JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 9 (ISSUE 1) JUNE 2025: 126 – 139
https://doi.org/10.21009/JSA.09111

ANALYSIS OF INDONESIAN STUDENTS' READING LITERACY USING THE SMOOTHLY CLIPPED ABSOLUTE DEVIATION (SCAD) PENALTY

Vera Maya Santi [1*], Ariq Muammar Riyantobi[1], Widyanti Rahayu[1]

[1]*Study Programme of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Rawamangun Muka, Jakarta, 13220, Indonesia*

Corresponding author's e-mail: * ***vmsanti@unj.ac.id***

ABSTRACT

Reading literacy is the ability to understand, use, evaluate, reflect on, and engage with texts to achieve one's goals, develop one's knowledge and potential, and participate in society. Reading literacy significantly impacts a country's educational level, making it crucial to further investigate this issue. Identifying factors that influence students' reading literacy, particularly in Indonesia, is a key area of exploration. PISA survey data, conducted every three years, is relevant for researching student proficiency. Each survey period focuses on one of three main topics: science literacy, mathematics literacy, and reading literacy. The 2018 PISA survey data is suitable for studying students' reading literacy, as the main topic that year was reading literacy. However, PISA survey data includes many strongly correlated independent variables, potentially violating the multicollinearity assumption. This research aims to investigate Indonesian students' reading literacy by employing the Smoothly Clipped Absolute Deviation (SCAD) penalty approach to enhance model estimation and variable selection. The SCAD (Smoothly Clipped Absolute Deviation) penalty function has proven effective in previous studies on PISA data. The model using the SCAD penalty function yielded excellent results, indicated by an Adjusted R^2 value of 0.967. Based on this model, three main factors influence students' reading literacy in Indonesia: learning facilities, general knowledge, and students' self-confidence.

Article History:
Received: May 21, 2025
Revised: June 25, 2025
Accepted: June 29, 2025
Published: June 30, 2025
Available online.

Keywords:
Reading literacy, PISA, Multicollinearity, SCAD

How to cite this article:
V.M. Santi, A.M. Riyantobi, W. Rahayu, "ANALYSIS OF INDONESIAN STUDENTS' READING LITERACY USING THE SMOOTHLY CLIPPED ABSOLUTE DEVIATION (SCAD) PENALTY", Jurnal Statistika dan Aplikasinya, vol. 9, iss. 1, pp. 126 – 139, June 2025

Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

A fundamental aspect in the development of the education sector is the issue of students reading proficiency. Students who have good reading skills are considered more likely to achieve better academic performance [1], [2]. Moreover, good reading skills can also help to enhance soft skills such as communication, presentation, leadership, and socialization abilities. Thereby making students more confident [3]. The concept of literacy has evolved from a simple notion of reading and writing abilities to the ability to apply various competencies and skills in life, which is crucial in keeping up with technological and socio-cultural developments in the 21st century [4]. Therefore, the issue of reading proficiency, commonly referred to as reading literacy, is a crucial point to be addressed.

One of the studies on reading literacy is the World's Most Literate Nations (WMLN) conducted by Central Connecticut State University. This study ranks countries not based on the reading ability of their population, but rather on their reading habits and supporting resources. The ranking is derived from five research indicators, namely libraries, newspapers, educational system inputs, educational system outputs, and the availability of computers [5].

The results of this reading literacy study showed that Indonesia's literacy level ranks second from the bottom, or 60th out of a total of 61 countries, indicating that the literacy level in Indonesia is still relatively low. This ranking, when detailed based on the five research indicators, can be seen in Table 1.

Table 1. Indonesia's Reading Literacy Ranking Based on WMLN Indicators

Indicators	Definition	Rank
Libraries	To measure the quantity and quality of public libraries available to the community, including book collections, the availability of libraries per capita, and the level of public access to libraries	36
Newspapers	To measure the circulation of print and digital newspapers per capita, including the diversity of news media available and accessed by the public.	55
Education System – Input	To measure the input of the education system, such as the enrollment rates in primary and secondary education, student-teacher ratios, teacher qualifications, educational facilities, and education funding	54
Education System – Output	To measure the outcomes of the education system, such as graduation rates, students' reading proficiency based on international assessments (e.g., PISA), and achievement in reading literacy.	45
Computers	To measure the level of ownership and access to computers and information technology infrastructure, both in schools and households.	60
Final Rank	It represents the overall literacy ranking based on a composite of all the above indicators.	**60**

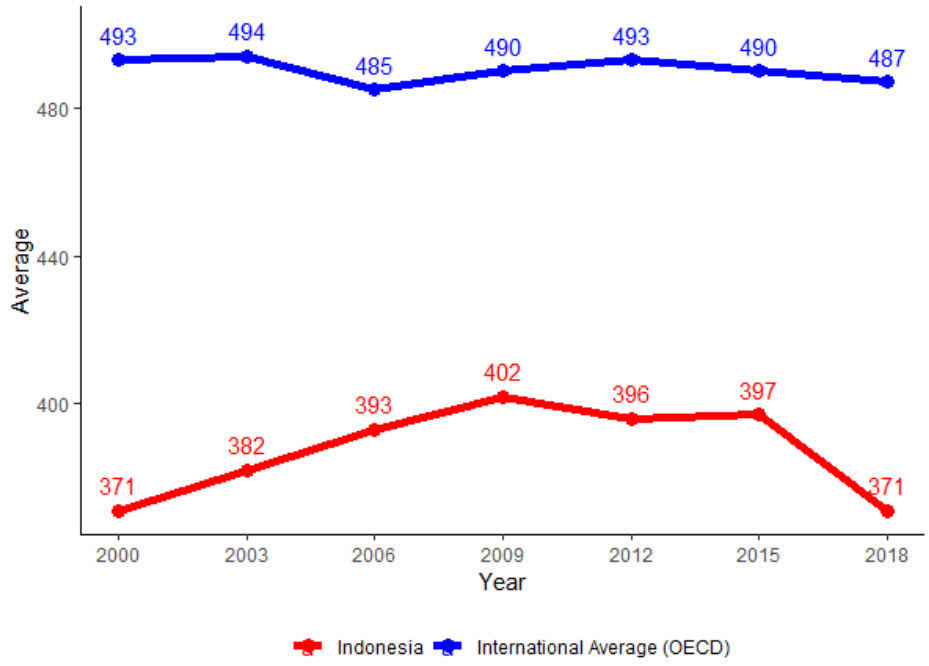
Source: https://www.ccsu.edu/wmln/rank.html

The education system in a country is closely related to its reading literacy level, as evidenced by the inclusion of two education-related indicators in this study. These indicators are also considered important in the ranking process. One of the indicators is the output of the education system, measured by the country's reading assessment scores on international assessments it has participated in, while the other indicator reflects the input of the education system [5].

A country's participation in such assessment program is also carried out as an effort to improve and develop the education system [6]. Both assessments focus on looking at international trends in comprehension in young students, TIMSS looks at trends in math science comprehension in fourth and eighth graders, while PIRLS assesses international trends in reading comprehension in students in their fourth year of primary school which is an important transition point in children's development as readers [7], [8].

In a further effort to develop the education system, Indonesia also collaborates with the Organization for Economic Co-operation and Development (OECD) on one of its assessments, the Programme for International Student Assessment (PISA). PISA is organized every three years, with three research focus topics, where the topics are core subjects in schools, namely reading, mathematics, and science, with the research target being students aged 15 years [9].

Between 2003 and 2018, Indonesia was able to add almost 1.8 million 15-year-old students to secondary school education without compromising the quality of education provided. This allowed Indonesia to increase its coverage of the 15-year-old population from 46% in 2003 to 85% in 2018, it can be said that PISA 2018 data can be used to describe the reading literacy level of Indonesian students [10], [11], [12].



Source: https://pisadataexplorer.oecd.org/

Figure 1. Line Graph of Indonesia's Mean PISA Reading Literacy Score vs International Average 2000 – 2018

The average reading literacy score throughout Indonesia's participation in the PISA assessment, from 2000 to 2018, is shown in Figure 1 The scores tend to stagnate and experience a negative trend in the last three rounds of PISA, and the scores are also always below the international average. Examining the average reading literacy score in 2018 is interesting because there was a significant decrease of 26 points when compared to the score in 2015. This result placed Indonesia 72nd out of a total of 79 participating countries in the last round of PISA [13].

The items in the questionnaire that relate to reading literacy are mentioned in the questionnaire compendium, with further explanation discussed in the PISA 2018 questionnaire framework [9]. Identifying the factors that influence student performance on reading literacy is important because the results from PISA 2018 suggest that more than ten million students represented by PISA were unable to complete the most basic reading test [14].

A study investigated the factors that have a statistically significant effect on reading literacy scores in PISA 2018 for students in Turkey, China, and Mexico. The results showed that 'Economic, social, and cultural index', 'Meta-cognition: assessing credibility', and 'Meta-cognition: summarizing' were the most influential factors on students' reading performance in Turkey, China, and Mexico [15]. Several studies have focused on examining the main topic of PISA 2018, namely the issue of reading literacy [16], [17], [18]. However, most of these studies discuss PISA with a qualitative approach, while there are still relatively few studies that use a quantitative approach.

Quantitative research needs to be done to obtain useful information from the data to be studied [19]. The use of analytical methods also needs to be adjusted to the objectives of the research to be carried

out to obtain appropriate results. In the case of this study, the objective is to find out about the factors that have the most influence on reading literacy scores. This can be done with an analysis in the quantitative approach, namely linear regression analysis.

Linear regression analysis is an analysis used to see the linear relationship between the independent variable (X) and the dependent variable (Y). Since there are several variables that will be used as independent variables, the analysis used is multiple linear analysis [20]. The independent variables are questionnaire items given to students related to the reading framework, so they are likely to have strong interrelationships among their variable groups. This causes the commonly used coefficient estimation, namely the ordinary least square method (OLS) to produce a singular matrix $(\boldsymbol{X^T X})$ so OLS cannot be applied [21]. An alternative solution is to use the penalized regression model with a penalty function to select the independent variables while estimating the parameters of the regression model that has been previously formed [22].

There are several penalty functions that are commonly used as variable selection methods in penalized regression analysis, namely Ridge Regression, LASSO, SCAD, MCP, and Adaptive LASSO. The suitability of a penalty function for describing PISA data has been explored in a study comparing four penalty functions used in convex penalized likelihood: Ridge Regression, LASSO, MCP, and SCAD, applied to 200 independent variables related to the 2015 PISA mathematics scores in Indonesia. The results showed that the SCAD method produced the simplest and nearly unbiased model [23].

SCAD works by shrinking the parameter until its value approaches or even becomes zero, so that the minimum variance estimation will be obtained [24]. It is very suitable to be applied to PISA data because it has quite a lot of research variables. In addition, SCAD has an uncomplicated explanation and is quite easy to use because it has the same local and global asymptotic properties [25].

SCAD has also been used in several studies, especially as a variable selection method. In one study, SCAD was applied to select fixed-effect variables in the AFT (Accelerated Failure Time) model with random-effect, the results obtained stated that SCAD identified important variables better than LASSO [26]. Recent research combined SCAD with Lavenberg-Marquardt Artificial Neural Network (LM-ANN) to select variables in the Quantitative Structure-Activity Relationship (QSAR) study. The results state that SCAD is very efficient to be used as a variable selection method and the combination of the two methods produces an accurate model [27]. SCAD was used as an approach to select variables in the hypertension component data in 2009 survey results entitled "Survey on Living with Chronic Diseases in Canada." The results state that the sample-based SCAD estimator is consistent in the variable selection process and parameter estimation [28].

Based on the explanations discussed in the previous paragraphs, to see the factors that affect the reading literacy scores of Indonesian students in PISA 2018 is to apply penalized regression with the SCAD penalty function to the PISA 2018 survey data. Applying SCAD to PISA 2018 data will select the independent variables and produce variables that significantly affect the reading literacy scores of Indonesian students.

2. METHODS

Material and Data

The data used in this study is secondary data, namely PISA data in 2018 obtained through the PISA database on the official website of the Organization for Economic Cooperation and Development (OECD), available on the OECD website: https://www.oecd.org/pisa/data/2018database/. Questionnaire data for students in Indonesia is used, which has a total of 2,759 observations with a total of 300 variables used. The variables used in this study are only variables related to reading literacy scores.

This research focuses on reading literacy in PISA 2018, so the reading literacy score will be used as the dependent variable (Y) and the independent variables (X) used are only variables in the questionnaire related to the reading framework. A detailed explanation of the questionnaire filled out by students participating in the PISA 2018 program is written in the PISA 2018 compendia.

Based on the PISA 2018 compendia, it is known that there are five sub-themes in the main questionnaire given to students, namely EC (Educational Carrer), ICT (Information Communication Technology), WB (Well-being), FL (Financial Literacy), and ST (Student). From these five sub-themes, 586 variables related to the reading framework were obtained and will be used as indicators that affect

reading literacy scores. However, for Indonesian students, the EC, ICT, and WB sub-themes were not applied to the national questionnaire. Only two sub-themes were applied to the main questionnaire for Indonesian students, but the FL sub-theme was not related to the PISA reading framework. Therefore, only variables within the ST sub-theme were used with a total of 300 independent variables.

Research Method

This research was conducted with the help of RStudio and Posit application with R version 4.3.2 to conduct data analysis. The research procedure is represented in Figure 2.

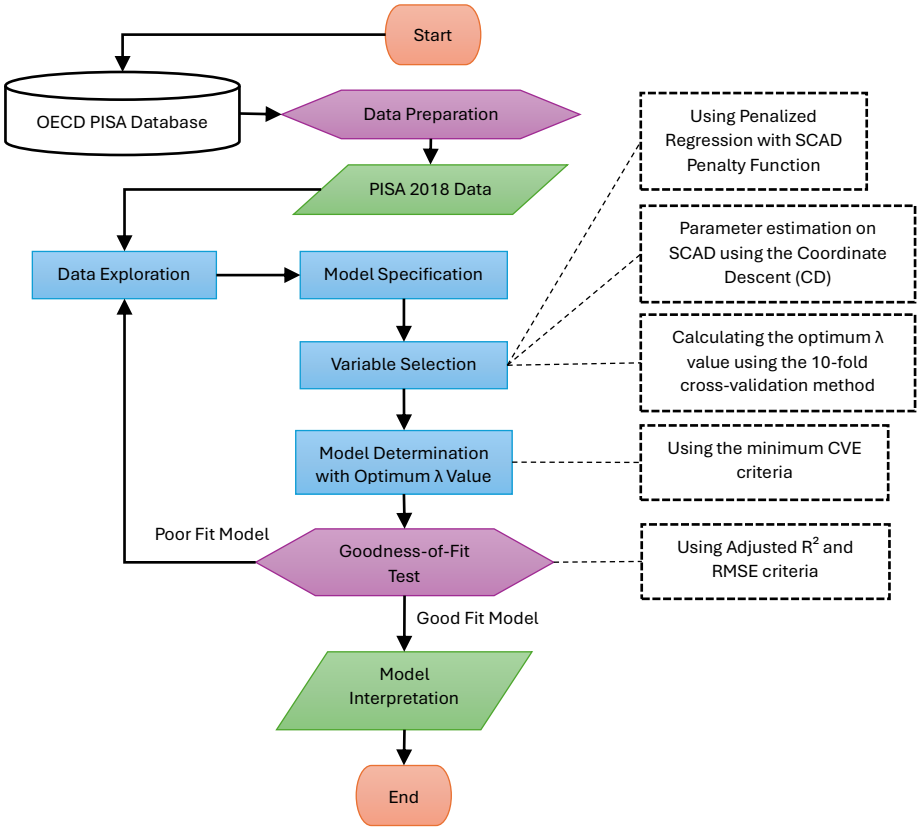


Figure 2. Flow Chart of Research Procedure

The data analysis procedure based on Figure 2 is as follows:

1. Data preparation, adjusting the variables used to only those related to the reading framework.
2. Importing data, the data that has been prepared is imported into R.
3. Exploring the data, using R to look at descriptive statistics of the variables related to the demographics of the PISA 2018 data respondents and see their relationship with literacy scores using frequency tables.
4. Create dummy variables for all categorical variables that have more than two categories, which results in a total of 1076 independent variables.
5. Conducting multicollinearity test, using poly-correlation and polyserial correlation, to see the correlation value between independent variables.

Drasgow [29] discusses polychoric and polyserial correlations, as follows:

$$\rho_{\text{pp}} = \frac{\rho_{\text{ps}}}{\sigma_{\text{d}}} \sum_{j=1}^{s-1} \phi(\tau_j)(d_{j+1} - d_j) \tag{1}$$

with σ_d^2 is variance of D and $\phi(\tau) = (2\pi)^{-1/2}\exp(-\tau^2/2)$ is normal distributed form.

$$\mathrm{P_r}(D = d_k|x_k) = \phi(\tau_j^*) - \phi(\tau_j^*), j = 1, \dots, s \quad (2)$$

where,

$$\tau_j^* = \frac{\tau_j - \rho z_k}{(1-\rho^2)^{1/2}} \quad , \qquad \phi(\tau_j) = \int_{-\infty}^{\tau} \phi(t)\, dt$$

6. Performing model specification, modeling is done using a linear regression model containing 300 independent variables, or after being converted into dummy variables, there are 1076 dummy variables on reading literacy scores.
7. Selecting variables, the variable selection process is carried out simultaneously with the determination of the optimum λ. This process is based on the Smoothly Clipped Absolute Deviation (SCAD) function as a variable selection method with the Coordinate Descent (CD) algorithm to help the parameter estimation process and 10-fold cross-validation for the calculation of the optimum lambda. In this process, j-iterations will be tried, where j is the number of λ to be tried on the model and in each iteration a different λ value will be used.
8. Final model, the final model chosen is the model with the optimum λ value. The optimum λ value is obtained based on the minimum CVE (Cross-Validation Error) value.
9. Model feasibility test, after obtaining the model at point (8), it is necessary to check the feasibility of the model with a goodness-of-fit test, namely with Adjusted R^2 and RMSE.
10. Model interpretation, by interpreting the final model obtained, it can be seen which variables, or in this case factors, have a significant effect on the dependent variable (Indonesian students' reading literacy scores).

3. RESULTS

Descriptive Analysis

Before analyzing the reading literacy scores of Indonesian students (*readscore*), it is important to note the characteristics of the PISA data, where the PISA data is the result of a questionnaire. Questionnaire data is closely related to demographic aspects. Therefore, it is necessary to look more deeply into the demographics of PISA 2018 respondents in Indonesia. Looking at the demographics of PISA 2018 respondents in Indonesia based on several variables, such as: Gender, Grade, and Parents' education.

Table 2. Student Demographics of 2018 PISA Respondents in Indonesia

Gender	Age	Grade						Students
		7th	8th	9th	10th	11th	12th	
Girl	15.848	2	33	464	1001	51	3	1554
Boy	15.847	5	38	418	699	33	12	1205

Based on Table 2, PISA 2018 respondents in Indonesia were dominated by female students at 56% and 44% for male students. PISA 2018 respondents in Indonesia had an average age of 15.848, which is in line with global PISA rules that targeted respondents to be 15-year-old students [9].

Students aged 15 years in Indonesia are in various grades, ranging from junior high school grade 7 to senior high school grade 12. However, 61.617% of PISA respondents were grade 10 students. Grade 10 students in Indonesia, if they follow the standards for international education programs issued by UNESCO, called ISCED (International Standard Classification of Education), are at ISCED level 3 [30].

PISA respondents in Indonesia are at ISCED level 3 or high school level and ISCED 2 or junior high school level. The division of ISCED levels for PISA respondents is shown in Table 3.

Table 3. ISCED Level of Students PISA 2018 Respondents

ISCED Level	Number of Students	Percentage	Mean *Readscore*
3 (high school)	1799	65%	445.538
2 (junior school)	960	35%	401.888

Meanwhile, one of the demographics that needs to be discussed is parental education. On average, the highest education attained by parents in Indonesia is ISCED level 4 or equivalent to a bachelor's degree. In addition, we can also roughly see that the effect of parental education has an influence on *readscore* in PISA 2018 [31], [32].

After discussing a little about the *readscore* compared to the PISA respondents' parents' education, we will discuss more about the *readscore*. *Readscore* data is continuous data that has a normal distribution, which can be seen in the histogram.

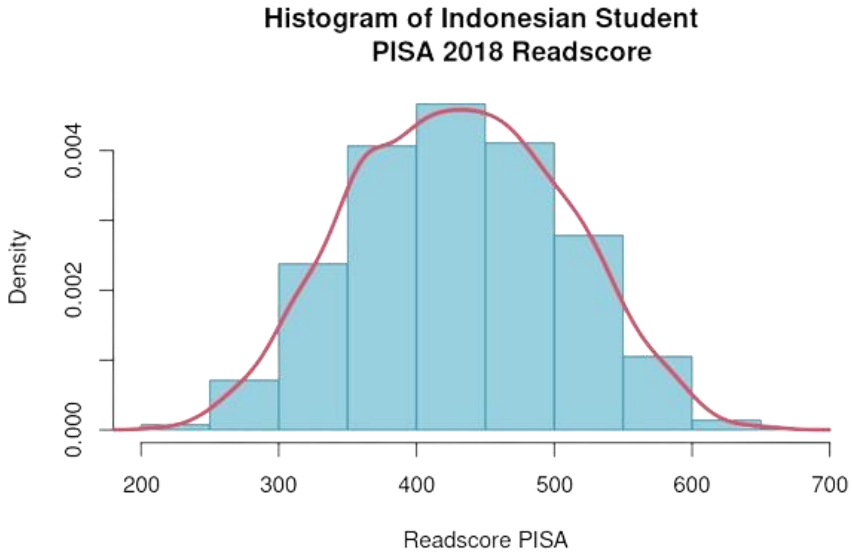


Figure 3. Histogram of PISA 2018 *Readscore* of Students in Indonesia

Based on the histogram in Figure 3, the PISA *readscore* data has a normal distribution, because it has a curve shape that tends to be symmetrical and resembles a bell. This indicates that the mean, median, and mode values are in the same interval. The interval is *readscore* 400 – 450 is the interval with the peak density, which is 0.004.

Before modeling the PISA data, it is necessary to check the independence assumption. This is done to ensure that there are no variables that have a perfect correlation, so that the resulting model will not be biased. Checking the independence assumption is done with poly-correlation to see the correlation between two categorical variables and polyserial correlation to see the correlation between categorical variables and continuous variables using equations (1) and (2). So that it can be seen whether there is an association or relationship between the independent variables used. A summary of the test results is shown in Table 4.

Table 4. Correlation between Independent Variables (X)

$X_{i,j}$	Correlation Value	$X_{i,j}$	Correlation Value
$X_{292,3}$	-0.920	$X_{41,39}$	-0.875
$X_{293,3}$	-0.952	$X_{42,39}$	-0.886
$X_{297,3}$	-0.824	$X_{41,40}$	0.959
$X_{6,4}$	0.886	$X_{42,40}$	0.912
$X_{9,4}$	0.876	$X_{42,41}$	0.904
$X_{296,4}$	-0.969	$X_{43,42}$	0.854

$X_{i,j}$	Correlation Value	$X_{i,j}$	Correlation Value
$X_{293,5}$	-0.924	$X_{44,42}$	0.800
$X_{297,5}$	-0.877	$X_{44,43}$	0.927
$X_{292,6}$	-0.854	$X_{249,248}$	0.901
$X_{296,6}$	-0.823	$X_{263,262}$	0.813
$X_{294,8}$	-0.932	$X_{265,264}$	0.904
$X_{295,8}$	-0.959	$X_{266,264}$	0.912
$X_{296,8}$	-0.817	$X_{266,265}$	0.909
$X_{297,8}$	-0.848	$X_{286,285}$	0.872
$X_{294,9}$	-0.980	$X_{287,285}$	0.847
$X_{295,10}$	-0.955	$X_{287,286}$	0.856
$X_{297,10}$	-0.961	$X_{290,289}$	0.977
$X_{294,11}$	-0.822	$X_{291,289}$	0.977
$X_{296,11}$	-0.825	$X_{296,292}$	0.884
$X_{29,16}$	-0.854	$X_{297,292}$	0.857
$X_{37,36}$	0.877	$X_{296,293}$	0.860
$X_{38,36}$	0.813	$X_{297,293}$	0.885
$X_{38,37}$	0.907	$X_{296,294}$	0.905
$X_{289,38}$	0.955	$X_{297,294}$	0.891
$X_{299,38}$	0.989	$X_{296,295}$	0.886
$X_{40,39}$	-0.878	$X_{297,295}$	0.908

The results shown in Table 4 are only variables that have a correlation value above 0.8 which indicates that there is a strong relationship between the independent variables [33]. When there is a strong relationship between independent variables, the assumption of independence cannot be fulfilled. Continuing the analysis with ordinary linear regression is not possible because it will produce a less accurate prediction model [34].

The process of selecting SCAD penalty variables to see which independent variables affect the *readscore* (Y) by overcoming the assumption of mutual independence. In this process, parameter estimation is performed simultaneously with the determination of the optimum lambda. The calculation is done using the CD (Coordinate Descent) algorithm method by trying different λ (lambda) values in each iteration.

One of the important things in the SCAD penalty is the optimum value of λ or in this case referred to as the tuning parameter of the model. Obtaining the optimum value uses the cross-validation method, in this case 10-fold cross-validation is used.

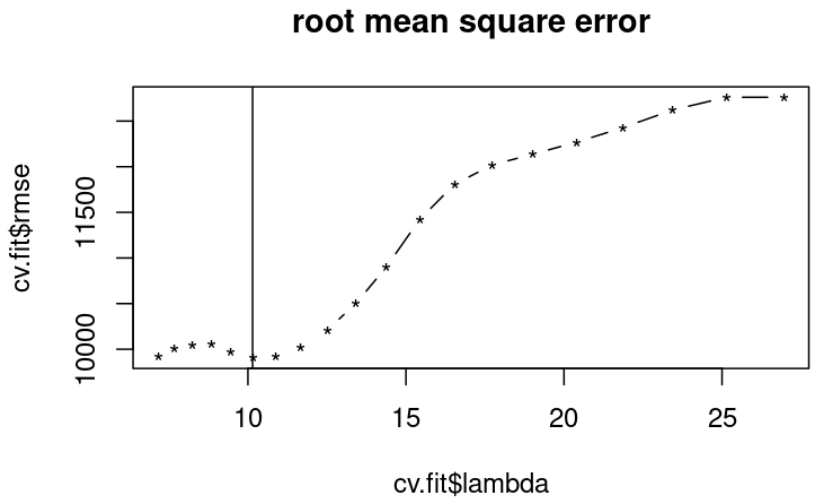


Figure 4. Cross-validated Error Curve for Lambda Value

Based on Figure 4, the determination of the optimum lambda is calculated using an iteration process and is carried out in stages. The optimum lambda value is obtained when it reaches the minimum RMSE value. In this case, the optimum lambda value or tuning parameter λ obtained is 10.147. The model formed when the lambda value is the best model based on cross-validation error. The SCAD equation with the lambda (λ) obtained and the natural parameter (a) value that has been given of 3.7 can be written as in the equation below.

$$P_{\lambda,a}^{SCAD}(\beta_j) = \begin{cases} 10.147\,|\beta_j| & ;0 \le |\beta_j| \le 10.147 \\ \dfrac{(3.7)(10.147)\,|\beta_j| - 0.5\left(|\beta_j|^2 + (10.147)^2\right)}{(3,7)-1} & ;10.147 < |\beta_j| \le (3.7)(10.147) \\ \dfrac{10.147^2(3.7^2+1)}{2(3.7-1)} & ;|\beta_j| > (3.7)(10.147) \end{cases}$$

Parameter Estimation Results

After determining the optimum lambda value for the model based on the minimum CVE results, we can see the parameter estimation results in the SCAD equation above for the model with the optimum lambda value. The results obtained are that there are 27 dummy variables that have a coefficient value of ≠ 0 or can be said to have an influence on the *readscore*. The 27 dummy variables are derived from 22 original variables. The description of the 22 variables that significantly affect the model will be explained further.

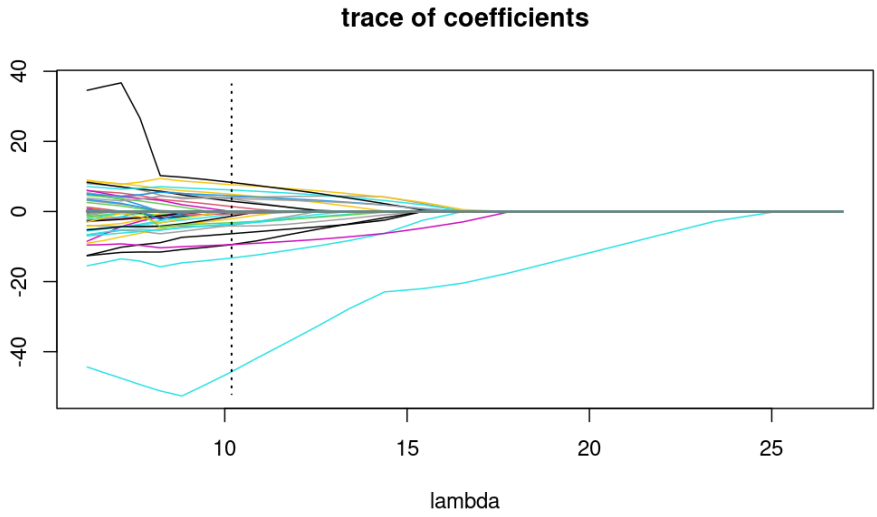


Figure 5. Trace of coefficient plot with SCAD penalty function

The parameter estimation process is depicted in the form of a plot in Figure 5. Each color line represents a change in the coefficient value of the variables that affect the model. The change in the coefficient value is based on the different lambda values that are tried to be applied to the model. There are only two variables that have significantly different coefficient values, when compared to the coefficient values of other variables, or can be said to be outliers.

However, overall, all coefficient values show a regular pattern, tend to converge, and do not fluctuate. Especially when the lambda value is in the range of 10, marked with a dotted line, which means there is a plateau where the coefficient remains relatively stable. Therefore, it can be said that the performance of the model, given the SCAD penalty function, is able to carry out an effective optimization process.

Based on the parameter estimation results, the final model for student *readscore* in Indonesia is written in the equation below as follows.

$$\hat{Y} = 440.226 - 6.366\,X_{28,1} - 45.929\,X_{29,1} + 4.034\,X_{29,1} - 4.151\,X_{56,1} + 1.464\,X_{56,3} - 3.958\,X_{96,3} - 0.366\,X_{104,3} + 4.953\,X_{107,5} - 0.838\,X_{128,6} - 9.491\,X_{130,6} - 0.249\,X_{132,2} - 2.132\,X_{133,2} + 3.664\,X_{136,1} + 2.963\,X_{138,1} + 6.161\,X_{153,4} + 0.039\,X_{158,1} + 7.662\,X_{158,2} - 1.418\,X_{166,4} - 3.697\,X_{187,1} - 13.315\,X_{191,1} - 9.478\,X_{191,2} - 0.685\,X_{196,2} + 4.505\,X_{243,4} + 8.348\,X_{263,1} - 3{,}201\,X_{263,3} - 3.530\,X_{280,1} + 3.705\,X_{282,4} + \varepsilon$$

Goodness-of-fit Test

After the final model is obtained, it is necessary to test the goodness-of-fit of the model to ascertain whether the resulting model is the best model. In this study, because there is only one model produced, the normality assumption and homogeneity assumption will be checked first before conducting a goodness-of-fit test by looking at the RMSE and Adjusted R^2 values.

RMSE and R^2

The resulting model has met the assumptions of normality and homogeneity, then to test the model is appropriate or not is done by looking at the RMSE and R^2 values for the linear model that has been given the SCAD penalty function displayed in Table 5.

Table 5. Goodness-of-Fit Test Result

Criteria	Linear Model-SCAD
R^2	0.974
Adjusted R^2	0.967
RMSE	61.888

The final model has an R^2 value of 0.974. This indicates that the model formed can explain 97.388% of the *readscore,* and the remaining 2.612% is explained by variables outside the model. This is in line with the Adjusted R^2 value of 0.967, which indicates that the resulting model is very good.

4. DISCUSSIONS

There are 27 dummy variables that have an influence on student *readscore* in Indonesia. When these variables are returned to their original constituent variables, there are 22 variables that have an influence on the *readscore*. The following will describe the variables that have a significant effect on student reading literacy scores in Indonesia, starting with variables that have a large influence, both positive and negative.

Grouping the variables for the effectiveness of interpreting the model, the 22 variables that have an influence on the *readscore* are grouped into eight group aspects, as follows.

Learning Facilities at Home

In this case, the learning facilities aspect includes the number of computers (desktop, laptop, or notebook) at home (X_{29}) and the number of cellphones with internet access at home (X_{28}). The aspect of learning facilities at home can influence students' reading literacy scores in Indonesia, in line with the findings in [35].

Teacher's Teaching Strategy

This aspect of the teacher's teaching strategy refers to the way the teacher teaches in class, especially during Indonesian language lessons. The variable that relates to this is "At the beginning of the lesson, the teacher gives a brief summary of the previous learning during Indonesian lessons" (X_{56}). The delivery of the previous learning summary can affect students' reading ability [36].

Reading Habits

The statement that reading books is a waste of time has a negative connotation. Students who agree with this can be interpreted as that these students are not fond of reading or do not have a high interest in reading. This is in line with the results obtained in this study, students who have the opinion that

reading books is a waste of time are considered to have a lower readscore or reading ability of 3.958 units when compared to students who disagrees [37].

Student Learning Strategies

Students who understand and remember reading by reading it aloud to others are considered to have a lower readscore or reading ability of 0.838 units when compared to students who do not do that. Reading aloud techniques are indeed one of the methods that are often applied to help students improve their reading skills [38], [39].

Student Awareness

The next attitude or response is about checking the official website of the cellular phone operator that made the invitation before replying to the email, or what can be called a background check. Students who feel it is important to do a background check are considered to have a higher readscore of 2.963 units when compared to students who do not do this. Background check is strongly related to the habit of reading carefully, which can improve reading ability [40].

Student Self-Confidence

The aspect of student self-confidence, related to the nature of students who are happy if they feel an increase in ability when completing assignments (X_{153}). Students who strongly agree with the opinion that the pleasure they get from doing an assignment is when their ability increases compared to before are considered to have a higher reading ability or *readscore* of 6.161 units when compared to disagree with this opinion. This argument indicates that students like to do assignments and reflect students' disciplinary attitude, where it can improve students' reading ability [41], [42].

General Knowledge

Students who had only heard of the topic and could not explain the causes of poverty were assessed to have a lower *readscore* of 0.685 when compared to students who could explain the outline of the topic. This is in line with research which states that students' understanding of a particular topic is based on their reading ability [43].

Social Empathy

Students who agreed that immigrant children should have the same opportunities for education as other children in the country were rated as having a higher *readscore* of 4.505 compared to students who disagreed with the statement. The statement leads to the empathetic nature of students, so it can be said that students who have high empathy values are also considered to have higher reading skills [44].

5. CONCLUSION

Estimating parameters with the SCAD penalty function, resulting in 22 variables that significantly affect students' reading literacy in Indonesia. These variables are grouped into eight groups of variables, namely: Home learning facilities, Teacher teaching strategies, Reading habits, Student learning strategies, Student awareness, Student confidence, Student general knowledge, and Student social empathy. The aspect of learning facilities at home is the most influential aspect on students' reading literacy in Indonesia. This is followed by students' general knowledge on topics such as: climate change; global warming; immigrants; and causes of poverty, which are considered to have positive

The use of SCAD penalty function is also considered very good in modeling PISA data, especially on PISA 2018 data. This is in line with the result of Adjusted R^2 obtained 0.967 for the model formed, these results indicate that the model formed is very good. Therefore, it can be said that the SCAD penalty function is suitable for modelling PISA data.

PISA survey data will also remain interesting data to be studied in every period, because each period has a topic with a different research focal point. For example, in 2015 the focus was on math literacy, while in 2018 it shifted to reading literacy. The latest PISA 2022 is also very interesting to study because it discusses the effect of the pandemic on changes in student learning performance. In addition, it is quite interesting to look more deeply at PISA data, especially financial literacy which is also studied in 15-year-old students and aspects of global competencies which in each period examine various things or topics about students' cognitive abilities.

The use of different analysis methods can also be tried to be applied to PISA data. Using analytical methods that can overcome multicollinearity problems, such as applying other penalty function groups like LASSO, adaptive LASSO, and Minimax Concave Penalty (MCP).

6. ACKNOWLEDGMENTS

We express our deepest gratitude to the Faculty of Mathematics and Natural Science, Universitas Negeri Jakarta through the competitive grant program, BLU FMIPA, which has funded this research.

7. REFERENCES

[1] R. R. Patria, "Why Indonesian Students Struggle in Reading Test? An Insight from PISA 2018 Result," in *Proc. Int. Conf. Educ. Assess. Policy (ICEAP 2020)*, vol. 545, pp. 29–40, 2021.

[2] A. Patterson, D. Roman, M. Friend, J. Osborne, and B. Donovan, "Reading for meaning: The foundational knowledge every teacher of science should have," *Int. J. Sci. Educ.*, vol. 40, no. 3, pp. 291–307, 2018.

[3] R. Rintaningrum, "Explaining the Important Contribution of Reading Literacy to the Country's Generations: Indonesian's Perspectives," *Int. J. Innov. Creat. Change*, vol. 5, no. 3, pp. 936–953, 2019.

[4] V. S. Damaianti, Y. Abidin, and R. Rahma, "Higher order thinking skills-based reading literacy assessment instrument: An Indonesian context," *Indones. J. Appl. Linguist.*, vol. 10, no. 2, pp. 513–525, 2020. doi: 10.17509/ijal.v10i2.28600.

[5] J. W. Miller and M. C. McKenna, *World Literacy: How Countries Rank and Why It Matters*, New York: Routledge, 2016. doi: 10.4324/9781315693934.

[6] I. Pratiwi, "Efek Program Pisa Terhadap Kurikulum Di Indonesia," *J. Pendidik. dan Kebudayaan*, vol. 4, no. 1, p. 51, 2019.

[7] I. V. S. Mullis, *TIMSS 2019 Assessment Frameworks*, M. O. Martin, Ed. Boston College: TIMSS & PIRLS Int. Study Center, 2017.

[8] I. V. S. Mullis, *PIRLS 2021 Reading Assessment Framework*, M. O. Martin, Ed. Boston College: TIMSS & PIRLS Int. Study Center, 2019.

[9] OECD, *PISA 2018 Assessment and Analytical Framework*, Paris: OECD Publishing, 2019, doi: 10.1787/b25efab8-en.

[10] V. M. Santi, K. A. Notodiputro, L. Indahwati, and B. Sartono, "Restricted Maximum Likelihood Estimation for Multivariate Linear Mixed Model in Analyzing Pisa Data for Indonesian Students," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 2, pp. 607–614, 2022. doi: 10.30598/barekengvol16iss2pp607-614.

[11] V. M. Santi, I. Hasari, and D. Handayani, "Linear Mixed Models to Analyze Indonesia's Pisa Reading Literacy Score," *J. Phys. Conf. Ser. AIP Conf. Proc.*, vol. 2982, 2024. doi: 10.1063/5.0183760.

[12] V. M. Santi, R. I. Hayati, and B. Sumargo, "The Group Selection of Variables That Effected Science Scores of Indonesia's Pisa Using Group LASSO," *J. Phys. Conf. Ser. AIP Conf. Proc.*, vol. 2982, 2024. doi: 10.1063/5.0183761.

[13] OECD, *PISA 2018 Results (Volume I): What Students Know and Can Do*, vol. I, Paris: OECD Publishing, 2019.

[14] A. Schleicher, *PISA 2018: Insights and Interpretations*, OECD, 2019. [Online]. Available: https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf

[15] İ. Koyuncu and T. Fırat, "Investigating reading literacy in PISA 2018 assessment," *Int. Electron. J. Elem. Educ.*, vol. 13, no. 2, pp. 263–275, 2020.

[16] L. Hewi and M. Shaleh, "Refleksi Hasil PISA (The Programme for International Student Assessment): Upaya Perbaikan Bertumpu Pada Pendidikan Anak Usia Dini," vol. 4, no. 1, pp. 30–41, 2020.

[17] S. P. Liestari and M. Muhardis, "Kemampuan Literasi Membaca Siswa Indonesia (Berdasarkan hasil UN dan PISA)," *Indones. J. Educ. Assess.*, vol. 3, no. 1, p. 24, 2020. doi: 10.26499/ijea.v3i1.53.

[18] F. Nur'Aini, I. Ulumuddin, L. Sulinar Sari, and S. Fujianita, *Risalah kebijakan nomor 3, April 2021: meningkatkan kemampuan literasi dasar siswa Indonesia berdasarkan analisis data PISA 2018*, 2021.

[19] M. Ansori and S. Iswati, *Metodologi Penelitian Kuantitatif*, 1st ed. Surabaya: Airlangga Univ. Press, 2009.

[20] Suyono, *Analisis Regresi untuk Penelitian*, 1st ed. Yogyakarta: Deepublish, 2015.

[21] R. H. Myers and J. S. Milton, *A First Course in the Theory of Linear Statistical Models*, Boston: PWS-KENT Publishing Company, 1991.

[22] L. Wang, Y. Kim, and R. Li, "Calibrating Nonconvex Penalized Regression in Ultra-high Dimension," *Ann. Stat.*, vol. 41, no. 5, pp. 2505–2536, 2013. doi: 10.1214/13-AOS1159.

[23] V. M. Santi, K. A. Notodiputro, and B. Sartono, "Variable selection methods applied to the mathematics scores of Indonesian students based on convex penalized likelihood," *J. Phys. Conf. Ser.*, vol. 1402, no. 7, 2019. doi: 10.1088/1742-6596/1402/7/077096.

[24] J. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *J. Am. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001. doi: 10.1198/016214501753382273.

[25] S. Kwon and Y. Kim, "Large sample properties of the SCAD-penalized maximum likelihood estimation on high dimensions," *Stat. Sin.*, vol. 22, no. 2, pp. 629–653, 2012. doi: 10.5705/ss.2010.027.

[26] E. Park, S. Kwon, J. Kwon, R. Sylvester, and I. Do Ha, "Penalized h-likelihood approach for variable selection in AFT random-effect models," *Stat. Neerl.*, vol. 74, no. 1, pp. 52–71, 2019. doi: 10.1111/stan.12179.

[27] Z. Mozafari, M. Arab Chamjangali, M. Arashi, and N. Goudarzi, "Performance of smoothly clipped absolute deviation as a variable selection method in the artificial neural network-based QSAR studies," *J. Chemom.*, vol. 35, no. 5, pp. 1–20, 2021. doi: 10.1002/cem.3338.

[28] C. Xu, J. Chen, and H. Mantel, "Smoothly clipped absolute deviation in analysis of survey data," in *Proc. Survey Methods Sect.*, pp. 1–7, 2010. [Online]. Available: https://ssc.ca/sites/default/files/data/Members/public/about/Awards/Survey/2010_Proceedings/SSC2010_XChen.pdf

[29] F. Drasgow, "Polychoric and Polyserial Correlations," *Wiley StatsRef: Statistics Reference Online*, Wiley, 2014. [Online]. Available: https://doi.org/10.1002/9781118445112.stat02493.

[30] UNESCO Institute for Statistics, *International Standard Classification of Education: ISCED 2011*, Montreal: UNESCO, 2012.

[31] E. Badroeni and N. Cahyati, "Kegiatan Home Literacy Dalam Mengembangkan Kemampuan Awal Membaca Anak Usia Dini di Masa Wfh," *Jurnal Golden Age*, vol. 4, no. 1, pp. 160–166, 2020.

[32] F. Niklas and W. Schneider, "Intervention in the home literacy environment and kindergarten children's vocabulary and phonological awareness," *First Language*, vol. 37, no. 5, pp. 433–452, 2017, doi: 10.1177/0142723717698838.

[33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, vol. 103. Springer, 2013, doi: 10.1007/978-1-4614-7138-7.

[34] T. T. Pardede, B. Sumargo, and W. Rahayu, "Penerapan Regresi Least Absolute Shrinkage and Selection Operator (LASSO) Untuk Mengidentifikasi Variabel yang Berpengaruh terhadap Kejadian Stunting di Indonesia," *Jurnal Statistika dan Aplikasinya*, vol. 6, no. 1, 2022.

[35] V. M. Santi, S. Azzahra, and D. Siregar, "Analisis Skor Literasi Membaca Siswa Indonesia Menggunakan Linier Mixed Models," *MUST: Journal of Mathematics Education, Science and Technology*, vol. 7, no. 2, p. 116, 2022, doi: 10.30651/must.v7i2.14420.

[36] M. H. Davis, J. M. McPartland, C. Pryseski, and E. Kim, "The effects of coaching on English teachers' reading instruction practices and adolescent students' reading comprehension," *Literacy Research and Instruction*, vol. 57, no. 3, pp. 255–275, 2018, doi: 10.1080/19388071.2018.1453897.

[37] L. Scott and E. Saaiman, "Promoting reading skills or wasting time? Students' perceived benefits of reading in an intermediary programme at the Vaal University of Technology," *Reading & Writing*, vol. 7, no. 1, 2016, doi: 10.4102/rw.v7i1.82.

[38] A. and M. Suryaman, "EYL's perceptions about the use of reading aloud technique for students' cognitive and affective in reading skill," *Journal of English Educational Study (JEES)*, vol. 5, no. 2, pp. 152–161, 2022.

[39] M. Khalid, M. Sajid, and H. Kassim, "The Effects of Reading Aloud Strategies on Text Level Difficulties, Reading Proficiency and Reading Comprehension Skill," *International Journal of Language Education and Applied Linguistics*, 2019. [Online]. Available: http://ijleal.ump.edu.my/.

[40] J. Levashina, J. A. Peck, and L. Ficht, "Don't Select until You Check: Expected Background Checking Practices," *Employee Responsibilities and Rights Journal*, vol. 29, no. 3, pp. 127–148, 2017, doi: 10.1007/s10672-017-9294-4.

[41] J. Ciarrochi, P. C. L. Heaven, and F. Davies, "The impact of hope, self-esteem, and attributional style on adolescents' school grades and emotional well-being: A longitudinal study," *Journal of Research in Personality*, vol. 41, no. 6, pp. 1161–1178, 2007, doi: 10.1016/j.jrp.2007.02.001.

[42] V. M. Santi, R. Kamilia, and F. Ladayya, "Multilevel regression with maximum likelihood and restricted maximum likelihood method in analyzing Indonesian reading literacy scores," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 4, pp. 1423–1432, 2022, doi: 10.30598/barekengvol16iss4pp1423-1432.

[43] E. D. Hirsch, "Reading Comprehension Requires Knowledge of Words and the World," *American Educator*, vol. 27, no. 1, pp. 10–13, 2003.

[44] S. A. Roza and S. R. K. Guimarães, "The relationship between reading and empathy: An integrative literature review," *Psicologia - Teoria e Prática*, vol. 24, no. 2, 2022, doi: 10.5935/1980-6906/eptppe14051.en.