Advances in Social Science, Education and Humanities Research, volume 474
Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)

Modelling of Poverty Percentage Based on Mean Years of Schooling in Indonesia Using Local Linear Estimator

N Chamidah[1,*], M F F Mardianto[1], E E Limanta[2], D R Hastuti[2]

[1]*Department of Mathematics, Faculty of Science and Technology, Airlangga University, Indonesia.*
[2]*Student of Statistics Study Program, Department of Mathematics, Faculty of Science and Technology, Airlangga University, Indonesia.*
**Corresponding author: nur-c@fst.unair.ac.id*

ABSTRACT
The United Nation as an intergovernmental organization confirms a Sustainable Development Goals (SDGs) otherwise known as the Global Goals. Out of the 17 goals, the Indonesian government is mainly focused on eradicating poverty and improving the quality of education. The opportunities for escaping poverty consistently increase with increasing levels of education. Therefore, education is the most important factor regarding poverty reduction. In this paper, we investigate the impact of mean years of schooling (MYS) on the percentage of poverty in Indonesia using local linear estimator of nonparametric regression and compare it with the parametric regression approach. Nonparametric regression is a method for analyzing the relationship between response and predictor variables by assuming no specific form of regression curve. One of the estimators frequently used in nonparametric regression is local linear estimator which is thought to be superior to kernel estimator. This estimator bases on locally fitting a line rather than a constant. Unlike kernel estimator, local linear estimation would have no bias if the true model were linear. In general, local linear estimation removes the bias term from the kernel estimator, which makes it has better behavior near the boundary of the x's and smaller MSE everywhere. The results show that mean square error (MSE) values are 22.2 for local linear nonparametric regression approach and 41.85 for linear parametric regression approach. It means that local linear nonparametric regression approach is better than the linear parametric regression approach to analyzing the impact of MYS on the percentage of poverty in Indonesia.

Keywords: *local linear; mean years of schooling; percentage of poverty.*

1. INTRODUCTION

The United Nations (UN) as an international organization established sustainable development goals (SDGs) as a global development agreement with the theme of "Changing Our World: Agenda 2030 for Sustainable Development". The policy was ratified as demand for changes in the economic, social, knowledge and technology environment in the era of industrial revolution 4.0. There are 17 goals and 169 targets of global actions, including alleviating all forms of poverty and ensuring quality education that is also in line with the focus of sustainable development in Indonesia.

Poverty is a fundamental problem that is being a concern of the government in any country. According to the Central Bureau of Statistics (BPS), poverty is an inability to fulfill basic needs. Poverty arises from differences in abilities, differences in opportunities, and differences in resources. In Indonesia, poverty is still a serious problem. Although various efforts have been made both by the central and regional governments to alleviate poverty, such as the Healthy Indonesia Card (KIS), Smart Indonesia Card (KIP), School Operational Assistance (BOS), People's Business Credit (KUR), Non-Cash Food Assistance (BPNT), conditional cash assistance, and many other programs, BPS noted that the number of poor people in Indonesia still reached 25.14 million or around 9.41% [1].

One of the reasons for the high rate of poverty is the low level of education [2]. Education makes people think creatively and able to follow changes such as the use of new innovations, the application of technology, and a mindset that are oriented towards development. Various studies about the impact of education on poverty have shown that education reduces poverty [3]. Another notable point is the consistent increase in chances of escaping poverty as the educational level increases. Therefore, education is the most important factor regarding poverty reduction [4]. Education level can be determined by mean years of schooling (MYS). The general assumptions are a higher level of education of a person, higher quality of a person, both the mindset and the pattern of actions.

Percentage of poverty tends to change in accordance with the mean years of schooling [5]. This will yield a good percentage of poverty if estimated locally. Data modeling will be more suitable if we use a nonparametric regression approach because it is more flexible than the parametric

regression approach [6]. Nonparametric regression is a method for analyzing the relationship between response and predictor variables by assuming no specific form of regression curve [7]. One of the estimators frequently used in nonparametric regression is local linear estimator, which is thought to be superior to kernel estimator. It bases on locally fitting a line rather than a constant. Unlike kernel estimator, local linear estimation would have no bias if the true model were linear [4]. In general, local linear estimation removes a bias term from the kernel estimator, which makes it has better behavior near the boundary of the x's and smaller MSE everywhere. Other advantages of this estimator are able to estimate the function at each point such that the model closes to the real pattern, and also no needs much data to estimate the model [8]. Some former researches that used local linear estimator were carried out by [9-13]. Data processing will be easier if we use the help of statistical software than manually. One of the open-source soft-wares that can be used in data processing is software R.

In this paper, we intend to investigate the relationship between MYS and percentage of poverty. The results can be used to give some recommendations to local and central governments to improve the quality of education in accordance with the current generation's atmosphere as well as planning material for policy making to realize the target SGDs of 2030 specifically in the fields of poverty and education in Indonesia.

2. MATERIALS AND METHODS

2.1. Nonparametric Regression

Nonparametric regression is a statistical method used to determine the relationship between response variables and predictor variables when the form of the regression curve between the response variable and predictor variable does not be assumed has a certain form or has no previous information about it. The nonparametric regression minimizes the assumptions of the form of the regression functions and lets the data itself look for the form of estimation. Nonparametric regression has high flexibility and has the assumption that the form of regression curve is smooth that is in a continuous sense and can be derived.

Next, suppose that the response variable Y and predictor variable X are obtained from paired observations (x_1, y_1), (x_2, y_2), …,(x_n, y_n) that follow the model:

$$y_i = g(x_i) + \varepsilon_i,\ i = 1,2,\dots,n \tag{1}$$

where ε_i is a random error which is assumed to be independent with a zero mean and variance of σ^2 , and g is an unknown regression function.

2.2. Local Linear Estimator

A local linear estimator is one of the estimators of the nonparametric regression model that uses smoothing techniques. The local linear estimator is considered as a special case of local polynomial estimator that has an order of 1. Local linear estimator has advantages over other estimators namely estimating functions at each point such that the estimated model is closer to the actual data pattern, and this estimator does not require large amounts of data in estimating model. In addition, the local linear estimator can be interpreted easily because it has met the parsimony principle or can explain the model well with a minimum number of parameters.

The local linear regression approach assumes that n paired observations (x_1, y_1), (x_2, y_2), … ,(x_n, y_n) follows local linear regression models which in matrix notation is $g(x) = \mathbf{X(x_0)\beta(x_0)}$. Next, to get the value of $\hat{\beta}(x_0)$ by using the local linear regression approach, we minimize the function of $Q(x_0) = (y - X(x_0)\beta(x_0))^T(y - X(x_0)\beta(x_0))$ so that we obtain the estimated model as follows:

$$\hat{g}(x) = \mathbf{X(x_0)\hat{\beta}(x_0)} \tag{2}$$

Hence, $\mathbf{\hat{\beta}(x_0)}$ can be obtained by taking n paired samples data of $\{X_i, Y_i\}_{i=1}^n$ that can be presented in equations as follows:

$$y_1 = \beta_o(x_o) + \beta_1(x_o)(x_1 - x_o) + \varepsilon_1$$
$$y_2 = \beta_o(x_o) + \beta_1(x_o)(x_2 - x_o) + \varepsilon_2$$
$$\vdots$$
$$y_n = \beta_o(x_o) + \beta_1(x_o)(x_n - x_o) + \varepsilon_n \tag{3}$$

Further, we can be expressed equations in (3) in matrix notation as follows:

$$\mathbf{y} = \mathbf{X_{(x)}\beta_{(x_o)}} + \boldsymbol{\varepsilon} \tag{4}$$

where

$$\mathbf{\hat{\beta}(x_0)} = \begin{pmatrix} \hat{\beta}_0(x_0) \\ \hat{\beta}_1(x_0) \end{pmatrix} \tag{5}$$

$$X_{(x_o)} = \begin{pmatrix} 1 & x_1 x_0 \\ 1 & x_2 x_0 \\ \vdots & \vdots \\ 1 & x_n x_0 \end{pmatrix}$$

(6)

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad (7)$$

Value of $\hat{\boldsymbol{\beta}}(\mathbf{x_0})$ can be obtained by using local linear estimator with a weight function of $\mathbf{K_h}(x_i - x_0)$. This weight function is called the kernel function. That weight function is determined by the kernel function, while the weight size is determined by the h parameter called as bandwidth.
Estimation of $\boldsymbol{\beta}(\mathbf{x_0})$ in equation (4) is used the weighted least square (WLS) method by minimizing the following equation:

$$Q(\mathbf{x_0}) = (y - \mathbf{X}(\mathbf{x_0}) - \boldsymbol{\beta}(\mathbf{x_0}))^T \mathbf{K_h}(\mathbf{x_0})(y - \mathbf{X}(\mathbf{x_0}) - \boldsymbol{\beta}(\mathbf{x_0})) \quad (8)$$

where $\begin{bmatrix} \mathbf{Kh}\,(x1 - x0) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Kh}\,(xn - x0) \end{bmatrix}$.

The estimated value for $\boldsymbol{\beta}(\mathbf{x_0})$ is $\hat{\boldsymbol{\beta}}(\mathbf{x_0})$ which is obtained through differentiating equations in (8) with respect to $\beta(x_0)$ such that we get:

$$\hat{\boldsymbol{\beta}}(\mathbf{x_0}) = (\mathbf{X^T}(x_0)\mathbf{K_h}(x_0)(\mathbf{X})(x_0))^{-1}\mathbf{X^T}(x_0)\mathbf{K_h}(x_0)y \quad (9)$$

Because of equations in (2) and (9), the local linear estimator of $\hat{g}(x)$ can be written as follows:

$$\hat{g}(x) = \mathbf{x}(\mathbf{x_0})(\mathbf{X^T}(x_0)\mathbf{K_h}(x_0)(\mathbf{X})(x_0))^{-1}\mathbf{X^T}(x_0)\mathbf{K_h}(x_0)y \quad (10)$$

Also, since equation (10), estimation in target point of $\mathbf{x} = \mathbf{x_0}$ is $\hat{g}(\mathbf{x_0})$. Thus, equation (10) can be expressed as follows:

$$\hat{g}(\mathbf{x_0}) = e_1(\mathbf{X^T}(x_0)\mathbf{K_h}(x_0)(\mathbf{X})(x_0))^{-1}\mathbf{X^T}(x_0)\mathbf{K_h}(x_0)y \quad (11)$$

Where $e_1 = (1\ 0)$.

2.3. Generalized Cross Validation for Determining Optimal Bandwidth

Bandwidth *(h)* is a smoothing parameter that controls the smoothness of the curve. The selection of bandwidth (h) is very important to get an estimate of the appropriate response variable. Generalized cross validation (GCV) method is one of the methods used to obtain the optimal bandwidth [12] which is defined as follows:

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{n,i}(t_i))^2 \quad (12)$$

The optimal bandwidth value is the value that produces the minimum GCV.

2.4. Goodness of Fit

MSE is obtained from the average expectation of the square of the difference estimator around the actual population parameter value, that is calculated through:

$$MSE(h) = n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{y}(x_i)\right)^2 \quad (13)$$

According to [13], the coefficient of determination (R^2) draws the accuracy of the regression curve in order to know the variation of the response variable (y) which can be explained by several predictor variables x. The coefficient of determination can be calculated by the following formula [14]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (14)$$

2.5. Data and Steps of Analysis

The data used in this paper is secondary data including MYS and the percentage of poverty in Indonesia per province in 2018 obtained from ([15], [16]). The number of data used in this study amounts to 34 data. The steps of analysis in accordance with the research objectives are as follows:

1. Modeling the poverty percentage over mean year of schooling using optimal bandwidth values and minimum value of GCV by using nonparametric regression approach based on local linear estimator, through the following steps:
   a. Testing independence is used the Run Test on Minitab software for data poverty percentage in Indonesia.
   b. Selecting the optimum bandwidth (h) value that has a minimum CV value based on equation (12) on the data on the poverty percentage and mean years of schooling in Indonesia.
   c. Using optimal bandwidth to model the poverty percentage uses a linear local estimator approach and known parameter estimators $\hat{\beta}$.

2. Interpreting the model of the poverty percentage based on mean years of schooling in Indonesia using nonparametric regression approach using linear local estimators

3. RESULTS AND DISCUSSION

There are two approaches on regression analysis, the first one is parametric regression approach and the second one is nonparametric regression approach. According to [9], nonparametric regression is more flexible than parametric regression. Comparison of MSE values between parametric regression approach and nonparametric regression approach is shown in Table 1.

Table 1. MSE Values Between Parametric and Nonparametric Regressions Approaches

	Parametric	**Nonparametric**
MSE	41.83	22.2

According to Table 1, it was found that the MSE value in the nonparametric regression approach was smaller than that in the parametric regression approach. Thus, the testing result using the nonparametric regression approach is better than that using parametric regression approach.

Next, we do testing independence that is data should be independent at each observation. For this reason, a run test was conducted by using Minitab software on percentage data of the illiterate population in Indonesia. The hypothesis used is H_0: percentage data of the poor is independent or mutually independent and H_1: percentage data of the poor is dependent. The critical area used is that H_0 is rejected if P-value is less than α. Based on the output of the independent run test, we decide that H_0 can be accepted because the P-value is not significant, namely the P-value= 0.787 that is greater than the significance level α = 0.05. Thus, the percentage of data of the poor is independent or mutually independent. Therefore, it can be continued to nonparametric regression analysis.

The selection of optimal bandwidth as smoothing parameters is very important to get an estimate of the appropriate response variable. The smoothing or bandwidth (h) parameter value can be determined based on the minimum GCV value calculated by using software R. Based on obtained R output, the value of optimal bandwidth is 0.5614768.

Next, we use optimal bandwidth to model the percentage of poverty by using local linear estimator of nonparametric regression. The plot of observed and estimated of poverty percentage versus mean years of schooling is given in Figure. 1.

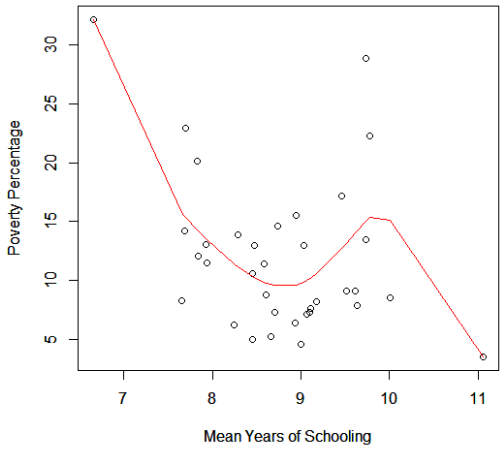

Figure 1. Plot of observed and estimated of poverty percentage versus mean years of schooling

Interpretation of the model of several data per province in Indonesia in 2018 are as follows:

1. Province of Papua
   The model of the percentage of poor people to the percentage of the average length of school for the Papua Province is as follows:

$$\hat{Y}_1 = 31.81021 - 15.117(x - 6.66) \quad (15)$$

   Equation in (15) presents that every one-year increase in the mean years of schooling results in a decrease in the percentage of poor people by 15.117% in the Papua Province. The decrease in poverty percentage looks high enough, so it can be concluded that education is very important in Papua Province to reduce poverty.
2. Province of DKI Jakarta
   The model of the percentage of poor people to the average length of school for DKI Jakarta Province is as follows:

$$\hat{Y}_2 = 3.60258 - 8.0167205(\mathrm{x} - 11.06) \quad (16)$$

   Equation (16) presents that every one-year increase in the mean years of schooling results in a decrease in the percentage of poor people by 8.0167205% in DKI Jakarta Province.
3. Province of Kalimantan Timur
   The model of the percentage of poor people to the average length of school for DKI Jakarta Province is as follows:

$$\hat{Y}_3 = 13.59733 + 5.66639695(\mathrm{x} - 9.03) \quad (17)$$

Equation (17) presents that every one-year increase in the mean years of schooling results in an increase in the percentage of poor people by 5.6663969% in Kalimantan Timur Province.

4. CONCLUSION

There is a fluctuating model of the percentage of poverty based on the mean years of schooling in 34 provinces of Indonesia. Also, we obtained the estimated model of the poverty percentage from the mean years of schooling in each province in Indonesia. A total of 18 provinces experienced a decrease in the percentage of poverty when there was a one-year increase in the mean years of schooling. However, as many as 16 provinces experienced a poverty percentage increase when there was a one-year increase in the mean years of schooling. It means that the local linear approach is better than the global approach to model the percentage of poverty based on the mean years of schooling in Indonesia.

REFERENCES

[1] Central Statistics Agency. (2018). Indonesia Central Statistics Agency's Profil.

[2] Thapa, S. B. (2013). Relationship between education and poverty in Nepal. Economic Journal of Development Issues. 148-161.

[3] Awan, M. S., Malik N., Sarwar, H., Waqas M.. (2011). Impact of Education on Poverty Reduction. International Journal of Academic Research. 3(1): 659–664.

[4] Olito, C., White, C. R., Marshall, D. J., Barneche, D. R. (2017). Estimating monotonic rates from biological data using local linear regression. Journal of Experimental Biology. 220(5): 759-764.

[5] Ferguson, H. B., Bovaird, S., Mueller, M. P. (2007). The Impact of Poverty on Educational Outcomes for Children. Paediatrics & Child Health. 12(8): 701-706.

[6] Hardle, W. (1999). Applied Nonparametric Regression. United Kingdom: Cambridge University Press.

[7] Hollander, Myless, Wolfe, Douglas, A., Chicken, E. (2004). Nonparametric Statistical Methods: Third Edition. Canada: New Jersey.

[8] Nottingham Q. J. Cook, D. F. (2011). Local Linear Regression for Estimating Time Series Data. Journal of Computational Statistics and Data Analysis. 37: 209–217.

[9] Chamidah, N., Rifada, M. (2016). Estimation of Median Growth Curves For Children up Two Years Old Based on Biresponse Local Linear Estimator. AIP Conference Proceedings. 1718 110001

[10] Chamidah, N. Tjahjono, E., Fadilah, A.R., Lestari, B,. (2018). Standard Growth Charts for Weight of Children in East Java Using Local Linear Estimator. Journal of Physics: Conference Series 1097 012092.

[11] Nidhomuddin, Chamidah N, Kurniawan N. (2019). Admission Test Modelling of State Islamic College in Indonesia Using Local Linear for Bivariate Longitudinal Data. IOP Conf. Series: Materials Science and Engineering. 546: 052047, doi:10.1088/1757-899X/546/5/052047.

[12] Ana, E., Chamidah, N., Andriani, P., Lestari, B. 2019.Modeling of hypertension risk factors using local linear of additive nonparametric logistic regression Journal of Physics: Conference Series. 1397 012067

[13] Puspitawati, A., Chamidah, N. (2019). Choroidal Neovascularisation Classification on Fundus Retinal Images Using Local Linear Estimator. IOP Conf. Series: Materials Science and Engineering 546 052056 doi:10.1088/1757-899X/546/5/052056.

[14] Gujarati, D. (2004). Basic Econometric Fourth Edition. The McGraw-Hill Companies.

[15] Gyorfi, L., Kohler, M., Krzyzak, A., Walk, H. (2002). A Distribution-Free Theory of Nonparametric Regression. New York : Springer.

[16] Greene, W.H. (2003). Econometric Analysis Fifth Edition. Pearson Education. Inc. New Jersey.

[17] Central Statistics Agency. (2018). Poverty Percentage. BPS RI Jakarta.

[18] Central Statistics Agency. (2018). Mean Years of Schooling. BPS RI Jakarta.