Modeling of composite stock price index (CSPI) using semiparametric regression truncated spline based on GUI R

Modeling of composite stock price index (CSPI) using semiparametric regression truncated spline based on GUI R

W Marbun, Suparti, and D A I Maruddani

Department of Statistics, Diponegoro University, Semarang, 50275, Indonesia

Corresponding author : wilsonmarbunlate@gmail.com

Abstract. The Composite Stock Price Index (CSPI) is widely used as the beginning of consideration for investors to invest, because from the CSPI it can be known the general situation of market conditions is whether stock prices are experiencing an increase or decrease. This condition is characterized by a decrease or increase in the price of the CSPI. The Amount of Money Supply and the NASDAQ Index is thought to affect the price movements of the CSPI. The Amount of Money Supply has a nonparametric relationship pattern to the CSPI and the NASDAQ Index has a parametric relationship pattern to the CSPI. The correct method for conducting modeling is use the semiparametric *spline truncated* regression. The semiparametric regression *spline truncated* coefficient is estimated using the *Ordinary Least Square* (OLS) method which is determined based on the polynomial degree, much and the location of the optimum *knot* point is seen from the criteria of *Mean Square Error* (MSE). This study uses (*Graphical User Interface*) GUI R with the intention of facilitating the analysis process. The data used are monthly data from June 2014 to March 2019. Based on the results of the analysis that has been done, the best semiparametric *spline truncated* regression model with order 3 with the optimal three *knots* is 4246.361, 4443.078, and 4730.38. Evaluation results the *in-sample* data model produces a coefficient of determination of 90.25%. The results of the performance evaluation of the *out sample* data model resulted in a MAPE value of 3770204% indicating the performance of the model was very good.

1. Introduction

The capital market is one of the supports for a country's economic sustainability [1]. Nowadays the capital market is one of the economic instruments that are developing significantly. This case is indicated by the increase in the stock price index which stock is one alternative that can attract the investors [2]. Composite Stock Price Index (CSPI) is a value that used as a tool to measure the combined performance of all company listed in the Indonesia Stock Exchange (IDX). Therefore, research on CSPI is important. CSPI is the beginning of consideration for investors to invest, because of CSPI can be seen the general situation of market conditions whether the stock price is experiencing an increase or decrease. There are theories and previous studies which concluded that the movement of the Composite Stock Price Index (CSPI) is infulenced by external factors from abroad and interal factors from within the country. NASDAQ index is often used in the study of external factors which infulence the price of the CSPI. Some internal or domestics factors that are often used to be macroeconomic factors which affect the CSPI movement some of them are the consumer Price Index (CPI), the Amount of Money Supply, BI rates, inflation, and others [3]. In this study, NASDAQ index will be used as an external factor and the Amount of Money Supply as an internal factor affecting the price of CSPI.

In the early analysis, the regression curve between each factor to the CSPI, it was seen that the regression curve between the NASDAQ index and the CSPI follows a linear function. Meanwhile, the Amount of Money Supply that the regression curve against the CSPI is not suitable if it follows a linear function, so is possible to approach a mixed regression function with a nonparametric function by using semiparametric regression. The nonparametric function that will be used is the spline truncated function. The use of GUI (Graphical User Interface) as a tool in statistics has been widely used. The GUI makes the statistisc analysis more interactive and easier to use. The GUI can be designed as needed. Therefore, researchers will model a Composite Stock Price Index (CSPI) using spline truncated semiparametric regression based on GUI R. Use of GUI R to help analysis to be more interactive and efficient. The problem that will be discussed in this study is how to model the Composite Stock Price Index (CSPI) using GUI R based spline truncated semiparametric regression for monthly IHSG data from June 2017 to March 2019. The purpose of this study was to obtain a Composite Stock Price Index (CSPI) model using semiparametric truncted spline regression based on GUI.

2. Literature review

2.1. The capital market and composite stock price index (CSPI)

In a country's economy, the capital market has an important role because the capital market as an intermediary for companies that will receive investment funds from the investor. One of the main indicators that reflect the performance of the capital market in Indonesia when it is experiencing an increase or is experiencing a decline is the Composite Stock Price Index (CSPI). CSPI is a value that is used as a tool to measure the combined performance of all company listed on the Indonesia Stock Exchange (IDX). The index serves as an indicator of market trends, meaning that the movement of the index describes the condition of the market at one time, whether the market is active or lethargic.

2.2. NASDAQ index

The NASDAQ index is one of the main indices in the United States. National Associenat of Sécurities Automated Quotations (NASDAQ) Composite Index is the first electronic's stock market in the world. NASDAQ index measures all the common domestic and international stocks listed on the NASDAQ stock market.

2.3. The amount of money supply

The Amount of Money Supply is the total value of money in the hands of the comunity[4]. The Amount of Money Supply in a narrow defiinition is the Amount of Money Supply which consists of currency and demand deposits. Menawhile, the Amount of Money Supply in a broad definition is the Amount of Money Supply in a narrow definition plus time deposits. The growth in the Amount of Money Supply enables the public to increase investment, including investment in shares or other investment instruments [5].

2.4. Semiparametric spline truncated regression

Semiparametric regression is a statistical method that consists of parametric and nonparametric regression. One of the most popular approach techniques in nonparametric regression is truncated spline [6]. Spline regression is a piece of polynomial has more flexible properties than ordinary polynomial, so that it is possible to adjust effectively to the local characteristics of the data. Spline is very dependent on the point of the knots. Knots are joint fusion points where there is a pattern of behavior change of a function at different interval. The semiparametric regression model with its parametric components having a simple linear pattern can be written as follows:

$$y_i = \sum_{j=0}^{l} \beta_j x_i + f(z_i) + \varepsilon_i \quad , i = 1,2,\dots,n \; ; j = 1,2,\dots,l \tag{1}$$

where y_i is i-th response variable, $\beta_j$is coefficient of parametric function, x_i is i-th predictor variable, $f(z_i)$ is nonparametric function, and ε_i is residual which has N (0, σ^2) distribution. The nonparametric function can be formed by spline approach. Spline function m order with r knots for one predictor variable can generally be written in the following form:

$$f(z_i) = \sum_{p=0}^{m-1} \delta_p z_i^j + \sum_{p=1}^{r} \delta_{p+m-1} (z_i - k_p)_+^{m-1} + e_i \tag{2}$$

with truncated function:

$$(z_i - k_p)_+^{m-1} = \begin{cases} (z_i - k_p)^{m-1}, & z \geq k_p \\ 0, & z < k_p \end{cases}$$

2.5. Semiparametric spline model estimation

Semiparametric spline model estimation can be done using OLS (Ordinary Least Square) by minimizing the Sum Square of Error (SSE). If written in matrix form, the Semiparametric Spline Truncated Regression equation can be written as follows:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}_{\prod}\boldsymbol{\delta}_{\prod} + \boldsymbol{\varepsilon} \tag{3}$$

Then, we determine $\boldsymbol{C} = [\boldsymbol{X} \quad \boldsymbol{Z}_{\prod}]$, $\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\delta}_{\prod} \end{bmatrix}$ so that the previous equation can be expressed as:

$$\boldsymbol{Y} = \boldsymbol{C\gamma} + \boldsymbol{\varepsilon} \tag{4}$$

Then, defined Sum Square Error (SSE) equation as follows:

$$\begin{aligned} SSE &= \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} \\ &= (\boldsymbol{Y} - \boldsymbol{C\gamma})^T(\boldsymbol{Y} - \boldsymbol{C\gamma}) \\ &= (\boldsymbol{Y}^T - \boldsymbol{\gamma}^T\boldsymbol{C}^T)(\boldsymbol{Y} - \boldsymbol{C\gamma}) \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{C\gamma} - \boldsymbol{\gamma}^T\boldsymbol{C}^T\boldsymbol{Y} + \boldsymbol{\gamma}^T\boldsymbol{C}^T\boldsymbol{C\gamma} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - (\boldsymbol{\gamma}^T\boldsymbol{C}^T\boldsymbol{Y})^T - \boldsymbol{\gamma}^T\boldsymbol{C}^T\boldsymbol{Y} + \boldsymbol{\gamma}^T\boldsymbol{C}^T\boldsymbol{C\gamma} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\gamma}^T\boldsymbol{C}^T\boldsymbol{Y} + \boldsymbol{\gamma}^T\boldsymbol{C}^T\boldsymbol{C\gamma} \end{aligned} \tag{5}$$

The condition for a function is a minimum value, i.e. (i) The first derivative is equated with vector **0** and (ii) the second derivative of the matrix element must be a positive definite matrix.

$\frac{\partial SSE}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} = -2\,\boldsymbol{C}^T\boldsymbol{Y} + 2\,\boldsymbol{C}^T\boldsymbol{C\gamma} = 0$, then $\boldsymbol{C}^T\boldsymbol{C}\hat{\boldsymbol{\gamma}} = \boldsymbol{C}^T\boldsymbol{Y}$

$\frac{\partial SSE}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} = 2\,\boldsymbol{C}^T\boldsymbol{C}$, the $\boldsymbol{C}^T\boldsymbol{C}$ matrix is a positive definite matrix[7]. Then $\hat{\boldsymbol{\gamma}}$ is the least squares estimator for $\boldsymbol{\gamma}$ with,

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{Y} \tag{6}$$

2.6. GUI (Graphical User Interface) from R and RS studio

Users of R software can not only do statistical computing, but can also visualize the results of their analysis in the form of a GUI (Graphical User Interface). Just like other computing, to make this interface needed R packages that are designed specifically for build a GUI, one of the most popular is shiny. The shiny package is an R package developed by RStudio developers to create GUI applications on R.

3. Research methods

The data used in the study are secondary data, The data used are monthly data from June 2014 to March 2019. The data is divided into two parts, namely in sample data from June 2014 to June 2017 and out sample data from July 2017 to March 2019. The variable used in this study is the Composite Stock Price Index (CSPI) as the response variable y with a predictor variable composed of 2 components, namely parametric component variables and non-parametric component variables. The predictor variable x or the variable used for the parametric component is the NASDAQ Index and the z predictor variable used for the non parametric component is the Amount of Money Supply. The steps of analysis are as follows:

1. Prepare the GUI R application
2. Determine the variable belongs to parametric or nonparametric component by the application after the data was inputed.
3. Determine the knot points for nonparametric component in the model by 1 knot, 2 knots, 3 knots combined for 2nd order, 3-rd order and 4-th order.
4. Determine the best model using knot points optimum based on the minimum MSE (Mean Square Error) value from each combination of experiment. The MSE method is generally defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

5. Estimating the parameter semiparametric regression for the best model by using the OLS (Ordinary Least Square) method.
6. Calculate the value of R^2 for in sample data as a criterion for the good of the model. The R^2 formula is generally defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

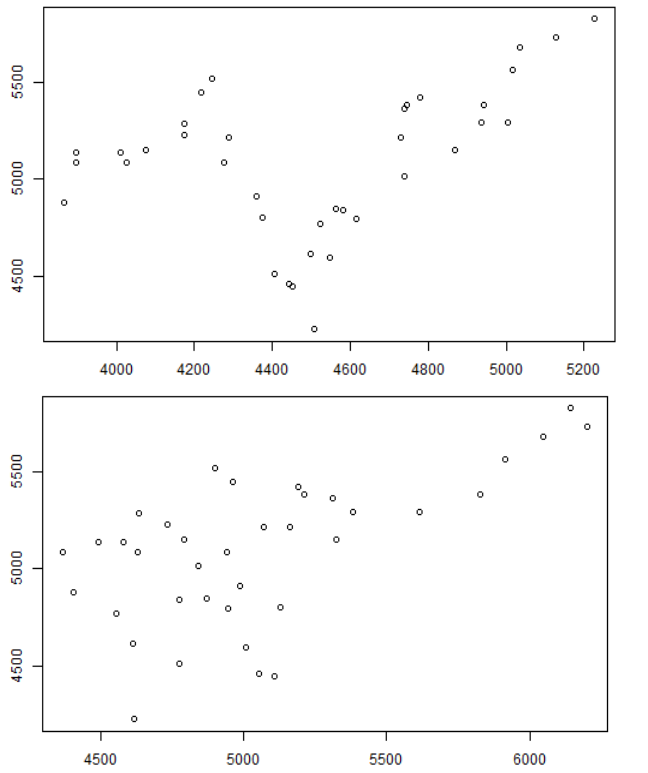
7. Calculate the value of MAPE for out sample data as a criterion for the accuration of the model. The MAPE formula is generally defined as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{9}$$

4. Results

4.1. Scatterplot analysis

In semiparametric regression, the determination of parametric and nonparametric components can be seen by looking at the pattern of the data. The following picture is scatterplots between the response variable CSPI (y) with each predictor variable NASDAQ index (x) and The Amount of Money Supply (z)



(a) (b)

Figure 1.(a) Scatterplot between y vs z, (b) Scatter plot between y vs x

In the Figure 1 (a), it can be seen that the scatterplot the relationship between the variable CSPI (y) and the Amount of Money Supply (z) shows a data pattern that showing irregular and fluctuating patterns. Therefore, the The Amount of Money Supply can be assumed as a nonparametric component. While in Figure 1 (b), there is a distribution of plot relationships between the variable CSPI (y) and NASDAQ index (x) directly proportional. In other word patterns can follow linear functions. Therefore NASDAQ index is assumed as the parametric component. That conclucion is supported by the correlation analysis that has been provided in the application. Based on the result of correlation analysis, can be seen that the correlation value of the Amount of Money Supply variable to the CSPI is 0.3730754 with a p-value = 0.02295 <α = 5% which means that the correlation between the two variables is significant. However, the correlation value is between 0.2-0.4 meaning a weak relationship between the two variables[8]. While the correlation value between the NASDAQ Index and CSPI is 0.6117764 with p-value = 5.743e-05. The p-value < α = 5% which shows that the correlation between the two variables is significant and based on the criteria in Usman (2008), the correlation value is between 0.6 -0.8 indicating a strong relationship between the two variables [8].

4.2. Optimum Knot point selection

Knot selection is important because it'll influence the result of model. The selection of the knots point is done by trial and error technique to obtain a minimum MSE value. In this study the choice of knots is chosen from the data value. That is, each non-parametric component predictor variable data as a candidate becomes a knot point except the maximum and minimum observed values of the variable. The following table will shows the combination of models tested by 1,2, and 3 knots point and 2,3, and 4 order which has minimum value of MSE.

Table 1. MSE value based on the combination of number of knots with order

No.	Order	Number of *Knot*	Value of *Knot*	MSE
1	2	1	4452.325	47748.35
2	3	1	4218.123	55407.78
3	4	1	4561.873	59014.96
4	2	2	4246.361 , 4404.085	20957.48
5	3	2	4443.078 , 4452.325	21175.52
6	4	2	4498.361 , 4868.651	35061.92
7	2	3	4246.361 , 4443.078 , 4730.38	13156.55
8	3	3	4443.078 , 4452.325 , 4868.651	16405.41
9	4	3	4373.208 , 4452.325 , 4498.361	14716.03

Based on the results in Table 1, it can be seen that the three-point knots experiment in second order produces the minumum MSE value of all experiments that is equal to 13156.55.

4.3. Spline truncated semiparametric regression parameter estimation based on optimum knot points
The next step is to find the estimated value of the spline semiparametric regression parameter based on the optimum order and knot points. Parameter estimation is done based on OLS (Ordinary Least Square) criteria, namely by minimizing the number of residual squares. The following table will shows the spline truncated semiparametric regression model parameters estimation.

Table 2. the spline truncated semiparametric regression model parameters estimation.

Variable	Parameter	Parameter Estimated Value
x	β_0	4564.817
	β_1	0.2764097
z	δ_0	-3429.495
	δ_1	0.6807526
	δ_2	-5.592681
	δ_3	7.487946
	δ_4	-2.132885

Based on the result of parameter estimation in Table 2, the semiparametric truncated spline regression model is written by,

$$\hat{y}_i = 4564.817 + 0.2764097x_{\mathrm{i}} - 3429.495 + 06807526z_i - 5.592681(z_i - 4246.361\ )_+ + 7.487946(z_i - 4443.078)_+ - 2.132885(z_i - 4730.38)_+$$

If the constants of the parametric component and the nonparametric component are combined, the model can be written as follows:

$$\hat{y}_i = 1135{,}322 + 0{,}2764097x_{\mathrm{i}} + 0{,}6807526z_i - 5{,}592681(z_i - 4246{,}361)_+ + 7{,}487946(z_i - 4443{,}078)_+ - 2{,}132885(z_i - 4730{,}38)_+$$

The model equation and its truncated form can be stated as follows:

$$\hat{y}_{\mathrm{i}} = \begin{cases} 1135.322+0.2764097x_{\mathrm{i}}+0.6807526z_{\mathrm{i}} \ . & z<4246.361 \\ 1135.322+0.2764097x_{\mathrm{i}}+0.6807526z_{\mathrm{i}}-5.592681\ (z_{\mathrm{i}}-4246.361)\ . & 4246.36\leq z<4443.078 \\ 1135.322+0.2764097x_{\mathrm{i}}+0.6807526z_{\mathrm{i}}-5.592681(z_{\mathrm{i}}-4246.361)+7.487946(z_{\mathrm{i}}-4443.078)\ . & 4443.078\leq z<4730.38 \\ 1135.322+0.2764097x_{\mathrm{i}}+0.6807526z_{\mathrm{i}}-5.592681(z_{\mathrm{i}}-4246.361)+7.487946(z_{\mathrm{i}}-4443.078)-2.132885(z_{\mathrm{i}}-4730.38) & z\geq 4730.38 \end{cases}$$

Based on the polynomial pieces, it is seen that CSPI is influenced by the Amount of Money Supply in 4 different segments and still continuous, i.e when the Amount of Money Supply is smaller than 4246.361, Total Money Supply between 4246.361 to 4443.078, Total Money Supply between 4443.078 to 4730.38, The Amount of Money Supply is greater than 4730.38.

4.4. Model evaluation

Based on the results of processing for in sample data, obtained R^2 value of 0.9075. That means, the predictor variables x and z affect the response variable y by 90.75% and 9.25% is influenced by other variables. Based on the results of processing in the GUI R program , the MAPE value for out sample data is 3.770204%. This value is at the MAPE interval <10%, which means that the performance of the model was very good.

5. Conclusion

Based on the discussion of the results of research that has been done, the best semiparametric regression model for CSPI data analysis with the factors that affect it by the NASDAQ Index as a parametric component and the Amount of Money Supply as a non parametric component is a spline truncated order 2 (linear) regression model with 3 optimum knots determined based on criteria the minimum MSE value is 13156.55. The optimum knots point is 4246.361, 4443.078, and 4730.38. The equation of the spline truncated semiparametric regression model obtained is as follows, $\hat{y}_i = 4564.817 + 0.2764097x_{\text{i}} - 3429.495 + 0.6807526z_i - 5.592681(z_i - 4246.361\ )_+ + 7.487946 - 2.132885(z_i - 4730.38)_+$
In evaluating of the in sample data model is obtained the R^2 value of 0.9075, which means that the predictor variable influences the response variable by 90.75% and 9.25% is influenced by other variables. Whereas in evaluating the performance of the out sample data model, the MAPE value of 3.770204% , which means that the spline truncated semiparametric regression model produced was very well used in forecasting.

References

[1] Kowanda D, Binastuti S, Rowland Pasaribu R B F and Ellim M 2014 *Jurnal Akutansi & Manajemen* **25**(**2**) pp. 79-88
[2] Sihombing T 2013 *Jurnal Ekonomis* **7**(2) pp 15-29
[3] Heriyanto & Chen M 2014 *Jurnal Nominal* **3**(2) pp 25-54
[4] Anggarini, D T 2016 *Moneter* **3**(2) pp 161-169
[5] Sudirman 2015 *Pasar Modal dan Manajemen Portofolio* (Gorontalo: Sultan Amai Press IAIN Sultan Amai)
[6] Budiantara I N 2005 *Berkala MIPA* **3**(15) pp 55-61
[7] Searle S 1971 *Linear Models s.l.*(:John Wiley & Sons., Inc)
[8] Usman H & Akbar P S 2008 *Pengantar Statistika. 2nd* ed. (Jakarta: Bumi Aksara)