Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12558
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison Of Lasso And Adaptive Lasso Methods In Identifying Variables Affecting Population Expenditure

Atika Rahayu[1)*] , Ismail Husein[2)]
[1,2)] Department of Mathematics, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[1)]atika0703191065@uinsu.ac.id [2)] husein_ismail@uinsu.ac.id

Submitted: Jun 15, 2023 | **Accepted**: Jun 20, 2023 | **Published**: Jul 1, 2023

Abstract: Since 2019, the difference in the increase in per capita expenditure of the population has continued to decline, and the most significant was only IDR 18,464 in 2021, indicating that the level of consumption of the population has not improved significantly, and the turnover of the community's economy is also not good. Multiple linear regression is more appropriate than other types of linear regression because it considers the influence of more than one independent variable on the dependent variable. However, problems may arise in the use of multiple linear regression, such as multicollinearity. To overcome this problem, other methods such as LASSO and adaptive LASSO should be used. Both methods have the ability to overcome multicollinearity between independent variables, thereby reducing the risk of misestimation. Nevertheless, the LASSO and Adaptive LASSO methods have differences in selecting important variables, so it is necessary to compare which method is better in terms of identifying influential variables. Based on the MSE and R-square comparison values, it is concluded that the Adaptive LASSO method model is the best model with a lower MSE value and a higher R-square value of 93%. The variable selection results of the Adaptive LASSO model are population size, number of households, average number of household members, constant price GDP, confirmed cases of COVID-19, human development index, percentage of the poor population, university student participation rate, and open unemployment rate.

Keywords: Regression, LASSO, Adaptive LASSO, Comparison, Variables

INTRODUCTION

Based on data from the Central Statistics Agency (BPS) of North Sumatra Province, it is recorded that the average per capita expenditure of residents in North Sumatra Province has increased every year, but even though it has increased, the difference in the increase has started to decrease every year since 2019, the average population expenditure has continued to decline and the most significant is in 2021, which is only 18,464 rupiah, while the difference from 2017 and 2018, namely before the covid-19 increase in the average per capita expenditure of North Sumatra residents reached 91,891 rupiah, this indicates that the population's consumption level has not increased well and the community's economic turnover is also not very good.

This problem needs to be known because if it is ignored, it will get worse, especially since population expenditure is related to the level of life and welfare of the community, although it is not the main factor, but the decline has been very significant, so it is important to know and understand what variables affect the average per capita expenditure of the population.

*name of corresponding author

Linear regression is a statistical method used to model the relationship between the dependent variable (the variable we are trying to influence) and the independent variable (the variable we are trying to predict). This method is used to determine the relationship between two or more variables using a linear regression line. The relationship between the independent variable and the dependent variable in regression is used to make predictions on various research problems in the field (Retnawati, 2017).

The type of linear regression used in finding influential variables is multiple linear regression because it can analyze more than one dependent variable, this regression each variable predictor has a coefficient that represents the contribution to the response. The coefficient can be used to assess the level of significance of each variable and determine whether the variable has a significant influence on the response. Multiple linear regression with OLS (*Ordinary Least Squares*) parameter estimation is the usual estimation to determine coefficients that minimize the difference between observed and predicted values. However, there are problems that can arise in the use of multiple linear regression, namely multicollinearity.

Multicollinearity is the correlation between independent variables. A good regression model should not have a correlation between independent variables. A widely used method to detect multicollinearity is the *variance infl*ation factor (*VIF*) (Retnawati, 2017). The LASSO (*Least Absolute Shrinkage and Selection Operator*) method and the *Adaptive* LASSO method are variable selection methods that can help the problem of multiple linear regression.

LASSO *(Least Absolute Shrinkage and Selection Operator)* applies regularization where some regression coefficients are reduced to zero. For feature selection, all coefficients with non-zero values are selected and the prediction error is minimized. When the value of this parameter is very high, the coefficient of the regression variable becomes zero (Melkumova & Shatskikh, 2017). Adaptive LASSO is a LASSO regression method with the oracle property. This property allows the regression method to select the correct subset of predictors that have optimal estimation rates. Adaptive LASSO is a regularization method that does not mask large regression coefficients and also provides subset selection of predictors by shrinking some coefficients to zero. (Qian & Yang, 2013) (Husein, Zein & Sabrina,2022). Both of these methods have the ability to overcome the problem of multicollinearity between independent variables, thereby reducing the risk of misestimation.

Even so, the LASSO and *Adaptive* LASSO methods have differences in choosing important variables, some of the differences between the two models are determining penalties and identifying different variable interactions, because of these differences the results of the analysis will be different. As in some previous studies that discussed the LASSO and Adaptive LASSO methods in 2021 with the title "Classification of Individual Working Status in Banten Province in 2020 Using the LASSO and *Adaptive* LASSO Methods" conducted by Pardomuan Robinson Sihombing, Khairil A. Notodiputro, and Bagus Sartono. The conclusion of this study with data in the form of binary logistic regression is that the best model is the LASSO model with 60:40 data simulation, this is by confirming the stability of the performance of the two models. Also research in 2020 with the title "Increasing the Precision of Grain Weight Estimation Through a Variable Selection Process in Statistical Machine Learning". Research conducted by Muhlis Ardiansyah, Khairil Anwar Notodiputro, and Bagus Sartono. Finding the best model in estimating grain weight based on 2019 Cropping Survey data in Central Kalimantan, the conclusion of this study is The results show that Adaptive LASSO- QRF provides the best performance with the smallest RMSE value.

From previous research, it indicates that both the LASSO and *Adaptive* LASSO methods have their respective advantages and disadvantages in determining different problems as well as different forms of regression so it is important to evaluate. In this study, we will compare the LASSO method and the *Adaptive* LASSO method in identifying variables that affect the average per capita expenditure of residents in the province of North Sumatra, a model of which method is better and what variables affect population expenditure generated by the model of the best method. *Means Square Error* (MSE) and *R-square* are comparative values that will be used to conclude which method produces the best model in identifying these influential variables.

*name of corresponding author

LITERATURE REVIEW

Previous research that discusses the LASSO and Adaptive LASSO methods in 2021 with the title "Classification of Individual Working Status in Banten Province in 2020 Using the LASSO and *Adaptive* LASSO Methods" Journal article research conducted by Pardomuan Robinson Sihombing, Khairil A. Notodiputro, and Bagus Sartono. Information The data used in this study came from the Banten Province National Labor Force Survey (Sakernas) for the August 2020 period, which was conducted by the Central Statistics Agency (2021). The total sample used was 11,469 respondents, of which 10.2 percent were not working, the remaining 89.8 percent were working and this study used a logistic regression model. The conclusion of this study is that the best model is the LASSO model with 60:40 data simulation, this is by confirming the stability of the performance of the two models. (Sihombing, Notodiputro, & Sartono,2021).

Another research that also uses the LASSO and *Adaptive* LASSO methods is in 2020 with the title "Increasing the Precision of Grain Weight Estimation Through the Process of Variable Selection in Statistical Machine Learning". The research was conducted by Muhlis Ardiansyah, Khairil Anwar Notodiputro, and Bagus Sartono. In this study, a new solution was sought to overcome the non-response problem, namely by estimating non-response data on grain weight based on several variables obtained from interviews with farmers after the social restriction policy ended. In the study, various variable selection methods were tried, namely *Stepwise*, LASSO, *Elastic Net*, Adaptive LASSO, and *Relaxed* LASSO to find the best model in estimating grain weight based on 2019 Cropping Survey data in Central Kalimantan, the conclusion of this study is he results show that Adaptive-QRF provides the best performance with the smallest RMSE value. The selected variables that significantly affect the grain weight of the Adaptive LASSO selection results are seed variety, planting method (monoculture or intercropping), planting system (jajar legowo or not), pest attack or not, planting location, and harvest month (Ardiansyah, Notodiputro, & Sartono, 2020).

Another research that uses the LASSO method is research in 2022 with the title "Determining Factors Identifying Poverty Rate Due To Covid-19 Pandemic In North Sumatra Using The Least Absolute Shrinkage And Selection Operator (LASSO) Method".This research was conducted by Nur Indah Sari, Hendra Cipta, Muhammad Fathoni. The conclusion of this study is that by using the LASSO method using the LARS algorithm, the factors that affect poverty due to the covid-19 pandemic are the open unemployment rate, human development index, district / city minimum wage and the number of unemployed 15 years and over (Sari, Cipta & Fathoni, 2022)

Research using the LASSO method is research in 2022 with the title "Application of Least Absolute Shrinkage And Selection Operator (LASSO) Regression to Identify Variables Affecting the Incidence of Stunting in Indonesia".This research was conducted by Tesa Trilonika Pardede, Bagus Sumargo, Widyanti Rahayu. least absolute shrinkage and selection operator (LASSO). This study was conducted by Tesa Trilonika Pardede with the Least Angle Regression (LAR) algorithm because in the stunting data in Indonesia there is a multicollinearity problem among the independent variables used. The conclusion of this study is based on the analysis that has been carried out, namely the exclusive breastfeeding variable (X1), protein consumption (X2), DPT-HB immunization (X5), maternal height (X8) and diarrhea (X9) affect stunting in Indonesia in 2018 (Pardede, Sumargo, & Rahayu, 2022)

Research that discusses the comparison of the LASSO method is entitled "Performance of Ridge Regression and Lasso Regression on Data with Multicollinearity" (2022). Research conducted by Fitri Rahmawati, Risky Yoga Suratman. In this study, researchers compared the LASSO method and the ridge method for data that contains multicollinearity, namely variables that affect the number of applicants at universities in the United States. In this study the data is divided into training data and testing data, as many as 388 observations as training data and 389 as testing data. The conclusion of this study is that ridge regression is superior to lasso regression in terms of MSE size. Ridge regression has a smaller MSE of 1107166 than the lasso regression of 1138842. However, in terms of variable selection and model interpretation, lasso regression is considered superior because model interpretation is more efficient by using only 4 predictor variables (Rahmawati & Suratman, 2022) .

Based on previous studies, it can be seen that adaptive LASSO is not always the best in overcoming cases, different forms of regression and the form of data that affects it, so researchers are interested in comparing the LASSO method with adaptive LASSO using the LASSO method.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12558
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

This study is a comparative study that aims to compare the LASSO and Adaptive LASSO methods in identifying variables that affect the average per capita expenditure of residents in North Sumatra Province. This research uses quantitative research type and the research method used is secondary data analysis method. Data analysis conducted in this study is multiple linear regression analysis using the LASSO method and the Adaptive LASSO method in selecting variables that affect population expenditure in North Sumatra province. This research uses the help of SPSS and Rstudio software with glmnet package. The research procedures that will be carried out in this study are, Perform descriptive statistics on the data, Modeling with OLS, Multicollinearity test, Finding the best lambda with cross validation for the LASSO model, Forming the LASSO regression model, Finding adaptive weights for the LASSO adaptive model, Finding the best lambda with cross validation for the LASSO adaptive model, Forming the LASSO adaptive regression model, Comparing with MSE and Rsquare values. The following is a flow chart of this research.

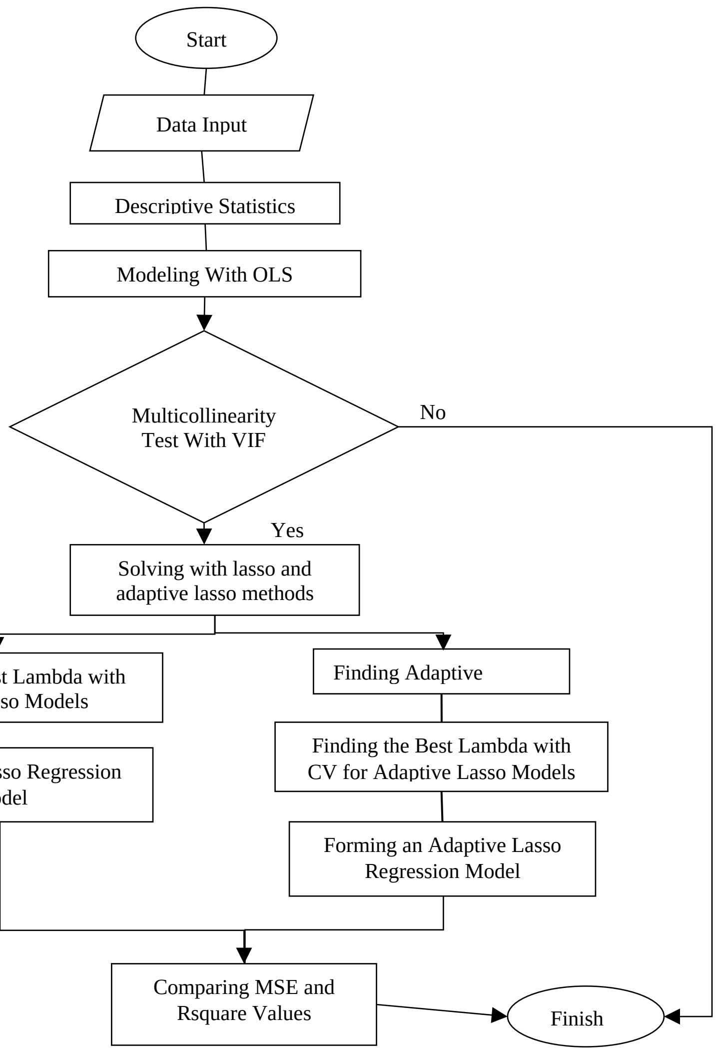

Figure 1. Research *flowchart* diagram

The research data used is secondary data based on existing data in 33 districts / cities of North Sumatra Province, all data used is data in 2021. the data needed comes from the official website of the central statistics agency (BPS) of North Sumatra province, the data used consists of 10 independent variables and 1 dependent variable, each variable consisting of 33 data. The following are the variables used in this study:

*name of corresponding author

 e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1. Research Variables

No.	Variables	Symbol	Unit
1	Population per capita expenditure	PPP	Rupiah
2	Total Population	JP	People
3	Number of Households	RT	Family
4	Average Number of Household Members	ART	Person per family
5	Gross Regional Domestic Product at Constant Prices	PDRB	Rupiah
6	Number of Covid-19 Confirmation Cases	KC19	People
7	Human Development Index	HDI	Percent
8	Percentage of Poor Population	PM	Percent
9	District/City Minimum Wage	UM	Rupiah
10	APK Higher Education Level	APKP	Percent
11	Open Unemployment Rate	PT	Percent

RESULT

Before conducting data analysis, descriptive statistics of the research variables used are displayed first. These descriptive statistics are useful for displaying an overview of the data used, a total of 363 data consisting of 330 data divided into 10 independent variables and 33 data from 1 dependent variable. The following are descriptive statistics of the data carried out with the help of SPSS *software*.

Table 2. Data Descriptive Statistics

Var	N	Min	Max	Mean	Standard Deviation
PPP (Y)	33	580838	1788156	1031794.363636	231756.714351
JP (X_1)	33	26941	1225201	227059.787879	256672.355720
RT (X_2)	33	12012	564619	105646.151515	120337.257877
ART (X_3)	33	3.73	5.07	4.391818	0.297620
PDRB (X_4)	33	11832505	64078946	30451109.333333	13392033.827061
KC19 (X_5)	33	165	49416	3175.757576	8721.364999
HDI (X_6)	33	61.99	81.21	71.064242	4.521144
PM (X_7)	33	4.01	26.42	11.085455	4.780783
UM (X_8)	33	0	3329867	2514564.545455	835991.600077
APKP (X_9)	33	9.2	39.76	18.300909	7.614253
PT (X_{10})	33	0.7	11.00	4.908182	2.877869

After displaying Table 2, then calculate the regression coefficient of OLS (*Ordinary Least Squares*), the results of which will be used as the value of the adaptive weight of the adaptive LASSO method. β_j' to calculate the *adaptive* weight of the *adaptive* LASSO method and the regression coefficient results from OLS are also used as a comparison with the LASSO and *Adaptive* LASSO methods. The following are the results of the regression coefficient of OLS processed using RStudio *software*.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12558
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3. OLS Regression Coefficients

Variables	OLS
Intercept	4.680411e+05
X_1	-2.788330e+00
X_2	4.601692e+00
X_3	-1.925217e+05
X_4	9.003546e-04
X_5	2.428526e+01
X_6	2.134603e+04
X_7	-5.569247e+03
X_8	-9.843553e-03
X_9	-3.707780e+03
X_{10}	1.815442e+04

Furthermore, multicollinearity test is conducted on the data, a widely used method to detect the presence of multicollinearity is the *variance inflation* factor (VIF). Multicollinearity occurs if the VIF value> 10 (Retnawati, 2017). The formula for finding VIF is :

$$VIF = \frac{1}{1-R_j^2} \quad (1)$$

With $j = 1, 2, 3, \ldots, k$. Where R^2 is the coefficient of determination between the independent variable X_j and variables explained by other independent variables in the regression model. The following Table 4 shows the results of the VIF value using the help of RStudio *software*.

Table 4. VIF Value

Variables	VIF
JP (X_1)	1814.925010
RT (X_2)	1750.767076
ART (X_3)	8.694380
PDRB (X_4)	2.478782
KC19 (X_5)	9.517564
HDI (X_6)	6.802797
PM (X_7)	4.999218
UM (X_8)	1.693739
APKP (X_9)	3.145217
PT (X_{10})	3.478893

From table 4 above, it can be seen that the JP and RT variables have VIF values of more than 10, so this indicates that there is a multicollinearity problem in the independent variables of the multiple linear regression, requiring the LASSO and Adaptive LASSO methods to overcome this multicollinearity problem.

Before looking for regression coefficients from LASSO, first determine the λ by using *K-Fold Cross Validation. Cross Validation* or what can be called rotation estimation is a model validation technique to assess how the results of statistical analysis will generalize to independent data sets, One technique of cross validation is *k-fold cross validation* which breaks the data into K parts of the data set of the same size. The use of *k-fold* = 10 *cross-validation* to eliminate bias in the data (Azis, 2020). .The formula for *K-fold cross validation* is:

$$CV_K = \frac{1}{k}\sum_{i=1}^{k}(evaluasi_i) \quad (2)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12558
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Where $(evaluasi_i)$ is the evaluation value at each iteration I, the evaluation value used is the MSE value. Here are the results λ The best results of the LASSO method using RStudio *software*.

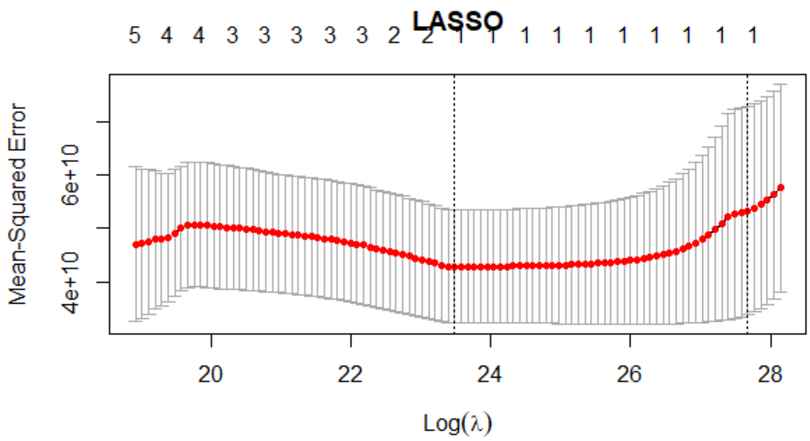

Figure 2. *Cross Validation* Results of LASSO Regression

By using RStudio *software*, namely the *glmnet* package, the best value is obtained. λ the best value in LASSO regression which has the lowest MSE value, namely λ = 15840564041, after the λ best value has been obtained then the regression coefficient produced by LASSO. The equation of the LASSO method is as follows:

$$\hat{\beta}^{lasso} = \arg\min\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (3)$$

Where, y_i is the response variable for the i-th observation, β_0 is a constant in the regression model, x_{ij} is the value of the jth variable predictor for the i-th observation, β_j is the regression coefficient for the jth variable predictor, λ is a penalization parameter that determines how much influence the variables have on the model. By using RStudio *software, the glmnet* package, the regression coefficients of LASSO are shown in table 5 below.

Table 5. LASSO Regression Coefficient

Variables	**LASSO**
Intercept	7.440007e+05
X_1	0
X_2	0
X_3	0
X_4	9.451007e-03
X_5	0
X_6	0
X_7	0
X_8	0
X_9	0
X_{10}	0

Unlike LASSO, before finding the λ *Adaptive* LASSO first calculates the value of the *adaptive* weight or ω_j of each variable. The formula for finding *adaptive* weights is as follows:

$$\omega_j = \frac{1}{|\beta_j'|^q} \quad (4)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12558
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Where β_j' is the initial value of the regression coefficient that has been calculated using OLS in the previous table 3 and q is a positive constant value or gamma value used for *adaptive* weight adjustment taken between 0.5 or 1 or also 2, in this calculation using q = 1. The following is the resulting *adaptive* weight value.

Table 6. *Adaptive* weight values

Variables	***Weight***
X_1	3.586376e-01
X_2	2.173114e-01
X_3	5.194220e-06
X_4	1.110673e+03
X_5	4.117724e-02
X_6	4.684712e-05
X_7	1.795575e-04
X_8	1.015893e+02
X_9	2.697032e-04
X_{10}	5.508301e-05

Furthermore, the results λ The best results of the *Adaptive* LASSO method using RStudio *software are* shown in Figure 3 below.

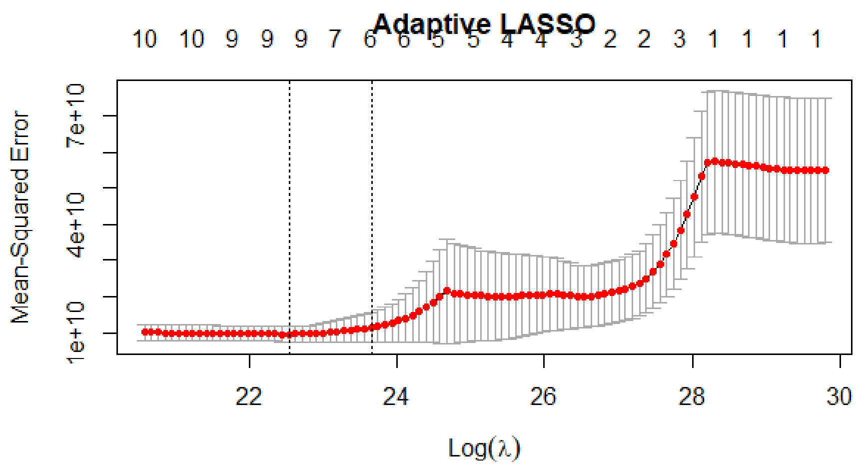

Figure 3. *Cross Validation* Results of *Adaptive* LASSO

Obtained value λ The best value for *Adaptive* LASSO which has the lowest MSE value is λ = 6160073748, after the best value of λ best value has been obtained then the regression coefficient produced by *Adaptive* LASSO. The equation of the *Adaptive* LASSO method is as follows:

$$\hat{\beta}^{lasso} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right\} \tag{5}$$

The difference between the LASSO method and *Adaptive* LASSO in equation (3) and equation (5) is that there is a value of ω_j in *Adaptive* LASSO. By using RStudio *software*, *the glmnet* package, the regression coefficients of *Adaptive* LASSO are shown in table 7 below.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12558
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 7. LASSO *Adaptive* Regression Coefficient

Variables	***Adaptive* LASSO**
Intercept	7.448707e+05
X_1	-5.897971e-01
X_2	6.745307e-04
X_3	-2.882030e+05
X_4	1.940082e-04
X_5	2.253078e+01
X_6	2.245532e+04
X_7	-1.577125e+03
X_8	0
X_9	-2.451876e+03
X_{10}	1.540926e+04

To compare the performance of the LASSO and *Adaptive* LASSO methods in determining influential variables using the *Means Square Error* (MSE) and *R-square* values as benchmarks. MSE is the average of the squared difference between the predicted value and the observed value. In general, the smaller the MSE value, the better. MSE is used when the modeling objective is to identify the model that most closely reproduces the true data generating distribution, hence the "best" model minimizes the cross entropy between the model predictions and the training data (Hodson, 2021) .The MSE value is calculated with the following equation.

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (6)$$

y_i is the outcome data value for the i-th observation, i = 1, 2,...,n and $\hat{y}_i$ is the predicted value for the i-th observation, n is the sample size. The lower the MSE value, the better the model built and the more accurate the prediction results produced.

R-square (R^2) or a small coefficient of determination means that the ability of the independent variables to explain the dependent variable is very limited, on the other hand, if the value is close to 1 (one) and away from 0 (zero) it means that the independent variables have the ability to provide all the information needed to predict the dependent variable. (Ghozali, 2016). .The equation for calculating R-square is :

$$R^2 = 1 - \frac{RSS}{TSS} \quad (7)$$

Where RSS is the *sum of squared residuals*, which is the sum of squares of the difference between the actual value and the predicted value by the regression model and TSS is the sum of squared total, which is the sum of squares of the difference between the actual value and the average value of the dependent variable. the greater the value of R^2, the better the linear regression model fits the data.

Using RStudio *software*, the following comparison of MSE and *R-square* values generated from the LASSO and *Adaptive* LASSO models is shown in Table 8.

Table 8. Comparison Values

Comparison Value	**LASSO**	**Adaptive LASSO**
MSE	34196515844	0.3492015
R-square	3636222801	0.9307986

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Number 3, July 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.12558 p-ISSN : 2541-044X

DISCUSSIONS

Based on the LASSO regression coefficients in Table 5 and the *Adaptive* LASSO regression coefficients in Table 7, it can be seen that the variables that experience multicollinearity are the variables JP (X_1) and RT (X_2) there is a shrinkage of the coefficient on the LASSO regression model to zero, which means that the variable X_1 and X_2 are no longer included in the LASSO regression model, therefore the multicollinearity problem in the independent variables is resolved. While in Table 7 the variables X_1 and X_2 did not shrink but increased compared to the regression coefficients from OLS in Table 3 and the coefficient value that shrinks exactly to 0 is X_8.

The following LASSO regression model is generated:

$$y = 7.440007\text{e} + 05 + 9.451007\text{e} - 03x_4$$

The following *Adaptive* LASSO regression model is generated:

$$\begin{aligned} y = 7.448707\text{e} + 05 + (-5.897971\text{e} - 01)x_1 + 6.745307\text{e} - 04x_2 + (-2.882030\text{e} + 05)x_3 \\ + 1.940082\text{e} - 04x_4 + 2.253078\text{e} + 01x_5 + 2.245532\text{e} + 04x_6 \\ + (-1.577125\text{e} + 03)x_7 + (-2.451876\text{e} + 03)x_9 + 1.540926\text{e} + 04x_{10} \end{aligned}$$

In Table 5 the regression coefficients of LASSO produce regression coefficients that shrink greatly to zero and only the variable X_4 Thus the only independent variable that affects the average per capita expenditure of the population generated by the LASSO method is GDP (Gross regional domestic product at constant prices), This result is in accordance with LASSO applying when the value of this parameter is very high, the coefficient of the regression variable becomes zero, this makes LASSO a widely accepted regression technique. (Melkumova & Shatskikh, 2017). Thus the results of the interpretation of the LASSO model become more efficient because it only produces 1 independent variable.

While in Table 7 only variables X_8 Thus, the independent variables that affect the average per capita expenditure of the population produced by the *Adaptive* LASSO method are population, number of households, average number of household members, gross regional domestic product at constant prices, covid-19 confirmation cases, human development index, percentage of poor people, college APK, and open unemployment rate. This result is consistent with *Adaptive* LASSO is a regularization method that does not mask large regression coefficients and also provides subset selection of predictors by reducing some coefficients to zero. (Qian & Yang, 2013) .Because it has the oracle property of different weights in each variable that makes *Adaptive* LASSO not easily shrink the coefficients.

The comparative values shown in Table 8 show that LASSO regression has an MSE of 34196515844, while *Adaptive* LASSO has an MSE value of 3636222801 which can be concluded that the MSE of

Adaptive LASSO is lower than LASSO regression. The *R-square* value of LASSO is 0.3492015, while the *R-square* of *Adaptive LASSO* is 0.9307986 which indicates that the *R-square* value of *Adaptive LASSO is* higher than LASSO regression. So overall the *Adaptive* LASSO method is better in terms of model performance and the ability of the independent variables to explain the dependent variable. This is in accordance with the explanation that there are several problems in applying the Lasso method. The main reasons are that the estimation results of the Lasso method are partial estimates and do not have philosophical properties (model fit and asymptotic normality of parameter estimates), and there is excessive variable compression (Wang et al., 2019).

CONCLUSION

Based on the results and discussion above, it is concluded that in terms of analyzing influential variables, the model generated by the *Adaptive* LASSO method is better than the model generated by the LASSO method, this is because the model from *Adaptive LASSO* has a lower MSE value of 3636222801 compared to LASSO which is 34196515844 and the model from *Adaptive LASSO* has a higher *R-square* value of 0.9307986 compared to LASSO which is 0.3492015, which can be interpreted that the model generated by *Adaptive LASSO* can 93% explain the variation in the

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Number 3, July 2023
DOI : https://doi.org/10.33395/sinkron.v8i3.12558
e-ISSN : 2541-2019
p-ISSN : 2541-044X

dependent variable using the independent variables in the model which makes the accuracy of the model generated by *Adaptive* LASSO better. The selection of influential variables resulting from *Adaptive* LASSO is population, number of households, average number of household members, gross regional domestic product at constant prices, covid-19 confirmation cases, human development index, percentage of poor people, college APK, and open unemployment rate.

REFERENCES

Ardiansyah, M., Notodiputro, K. A., & Sartono, B. (2020). Peningkatan Presisi Dugaan Berat Gabah Melalui Proses Seleksi Peubah Dalam Pembelajaran Mesin Statistika. *Jurnal Seminar Nasional Varians*, 171–183.

Azis, H., Purnawansyah, Fattah, F., & Putri, I. P.(2020). Performa Klasifikasi K-NN Dan Cross-Validation Pada Data Pasien Pengidap Penyakit Jantung. ILKOM Jurnal Ilmiah. 12 (2), 81-86.

Berrar, D. (2018). Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vols. 1–3). https://doi.org/10.1016/B978-0-12-809633-8.20349-X

BPS. (2022). *https://www.bps.go.id/subject/5/konsumsi-dan-pengeluaran.html*.

BPS Provinsi Sumatera Utara. (2022). *https://sumut.bps.go.id/subject/154/konsumsi-dan-pengeluaran.html*.

Ghozali, I. (2016). Aplikasi Analisis Multivariete dengan Program IBM SPSS 23 (VIII). Semarang: Badan Penerbit Universitas Diponegoro. *Cetakan Ke VIII. Semarang: Badan Penerbit Universitas Diponegoro*, *14*(July).

Husein, Ismail., Zein, Achyar., Sabrina, Ladia. (2022). Mustahiq *(Zakat Recipient*) Determination Application With Analytic Hierarchy Process Model. Sinkron : Jurnal dan Penelitian Teknik Informatika. 7(4). 2462-2468. https://doi.org/10.33395/sinkron.v7i4.11825

Hodson, T. O. (2021). Mean Square Error Deconstructed. *James: Journal Of Advances In Modeling Earth Systems*.

Kuncoro, D. (2022). *Kecerdasan Komputasional*. Teknik Informatika: Universitas Pancasakti Bekasi.

Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering, 201*. https://doi.org/10.1016/j.proeng.2017.09.615

Paiman. (2019). *Teknik Analisis Korelasi dan Regresi Ilmu*. UPY Press.

Pardede, T. T., Sumargo B., & Rahayu W.(2022). Penerapan Regresi *Least Absolute Shrinkage And Selection Operator (LASSO)* Untuk Mengidentifikasi Variabel yang Berpengaruh terhadap Kejadian *Stunting* di Indonesia**.** Jurnal Statistika dan Aplikasinya. 6(1),37-48.

Qian, W., & Yang, Y. (2013). *Model Selection Via Standard Error Adjusted Adaptive Lasso*. *65*(2), 295–318.

Rahmawati, F., & Suratman, R. Y. (2022). PERFORMA REGRESI RIDGE DAN REGRESI LASSO PADA DATA DENGAN MULTIKOLINEARITAS. *Leibniz: Jurnal Matematika*, *2*(2).

Retnawati, H. (2017). *Pengantar Analisis Regresi Dan Korelasi*. FMIPA Pendidikan Matematika.

Sari, Nur Indah., Cipta, Hendra., & Fathoni, Muhammad. (2022). Determining Factors Identifying Poverty Rate Due To Covid-19 Pandemic In North Sumatera Using The Least Absolute Shrinkage And Selection Operator (LASSO) Method. IOCS : Jurnal Mantik, 6 (1), 271-277.

Sethi, J. K., & Mittal, M. (2021). An efficient correlation based adaptive LASSO regression method for air quality index prediction. *Earth Science Informatics*, *14*(4), 1777–1786. https://doi.org/10.1007/s12145-021-00618-1

Sihombing, P. R., Notodiputro, K. A., & Sartono, B. (2021). Klasifikasi Status Bekerja Individu di Provinsi Banten Tahun 2020 dengan Menggunakan Metode LASSO dan Adaptive LASSO. *STATISTIKA Journal of Theoretical Statistics and Its Applications*, *21*(1). https://doi.org/10.29313/jstat.v21i1.7810

Wang, W., Wang, W., & Ding, E. (2019). Identification of Critical-to-quality Characteristic in Complex Products Based on the Adaptive-Lasso Method. *Proceedings of 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2019*. https://doi.org/10.1109/IAEAC47372.2019.8998012

*name of corresponding author