Proceeding of ICSA 2019, p: 96-107
ISBN 978-979-19256-3-1 (PDF)

Estimation Curve Semiparametric Regression with the Spline Linear Approach to Poverty Data in Bali Province

I Wayan Sudiarsa[*]

Department of Mathematics Education, Faculty of Mathematics Education and Natural Sciences, IKIP PGRI BALI, Denpasar Utara, 80239, Indonesia

[*)]Corresponding author: wayansudiarsa1804@gmail.com / wsudiarsa72@yahoo.com

Abstract

In semiparametric regression, nonparametric components can be approached by spline. Splines are pieces of polynomial that are segmented and continuous. The one advantages of spline is the presence of knot points that indicate changes in the pattern of data behavior. This research purposetoobtain semiparametric regression curve estimation with linear spline approach. The method of optimization approach used by ordinary least square (OLS). Based on thisresearch, there are two variables that have a significant effect on the percentage of poor people in Bali Province, namely the Open Unemployment the rate of economic growth. The total variance of response that can be explained by predictor in this model is 67.97% with MSE of 9,7854.

Keywords: poverty, semiparametric regression, splinelinear

Introduction

Poverty is a fundamental problem and is a serious concern from the government. BPS uses the concept of the ability to meet *basic needs approaches* in measuring poverty. With this approach, poverty is seen as an inability from the economic side to meet basic food and non-food needs measured from the perspective of income. The issue of poverty is important in relation to the main priorities of the MDGs (*Millennium Development Goals*), namely tackling poverty and hunger as an effort to global commitment throughout the nation. Poverty is caused by many factors and poverty is rarely found due to a single factor (Suharto, 2006).

Whittaker in 1923 first introduced spline as a data pattern approach. Next Reinsch in 1967 developed a spline based on his optimization problems (Lin & Carrol, 2001). The advantage of Spline is that it can describe changes in behavior patterns and functions at certain sub intervals. In addition, spline has the advantage of handling data patterns that show sharp up / down with the help of knot points and the resulting curve is relatively smooth (Chen & Jin, 2006); (Eubank, 1988). This study uses semiparametric regression with a linear spline approach. One of the advantages of using a linear spline regression approach is its mathematical ease and simplicity. The method optimization approach *Ordinary least square* (OLS) is used to obtain a linear spline estimator in semiparametric regression. The OLS method optimization approach was chosen because it is mathematically easier, simpler, and better in getting statistical inference.

This study aims to obtain an estimate of the semiparametric regression curve with a linear spline approach. The next objective is to apply a semiparametric linear spline regression model on poverty data in Bali Province.

Literature Review

A. Spline Approach to Semiparametric Regression

Undoubtedly, in regression analysis, parametric regression modeling is a common approach. However, there are other methods used related to the conformity of the shape of the data curve, namely the nonparametric regression method. Nonparametric regression methods have high flexibility and are generally free from assumptions commonly used in parametric regression approaches (Hardle, 1994). The nonparametric regression approach that is quite popular is spline. Spline has a very good ability to handle data whose behavior changes at sub-specified intervals.

The combination of parametric regression with nonparametric regression is called semiparametric regression (Hardle, 1994); (Draper & Smith, 1992). For example the following semiparametric regression model is given:

$$yi = \sum_{s=1}^{S} g_s(x_{is}) + \sum_{l=1}^{L} f_l(t_{ll}) + \varepsilon_i \quad (1)$$

Where $g_s(x_{is})$ is a parametric component and $f_l(t_{ll})$ is a nonparametric component with being the number of research units. The component ethics $g_s(x_{is})$ curve calculated by linear regression with S showing the number of parametric x components. The nonparametric component curve f_l (t_{ll}) is modeled linear spline with knots k_1, k_2, k_r while the symbol L shows the number of nonparametric components t.

$$g_s(x_{is}) = \alpha_s x_{is} \quad (2)$$

$$f_l(t_{il}) = \beta_l t_{il} + \sum_{u=1}^{r} \gamma_{ul}(t_{il} - k_{ul})_+ \quad (3)$$

Thus, the semiparametric regression model can be written as follows:

$$y_i = \sum_{s=1}^{S} \alpha_s x_{is} + \sum_{l=1}^{L} \left(\beta_l t_{il} + \sum_{u=1}^{r} \gamma_{ul}(t_{il} - k_{ul})_+ \right) + \varepsilon_i \quad (4)$$

In nonparametric and semiparametric regression with a spline approach, the important thing that plays a role in obtaining the spline estimator is the selection of the optimal knot point. One method that is often used in selecting the optimal knot point is *Generalized Cross Validation* (GCV). GCV is the development of CV (Wahba, 1990). Theoretically, Wahba has shown that the GCV method has asymptotic optimal properties, which other methods do not have. The full theory of GCV can be seen in Wahba. The GCV function for selecting the optimal knot point can be shown in the following equation:

$$GCV(K_1, K_2, ...K_r) \frac{MSE(K_1, K_2, ...K_r)}{(n^{-1} trace[l - A(K_1, K_2, ...K_r)])^2} \quad (5)$$

With

$$MSE(K_1, K_2, ...K_r) = n^{-1} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \quad (6)$$

B. Model Parameter Testing

Testing parameters in multiple regression as well. Semiparametric regression with a linear spline approach performed both simultaneously and partially. Simultaneous tests were conducted to determine whether predictor variables simultaneously had a significant effect on the model or not. Simultaneous tests are carried out on the following hypotheses:

H_0: $\alpha_s = \beta_l = \gamma_{ul} = 0$

H_1: there is at least one $\alpha_s \neq 0$, $\beta_l \neq 0$ or $\gamma_{ul} \neq 0$

The F test is used as a test statistic, which is as follows :

$$F = \frac{MS_{Regresi}}{MS_{Residusi}} \quad (7)$$

MS Regression and MS Error were obtained and Variance Analysis (ANOVA) as indicated by Draper and Smith.

The partial test is carried out on the following hypothesis:

H_0: $\alpha_s = \beta_l = \gamma_{ul} = 0$

$H_1: \alpha_s \neq 0, \beta_l \neq 0, \gamma_{ul} \neq 0$

The *t* test is used as a test statistic in partially testing parameters, namely:

$$t_{hitung} = \frac{\alpha_s}{SE(\alpha_s)}, \; t_{hitung} = \frac{\beta_l}{SE(\beta_l)}, \; t_{hitung} = \frac{\beta_{ul}}{SE(\beta_{ul})} \tag{8}$$

with the conclusion reject H_0 if $|t_{hitung}| > t_{\left(\frac{\alpha}{2}, n\{S+L(r+1)\}-1\right)}$, where *n* is the unit of observation and *p* is the number of parameters or if the *p-value*< α (Draper & Smith, 1992).

C. Research Test

In addition to *scatterplot*, linearity tests can be done to determine the relationship pattern of each predictor variable with the response variable. Ramsey's RESET test is a nonlinearity detection test proposed by Ramsey. The theory related to the Ramsey's RESET Test in full can be seen in (Ramsey, 1969) and (Gujarati, 2004).

Methods

A. Data source

This study uses secondary data. In 2012, publications published by BPS in Bali Province with observation units were all regencies, namely 9 regencies in Bali Province.

B. Research Phase

The first objective of this research was to obtain an estimate of the semiparametric regression curve with the spline approach. To complete the first goal, the steps are as follows:

1. Given the following semiparametric regression model:

$$y = \sum_{s=1}^{S} g_s(x_s) + \sum_{l=1}^{L} f_l(t_l) + \varepsilon \tag{9}$$

Where $\sum_{s=1}^{S} g_s(x_s)$ are parametric and components $\sum_{l=1}^{L} f_l(t_l)$ is a nonparametric component.

2. Presenting the semiparametric regression model with the linear spline approach as follows:

$$y_i = \sum_{s=1}^{S} \alpha_s x_{is} + \sum_{l=1}^{L} \left(\beta_l t_{il} + \sum_{u=1}^{r} \gamma_{ul} (t_{il} - k_{ul})_+ \right) + \varepsilon_i \quad (10)$$

3. Getting estimates of α and β uses the method *Ordinary Least Square*(OLS) by completing optimization:

$$\underset{\alpha\theta}{Min} \{(\boldsymbol{y} - \mathbf{X\alpha} - \mathbf{T\theta})'(\boldsymbol{y} - \mathbf{X\alpha} - \mathbf{T\theta})\} \quad (11)$$

4. Obtain an estimate of the semiparametric regression curve with a linear spline approach:

$$y_i = \sum_{s=1}^{S} \alpha_s x_{is} + \sum_{l=1}^{L} \left(\beta_l t_{il} + \sum_{u=1}^{r} \gamma_{ul} (t_{il} - k_{ul})_+ \right) + \varepsilon_i \quad (12)$$

The second objective of this research was to apply a semiparametric regression model with linear spline to the percentage of poverty data in Bali Province. To complete the second objective, the steps are as follows:

1. Create *scatter plots* and perform linearity tests as initial identification to determine which data patterns are classified as parametric and nonparametric components according to *scatter* interpretation *plot* and linearity test.
2. Approach the nonparametric component curve with a linear spline with various knots (one knot, two knots, three knots and a combination of all three knots).
3. Perform hypothesis testing for parameters simultaneously and partially.
4. Calculate the MSE and R^2 as part of the model goodness criteria.
5. Draw conclusions.

Results and Discussion

A. Estimation of Linear Spline Semiparametric Regression Curves

Given the semiparametric linear spline regression model as follows.

$$y_i = \sum_{s=1}^{S} g_s(x_{is}) + \sum_{l=1}^{L} f_l(t_{il}) + \varepsilon_i \quad (13)$$

with $i = 1,2,..., n$ where y_1 is the response variable, $\sum_{s=1}^{S} g_s(x_{is})$ is a parametric component with the number of predictor variables of parametric components as much as $s = 1,..., S$, $\sum_{l=1}^{L} f_l(t_{il})$ nonparametric component with a number of nonparametric component predictor variables of $l=1, ...,L$ and *error* ε_i, $i=1,2,..., n$ saling independent. *are* mutually independent. The parametric component curve $g_s(x_{is})$ approximated by a linear function, namely:

$$g_s(x_{is}) = \alpha_0 + \alpha_s x_{is},\ S = 1,2\ldots,\ S \quad (14)$$

Where α_s, is a linear parametric component estimator, while s is the number of predictor variables for parametric components. The nonparametric component curve is approximated by a linear spline function, with knots $k_1, k_2, ... k_r$ following:

$$f_l(t_{il}) = \beta_l t_{il} + \sum_{u=1}^{r} \gamma_{ul}(t_{il} - k_{ul})_+ \quad (15)$$

Thus, the semiparametric regression model can be written in general can be presented in the form of a matrix in the following forms:

$$\boldsymbol{y} = \mathbf{X\alpha + T\theta + \varepsilon,} \quad (16)$$

The estimator for the parameters α and θ is obtained using the *Ordinary Least Square* (OLS) method. Based on the model:

$$\boldsymbol{y} = \mathbf{X\alpha + T\theta + \varepsilon,} \quad (17)$$

obtained equation:

$$\varepsilon = \boldsymbol{y} - \mathbf{(X\alpha + T\theta) = y - X\alpha + T\theta.} \quad (18)$$

To get an estimator of parameters α and θ, it is done with partial derivatives. Partial derivatives Q (α, θ) to α given by:

$$\frac{\partial Q(\alpha,\theta)}{\partial \alpha} = \frac{\partial(y'y - 2\alpha'X'y - 2\theta'T'y + 2\alpha'X'T\theta + \alpha'X'X\alpha + \theta'T'T\theta)}{\partial \alpha}.$$

$$= \mathbf{-2X'}\,\boldsymbol{y} \mathbf{+ 2X'T\theta + 2X'X\alpha} \quad (19)$$

The estimator α is given by:

$$\mathbf{A = (X'X)^{-1}\,(X'}\boldsymbol{y} \mathbf{- X'T\theta)} \quad (20)$$

Next, to get the estimator and parameter θ, partial derivatives are carried out Q (α, θ) against θ as follows:

$$\frac{\partial Q(\alpha,\theta)}{\partial \theta} = \frac{\partial(y'y - 2y'X\alpha - 2\theta'T'y + 2\theta'T'X\alpha + \alpha X'X\alpha + \theta'T'T\theta)}{\partial \theta}$$

$$= \mathbf{-2X'}\,\boldsymbol{y} \mathbf{+ 2X'T\theta + 2X'X\alpha} \quad (21)$$

So that it is obtained:

$$\boldsymbol{\theta} \mathbf{= (T'T)^{-1}\,(T'}\boldsymbol{y} \mathbf{- T'X\alpha)} \quad (22)$$

The estimator α in equation (4) still contains the estimator θ. Similarly, the estimator in equation (6) still contains the α estimator. To obtain an α estimator that is free from the estimator θ, substitution of equation (6) is done in equation (4). Thus, α can be stated as follows:

$$\boldsymbol{\alpha} \mathbf{= M\,(k)}\,\boldsymbol{y} \quad (23)$$

Where

$$\boldsymbol{M\,(k) = U(X'X)^{-1}\{X' - X'T\,(T'T)^{-1}T'\}} \quad (24)$$

M (k)is a matrix that contains points of knots in nonparametric components. Next, to get θ which is free from the α estimator, substitute equation (4) in equation (6) as follows.

$$\boldsymbol{\theta} \mathbf{= (T'T)^{-1}[T'}\boldsymbol{y} \mathbf{- T'X\{(X'X)^{-1}\,(X'}\boldsymbol{y} \mathbf{- X'T\theta)\}]}$$

$$\mathbf{= (T'T)^{-1}\,T'}\boldsymbol{y} \mathbf{- (T'T)^{-1}\,T'X(X'X)^{-1}\,X'}\boldsymbol{y} \mathbf{+ (T'T)^{-1}\,T'X(X'X)^{-1}\,X'T\theta}$$

Next, the tribe containing the estimator is grouped in one segment. Then obtained:

$$\boldsymbol{\theta} = \mathbf{\{I-(T'T)^{-1}T'X(X'X)^{-1}X'T\}^{-1}\ \{(T'T)^{-1}T'}\boldsymbol{y}\mathbf{-(T'T)^{-1}T'X(X'X)^{-1}X'}\boldsymbol{y}\mathbf{\}}$$

if it is defined that $\mathbf{V = \{I-(T'T)^{-1}T'X(X'X)^{-1}X'T\}^{-1}}$, then

$$\boldsymbol{\theta} = \mathbf{V\{(T'T)^{-1}T'}\boldsymbol{y}\mathbf{-(T'T)^{-1}T'X(X'X)^{-1}X'}\boldsymbol{y}\mathbf{\}}$$

$$= \mathbf{V(T'T)^{-1}\ \{T'- T'X(X'X)^{-1}X\}}\ \boldsymbol{y} \quad (25)$$

Thus, θ can be stated as follows:

$$\boldsymbol{\theta} = \mathbf{N\ (k)}\ \boldsymbol{y} \quad (26)$$

Where

$$\mathbf{N\ (k) = V(T'T)^{-1}\{T' – T'X(X'X)^{-1}X'} \quad (27)$$

After obtaining an estimator for parametric and nonparametric components, the next step is to determine the semiparametric linear spline regression estimator.

$$\boldsymbol{y} = \mathbf{f(x,t)}$$

$$= \mathbf{X\alpha + T\theta}$$

$$= \boldsymbol{\omega}\mathbf{(k)}\boldsymbol{y} \quad (28)$$

With$\boldsymbol{\omega}\mathbf{(k) = XM\ (k) + TN\ (k)}$. element-*i*, i=1,2,…,n from the vector y in equation (32) can be presented as follows.

$$y_i = \sum_{s=1}^{S} \alpha_{js} x_{is} + \sum_{l=1}^{L} \left(\beta_l t_{il} + \sum_{u=1}^{r} \gamma_{ul} (t_{il} - k_{ul})_+ \right) + \varepsilon_i \quad (29)$$

Estimator α_s,s =1,2,..,S β_l, l = 1,2,...,L and γ_{ul}, u=1,2…,r obtained from the equation:

$$\boldsymbol{\alpha} = \mathbf{M\ (k)}\ \boldsymbol{y}\mathbf{,} \quad (30)$$

where

$$\mathbf{M\ (k) = U\ (X'X)^{-1}\ \{X' – X'T(T'T)^{-1}\ T'\},}$$

And

$$\boldsymbol{\theta} = \mathbf{(\beta'\ \gamma') = N\ (k)}\ \boldsymbol{y} \quad (31)$$

where

$$\mathbf{N(k) = V(T'T)^{-1}\{T' - T'X(X'X)^{-1}X'\}.}$$

B. Application of Linear Semiparametric Model on Poverty Data in Bali Province

Response variables with predictor variables are modeled using semiparametric linear spline regression. *Scatter* used *plot* and linearity test to determine the pattern of relationships formed.

Scatterplot of poor vs. AMH, unemployment, agricultural sector, LPE, RLS

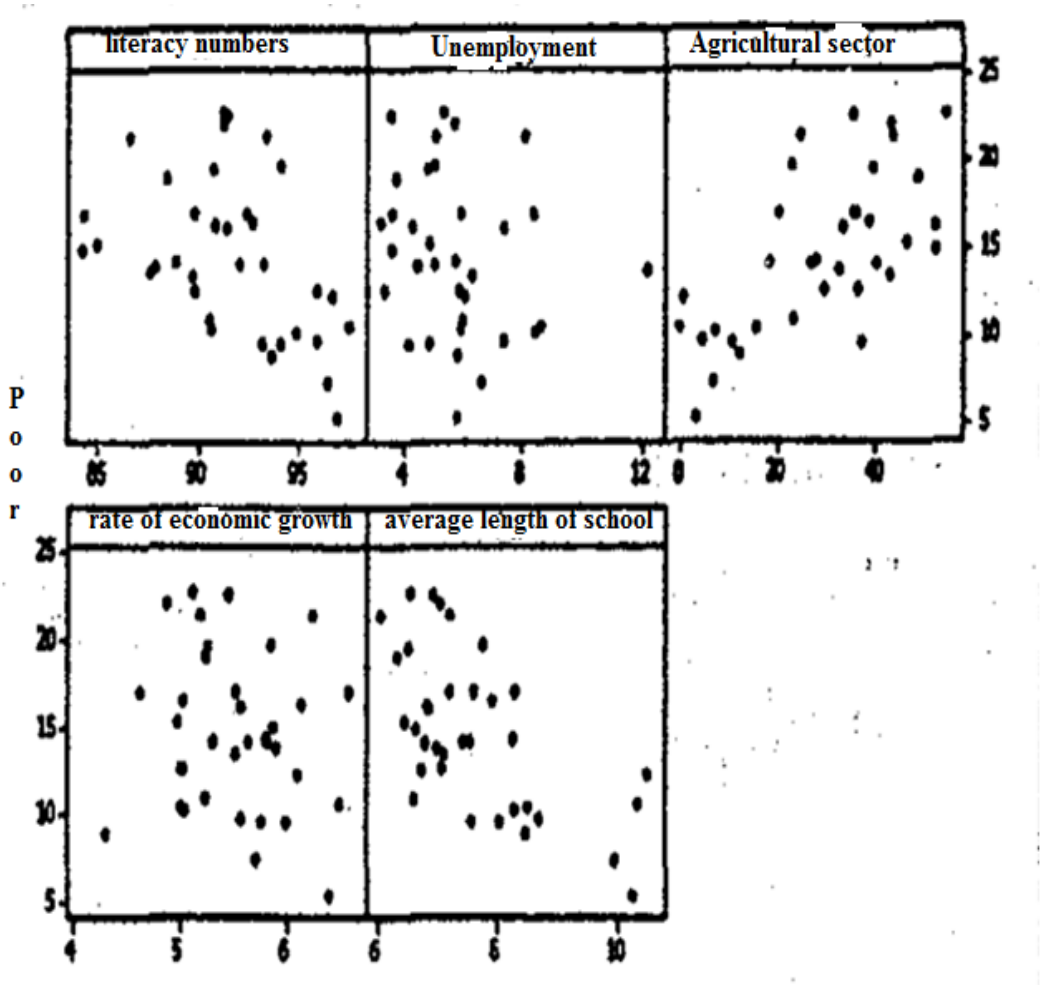

Figure 1. Scatterplot between y with x

Based on the *scatterplot*, it can be seen that there is a tendency for some data plots to form certain patterns (linear patterns) and some other data plots do not show a particular pattern.

The selection of the best model is done by comparing R^2 and MSE between the results of the *scatterplot* and the linearity test. In the previous explanation, in the know that the value of the minimum GCV obtained at three knots duck t test results for both models assuming a residual, R^2 and MSE both models can be see in Table 7.

Table 7. Comparison of Model Goodness Measures

Determination Basis Parametric Components and Nonparametric	**Residual Assumption**	**R^2 (%)**	**MSE**
Scatter plot	Fulfilled	67.97	9,785
Linearity Test	Fulfilled	56.3	11,806

Based on the values of R^2 and MSE in Table 7, a better model is the model obtained from the use of *scatter plots*.

Conclusion

The conclusions obtained based on the research that has been done are:

1. The semiparametric linear spline regression curve estimation is obtained from optimization:

   $$\underset{\alpha\theta}{Min}\ \{(y - X\alpha - T\theta)'\ (y - X\alpha - T\theta)\}$$

   This optimization produces an estimator for the following linear spline semiparametric curves.

$$\boldsymbol{y = \omega(k)y}$$

with $\boldsymbol{\omega(k) = XM\,(k) + TN\,(k)}$.

2. The semiparametric linear spline regression model that uses *scatter plots* as the basis for determining parametric and nonparametric components produces a model that is better than the use of the RESET Test.

References

Chen, K., & Jin, Z. (2006). Partial Linear Regression Models for Clustered Data. *Journal of the American Statistical Association, 101*(473), 195-204.

Draper, N. R., & Smith, H. (1992). *Applied Regression Analysis.* Jakarta: PT Gramedia Pustaka Utama.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker, Inc.

Gujarati, D. N. (2004). *Basic Econometrics* (5 ed.). New York: McGraw Hill International.

Hardle. (1994). *Applied Nonparametric Regression.* New York: Cambridge University Press.

Lin, X. H., & Carrol, R. J. (2001). Semiparametric Regression for Clustered Data. *Biometrics Trust, 88*(4), 1179-1185.

Ramsey, J. B. (1969). Test for Specification of Error in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society, Series B, 31*, 350-371.

Suharto, E. (2006). *Poverty and Social Protection in Indonesia: Initiating the Universal Social Security Model for Health.* Bandung: Alfabeta.

Wahba, G. (1990). *Spline Models For Observasion Data.* Pensylvania: SIAM.