Analysis of Factors that Influence Life Expectancy in East Java (Indonesia) Using Semiparametric Spline Regression Approach

[1]I Nyoman Budiantara, [2]Ayuk Putri Sugiantari, [1]Vita Ratnasari, [1]Madu Ratna, [1]Ismaini Zain
[1]Lecturer of Statistics Department, Sepuluh Nopember Institute of Technology,
[2]Student of Statistics Department, Sepuluh Nopember Institute of Technology,
ITS campus, Sukolilo, Surabaya - 60111

***Abstract*— Life expectancy is one of the indicators used to assess quality of health of society. Based on Statistics Indonesia, life expectancy at birth is average years of life that will be lived by a newborn in a given year. Life Expectancy in a region different from other regions depends on the quality of life that can be achieved by the resident. Many factors affect the life expectancy in East Java (Indonesia), thereby modeling needs to be done to determine the factors that affect life expectancy significantly. This study uses six factors which are suspected affect life expectancy in East Java include social, economic, health, and education factor. Data of Life Expectancy and 6 factors were recorded in 2010 obtained from the Central Bureau of Statistics of East Java. The method used to model the life expectancy is semiparametric spline regression. The variables that have a significant impact are infant mortality rate, percentage of infants aged 0-11 months who were breastfed for 4-6 months, and the percentage of infants aged 1-4 years who received complete immunization.**

***Index Term*— life expectancy, semiparametric regression, spline regression, knot points**

I. INTRODUCTION

The high quality of society's health can be used as the indicator of success of health program and development of social and economic program that can increase life expectancy indirectly. Based on Statistics Indonesia, life expectancy at birth is average years of life that will be lived by a newborn in a given year. Life expectancy of Indonesia's resident in 2010 according to the Health Department of the Republic of Indonesia is 69.43 years. While the life expectancy of East Java's resident based on the National Socio-Economic Survey [1] were increasing continuously since 2007 at 68.9 years to 2010 at 69.6 years. Although life expectancy in East Java relatively have increased, but there are 9 districts in East Java, which has life expectancy below 65 years. The increase or decrease in life expectancy cannot be separated from various factors, so it needs to be investigated which factor that influence life expectancy by using regression modeling.

Some previous studies of life expectancy had done by Cleries et al [2] to determine trend of Spanish resident mortality rate in 1977 until 2001 and their impact on life expectancy using Bayesian Age Period Cohort (APC). Lusi [3] also conducted a study for life expectancy modeling in East Java and Central Java using Geographically Weighted Regression method (GWR). Meanwhile, Rakhmawati [4] conducted a study about life expectancy in West Java using panel regression. Halicioglu [5] conducted a study to determine the factors that influence life expectancy in Turkey using Autoregressive Distributed Lag (ARDL) approach.

Another method that can be used to model the life expectancy is semiparametric spline regression. Spline is used because it has several advantages such as spline has high flexibility, obtained from the optimization Penalized least squares (PLS), and spline able to handle behavioral pattern of data in different subintervals [6]. There are some researchers who apply semiparametric spline regression methods, such as Gilboa et al [7] who conducted a study on the relationship between maternal prepregnancy body mass index and adverse birth outcomes in the Baltimore-Washington infant study. Asmin [8] conducted a study to model national final exam scores of natural sciences majors at 1 Grati senior high school in Pasuruan (Indonesia). Kim et al [9] conducted a study on childhood aseptic meningitis. Bandyopadhyay and Maity [10] also use semiparametric spline regression approach to model average annual water flow in Sabine River.

In this paper will be using semiparametric spline regression approach for life expectancy modeling in east java so that obtained an appropriate model to explain life expectancy phenomena in east java.

II. THEORITICAL PROPERTIES

Nonparametric regression is a method for modeling the behavior of data when only few information available on the form of regression curve [11]. Regression curve only assumed smooth that contained in a certain function space. One of nonparametric regression is truncated spline. If given pairs of data $\{(t_1, t_2, \cdots, t_Q, y)\}$ then nonparametric truncated spline regression model can be written by equation (1).

$$y_i = \sum_{q=1}^{Q} f(t_{iq}) + \varepsilon_i \qquad ;i = 1,2,\cdots,n \tag{1}$$

with $f(t_{iq})$) is spline function of order p with knot points $k_1, k_2, \cdots, k_r$ that defined as

$$f(t_{iq}) = \sum_{j=0}^{p} \gamma_i t_{iq}^j + \sum_{l=1}^{r} \gamma_{p+l} (t_{iq} - k_l)_+^p \tag{2}$$

$(t_{iq} - k_l)_+^p$ is truncated function that can be defined by

$$(t_{iq}-k_l)_+^p=\begin{cases}(t_{iq}-k_l)^p & ,t_{iq}\geq k_l\\ 0 & ,t_{iq}<k_l\end{cases} \quad (3)$$

If equation (2) is substituted into equation (1) then obtained nonparametric spline regression model as follow.

$$y_i=\sum_{q=1}^{Q}\left(\sum_{j=0}^{p}\gamma_i t_{iq}^j+\sum_{l=1}^{r}\gamma_{p+l}(t_{iq}-k_l)_+^p\right)+\varepsilon_i \quad (4)$$

The best spline model is obtained from optimal knot points. Knot point is a blend knot where there is a change of behavior pattern of data or curve. Optimal knot point is obtained by using Generalized Cross Validation (GCV) method.

$$GCV(k)=\frac{MSE(k)}{\left[n^{-1}trace(\mathbf{I}-\mathbf{A}(\mathbf{k}))\right]^2} \quad (5)$$

with $MSE(k)=n^{-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$, and

$$\mathbf{A}(\mathbf{k})=\mathbf{X}(\mathbf{k})\left(\mathbf{X}(\mathbf{k})'\mathbf{X}(\mathbf{k})\right)^{-1}\mathbf{X}(\mathbf{k})'$$

Semiparametric regression is one of statistics method to determine the pattern of relationship between respon variable and predictor variable where part of shape of its pattern are known and partly unknown. Semiparametric spline regression model can be defined as follow.

$$y_i=\sum_{k=1}^{m}(x_{ik}'\beta_k)+\sum_{q=1}^{Q}(f(t_{iq}))+\varepsilon_i \quad ,t_{iq}\in[a,b] \quad (6)$$

with $i=1,2,\cdots,n$ and regression curve f approached by using spline function in equation (2).

III. MHETODOLOGY

Data used in this study were recorded in 2010 obtained from the Central Bureau of Statistics of East Java. Observation unit used are 38 regencies/cities in East Java. Variable used in this study are life expectancy (y), infant mortality rate (x_1), illiteracy rate of people aged 10 above (x_2), percentage of infants aged 0-11 months who were breastfed for 4-6 months (t_1), economic growth (t_2), the percentage of infants aged 1-4 years who received complete immunization (t_3), and labor force participation rate (t_4).

Steps of analysis performed in this study about life expectancy in East Java as follows:

1. Creating Descriptive Statistics of each variable to determine the characteristics of each cities / regencies in East Java.
2. Creating scatter plot between respon variable and predictor variable to determine the behavioral pattern data.
3. Modeling life expectancy in East Java using linear spline of 1 knot, 2 knot, and 3 knot.
4. Choosing the optimal point knots using GCV method whereby optimal point knots associated with the smallest GCV.
5. Modeling life expectancy in East Java using the best spline.
6. Parameters test with simultaneous test and individual test.
7. Goodness of fit test of residual (identical test, independent test, and normality test).
8. Making conclusion.

IV. CHARACTERISTIC OF LIFE EXPECTANCY IN EAST JAVA

Characteristics of life expectancy in East Java province visualized using a bar chart shown in Figure 1.

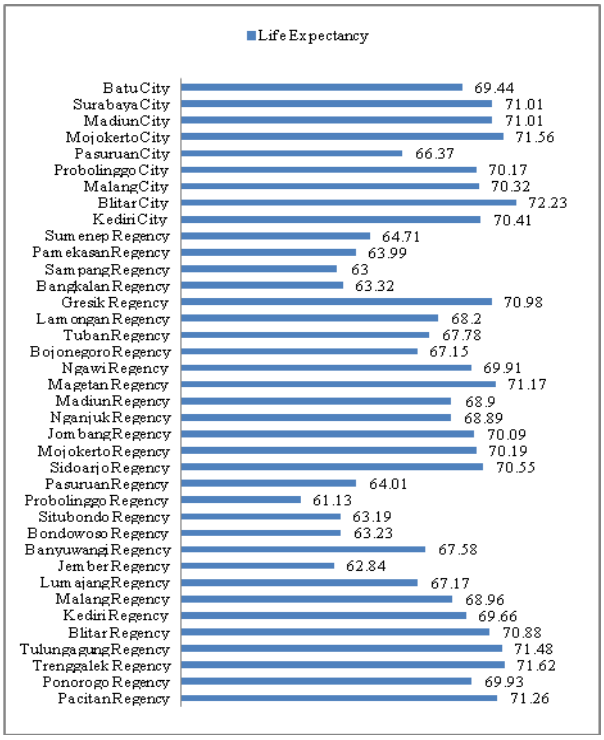

Fig 1. Bar Chart of Life Expectancy in Each Regencies/Cities in East Java Province

Figure 1 shows that the city / regency that has the highest life expectancy is Blitar City at 72.23 years. While that has the lowest life expectancy is Probolinggo regency at 61.13 years. There are 2 cities and 17 regencies in East Java that has a lower life expectancy than the life expectancy of East Java province (69.6), among others Batu City, Pasuruan City, Sumenep Regency, Pamekasan Regency, Sampang Regency, Bangkalan Regency, Lamongan Regency, Tuban Regency, Bojonegoro Regency, Madiun Regency, Nganjuk Regency, Pasuruan Regency, Probolinggo Regency, Situbondo Regency, Bondowoso Regency, Banyuwangi Regency, Jember Regency, Lumajang Regency, and Malang Regency.

V. MODELING LIFE EXPECTANCY IN EAST JAVA

The pattern of relationship between life expectancy with x_1 and x_2 variables shown in Figure 2.

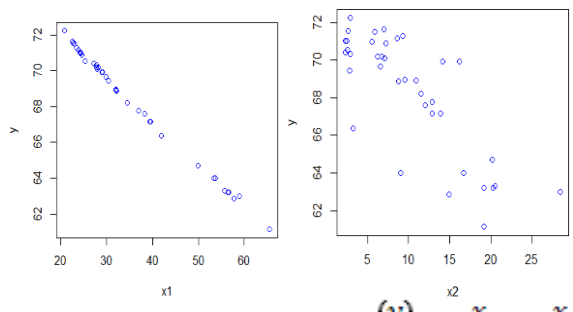

Fig. 2. *Scatterplot* between Life Expectancy (y) with x_1 and x_2 variables

Figure 2 shows that relationship between life expectancy with x_1 dan x_2 variables tend to be linear.

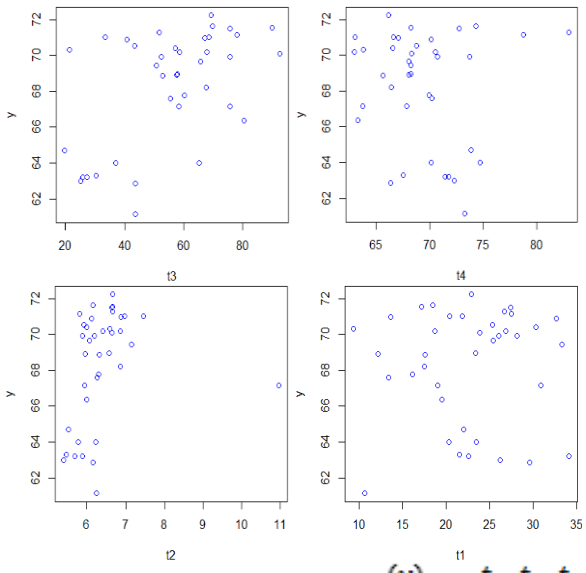

Fig. 3. Scatterplot between Life Expectancy (y) with t_1, t_2, t_3 and t_4 Variables

Figure 3 shows the pattern of relationship between life expectancy with t_1, t_2, t_3 and t_4 variables tend not to form a specific pattern. This indicates that there is a component nonparametric.

Semiparametric spline regression model with one knot points is given by

$$y_i = \varphi + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 t_{i1} + \gamma_2 (t_{i1} - k_1)_+ + \alpha_1 t_{i2} + \alpha_2 (t_{i2} - k_1)_+ + \delta_0 + \delta_1 t_{i3} + \delta_2 (t_{i3} - k_1)_+ + \tau_0 + \tau_1 t_{i4} + \tau_2 (t_{i4} - k_1)_+ + \varepsilon_i . \quad (7)$$

Semiparametric spline regression model with two knot points is difined by equation as follow

$$y_i = \varphi + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 t_{i1} + \gamma_2 (t_{i1} - k_1)_+ + \gamma_3 (t_{i1} - k_2)_+ + \alpha_1 t_{i2} + \alpha_2 (t_{i2} - k_1)_+ + \alpha_3 (t_{i2} - k_2) + \delta_1 t_{i3} + \delta_2 (t_{i3} - k_1)_+ + \tau_1 t_{i4} + \tau_2 (t_{i4} - k_1)_+ + \tau_3 (t_{i3} - k_2)_+ + \varepsilon_i . \quad (8)$$

Semiparametric spline regression model with three knot point is given as follow

The selection of optimal knot points obtained by GCV method. GCV value which obtained by using 1 knot points, 2 knot points, 3 knot points, and combination knot shown in Table I.

Table I
GCV Valiu Using 1 Knot Point, 2 Knot Point, 3 Knot Point, and Combination Knot

No.	Knot	GCV
1	1 Knot Point	0,039636
2	2 Knot Point	0,027702
3	3 Knot Point	0,02941
4	Combination Knot	**0,027599**

Bold– Knot value that result smallest GCV.

Table I shows that the minimum GCV value is produced when using combination knot with GCV at 0,027599. minimum GCV is associated with combination knot as follow . There are three knots at t_1 variable that is $t_1 = 25{,}5237$, $t_1 = 26{,}0282$, and $t_1 = 30{,}0641$. There are three knots at t_2 variable that is $t_2 = 9{,}0375$, $t_2 = 9{,}1512$, and $t_2 = 10{,}0606$. There are also three knot at t_3 variable that is $t_3 = 67{,}2314$, $t_3 = 68{,}7143$, and $t_3 = 80{,}5771$. Meanwhile, at t_4 variable only contained one knot at $t_4 = 67{,}4898$. Spline model of life expectancy with optimal knot points yields 99.89% of R^2 that given by:

$$\hat{y} = 78{,}6496 - 0{,}2355x_1 - 0{,}0195x_2 - 0{,}0089t_1 - 0{,}1149t_2 + - 0{,}0029t_3 - 0{,}0131t_4 + 0{,}9086(t_1 - 25{,}5237)_+ + - 0{,}9906(t_1 - 26{,}0282)_+ + 0{,}0943(t_1 - 30{,}0641)_+ + 0{,}1322(t_2 - 9{,}0375)_+ + 0{,}1244(t_2 - 9{,}1512)_+ + 0{,}0622(t_2 - 10{,}0606)_+ + 0{,}3067(t_3 - 67{,}2314)_+ + - 0{,}3707(t_3 - 68{,}7143)_+ + 0{,}1240(t_3 - 80{,}5771)_+ + 0{,}0189(t_4 - 67{,}4898)_+ . \quad (10)$$

There are 2 parameters test that is silmutaneous test and individual test. The result of simultaneous test is presented in Table II.

Table II
ANOVA of Simultaneous Test

Source	df	SS	MS	F	*P-value*
Regression	16	368,645	23,040	1259,325	0,00
Error	21	0,3842	0,0183		
Total	37	369,029	-		

Table 2 present that value of F test statistic is 1259,325 with p-value of 0,00. If p-value compared to significance level used (5%), the decision was taken to reject H_0. It can be concluded that there is at least one significant variable in the model.

Reject H_0 occurs indicates individual test needs to be done to determine which variables are effect on the model significantly. Individual test results are presented in Table III.

Table III
Individual Test

Variable	Parameter	Coeffisient	t_{value}	*P-value*
-	φ	78,6496	51,4243	0,0000
x_1	β_1	-0,2355	-60,0304	0,000*
x_2	β_2	-0,0195	-1,9770	0,0613
t_1	γ_1	-0,0089	-1,3393	0,1948
	γ_2	0,9086	3,1878	0,004*
	γ_3	-0,9906	-3,0656	0,006*
	γ_4	0,0943	1,1575	0,2601
t_2	α_1	-0,1149	-1,4573	0,1598
	α_2	0,1322	1,1233	0,2739
	α_3	0,1244	1,1233	0,2739
	α_4	0,0622	1,1233	0,2739
t_3	δ_1	-0,0029	-1,2852	0,2127
	δ_2	0,3067	4,2110	0,000*
	δ_3	-0,3707	-4,4350	0,000*
	δ_4	0,1240	4,0525	0,000*
t_4	τ_1	-0,0131	-0,6104	0,5481
	τ_2	0,0189	0,7303	0,4733

*variable have significant impact

Table III shows that there are 6 parameters that produce p-value less than 0.05 that is parameter of infant mortality rate, percentage of infants aged 0-11 months who were breastfed for 4-6 months, and percentage of infants aged 1-4 years were immunized complete. So that these three variables have a significant influence on the model.

Residual test which are conducted that is identical test, independent test, and normality test. Identical test results using Glejser test can be seen in table IV.

Table IV
ANOVA of Glejser Test

Source	df	SS	MS	F	*P-value*
Regression	16	0,05131	0,00321	1,0796	0,4275
Error	21	0,06237	0,00297		
Total	37	0,11368	-		

Value of F test statistic that produced is 1,0796. P-value generated in Glejser test is 0,4275. It shows identical assumptions on residuals are met.

Independent residual test results visualized in Figure 4

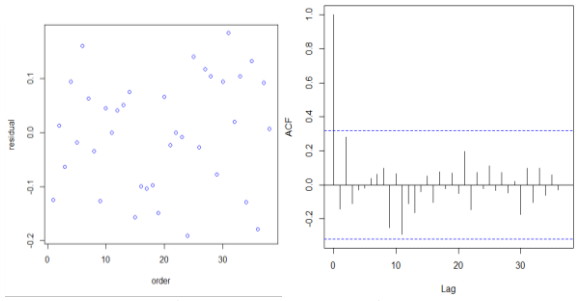

Fig. 4. Scatterplot of Residual vs Order (Left) and ACF Plot (Right)

Figure 4 (left) shows that there is no particular pattern that is formed on the distribution plot. Figure 4 (right) shows that the first lag to 38th lag are within tolerable limits indicating that the assumption of independent residuals satisfy.

Normality test result is shown by QQ Plot that can be seen in Figure 5.

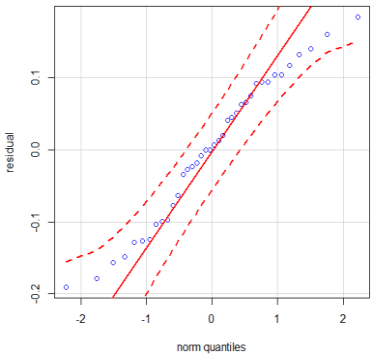

Fig. 5. QQ Plot of Residual

Figure 5 shows that the distribution of plots tend to follow a straight line (linear), which indicates that the residuals were normally distributed. Similar results were also described by the statistical test on the Shapiro Wilk test of 0.9675 and p-value of 0.3299. The resulting P-value is far greater than the significance level used that is 0.05. So it can be drawn a conclusion that the residuals are statistically normal distribution.

VI. CONCLUSION

The variables that significantly affect life expectancy in East Java is infant mortality rate (x_1), percentage of infants aged 0-11 months who were breastfed for 4-6 months (t_1), and percentage of infants aged 1-4 years who were given full immunization (t_2). Spline regression model produces 99.89% of R^2.

UCKNOWLEDGEMENT

The author would like to thank the Central Bureau of Statistics of East Java and Sepuluh Nopember Institute of Technology for the help and motivation given.

REFERENCES

[1] Badan Pusat Statistik Jawa Timur, 2011, *Survey Sosial Ekonomi Nasional Jawa Timur*, Badan Pusat Statistik Provinsi Jawa Timur Surabaya.

[2] Cleries, R., Martinez, J.M., Valls, J., Pareja, L., Esteban, L., Gispert, R., Moreno, V., Ribes, J., and Borras, J.M., 2009, Life Expectancy and Age-Period-Cohort Effects : Analysis and Projections of Mortality in Spain between 1977 and 2016, *Public Health*, 123, 156-162.

[3] Firdial, L., 2010, *Pemodelan Angka Harapan Hidup di Propinsi Jawa Timur dan Jawa Tengah dengan Metode Geographically Weighted Regression (GWR)*, Institut Teknologi Sepuluh Nopember Surabaya.

[4] Rakhmawati, D.P., 2011, *Analisis Faktor-Faktor yang Mempengaruhi Angka Harapan Hidup di Provinsi Jawa Barat*, Universitas Gadjah Mada.

[5] Halicioglu,F., 2011, Modeling Life Expectancy in Turkey, *Economic Modelling, 28*, 2075-2082.

[6] Budiantara, I.N., Ratna, M., Zain, I., and Wibowo, W., 2012, Modeling the Percentage of Poor People in Indonesia Using Spline Nonparametric Regression Approach, *International Journal of Basic & Applied Sciences IJBAS-IJENS*, 12, 119-124.

[7] Gilboa,M.S., Correa,A., and Alverson,J.C., 2008, Use of Spline Regression in an Analysis of Maternal Prepregnancy Body Mass Index and Adverse Birth Outcomes: Does It Tell Us More Than We Already Know ?, *Ann Epidemiol*, 18, 196-205.

[8] Asmin, S., 2010, *Pemodelan Nilai Unas IPA dengan Pendekatan Regresi Semiparametrik Spline di SMAN 1 Grati Pasuruan*, Institut Teknologi Sepuluh Nopember Surabaya.

[9] Kim,I., Cheong, H., and Kim,H., 2011, Semiparametric Regression Models for Detecting Effect Modification in Matched Case-Crossover Studies, *Statistics in Medicine*.

[10] Bandyopadhyay,S., and Maity,A.,2011, Analysis of Sabine River Flow Data using Semiparametric Spline Modeling, *Journal of Hydrology*, 399, 274-280.

[11] Ryan,T.P., 1997, *Modern Regression Methods*, John Willey and Sons,Inc., United States of America.