Variable that influence achievement of indonesian students in the program international student assessment (PISA) 2015 using structural equation modelling (SEM)

V Eminita¹*, K A Notodiputro² and B Sartono²

¹Program Study of Mathematic Education, Universitas Muhammadiyah Jakarta, Jl. KH. Ahmad Dahlan, Tangerang Selatan 15419, Indonesia
²Statistics Department, Institut Pertanian Bogor, Jl. Raya Dramaga, Bogor 16680, Indonesia

*Corresponding author’s email: phiartea.emn@gmail.com

Abstract. The Program for International Student Assessment (PISA) is an examination system that evaluates the education system every 3 years to 15-year-old students randomly selected. Indonesia ranking was generally still below the average of the Organisation for Economic Cooperation and Development (OECD). In addition, the results of the PISA study are not only interpreted as ranking, but also important to carry out follow-up based on diagnoses produced from the PISA diagnostic survey to determine the weaknesses and strengths of Indonesian students and evaluate the education system in Indonesia. It is important to consider the use of the right and simple model to see the relationship between the diagnostic results of the performance of students' PISA achievements in 2015. The purpose of this study was to select variables that significantly influence PISA achievement in 2015 using the SEM regularization model using The Least Absolute Shrinkage and Selection Operator (LASSO) method. The results of the analysis show that the variables that significantly effect students' PISA results are the students' insight into the environmental, the students' perceptions of science truth, and the students' participation in the International Standard Classification of Education level 0 (ISCED 0).

1. Introduction

Programme for International Student Assessment (PISA) is a test system that evaluates the education system from 72 countries around the world. The specific objectives of PISA in 2015 not only assess what students know about science, but also what they can do with what they know, and how they can creatively apply their scientific knowledge [1]. PISA is held by the Organization for Economic Cooperation and Development (OECD) every 3 years for 15-year students randomly selected. Indonesian students participated in the PISA since 2000. Based on the PISA student ranking results, Indonesia is at the bottom, above Brazil, Peru, Lebanon, Tunisia, Kosovo, Algeria and the Dominican Republic. The 2015 PISA result for Indonesia are ranked 65th grade Mathematics, 66th grade Reading literacy, 64th grade science.

In 2015 Indonesia was claimed to have experienced a significant increase in achievement which brought Indonesia 6 ranked upwards compared to the previous year, but this achievement was generally still below the average of the OECD. In addition, the results of the PISA study are not only interpreted as rankings, but also important to carry out follow-up based on diagnoses produced from
the PISA diagnostic survey. It aims to find out the weaknesses and strengths of Indonesian students and evaluate the education system in Indonesia [2]. Identity factors, socioeconomic and cultural conditions, computer ownership, and books are the main factors that are significantly affect the achievement of mathematics literacy in Indonesian students at PISA 2012 [3].

The diagnostic survey assessed in PISA contains many latent variables that measure students' background and characteristics, so it is important to consider the use of appropriate and simple models to see the relationship between student background and characteristics of the 2015 PISA performance. Least Absolute Shrinkage and Selection Operator (LASSO) method has been popularly used to obtain a simpler model if there are many variables, especially if \( n \ll p \) [4]. However, this method cannot be applied to data containing latent variables. The estimation methods of possible penalties on structural equation models (SEM) had been developed to handle it, namely Regularized SEM [5]. This method minimizes certain parameters in SEM so that the model obtained is easier to understand and simpler [6]. This method is expected to provide a description of the relationship between students 'backgrounds and characteristics of students' PISA performance simpler so that they can be interpreted and are more efficient in terms of evaluating education in Indonesia. The purpose of this study is to select variables, both exogenous latent variables and observed variables that influence the performance of 2015 PISA achievement students using Regularized SEM.

2. Methods

2.1 Data

The data in this study are the results of the 2015 PISA diagnostic survey, especially participants of 15-year-old Indonesian students with a sample of 6513 students. This data is obtained from http://www.oecd.org/pisa/data/. The missing observations in this study were issued, so the amount of data was reduced to 4,522. The response variable observed in this study is the result of PISA achievement, namely the value of mathematics, science, and reading. Furthermore, the explanatory variable is the student's background and the characteristics of students especially in the field of science (Table 1).

| No | Variable | Description |
|----|----------|-------------|
| 1  | Gender   | Student Gender |
| 2  | ISCED0   | Following ISCED 0 |
| 3  | Discipline | Students’ discipline in the learning process |
| 4  | Learning Science | The process of learning science in class |
| 5  | Insight  | Students’ insight into the environment |
| 6  | Interest | Students’ interest in learning science |
| 7  | Benefits | Students' perceptions of the benefits of learning science |
| 8  | Science performance | Students’ performance in science activities |
| 9  | Truth    | Students' perceptions of science truth |

2.2 Model and Analysis

In Table 1, it can be seen that the independent variables in this study consist of two observed variables and nine latent variables, so that the linear model used in describing the relationship between the response variables with independent variables is SEM. In this study it is expected that the model obtained is a model that is easier to understand and simpler. SEM Regularization is one method that can be used to achieve this goal.

The model used in this study is SEM, i.e.

\[
\eta = B\eta + \Gamma\xi + \zeta
\]
which is a structural model, where $\eta_{3x1}$ is a vector of endogenous latent variables, $\xi_{9x1}$ is a vector of exogenous latent variables, $\zeta_{3x1}$ is a structural error matrix, $B_{3x3}$ is a coefficient matrix for endogenous latent variables, $\Gamma_{3x9}$ coefficient matrix for exogenous latent variables.

$$x = \Lambda_x \xi + \delta,$$
$$y = \Lambda_y \eta + \varepsilon$$

(2)

is a measurement model for manifest $x$ and $y$ variables that are $q \times 1$ and $p \times 1$ respectively $\delta_{q \times 1}$ is a measurement error of $x$ and $\varepsilon_{p \times 1}$ is a measurement error of $y$. $\Lambda_x$ and $\Lambda_y$ are the matrix loading ($\lambda$) factors. This model assumes that $E(\eta) = 0$, $E(\xi) = 0$, $E(\zeta) = 0$, $E(\delta) = 0$, $E(\varepsilon) = 0$, $\zeta$ and $\xi$ does not correlate, $\varepsilon$, $\delta$, and $\zeta$ does not correlate, $\delta$ and $\xi$ do not correlate, $\varepsilon$ and $\eta$ do not correlate [7].

SEM generally implements the use of a RAM matrix that requires three matrices to specify SEM as a whole, namely filters $(F_{px(\rho+1)})$, asymmetric (A), and Symmetric (S). The F matrix contains one that corresponds to the manifest and zero variables except that $p$ is the number of manifest variables and $l$ is the number of latent variables. A matrix contains a loading factor which is a regression between manifest variables with latent variables. The S matrix is a square matrix that contains a variety or uniform for each manifest variable relationship with a latent variable. The three matrices are used to calculate the uniform matrix derived from the model, i.e.

$$\Sigma = F(I - A)^{-1}S(I - A)^{-1}F'$$

(3)

The matrix is then entered into the loss function, for example maximum likelihood (ML).

$$F_{ML} = \log(|\Sigma|) + \text{tr}(C^{-1} \Sigma^{-1}) - \log(|C|) - p$$

(4)

where $C$ is the observed covariance matrix, and $p$ is the expected number of parameters.

LASSO is one of the general methods of regularization in regression which suppresses non-essential variables to zero [4]. The LASSO estimation defined by equation (3) uses l1-norm to press $\beta_j$ to 0 with $\lambda$ as the selection parameter.

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

(5)

The best selection of $\lambda$ can use Akaike information criteria (AIC), Bayesian information criteria (BIC), or cross-validation (CV) [8].

SEM regularization (regSEM) generally implements the use of a RAM matrix which has the main benefit of only one matrix needed to capture direct influence as a natural parameter for regularization. RAM makes formulations simpler than having to add penalties to more than one matrix. SEM Regularization is built by adding elements to suppress certain model parameters, namely:

$$F_{\text{regSEM}} = F_{ML} + \lambda P(\cdot)$$

(6)

with $\lambda$ is the regularization parameter with $\lambda \geq 0$. When $\lambda$ is zero, the results obtained are the same as MLE, and when $\lambda$ is very large, all of the marginalized parameters will be reduced to zero. $P(\cdot)$ is a function of the sum of one or more values of the model parameter matrix. Form $P(\cdot)$ in LASSO (l1), which terminates the absolute number of parameters.

RegSEM specifies which parameters are penalized in the SEM model in general. Penetrating parameters in the regSEM in RAM matrix notation, only terminates the A or S matrix (not both) in the model. This is because penalizing parameters of one matrix means that it will also suppress other matrix values that are compatible.

2.3 Stages of Analysis

The data analyzed using Software R.3.5.2 uses the Package "lavaan" and "regsem" with the steps of the research, which first changes the category variable into a numerical variable, namely the dummy variable and Method Of Successive Interval (MSI) [9], this is important in linear modeling. Then the
second step standardizes each data first before being analyzed so that the LASSO method can work properly. The standardization of this data is done so that the data has mean 0 and variance 1 [10]. The third step in this research is to model Y with X using SEM Regularization using selected lambda values with CV regularization SEM results. The choice of lambda is important to get a more concise model. The final step is the interpretation of results, which are the variables that have an important influence on student PISA achievement.

3. Result and Discussion

3.1 Choosing of Selection Parameters (λ)

The choice of the value of λ is very important in obtaining a model of sparsity, which functions to suppress non-essential coefficients to zero [11]. The general technique that is usually used to select the value of λ is cross-validation (CV) based on the Bayesian Information Criteria (BIC) and Root Mean Square Error of Approximation (RMSEA) values. Previously, parameter estimation was done on the SEM model using the ML method, then the estimation results were used as the initial model in SEM regularization through the CV method to choose the right λ value. Figure 1(a) shows that the lambda value that produced the smallest BIC value is zero. This means that all variables are not selected, so CV result needed to be analysis further to get the best lambda value used in modelling.

Figure 1.(a)lambda with Cross-Validation Regularization SEM, (b) The BIC value for a particular λ

Figure 1(b) shows that the BIC value always rises if the lambda value increases. So that in this study based on Figure 1 and Figure 2 the value of λ, the value of which rises the most after, is chosen. Based on this, the lambda value used to model the data is $\lambda = 0.14$ and $\lambda = 0.18$. Of the three SEM Regularization models with different $\lambda$ analyzed further about the goodness of the model not only from BIC, but also from RMSEA.

Table 2 is a comparison of the RMSEA and BIC values for the SEM model and SEM Regularization with different $\lambda$. The results in Table 2 were consistent with the results in Figure 2, namely the ML method or $\lambda = 0$ produced the smallest RMSEA and BIC values, but not too significantly different for RMSEA, as well as BIC.
### Table 2. Table of RMSEA and BIC values of SEM regularization models

| Estimation Method | RMSEA  | BIC (i)     | BIC (i+1)   |
|-------------------|--------|-------------|-------------|
| SEM               | 0.0461 | 535682.5    |             |
| RSEM.14 (λ = 0.14)| 0.0465 | 535878.6    | 535931.7    |
| RSEM.18 (λ = 0.18)| 0.0467 | 535964.6    | 535995.1    |

#### 3.2 The SEM Model and The SEM Regularization Model

The SEM regularization model with the LASSO method produced a more concise model than the SEM model. Penalty LASSO in the SEM regularization model was only given to the regression coefficient, namely the coefficient on the structural model in the SEM model. This can be seen in Figure 2, which is the estimated result of the coefficient parameters of the SEM, RSEM.14, and RSEM.18 models. Note that the SEM Model included all the variables in the model, because there was no penalty given to the regression coefficient, so the non-essential variables remain in the model. Therefore, it needed a method that can produce a simpler and more concise model, namely a model that only included important variables or which had a significant effect.

The expected results using RSEM.14, there were 5 variables which were shrunk to 0 with 1 negative variable, while RSEM.18 as many as 6 variables were shrunk to zero. Note also that the selected variables in the SEM regularization model were indeed variables with insignificant effects in the SEM model. Therefore RSEM.18 was a SEM regularization model with λ = 0.18 having a BIC value (i) which had the largest difference with BIC thereafter (i + 1) and produced a more concise model than the regularization model with λ = 0.14, then the selection of explanatory variables which affects the achievement of Indonesian students’ PISA in 2015 is the SEM regularization model with λ = 0.18. The results show in Figure 2.

![Figure 2. The coefficient values of the latent variables and the observed variables](image)

#### 3.3 Model Interpretation

Table 3 was estimated results of parameters using the SEM and SEM regularization method. Based on the results in Table 3, the SEM regularization method with λ = 0.18 produced a model consisting of variables that have a considerable effect if returned to the SEM model, so the variables that significantly affect the results of student PISA were student’s insight into the environmental issues, student’s perceptions of science truth, and the student’s participation in ISCED 0. Students’ insight into the environment had the greatest effect, which was equal to 0.338. This meant that the better the student's insight into environmental issues, the higher the PISA achievement obtained. While the
students’ perceptions of the truth of science had an effect of 0.195, which was increasingly agreeing to the truth of science, hence increasing the PISA achievement obtained. The participation of students for one year in ISCED 0 tended to give a higher score of PISA results of 0.110 times than those who did not attend ISCED 0. Meanwhile, the participation of students for more than one year in ISCED 0 tended to give PISA results higher 0.217 times than not following ISCED 0.

Table 3. Estimated results of parameters using the SEM method and SEM Regularization with $\lambda = 0.14$ and $\lambda = 0.18$

| Variable                  | SEM     | RSEM.14  | RSEM.18  |
|---------------------------|---------|----------|----------|
| PISA manifest variable    |         |          |          |
| Mathematics               | 0.733   | 0.794    | 0.807    |
| Science                   | 0.765   | 0.833    | 0.846    |
| Read                      | 0.688   | 0.746    | 0.757    |
| Latent Variable           |         |          |          |
| Discipline                | -0.091  | 0        | 0        |
| Science Learning          | -0.125  | -0.021   | 0        |
| Environment Insight       | 0.458   | 0.355    | 0.338    |
| Benefit of Learning       | 0.048   | 0        | 0        |
| Science Interest          | -0.07   | 0        | 0        |
| Science Performance       | -0.029  | 0        | 0        |
| Science Truth             | 0.351   | 0.220    | 0.195    |
| Observed Variable         |         |          |          |
| Gender                    | -0.022  | 0        | 0        |
| ISCED0.2                  | 0.247   | 0.110    | 0.075    |
| ISCED0.3                  | 0.35    | 0.217    | 0.183    |

4. Conclusion
The SEM regularization method provided predictive results that are simpler and easier to understand than the predicted SEM model. This was because the model produced only consists of the variables that had the greatest influence or the most important. SEM regularization using the LASSO method with $\lambda = 0.18$ was quite good in selecting variables that have very little effect. In addition, the resulting model was also simpler and easier to interpret. The variables that significantly influence the results of student PISA are student’s insight into the environmental issues, student’s perceptions of science truth, and the student’s participation in ISCED 0. In this study only included variables that only use a few background variables and student characteristics, there were many other variables had not been included which can affect student achievement of PISA. This could be used as the basis of further research.

5. References
[1] OECD 2016 Volume 1: Excellence and Equity Education PISA 2015 Result Paris: OECD Publisher
[2] Rahmawati 2018 Interpreting PISA Results through Sample Weaknessess Perspective: It’s Time to Move from Rank Paradigm Proceedings National Conference on Educational Assessment and Policy Jakarta 79–88
[3] Pakpahan R 2016 Factors Affecting Literacy Mathematics Achievement of Indonesia Student in PISA 2012 J. Pendidik. dan Kebud. 1 331–48
[4] Tibshirani R 1996 Regression Shrinkage and Selection via The LASSO R. Stat. Soc. 58 267–88
[5] Jacobucci R 2017 Regsem: Regularized Structural Equation Modeling
[6] Jacobucci R, Grimm K J and McArdle J J 2016 Regularized Structural Equation Modeling Struct. Equ. Model. A Multidiscip. J. 23 555–66
[7] Joreskog K and Sorbom D 1996 LISREL 8: User’s Reference Guide Chicago: Scientific Software International Inc
[8] Homrighausen D and McDonald D 2013 The LASSO Persistence and cross-validation Proceeding of Machine Learning Research Georgia pp 1031–9
[9] Edwards A L and Gonzalez R 1993 Simplified Successive Intervals Scaling Appl. Psychol. Meas. 17 21–7
[10] Tibshirani R Michael S Rosset S Zhu J and Knight K 2005 Sparsity and Smoothness Via The Fused LASSO R. Stat. Soc. 67 91–108
[11] Yuan M and Lin Y 2006 Model Selection and Estimation in Regression with Grouped Variables R. Stat. Soc. 68 49–67