. According to some research, Sparse Least Trimmed Squares (Sparse LTS) is the suggested approach [8].

This research aims to solve the problem of high-dimensional data and outliers. Using the Sparse LTS approach, we succeeded in selecting independent variables from 34 to 14 independent variables included in the Sparse LTS modeling, and the results of calculating $R^2$ with R-Studio show that $R^2$ for Sparse LTS is 93.75% and for LASSO is -62.4%. It shows that Sparse LTS is better to use than the LASSO method. The findings of this study can be a useful resource for governments to focus on issues that significantly influence the proportion of measles cases.

**Sparse Least Trimmed Squares (Sparse LTS)**

Sparse least trimmed squares combine the strong method and the sparse estimation method. Sparse LTS can control data that is high in dimension and has outliers. High-dimensional data has greater explanatory variables than observations [8]. Consider a regression of the response y on a matrix design assuming a linear relationship between the explanatory variable $X \in \mathbb{R}^{n\times p}$ and the response variable y $y \in \mathbb{R}^{n\times p}$,

$$y = X\beta + \varepsilon \quad (1)$$

Where the regression coefficient is $\beta = (\beta_{1,,}\beta_2, \dots, \beta_p)^T$ and ε is an error that has a zero expectation value. With the penalty parameter α, the LASSO estimate of β is as follows:

$$\hat{\beta}_{\text{Lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - X_i'\beta)^2 + n\alpha \sum_{j=1}^{p}|\beta_j| \quad (2)$$

LASSO regression uses the $L_1$normalization technique to estimate regression coefficients, which can shrink the regression coefficients of variables that have a high correlation with error, with the aim of the regression coefficient being close to zero or equal to zero. So, the LASSO method can play a role in variable selection while overcoming multicollinearity [7]. However, the LASSO regression is not resistant to outlier data. An outlier is an observation whose observation point deviates from the data pattern. The presence of outliers can cause large residuals. So, a robust regression method is needed to handle this case [10].

The Least Trimmed Squares (LTS) method is a High Breakdown Value method, an alternative method to overcome the weaknesses of the Ordinary Least Squares (OLS) method. A robust regression indicator estimation technique is the LTS approach. The most often used robust regression estimator, this estimate has a straightforward specification and may be computed quickly [11].

$$\hat{\beta}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h}(r_{(i)}^2(\beta))_{i:n} \quad (3)$$

Where $r_{(i)}^2$is the squared residual ordered from smallest to largest $(r_{(1)}^2 \le r_{(2)}^2 \le \cdots \le r_{(n)}^2$and $h \le n,$, for $r_i = (y_i - x_i^T\beta),\ i = 1, \dots, n$. LTS has the same principle as the OLS method in estimating regression parameters, namely minimizing the number of residuals. However, the LTS method does not use all observations in its calculations but only minimizes the sum of residual squares from a subset of data of size h. Observations with the smallest residual squares only work $h < n$ or $n > p$. Thus, if an observation contains $n < p$, it is proposed to continue the fast-LTS algorithm for sparse data by adding a penalty $l_1$ with parameter α to the LTS estimation coefficient, which leads to sparse LTS estimation [8]. The form of the sparse least trimmed squares equation is as follows:

$$\hat{\beta}_{sparseLTS} = \arg\min_{\beta}\{\sum_{i=1}^{h} r_{(i)}^2(\beta) + h\alpha\|\beta\|_1\} \quad (4)$$

For $h \le n$ and tuning parameter $\alpha \ge 0$

**Criteria for Types of Data Analysis Techniques**

The coefficient of determination $(R^2)$ and RMSE (Root Mean Square Error), used as comparison parameters to choose the best model, are the parameters for the data analysis technique used. The value of the $R^2$ coefficient of determination and the RMSE determines the best model. The form of the RMSE and $R^2$ equations is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_1 - \hat{Y}_i)^2} \quad (5)$$

$$R^2 = \frac{\sum_{i=1}^{n}(Y_1 - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_1 - \bar{Y}_i)^2} \quad (6)$$

Where n is the number of observations, $\hat{Y}_i$ is the prediction of the ith response, $\hat{Y}_i$ is the value of the ith response variable, and $\hat{Y}_i$ is the average value of the response variable. The best model if it has a coefficient of determination $(R^2)$ that is more significant and has a smaller value on RMSE.

**RESEARCH METHODS**

This research begins with an initial literature study, namely data collection on the proportion of measles cases in North Sumatra. It then begins with applying LASSO and compares it with LTS sparse. R and SPSS software were used to analyze data in this research. The next stage is dataset detection using boxplots. Then, use 5-fold cross-validation to observe the predicted model parameters derived from the data on the proportion of measles cases. Then, conduct LASSO regression analysis to examine the sparse LTS model parameter estimates to choose the optimal model. λ values for LTS rarely use 3-fold Cross-Validation. Next, the final stage compares the LASSO and Sparse LTS estimation results by calculating. $R^2$ and RMSE for each estimate and then concluding the results of the analysis carried out. The steps in this research can be explained as follows.

The percentage of measles cases in North Sumatra in 2022 was the subject of data analysis. Thirty-four independent variables and 33 observations make up the dataset, which has one dependent variable. Statistics on the prevalence of measles connected to economic, environmental, human resource, and health statistics are all included in the data structure. The dataset was obtained through a documentation study using official documents from

the North Sumatra Health Office and from the official website of the North Sumatra Central Statistics Agency (BPS).

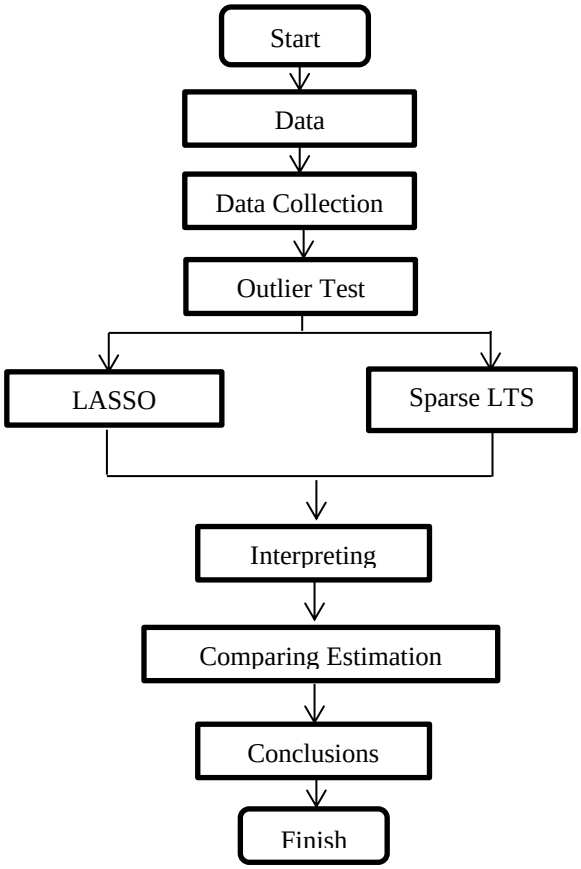


Figure 1. Research procedure

Table 1. Data Description

| Variable | Explanation | Denomination |
|---|---|---|
| $Y$ | Proportion of Measles Cases | Percent |
| $X_1$ | Measles Immunization Coverage Percentage | Percent |
| $X_2$ | Prevalence of Stunting Toddlers | People |
| $X_3$ | Number of Children Getting Coverage of Vitamin A Aged 6-11 Months | People |
| $X_4$ | Number of Children Getting Coverage of Vitamin A Aged 12-59 Months | People |
| $X_5$ | Percentage of Province Area | Percent |
| $X_6$ | Number of Babies Born | People |
| $X_7$ | Low Weight Baby (LWB) | Percent |
| $X_8$ | Number of General Hospitals | Unit |
| $X_9$ | Number of Specialized Hospitals | Unit |
| $X_{10}$ | Number of Clinics | Unit |
| $X_{11}$ | Number of Community Health Centers | Unit |
| $X_{12}$ | Number of Integrated Healthcare Center | Unit |
| $X_{13}$ | Number of Doctors | People |
| $X_{14}$ | Number of Midwives | People |
| $X_{15}$ | Number of Nurses | People |
| $X_{16}$ | Number of pharmacists | People |
| $X_{17}$ | Number of Nutritionists | People |
| $X_{18}$ | Number of Public Health Workers | People |
| $X_{19}$ | Number of Environmental Health Workers | People |
| $X_{20}$ | Number of Biomedical Engineering Personnel | People |
| $X_{21}$ | Long School Expectations | Index |
| $X_{22}$ | Gross Regional Domestic Product (GRDP) | Billion Rupiah |

| | | |
|---|---|---|
| $X_{23}$ | GRDP Growth Rate | Index |
| $X_{24}$ | Percentage of Poor Population | Percent |
| $X_{25}$ | Diphtheria, Pertussis and Tetanus (DPT) Immunization Percentage | Percent |
| $X_{26}$ | Open Unemployment Rate (TPT) | People |
| $X_{27}$ | Labor Force Participation Rate (TPAK) | People |
| $X_{28}$ | Human Development Index | Index |
| $X_{29}$ | Number of Villages/Subdistricts | Unit |
| $X_{30}$ | Percentage of Households Having Access to Adequate Sanitation | Percent |
| $X_{31}$ | Population density | People/$m^2$ |
| $X_{32}$ | Percentage of Babies Who Get Exclusive Breast Milk | Percent |
| $X_{33}$ | Number of Districts | Unit |
| $X_{34}$ | Percentage of Households that Have Access to an Improper Source of Drinking Water | Percent |

## RESULT AND DISCUSSION

### Descriptive Statistics

All of the variables utilized in this investigation are shown in Figure 2, along with descriptive statistics and the Pearson correlation matrix. Where the value that is highlighted is considered to be the true correlation value and whose value is below the 0.05 level of significance. It can be seen that numerous explanatory variables have relationships with a p-value of less than 0.05, indicating that the connection between these variables is significant. As a result, the data set employed has a multicollinearity issue.

| | Mean | SD | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | X13 | X14 | X15 | X16 | X17 | X18 | X19 | X20 | X21 | X22 | X23 | X24 | X25 | X26 | X27 | X28 | X29 | X30 | X31 | X32 | X33 | X34 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 78,6 | 15,3 | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X2 | 23,5 | 7,4 | -0,2 | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X3 | 1106,2 | 728,6 | .460** | -0,3 | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X4 | 18175,7 | 19381,8 | .352* | -0,3 | .793** | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X5 | 3,0 | 2,5 | 0,3 | 0,3 | 0,3 | 0,1 | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X6 | 8427,3 | 8690,9 | .392* | -0,3 | .725** | .948** | 0,2 | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X7 | 35,4 | 48,8 | 0,1 | 0,0 | .418* | 0,1 | 0,3 | 0,1 | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X8 | 5,5 | 9,5 | 0,2 | -.394* | 0,3 | .785** | -0,2 | .758** | 0,0 | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X9 | 0,5 | 2,4 | 0,1 | -0,2 | 0,1 | .629** | -0,2 | .583** | -0,1 | .932** | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | | |
| X10 | 33,9 | 59,7 | 0,3 | -.372* | .558** | .909** | -0,1 | .916** | 0,0 | .921** | .798** | 1,0 | | | | | | | | | | | | | | | | | | | | | | | | |
| X11 | 18,6 | 10,1 | .480** | -0,2 | .627** | .656** | .515** | .732** | .439* | .523** | .419* | .588** | 1,0 | | | | | | | | | | | | | | | | | | | | | | | |
| X12 | 476,2 | 388,3 | .475** | -0,3 | .707** | .781** | .476** | .884** | 0,3 | .604** | .419* | .741** | .862** | 1,0 | | | | | | | | | | | | | | | | | | | | | | |
| X13 | 276,1 | 695,4 | 0,1 | -0,3 | 0,2 | .702** | -0,2 | .671** | 0,0 | .969** | .989** | .861** | .478** | .516** | 1,0 | | | | | | | | | | | | | | | | | | | | | |
| X14 | 796,2 | 502,9 | .411* | -0,2 | .595** | .766** | .520** | .872** | 0,2 | .666** | .530** | .747** | .849** | .898** | .614** | 1,0 | | | | | | | | | | | | | | | | | | | | |
| X15 | 757,7 | 1266,0 | 0,2 | -0,3 | 0,2 | .713** | -0,1 | .695** | 0,0 | .972** | .975** | .865** | .539** | .546** | .990** | .657** | 1,0 | | | | | | | | | | | | | | | | | | |
| X16 | 112,4 | 208,5 | 0,2 | -0,3 | 0,3 | .731** | -0,2 | .707** | 0,0 | .976** | .976** | .883** | .507** | .534** | .992** | .642** | .993** | 1,0 | | | | | | | | | | | | | | | | | |
| X17 | 45,7 | 53,9 | 0,2 | -0,3 | 0,3 | .763** | -0,1 | .738** | 0,0 | .957** | .940** | .888** | .538** | .566** | .962** | .658** | .965** | .978** | 1,0 | | | | | | | | | | | | | | | | |
| X18 | 94,5 | 78,3 | 0,3 | -0,2 | .502** | .844** | 0,1 | .880** | 0,0 | .837** | .716** | .883** | .618** | .717** | .776** | .827** | .803** | .814** | .805** | 1,0 | | | | | | | | | | | | | | | |
| X19 | 29,4 | 40,4 | 0,2 | -0,3 | 0,3 | .712** | -0,1 | .685** | 0,0 | .943** | .961** | .851** | .536** | .536** | .971** | .626** | .965** | .968** | .966** | .755** | 1,0 | | | | | | | | | | | | | | |
| X20 | 89,5 | 214,5 | 0,1 | -0,3 | 0,2 | .701** | -0,2 | .663** | 0,0 | .968** | .988** | .859** | .466** | .496** | .998** | .594** | .990** | .995** | .970** | .773** | .969** | 1,0 | | | | | | | | | | | | | |
| X21 | 13,3 | 0,6 | -0,1 | 0,0 | -0,1 | 0,2 | -0,2 | 0,1 | -0,2 | .439* | .484** | 0,3 | -0,1 | -0,1 | .467** | 0,1 | .452** | .473** | .422* | .359* | .387* | .471** | 1,0 | | | | | | | | | | | | |
| X22 | 17637,4 | 30124,0 | 0,3 | -.352* | .465** | .868** | 0,0 | .859** | 0,0 | .959** | .899** | .950** | .609** | .722** | .942** | .754** | .940** | .945** | .931** | .864** | .923** | .934** | 0,3 | 1,0 | | | | | | | | | | | |
| X23 | 4,2 | 0,6 | 0,3 | 0,1 | .454** | .438* | .458** | .428* | 0,0 | 0,3 | 0,2 | .352* | 0,3 | .488** | 0,2 | .455** | 0,2 | 0,2 | 0,2 | .366* | 0,3 | 0,2 | 0,1 | .373* | 1,0 | | | | | | | | | | |
| X24 | 10,3 | 4,5 | -0,3 | 0,0 | .464* | -.413* | -0,3 | -.383* | 0,0 | -0,3 | -0,1 | -.350* | -0,2 | -.366* | -0,2 | -0,3 | -0,2 | -0,2 | -0,2 | -.353* | -0,2 | -0,2 | -0,3 | -0,3 | .795* | 1,0 | | | | | | | | | |
| X25 | 77,3 | 15,3 | .811** | -0,1 | 0,2 | 0,1 | 0,1 | 0,1 | -0,1 | 0,1 | 0,1 | 0,1 | 0,3 | 0,2 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,0 | 0,1 | 0,0 | -0,3 | 0,1 | 0,1 | 0,0 | 1,0 | | | | | | | | |
| X26 | 4,6 | 2,8 | 0,2 | -0,3 | .374* | .495** | 0,0 | .509** | 0,0 | .492** | 0,3 | .504** | .349* | .422* | .377* | .536** | .418* | .412* | .426* | .567** | .404* | .368* | 0,3 | .453** | 0,3 | -.389* | -0,1 | 1,0 | | | | | | | |
| X27 | 73,3 | 8,1 | 0,0 | .485** | -0,3 | -.409* | 0,0 | -.376* | 0,0 | .454* | -0,3 | -.417* | -0,2 | -0,3 | -0,3 | -.382* | -0,3 | -0,3 | -0,3 | -.424* | -0,3 | -0,3 | -0,3 | -.401* | -.350* | .391* | 0,2 | .754* | 1,0 | | | | | | |
| X28 | 71,8 | 4,4 | 0,1 | -0,3 | 0,3 | .436* | -0,2 | .380* | -0,1 | .561** | .451** | .519** | 0,2 | 0,3 | .487** | 0,3 | .468** | .504** | .503** | .448** | .495** | .485** | .554** | .486** | .555** | .738* | 0,0 | .474** | .490* | 1,0 | | | | | |
| X29 | 185,2 | 126,2 | .353* | 0,2 | .545** | .352* | .696** | .426* | .529** | 0,0 | 0,0 | 0,2 | .703** | .549** | 0,0 | .586** | 0,0 | 0,0 | 0,1 | 0,3 | 0,1 | 0,0 | -0,3 | 0,1 | 0,1 | -0,1 | 0,2 | 0,1 | 0,1 | -0,2 | 1,0 | | | | |
| X30 | 73,5 | 23,6 | 0,1 | -0,2 | 0,3 | 0,3 | 0,0 | 0,3 | -0,1 | 0,3 | 0,2 | 0,3 | 0,1 | 0,3 | 0,2 | 0,1 | 0,2 | 0,2 | 0,2 | 0,1 | 0,2 | 0,2 | 0,1 | 0,3 | .544** | .620* | 0,1 | 0,0 | -0,1 | .632** | -0,2 | 1,0 | | | |
| X31 | 1044,1 | 1933,7 | -0,2 | -.378* | -0,1 | 0,3 | .548* | 0,3 | -0,2 | .742** | .735** | .559** | 0,0 | 0,1 | .732** | 0,2 | .730** | .733** | .698** | .454** | .698** | .746** | .479** | .589** | 0,0 | -0,1 | -0,2 | .494** | -.435* | .596** | -.399* | 0,2 | 1,0 | | |
| X32 | 46,6 | 20,3 | 0,0 | 0,2 | 0,1 | 0,0 | 0,0 | -0,1 | 0,2 | -0,1 | -0,1 | -0,1 | 0,0 | -0,2 | -0,1 | -0,1 | -0,1 | -0,1 | 0,0 | 0,0 | 0,0 | -0,1 | 0,0 | -0,2 | 0,1 | -0,2 | -0,1 | 0,0 | 0,0 | 0,1 | 0,1 | 0,0 | -0,1 | 1,0 | |
| X33 | 13,8 | 7,9 | .370* | 0,0 | .534** | .441* | .589** | .525** | .467** | 0,2 | 0,2 | 0,3 | .896** | .710** | 0,2 | .691** | 0,3 | 0,2 | 0,3 | .424* | 0,3 | 0,2 | -0,3 | 0,3 | 0,2 | -0,1 | 0,2 | 0,2 | 0,0 | -0,1 | .852** | -0,1 | -0,2 | 0,1 | 1,0 |
| X34 | 86,5 | 12,6 | 0,2 | -.434* | .391* | .368* | 0,0 | 0,3 | 0,0 | .352* | 0,2 | .378* | 0,2 | .389* | 0,3 | 0,2 | 0,2 | 0,3 | 0,3 | 0,2 | 0,3 | 0,2 | 0,0 | .352* | .467** | -.579* | 0,1 | 0,3 | -.363* | .695** | -0,1 | .802** | .351* | 0,1 | 0,0 | 1,0 |

Figure 2. Descriptive Statistics and Pearson Correlation

### Outliers Detection

Box plot analysis is a way to detect outliers that are depicted graphically from numerical data. The following is a box plot analysis with the help of SPSS software. The box plot output is shown in Figure 3, which shows the existence of outlier and extreme data. Data number 19, marked with a circle (°), is an outlier data. While 12 and 30, with an asterisk (*), are extreme (Figure 3).

### Analysis of LASSO

Before looking for the LASSO regression coefficient, determine $\lambda$ using K-Fold Cross Validation. A model validation technique called cross-validation, known as rotation estimation, determines how well statistical analysis results generalize to different data sets. One such method is k-fold cross-validation, which divides the data into K equal-sized parts (Rahayu & Husein, 2023). The K-fold cross-validation formula is:

ISSN 1907-1744 (Print)
ISSN 2460-1500 (Online)

$$CV_K = \frac{1}{k}\sum_{i=1}^{k}(evaluasi_i) \quad (7)$$

Where is the evaluation value at iteration *i*, the MSE value is used as the evaluation value, and with the help of RStudio, a plot is displayed depicting the mean squared error (MSE) against Log(λ) in LASSO estimation. 5-fold cross-validation was carried out to select λ with the lowest MSE, namely λ=0.3031107.

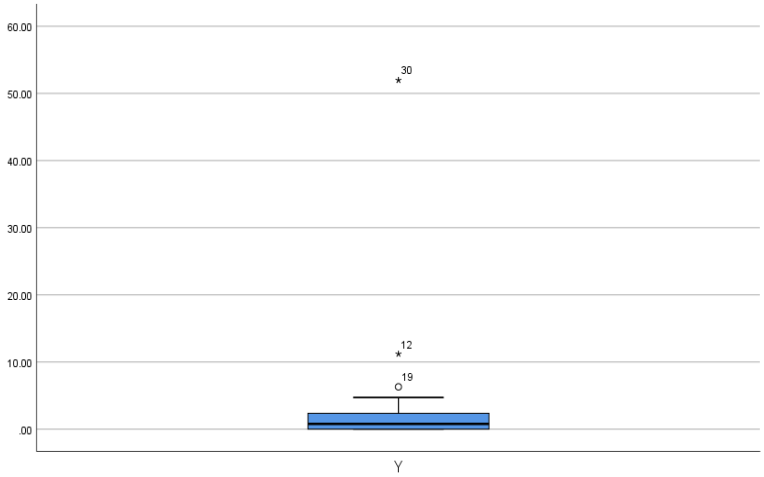


Figure 3. Boxplot for response variable (proportion of Measles cases)

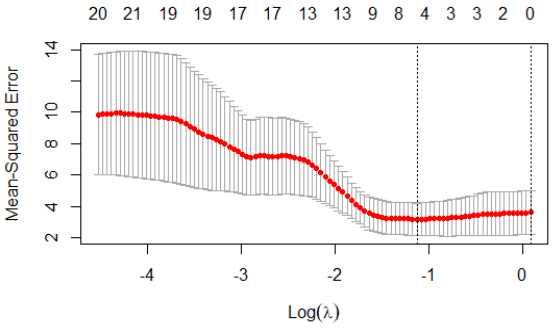


Figure 4. Cross-validation estimates of LASSO mean squared errors

Next, calculate and visualize the best model coefficients for each lambda value evaluated during cross-validation. By conducting regularization path analysis using the LASSO (L1) penalty method, we will calculate the coefficients of the linear regression model with various levels of regularization parameter λ. Figure 5 demonstrates how the L1-norm tends to drop as the LASSO regression seeks to bring the regression coefficient to zero. The total sum of the non-zero coefficients is known as the L1-norm.

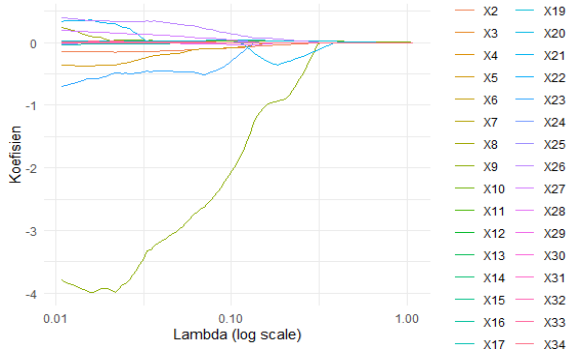


Figure 5. Regression coefficients of LASSO plot

Then, using LASSO regression, the regression coefficient is reduced to zero while explanatory factors are chosen, ensuring that only significant explanatory variables are incorporated into the regression model. Outcomes from many variables influencing the percentage of measles cases in North Sumatra were obtained after LASSO analysis. It can be seen from the variable coefficient. Variables that have non-zero coefficients are variables that influence the proportion of measles cases in North Sumatra. With the help of RStudio software, The following table shows the values of each explanatory variable's coefficients as determined by the LASSO analysis.

Table 2. Regression Coefficient LASSO Estimation Results

| Variable | Coefficient |
|---|---|
| Intercept | 1.370224e+00 |
| $X_2$ | -1.368564e-03 |
| $X_{10}$ | 1.919709e-02 |
| $X_{19}$ | 2.093945e-02 |
| $X_{21}$ | -9.571121e-02 |
| $X_{22}$ | 3.430933e-05 |
| $X_{26}$ | 3.276383e-08 |

**Analysis of Sparse Least Trimmed Squares**

LASSO regression is paired with one of the most well-liked strong regression estimators to handle high-density data and outliers simultaneously, namely Least Trimmed Squares (LTS), to form the Robust LASSO estimator or what is known as Sparse Least Trimmed Squares. It is using RStudio results in a sparse analysis of LTS model data. The following LTS sparse analysis results can be seen in Table 3. There is a linear regression coefficient that shrinks to zero. A variable that does not significantly affect the response variable is an explanatory variable with a value of zero. Fourteen variables in the LTS sparse

ISSN 1907-1744 (Print)
ISSN 2460-1500 (Online)

model can be used to clarify the percentage of measles cases in North Sumatra.

Table 3. Sparse LTS Estimation Coefficient of Regression

| Variable | coefficient |
|---|---|
| Intercept | 8.219036e-01 |
| $X_2$ | -2.739158e-02 |
| $X_3$ | 1.187109e-04 |
| $X_9$ | -1.228508e+00 |
| $X_{10}$ | -1.926760e-04 |
| $X_{11}$ | -1.576392e-03 |
| $X_{16}$ | -3.559100e-03 |
| $X_{18}$ | 7.765488e-03 |
| $X_{19}$ | 2.617175e-02 |
| $X_{20}$ | -1.638111e-04 |
| $X_{22}$ | -3.069781e-05 |
| $X_{26}$ | 7.214168e-02 |
| $X_{30}$ | -5.637282e-03 |
| $X_{32}$ | 1.133120e-02 |
| $X_{33}$ | -4.564364e-02 |

Furthermore, table 4 shows the selection of λ values in the LTS sparse model through 3-fold cross-validation. Nine values of λ are evaluated to choose the optimal value that gives the minimum prediction error.

Table 4. Results via 3-fold cross-validation of λ and RMSE values

| | Lambda (λ) | *RMSE* |
|---|---|---|
| 1 | 0.01000000 | 7.687137 |
| 2 | 0.03162278 | 7.484293 |
| 3 | 0.10000000 | 7.524574 |
| 4 | 0.31622777 | 7.354384 |
| 5 | 1.00000000 | 6.954277 |
| 6 | 3.16227766 | 6.954277 |
| 7 | 10.00000000 | 6.954277 |
| 8 | 31.62277660 | 6.954277 |
| 9 | 100.00000000 | 6.954277 |

Figure 6 is a plot that shows the relationship between the log10 value of the λ parameter and the RMSE (Root Square Error) value using the sparse LTS method through 3-fold cross-validation. 9 λ values are obtained, which are evaluated to select the optimal value that gives the minimum prediction error. The Log(λ) value chosen with a low RMSE is log(10) = 1 with the lowest RMSE being 6.954277.

Then, perform residual analysis of the regression model on the LTS sparse model, which produces a plot showing the standardized residuals vs. the fitted values calculated by the LTS sparse model. Based on observations from the plot, observations 2, 7, 13, 18, 19, and 20 were identified as potential outliers (Figure 7).

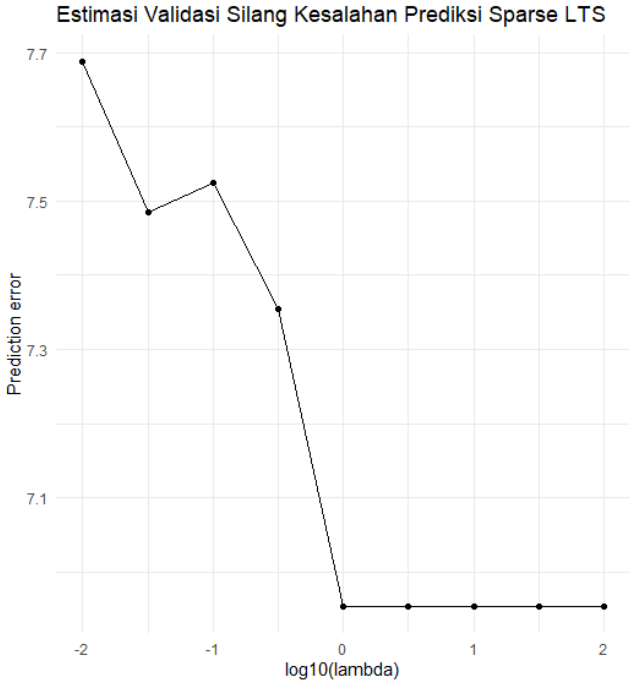


Figure 6. Cross Validation Estimation of Sparse LTS Prediction Error

**Evaluation of the Goodness of Fit Model**

The KPI (Key Performance Indicator) value in the form of $R^2$ and RMSE from the LASSO and sparse LTS models is used to select the best model. The following KPIs from the LASSO and sparse LTS models are shown in Table 5.

Table 5. Results of Evaluation of the LASSO and Sparse LTS Models

| KPI | LASSO | Sparse LTS |
|---|---|---|
| $R^2$ | -62.4% | 93.75% |
| RMSE | 2.1734 | 0.2933 |

According to Table 5, it can be seen that the sparse LTS model has a coefficient of determination. ($R^2$) that is greater than the LASSO model, namely 93.75%. Thus, if you look at the coefficient of determination ($R^2$), Utilizing the sparse LTS model rather than the LASSO model is preferable. Furthermore, if you look at the RMSE, the Sparse LTS model has a smaller value with an RMSE value of 0.2933. Based on the evaluation results, The most effective model for predicting the percentage of measles cases in North Sumatra in 2022 is the sparse LTS model.

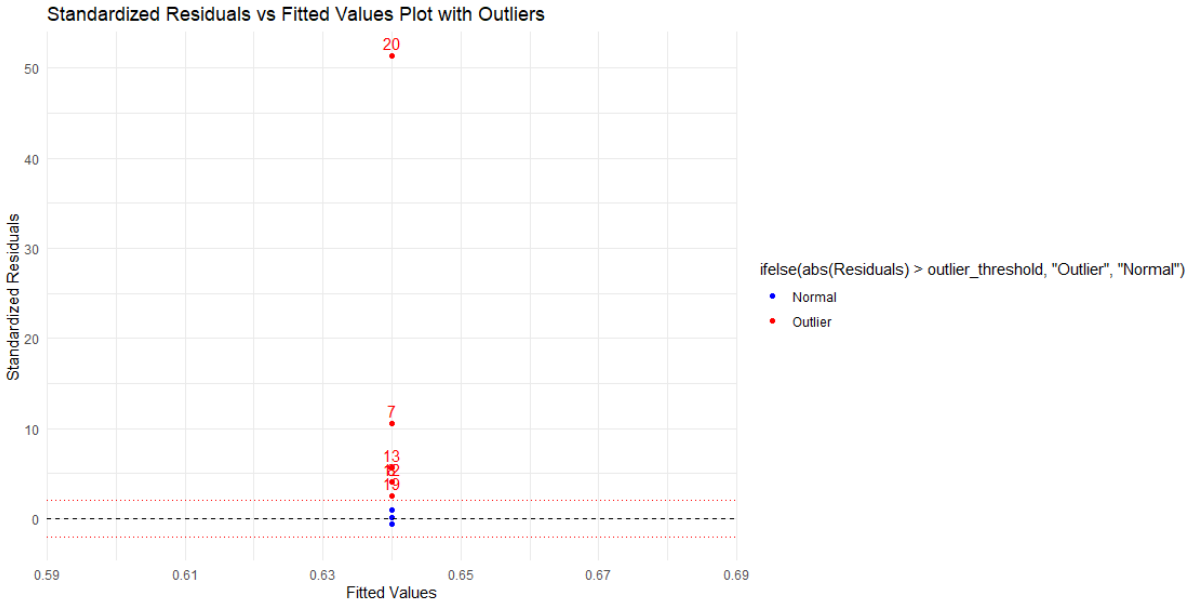


Figure 7. Standardized Residuals vs Fitted Value Model Sparse Lts

## CONCLUSION

This research resulted in the LTS sparse approach used in this study`s modeling to predict the percentage of measles cases in North Sumatra in 2022. Sparse LTS aims to provide a more robust and straightforward model that can efficiently make predictions with less justification. Compared to the traditional LASSO estimator based on R-square and RMSE values, the sparse LTS model successfully chose 14 out of 34 variables, reducing the number of explanatory variables required while maintaining model explanation. Meanwhile, there are 20 variables used, namely Measles Immunization Coverage Percentage ($X_1$), Number of Children Getting Coverage of Vitamin A Aged 12-59 Months ($X_4$), Percentage of Province Area ($X_5$), Number of Babies Born ($X_6$), Low Weight Baby (LWB) ($X_7$), Number of General Hospitals ($X_8$), Number of Integrated Healthcare Center ($X_{12}$), Number of Doctors ($X_{13}$), Number of Midwives ($X_{14}$), Number of Nutritionists ($X_{15}$), Number of Nurses ($X_{17}$), Long School Expectations ($X_{21}$), GRDP Growth Rate ($X_{23}$), Percentage of Poor Population ($X_{24}$), Diphtheria, Pertussis and Tetanus (DPT) Immunization Percentage ($X_{25}$), Labor Force Participation Rate ($X_{27}$), Human Development Index ($X_{28}$), Number of Villages/Subdistricts ($X_{29}$), Population density($X_{31}$), and Percentage of Households that Have Access to an Improper Source of Drinking Water) ($X_{34}$). These variables have a high influence on the correlation value of the Pearson variable with other variables. Make these variables not included in the model. According to these variables, the government may use it as a guide to reduce the number of measles cases in North Sumatra.

## REFERENCES

[1] Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares of analyzing high-dimensional large data sets. *Ann. Appl. Stat, 7*(1), 226-248.

[2] Ardhiansyah, F., Rahardjani, K. B., Suwondo, A., Setiawati, M., & Apoina, K. (2019). Faktor Risiko Campak Anak Sekolah Dasar pada Kejadian Luar Biasa di Kabupaten Pesawaran Provinsi Lampung. *JEKK: Jurnal Epidemiologi Kesehatan Komunitas, 4*(2), 64-72.

[3] Bottmer, L., Croux, C., & Ines, W. (2022). Sparse Regression for Large Data Sets With Outlier. *ELSEVIER: European Journal of Operational Research, 297*(2), 782-794.

[4] BPS Provinsi Sumatera Utara. (2023). *Provinsi Sumatera Utara Dalam Angka 2023.* (M. J. Guning, & A. O. Sihombing, Penyunt.) Sumatera Utara: CV. E'Karya.

[5] Hulu, V. T., Salman, Supinganto, A., Amalia, L., Sianturi, K. E., Nilasari, et al. (2020). *Epidemiologi Penyakit Menular: Riwayat, Penularan dan Pencegahan.* Medan: Yayasan Kita Menulis.

[6] Kemenkes RI. (2021). Dalam *Profil Kesehatan Indonesia Tahun 2021.* Jakarta: Kementerian Kesehatan Republik Indonesia.

[7] Kemenkes RI. (2022). Profil Kesehatan Indonesia Tahun 2021. Jakarta: Kementerian Kesehatan Republik Indonesia .

[8] KPAI. (2023, Januari 31). Dipetik Febuari 9, 2023, dari Kasus Campak Tinggi: KPAI Dorong KEMENKES Segera Lakukan Upaya Percepatan Layanan Imunisasi: https://www.kpai.go.id/publikasi/kasus-campak-tinggi-kpai-dorong-kemenkes-segera-lakukan-upaya-percepatan-layanan-imunisasi

[9] Monti, G. S., & Filzmoser, P. (2021). Sparse Least Trimmed Squares Regression With Compositional Covariates for High-Dimensional Data. *Oxfod: Bioinformatics, 37*(21), 3805-3814.
[10] Rahayu, A., & Husein, I. (2023). Comparison Of Lasso And Adaptive Lasso. *Sinkron*, 1435-1445.
[11] Randa, T. M., Tinungki, G. M., & Sunusi, N. (2022). Modeling the Proportion of Tuberculosis Cases in South Sulawesi using Sparse Least Trimmed Squares. *ESAKTA: Journal of Sciences and Data Analysis, 3*(2), 103-112.
[12] Kemenkes RI. (2018, April 24). *Situasi Campak dan Rubella di Indonesia*. Dipetik Febuari 2023 9, dari https://www.kemkes.go.id/article/view/18110600003/situasi-campak-dan-rubella-di-indonesia.html
[13] Sari, E. A., Rahma, H. I., Firdaus, M. R., Winarto, W., Indiyani, Y., & Nooraeni, R. (2020). Perbandingan Regresi OLS dan Robust MM- Estimation Dalam Kasus DBD di Indonesia 2018. *Jurnal Education and Development: Institut Pendidikan Tapanuli Selatan, 8*(2), 68-74.
[14] Sartika, I., Debataraja, N. N., & Imro'ah, N. (2020). Analisis Regresi Dengan Metode Least Absolute Shrinkage and Selection Operator (LASSO) dalam Mengatasi Multikolinearitas. *Bimaster: Buletin Ilmiah Mat. Stat. dan Terapannya, 9*(1), 31-38.
[15] Shodiqin, A., Aini, A. N., & Rubowo, M. R. (2018). Perbandingan Dua Metode Regresi Robust Yakni Metode Least Trimmed Squares (LTS) Dengan Metode Estimator-MM (Estimasi-MM) (Studi Kasus Data Ujian Tulis Masuk Terhadap Hasil IPK Mahasiswa UPGRIS),. *Jurnal Ilmiah Teknosains, 4*(1), 35-42.
[16] Sugiyono. (2019). *Metode Penelitian Kuantitatif Kualitatif dan R & D.* Bandung: Alfabeta.
[17] Suyono. (2018). *Analisis Regresi untuk Penelitian* . Yogyakarta: Deepublish.
[18] Syah, M. F. (24, Januari 2019). *Penyakit Campak Rubella (MR)*. Dipetik Febuari 9, 2023, dari https://dinkes.sarolangunkab.go.id/berita-penyakit-campak-rubella-mr.html
[19] Yahmal, P. N. (2021). Faktor-faktoryang Berhubungan Dengan Kejadian Campak. *JMH: Jurnal Medikal Hutama, 3*(10), 1612-1615.