Multivariable Truncated Spline Nonparametric Regression in Modeling Human Development Index (HDI) of Southeast Sulawesi Province

I Nyoman Sudiasa[1], Nurjeni Yanti[2], Baharuddin[3], Andi Tenri Ampa[4], Makkulau[5], Agusrawati[6], Andi Tenri Panngangareng Makkulau[7]

[1] Civil Engineering Department, Institut Teknologi Nasional Malang,Indonesia
[2,3,4,5,6,7] Statistics Department, Universitas Halu Oleo, Indonesia
anditenri.ampa@uho.ac.id

Abstract. The aim of this research was to determine the modelling of human development index (HDI) data of Southeast Sulawesi Province using the Spline Truncated Nonparametric Regression analysis method and also to determine the factors that influence it. Spline regression is able to overcome data patterns that show sharp increases or decreases with the help of knot points, and the resulting curve is relatively smooth. The data used in this study is the Human Development Index (HDI) data in 2022 along with the factors that are thought to affect it obtained from the Central Statistics Agency (BPS). Based on the test results that have been carried out, the best model using spline truncated nonparametric regression for the human development index (HDI) of Southeast Sulawesi Province is a model with three knot points with a minimum generalised cross validation (GCV) value of 5,93. The predictor variables used, namely the labour force participation rate (X_1), school enrolment rate (X_2), and morbidity rate (X_3) have a significant effect on the human development index (HDI) in Southeast Sulawesi Province with a coefficient determination (R^2) of 72,66%.

Keywords: Spline Truncated Nonparametric Regression, HDI, knot point, GCV.

1 Introduction

Spline regression is a nonparametric regression method that aims to minimize diversity and estimate the behavior of data that tends to be different. The Spline approach is able

N. Djam'an et al. (eds.), *Proceedings of the 5th International Conference on Statistics, Mathematics, Teaching, and Research 2023 (ICSMTR 2023)*, Advances in Computer Science Research 109,
https://doi.org/10.2991/978-94-6463-332-0_10

to overcome data patterns that show sharp rises or falls with the help of knot points, and the resulting curve is relatively smooth. Knot points are joint points that indicate changes in data behavior patterns [1]. A spline is a polynomial section of order q and has continuous derivatives with knots up to order (q-1) [2].

The truncated spline function is obtained based on the addition of the polynomial function and the truncated function. The regression curve $f_j(x_j)$ is assumed to be contained in a spline space of order m with knot points $K_{1j}, K_{2j}, \dots, K_{rj}$; $j = 1,2, \dots, p$ is given by the following equation.

$$f_j(x_j) = \sum_{k=0}^{m} \beta_{jk} x_j^k + \sum_{u=1}^{r} \beta_{j(m+u)} (x_j - K_{ju})_+^m \quad (1)$$

so we get the multivariable truncated spline nonparametric regression model equation:

$$y_i = \sum_{j=1}^{p} \sum_{k=0}^{m} \beta_{jk} x_{ji}^j + \sum_{j=1}^{p} \sum_{u=i}^{r} \beta_{j(m+u)} (x_{ji} - k_{ju})_+^m + \varepsilon_i , \quad (2)$$

The function $(x_{ji} - k_{ju})_+^m$ is a truncated function given by:

$$(x_{ji} - k_{ju})_+^m = \{(x_{ji} - k_{ju})^m \quad ; x_{ji} \geq k_{ju}\} 0 \qquad ; x_{ji} \leq k_{ju} \quad (3)$$

where:

β_{jk} : polynomial model parameters, $k = 1, 2, \dots, m$; $j = 1, 2, \dots, p$

x_{ji} : j ith predictor variable; $i = 1,2,..,n$

$\beta_{j(m+u)}$: parameters in the truncated component, $u = 1, 2, \dots, r$

K_{ju} : knot points; r: number of knots; m: number of orders; p : number of iterations

The parameters of the spline nonparametric regression model can be estimated using the weighted least squares (WLS) method. The nonparametric spline regression equation can be written as the following equation [3].

$$Y = X\beta + \varepsilon \quad (4)$$

Equation (4) can be presented in matrix notation as follows.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; X = \begin{bmatrix} X_1[k] & 0 & 0 & 0 \\ 0 & X_2[k] & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & X_n[k] \end{bmatrix}; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

By using the weight W, the estimate of β in equation (4) can be obtained by completing the optimization as follows.

$$\min_{\beta} \{(Y - X\beta)^T W(Y - X\beta)\} \quad (5)$$

with W the matrix as follows:

$$W = \begin{bmatrix} W_1 & 0 & 0 & 0 \\ 0 & W_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & W_n \end{bmatrix} \text{ where } W_n = \begin{bmatrix} W_{k_n} & 0 & \cdots & 0 \\ 0 & W_{k_n} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_{k_n} \end{bmatrix}; i = 1,2,..,n$$

The solution to equation (3) is carried out with the following explanation.

$$(Y - X\beta)^T W(Y - X\beta) = Y^T WY - 2\beta^T X^T WY + \beta^T X^T WX\beta$$

let $L = Y^T WY - 2\beta^T X^T WY + \beta^T X^T WX\beta$ so $\frac{\partial L}{\partial \beta} = -2X^T WY + 2\beta^T WX\beta$

and the minimum value of β is obtained $\frac{\partial L}{\partial \beta} = 0$ so

$$-2\beta^T X^T WY + 2\beta^T X^T WX\beta = 0$$
$$-X^T WY + X^T WX\beta = 0$$
$$X^T WX\beta = X^T WY \quad (4)$$

In equation (4) if both sides are multiplied by $(X^T WX)^{-1}$ then we get:

$(X^T WX)^{-1} X^T WX\beta = (X^T WX)^{-1} X^T WY$ so

$\hat{\beta} = (X^T WX)^{-1} X^T WY$

(5)

The spline approach has the advantage of dealing with data patterns that show sharp rises or falls with the help of knot points. Knot points are points that occur when the data pattern changes. In determining the optimal knot point, there are several methods used, one of which is the Generalized Cross Validation (GCV) method. The following is the equation for the GCV value:

$$\text{GCV} = \frac{MSE}{\left[I - \frac{C(M)}{n}\right]^2} = \frac{\frac{i}{n}\sum_{i=1}^{n} [y_i - f_M(x_i)]^2}{\left[I - \frac{C(M)}{n}\right]^2}$$

$$\text{MSE} = \frac{i}{n}\sum_{i=1}^{n} [y_i - f_M(x_i)]^2 \quad (6)$$

where:

y_i : ith predictor variable
M : number of basis functions
x_i : ith predictor variable
n : number of observations
f_M : estimation function
$C(M) : Trace\left[B(B'B)^{-1} - B'\right] + 1$

The best truncated spline model can be seen from the smallest GCV value. If the data has the same smallest GCV value, then it can be seen from considering the smallest MSE value.

In 1990, the United Nations Development Program (UNDP) introduced an indicator called the Human Development Index (HDI), which is an indicator that can describe the development of human development in a measurable and representative manner in a region. HDI is a more comprehensive indicator because apart from being able to measure economic development, it is also able to measure the development of social aspects and human welfare compared to other indicators. HDI is a composite index calculated as a simple average of the life expectancy index, education index and living standard index [8], [9], [10], [11], [12].

After being depressed in 2020 due to the Covid-19 pandemic, Indonesia's HDI in 2021 began to experience improvement. Indonesia's HDI in 2021 grew by 0,49 percent, higher than the previous year which grew by 0.03 percent, but still lower than in 2019 which grew by 0,74 percent. Meanwhile, the human development index (HDI) in Southeast Sulawesi itself has increased from 65,99 percent in 2010 to 71,66 percent in 2021. During this period, Southeast Sulawesi's HDI grew on average 0,75 percent per year and increased from the previous level. has been high since 2018. However, the Covid-19 pandemic has caused slight changes in the HDI achievement in Southeast Sulawesi. Southeast Sulawesi's HDI in 2020 was recorded at 71,45 or grew 0,35

percent, slowing down compared to growth in 2019. Meanwhile in 2021, Southeast Sulawesi's HDI increased again as Southeast Sulawesi's economy began to recover after the Covid-19 pandemic [4] (BPS, 2022) . In this research, the data used is the human development index (HDI) data for Southeast Sulawesi Province in 2022, which consists of 15 districts and 2 cities which have different data characteristics so that the shape of the data pattern will be random and not form a particular pattern. The appropriate regression analysis to be used in this research is nonparametric regression analysis because nonparametric regression can be used on data that does not follow a certain pattern. One nonparametric regression analysis that can be used is truncated spline nonparametric regression [1], [2], [3], [7].

2 Research Methods

This research uses secondary data obtained from the Website of the Central Statistics Agency for Southeast Sulawesi Province in 2022. The research observation data is 17 observations. This research uses 4 variables, namely the human development index variable for Southeast Sulawesi Province in 2022 (Y), the labor force participation rate variable for Southeast Sulawesi Province in 2022 (X_1), the school participation rate variable for Southeast Sulawesi Province in 2022 (X_2), the morbidity rate variable Southeast Sulawesi Province in 2022 (X_3).

The procedure for this research is as follows:

1. Review the literature on truncated spline nonparametric regression and the human development index (HDI).
2. Conduct descriptive analysis between the response variable and each predictor variable.
3. Create a scatter diagram between the response variable and each predictor variable to determine the data pattern.
4. Model the data using a truncated spline nonparametric regression model with various combinations of knot points.
5. Select the optimal knot point based on the smallest GCV value.
6. Carry out simultaneous and partial parameter significance tests on the best model.
7. Interpret the best model.

3 Results and Discussion

3.1 Descriptive Statistics

This data analysis consists of several stages. In the first stage, descriptive statistical analysis will be carried out on the response variable (Y), namely the human development index and predictor variables (X), namely the labor force participation rate (X_1), school participation rate (X_2), and morbidity rate (X_3) in 2022. The following are descriptive statistics for each research variable [4].

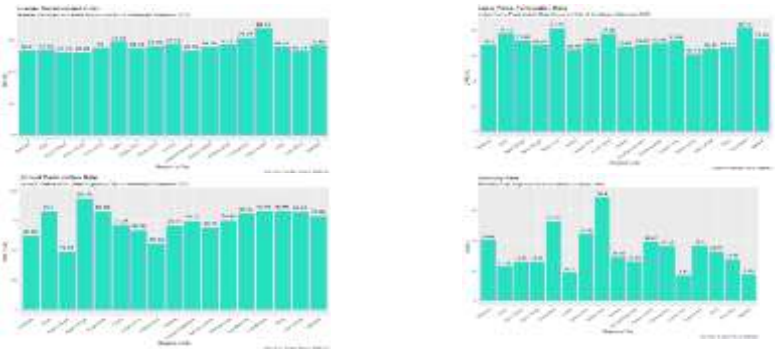

Fig. 1. Regency/City HDI of Southeast Sulawesi Province in 2022

Based on data from the Central Statistics Agency (BPS) in 2022, the HDI in Southeast Sulawesi was 72,23%, up 0,57% from 2021. This shows that the government has succeeded in increasing the HDI after experiencing a slowdown in 2020 due to the Covid-19 pandemic. In Figure 1, it can be seen that the lowest HDI in Southeast Sulawesi Province is in West Muna Regency with a HDI value of 66,21%, this is due to the slow economic recovery after the Covid-19 pandemic and also low public awareness about sending children to secondary education. associated with high poverty rates. Meanwhile, the highest HDI value is in Kendari City with an HDI value of 84,51%. Kendari City has the highest HDI value in Southeast Sulawesi Province due to the many facilities and infrastructure that are well used so that the HDI can increase every year. Bau-bau Island City has the lowest labor force participation rate, namely 61,14 percent, and West Muna Regency has the highest labor force participation rate, namely 82,12 percent. The school enrollment rate in Southeast Sulawesi Province is considered sufficient. South Buton Regency has the lowest school enrollment rate, namely 48,63 years. Meanwhile, Central Buton Regency is the region with the highest school enrollment rate in Southeast Sulawesi Province, namely 93,46 years. Bau-bau City has the lowest morbidity rate, namely 8,43 percent, and North Kolaka Regency has the highest morbidity rate, namely 34,4 percent.

3.2 HDI Relationship Pattern and Influencing Factors

Identification of the relationship pattern between predictor variables and the response variable is carried out to find out the relationship pattern of each predictor variable with the response variable. So, if the relationship pattern is known, you can determine the appropriate method for modeling. An easy check for relationship patterns is to look visually. The scatter diagram shows the relationship between two different variables and can be seen whether a certain pattern is formed or not.

A scatter diagram between the response variable, namely the human development index, and the predictor variables X_1, X_2, and X_3 do not follow or do not form a particular pattern. This is because the unit of analysis used is district/city, resulting in varied data. Therefore, model estimation cannot be carried out using a parametric regression approach so that the predictor variables used are non-parametric components.

3.3 Optimal Knot Point Selection

Knot points are obtained by looking at the points contained in changes in the data pattern. The knot points used in this research are one knot point, two knot point and three knot point.

Selection of optimal knot points for factors influencing HDI in Southeast Sulawesi Province using the GCV method starts with one knot point. The truncated spline nonparametric regression model formed with one knot point is as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (x_1 - K_1)_+^1 + \hat{\beta}_3 x_2 + \hat{\beta}_4 (x_2 - K_2)_+^1 + \hat{\beta}_5 x_3 + \hat{\beta}_6 (x_2 - K_2)_+^1$$

The following are the results of the minimum GCV value with one knot point for the truncated spline nonparametric regression model [2].

Table 1. GCV value for one knot point

Knot Point			GCV Value
K_1	K_2	K_3	
77,41	83,39	28,57	25,01

The minimum GCV value for one knot point is 25,01 with the optimal knot point for each variable written as follows.

X_1 : K_1 = 77,41; X_2 : K_2 = 83,39; and X_3 : K_3 = 28,57

Next, the optimal knot point is selected using two knot points. The nonparametric spline regression model formed with two knot points is as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (x_1 - K_1)_+^1 + \hat{\beta}_3 (x_1 - K_2)_+^1 + \hat{\beta}_4 x_2 + \hat{\beta}_5 (x_2 - K_3)_+^1 + \hat{\beta}_6 (x_2 - K_4)_+^1 + \hat{\beta}_7 x_3 + \hat{\beta}_8 (x_3 - K_5)_+^1 + \hat{\beta}_9 (x_3 - K_6)_+^1$$

The following are the results of the minimum GCV value with two knot points for the truncated spline nonparametric regression model.

Table 2. GCV value for two knot point

Knot Point			GCV Value
K_1	K_2	K_3	
67,56	62,35	16,38	14,91
67,99	63,26	16,91	

The minimum GCV value for two knot point is 14,91 with the optimal knot point for each variable written as follows.

X_1 : (K_1 = 67,56 K_2 = 67,99); X_2 : (K_3 = 62,35 K_4 = 63,26); and

X_3 : (K_5 = 16,38 K_6 = 16,91)

After obtaining the optimal knot with a minimum GCV value from one knot point and two knot points, the next step is to select the optimum knot point using three knot points. The truncated spline nonparametric regression model formed with three knot points is as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (x_1 - K_1)_+^1 + \hat{\beta}_3 (x_1 - K_2)_+^1 + \hat{\beta}_4 (x_1 - K_3)_+^1 \hat{\beta}_5 x_2 + \hat{\beta}_6 (x_2 - K_4)_+^1 + \hat{\beta}_7 (x_2 - K_5)_+^1 + \hat{\beta}_8 (x_2 - K_6)_+^1 + \hat{\beta}_9 x_3 + \hat{\beta}_{10} (x_3 - K_7)_+^1 + \hat{\beta}_{11} (x_3 - K_8)_+^1 + \hat{\beta}_{12} (x_3 - K_9)_+^1$$

The following are the results of the minimum GCV value with three knot points for the truncated spline nonparametric regression model.

Table 3. GCV values for three knot points

Knot Point			GCV Value
K_1	K_2	K_3	
64,99	56,86	13,2	5,93
80,40	89,80	32,28	
81,26	91,63	33,34	

The minimum GCV value for three knot point is 5,93 with the optimal knot point for each variable written as follows.

X_1 : $(K_1 = 64{,}99 \;\; K_2 = 80{,}40 \; K_3 = 81{,}26)$; X_2 : $(K_4 = 56{,}86 \;\; K_5 = 89{,}80 \;\; K_6 = 91{,}63)$; X_3 : $(K_7 = 13{,}2 \;\; K_8 = 32{,}28 \; K_9 = 33{,}34)$

Table 3 shows the results of the minimum GCV value for each knot. Based on the results in Table 3, the selection of knot points with three knot points has a minimum GCV compared to other knot points with a minimum GCV value of 5,93. Therefore, the truncated spline nonparametric regression model used to model the human development index (HDI) in Southeast Sulawesi Province is three knot points. The best truncated spline nonparametric regression model can be written as follows.

3.4 Selection of the Best Model

After conducting truncated spline nonparametric regression modeling using one knot point, two knot points, and three knot points, a comparison of the minimum GCV values will be seen to determine the best model. The following is a comparison of the GCV values for one knot point, two knot points and three knot points.

Table 4. GCV values for each knot points

Number of knots	GCV value
One knot	25,01
Two knot	14,91
Three knot	5,93

Table 4. shows the results of the minimum GCV value for each knot. Based on the results in table 4, the selection of knot points with three knot points has a minimum GCV compared to other knot points with a minimum GCV value of 5,93. Therefore, the truncated spline nonparametric regression model used to model the human development index (HDI) in Southeast Sulawesi Province is three knot points. The best truncated spline nonparametric regression model can be written as follows.

$$\begin{aligned}\hat{y} = {} & 0{,}0070 + 0{,}2818x_1 + 0{,}4513(x_1 - 64{,}9934)_+^1 + 0{,}1916(x_1 - 80{,}4037)_+^1 \\ & - 0{,}1785(x_1 - 81{,}2636)_+^1 - 0{,}2876x_2 \\ & - 0{,}2937(x_2 - 56{,}8640)_+^1 + 0{,}0486(x_2 - 89{,}8004)_+^1 \\ & - 0{,}1846(x_2 - 91{,}6302)_+^1 - 0{,}1976x_3 + 0{,}0981(x_3 - 13{,}2)_+^1 \\ & - 0{,}0369(x_3 - 32{,}28)_+^1 - 0{,}0044(x_3 - 33{,}34)_+^1\end{aligned}$$

3.5 Parameter Significance Testing

a. Simultaneous test (F test)

Simultaneous tests were carried out to determine the significance of the regression model parameters on the response variables together. The hypothesis for the simultaneous test is as follows.

$H_0 : \beta_0 = \beta_1 = \ldots = \beta_{12} = 0$ and H_1 : there is at least one $\beta_q \neq 0$, $q = 0, 1, \ldots, 12$

Test statistics use the F test with $F = \frac{KTR}{KTG}$

Table 5. Simultaneous test

Source	DF	SS	ST	F	p value
Regression	12	258,66	21,56	0,66	0,04
Error	3	97,31	32,44		
Total	16	355,98			

Table 5 shows the ANOVA table for the simultaneous test where the decision to reject H_0 is if the calculated F value is greater than the F_table value or p-value < α. In table 4.5 it can be seen that the p-value is 0,04 > 0,05 so that a decision can be made to reject H_0, which means there is at least one parameter that has a significant effect on the HDI.

b. Partial Test (t Test)

Partial Test (t Test) is used to determine the significance of regression parameters. Parameter testing with partial regression is carried out one by one on the model. The hypothesis used to partially test significance is as follows:

$H_0 : \beta_0 = \beta_1 = \ldots = \beta_{12} = 0$ and H_1 : there is at least one $\beta_q \neq 0$, $q = 0, 1, \ldots, 12$

The test statistics used for the t test are as follows $t = \frac{\widehat{\beta_q}}{SE(\widehat{\beta_q})}$

The result of partial model parameter testing. Based on the table, there are 10 parameters that have a significant effect on the model. Meanwhile, 3 parameters do not have a significant effect on the model. However, even though there are parameters that are not significant, these variables are still used because at least in one variable there is one parameter that is significant, so that the predictor variables X_1, AK) has a significant influence on the human development index in Southeast Sulawesi Province.

The coefficient of determination is a quantity that can explain the contribution of the predictor variable to the response variable. The higher the R^2 value produced by a model, the better the predictor variables in the model are in explaining the variability of the response variable. The following is the R^2 value from the regression model.

The truncated spline nonparametric regression model with three knots produces a coefficient of determination (R^2) of 72,66%, which means that the predictor variables used are able to explain 72,66% of the model and the rest is explained by other variables not studied.

4. Conclusion

Based on the results of the analysis and discussion that has been carried out, the conclusions obtained are as follows.

1. The best nonparametric spline truncated regression model for the human development index (HDI) of Southeast Sulawesi Province is a model with three knot

points with a minimum generalized cross validation (GCV) value of 5.9239. The following truncated spline nonparametric regression model.

2. Based on the results of the tests that have been carried out, the predictor variables used are the labor force participation rate (X_1), school participation rate (X_2), and morbidity rate (X_3) have a significant effect on the human development index (HDI) with the coefficient of determination(R^2) amounted to 72,66%, so it can be concluded that the factors that influence the human development index (HDI) of Southeast Sulawesi Province are the labor force participation rate, school participation rate, and morbidity rate.

References

1. Budiantara, I.N.: Truncated Polynomial Spline Family Models in Semiparametric Regression. Department of Statistics FMIPA-ITS, 15(3) (2005).
2. Ampa, A.T., Budiantara, I.N. Zain, I.: Selection of Optimal Smoothing Parameters in Mixed Estimator of Kernel and Fourier Series in Semiparametric Regression. Journal of Physics: Conference Series. **2123**(2021) 012035.
3. Bintariningrum, M.F. Budiantara, I.N.: Truncated Spline Nonparametric Regression Modeling and Its Application to Crude Birth Rates in Surabaya. Pomits Journal of Science and Arts, 3(1). 2337-3520 (2014).
4. Makkulau, Cahyono, E., Mukhsar, Sani, A., Saidi, L.O. and Ampa, A.T.: Outlier Detection in Multivariate Linear Models Using Lagrange Multipliers, Global Journal of Pure and Applied Mathematics/GJPAM, Vol. **13**, No. 6, pp. 2563-2578 (2017).
5. BPS: IPM Kabupaten/Kota Provinsi Sulawesi Tenggara Tahun 2022 (2023).
6. Maulana, A., Meilawati, R., and Widiastuti, V.: New Method of Human Development Index (HDI) Modeling by Province in 2015 Using Geographically Weighted Regression (GWR). Indonesian Journal of Applied Statistics, 2(1). 2621-086X. 2019.
7. Meilinda, Sudarmin. Nusrang, M.: Nonparametric Spline Truncated Regression Approach on the Human Development Index in South Sulawesi. Variance: Journal of Statistics and Its Application on Teaching and Research, 3(2). 2684-7590 (2021).
8. Putra, D.M. & Ratnasari, V.: Modeling the Human Development Index for East Java Province Using the Ridge Logistic Regression Method. ITS Science and Arts Journal, 4(2). 2337-3520 (2015).
9. Arofah, I. & Rohimah, S.: Path Analysis for the Effect of Life Expectancy, Expected Years of Schooling, Average Years of Schooling on the Human Development Index through Real Expenditures Per Capita in East Nusa Tenggara Province. Unpam Scientific Journal, 2(1), 2621-7856 (2019).
10. Bhakti, N.A.: Analysis of Factors Affecting the Human Development Index in Indonesia for the 2008-2012 Period. Equity: Journal of Economics and Finance, 18(4), 1411-0393 (2014).
11. Mahya, A. & Jauhar, W.: The Influence of Expected Years of Schooling, Average Years of Schooling, and Per Capita Expenditures on the Human Development Index. Journal of Mathematics Education and Research, 3(2), 2656-4181 (2021).
12. Fajri, R.H.: Analysis of Factors Affecting the Human Development Index in Riau Province. Journal of Economics, Accounting and Business, 1(1) (2021).