Comparison of generalized cross validation and unbiased risk method for selecting optimal knot in spline truncated

A R Devi¹, R F W Pratama² and Suparti¹

¹Department of Statistics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia
²Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya
E-mail: alvita.erdhe@gmail.com

Abstract. Gini ratio is an indicator to measure income inequality. Gini ratio of Indonesia in 2017 is 0.391, still far from Gini ratio target by Bappenas in 2019, that is 0.36. The Gini ratio modeling in this study uses a nonparametric regression approach because the form of the regression curve between the Gini ratio and its predictive variables is unknown. One of the estimators in nonparametric regression is spline truncated. Spline truncated has a knot that adjusts to the local characteristics of a function or data more effectively. The number of knots and their location affect the form of regression curve estimation, so it’s important to obtain optimal knot. There are methods for selecting optimal knots, such as Generalized Cross Validation (GCV) and Unbiased Risk (UBR). This study compares GCV and UBR in selecting optimal knots on Gini Ratio data in Indonesia 2017. The criteria of the best model are based on Mean Squared Error (MSE) and $R^2$ values. From the result, the optimal knot from GCV was a combination of 3-2-2-3 knot with MSE of 0.00085 and $R^2$ of 79.18%. Meanwhile, by using UBR, the optimal knot is three knots with MSE of 0.00095 and $R^2$ of 66.42%. In conclusion, GCV generated better model than UBR.

1. Introduction
Gini ratio is one indicator to measure income inequality. Income inequality is a condition where the income distribution in society is not evenly distributed. High income inequality between groups of people or regions can lead to social jealousy, vulnerability to regional disintegration and wider economic disparities [1]. In 2017, the Gini ratio in Indonesia is 0.391, which means the income inequality in Indonesia is at a moderate level. This number is still far from the Gini ratio target by the Indonesian government in 2019, that is 0.36. The income inequality in Indonesia needs to be reduced, by the president direction listed in the National Medium Term Development Plan 2015-2019.

Several studies on Gini ratios and factors that affect them have been conducted by several researchers. The percentage of households with PLN electricity sources is used as variables that affect Gini ratio on research of spline truncated on longitudinal data [1]. Human development index, the unemployment rate and economic growth also have a significant effect on the inequality of income distribution. By considering these studies and data availability, this study used the percentage of households with PLN electricity sources, human development index, unemployment rate and economic growth as predictor variables.

The result of an initial exploratory study by using scatter plot showed a relationship pattern between response and predictors that did not form a particular pattern. If the pattern of regression curve is unknown, it is recommended to use nonparametric regression [2, 5]. Nonparametric
regression does not depend on an assumption of the form of the regression curve. The estimated form of regression curve is expected to adjust to data without being influenced by the researcher’s subjectivity [2]. There are some method in nonparametric regression that is commonly used, such as Kernel [3, 4, 5, 6], Spline [3, 5, 7, 8, 9], Local Polynomial [10, 11] and Fourier Series [12, 13, 14]. Among estimators that previously mentioned, spline method is quite popular among researchers because it has several advantages, such as having a good visual interpretation, flexible, and can describe behavioral changes of data at certain sub-intervals [2, 8].

The spline is a polynomial that has segmented properties, which provides more flexibility than ordinary polynomial, making it possible to adjust more effectively to local characteristics of a function or data [15]. This segmented nature also provides advantages in overcoming data patterns that show a sharp rise or fall with the help of knot points [5].

There are some things to consider in forming a spline regression model, such as determining the order of the model, the number of the knot and the location of the knot [16]. Thus, determining the optimal knots is very important to obtain the best model. Generalized Cross Validation (GCV) is one of the models that is often used in selecting the optimal knot. The GCV method provides several advantages over other methods, including asymptotic optimal, invariance to transformation and it doesn’t need a known population variance [9]. Another method for selecting knots is Unbiased Risk (UBR). UBR is a smooth parameter selection method that requires an estimated value of the known error variance [9, 17]. The UBR method is used to select the smoothing parameter for spline estimation with non-Gaussian data [18, 19]. From simulation performed by [20], the UBR method estimated knots which were close with the original knot, so the UBR method can be taken into consideration in choosing optimal knot. Numerical experiments conducted by [21] show that the estimation of GCV and UBR is basically the same in generating experimental accuracy when the same variants are used to produce experimental data [9]. This study aims to compare the optimal knot selection method of GCV and UBR on spline truncated nonparametric regression on Gini ratio data in Indonesia. The best model is determined by MSE criteria, $R^2$ criteria and simultaneous and partial tests on each model.

2. Literature review

2.1. Nonparametric regression

Regression analysis is a dependency analysis between response variables and predictor variables. The relation of response and predictor variables for $n$ observations $(x_i,y_i)$, $i = 1,2, ..., n$ is:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1,2, ..., n$$  \hspace{1cm} (1)

where $y_i$ is response variables of $y$ on $i$-th observation, $x_i$ is predictor variables of $x$ on $i$-th observation, $\varepsilon_i$ is an error on $i$-th observation which is independent random variable with mean 0 and variance constant $\sigma^2$, $f(x_i)$ is the regression curve at point $x_i$ [2].

Parametric regression is the regression that has a known form of regression curve $f(x_i)$. Meanwhile, nonparametric regression is used when the form of regression curve $f(x_i)$ is unknown. Nonparametric regression has high flexibility because the form of regression curve adjusts to its data without being influenced by the subjectivity of the researcher [2]. Some of the approach in nonparametric regression are kernel regression, local polynomial, wavelet, fourier and spline.

2.2. Multivariable spline truncated regression

Suppose that data $(x_{1i},x_{2i}, ..., x_{pi},y_i)$ where the relation between $(x_{1i},x_{2i}, ..., x_{pi})$ and $y_i$ is assumed to follow nonparametric regression model, the model was expressed in equation (2):

$$y_i = f(x_{1i},x_{2i}, ..., x_{pi}) + \varepsilon_i, i = 1,2, ..., n$$  \hspace{1cm} (2)

where $f$ is the unknown form of regression curve. If the regression curve $f$ is an additive model and is approached by spline function, then regression model was expressed in equation (3):
\[ y_i = \sum_{j=1}^{p} f(x_{ji}) + \varepsilon_i, i = 1, 2, ..., n \] (3)

where

\[ f(x_{ji}) = \beta_{j0} + \sum_{l=1}^{m} \beta_{jl}x_{ji}^l + \sum_{k=1}^{r} \beta_{j(m+k)}(x_{ji} - K_{jk})^m_+, j = 1, 2, ..., p \] (4)

with

\[ (x_{ji} - K_{jk})^m_+ = \begin{cases} (x_{ji} - K_{jk})^m, & x_{ji} \geq K_{jk} \\ 0, & x_{ji} < K_{jk} \end{cases} \] (5)

where the function has a number of \( r \) knot \( K_{j1}, K_{j2}, ..., K_{jr} \).

2.3. Selecting optimal knot
Knot is the point where there is a change of pattern behavior at different intervals [2]. The form of spline estimator is influenced by the location and the number of knot, therefore, the selection of optimal knot is necessary to obtain a spline estimator according to the data.

GCV is a method that is commonly used to select the optimal knot. The GCV method provides several advantages over other methods, including asymptotic optimal, invariance to transformation and it doesn’t need a known population variance [9]. GCV function was given by equation (6):

\[ \text{GCV(K)} = \frac{n^{-1} y^T (I-A(K))^2 y}{(n^{-1} \text{trace}[I-A(K)])^2} \] (6)

Meanwhile, UBR is a method of selecting a smoothing parameter that requires an estimated value of error variance [9][19]. UBR gives the optimal result for selecting the smoothing parameter for spline estimation with non-Gaussian data [18][19]. UBR function was given by equation (7):

\[ \hat{R}(K) = \frac{1}{n} \| (I-A(K))y \|^2 - \frac{\sigma^2}{n} \text{tr} \left( I-A(K) \left( I-A(K) \right)^{-1} \right) \text{tr} (A(K)) (A(K)) \] (7)

where

\[ A(K) = X(K) (X'(K) X(K))^{-1} X'(K) \] (8)

which is symmetrical and semi-definite positive and \( X(K) \) is an \( n \times (1 + p(m + r)) \) matrix of model that forms \( f_K \) estimator and depends on knot.

3. Research methodology
The purpose of this study is to compare GCV and UBR method for selecting optimal knot on spline truncated nonparametric regression with a case study of multivariable data of Gini ratio in Indonesia 2017. The initial step before data modeling is to determine the pattern of relations between variables with a scatter plot between each predictor and response. To obtain further identification of data pattern, perform a formal linearity test by using Ramsey’s Reset test. If the pattern of the relationship between response and predictors is linear, modeling is performed by using linear regression analysis, while spline truncated nonparametric regression was performed if the pattern is nonlinear.

The modeling data by using spline truncated in this study using a linear model with one, two, three and a combination of knots. The selection of the optimal knot is performed by using GCV and UBR method. The best model is obtained by MSE and \( R^2 \) criteria. The next step is to test the parameter significance simultaneously and partially.

4. Result and discussion
Gini ratio is one of indicator to measure income inequality. This indicator is also used as a benchmark to see the development success in economy through income inequality. The Gini ratio of Indonesia in 2017 is 0.391, which is still in the moderate category. The highest Gini ratio is 0.44 in Yogyakarta Province, while the lowest is in Bangka Belitung Islands Province, which is 0.276.
In this study, the predictors for modeling the Gini ratio are Percentage of Households with PLN Electricity Sources \((x_1)\), Human Development Index \((x_2)\), Unemployment Rate \((x_3)\), and Economic Growth \((x_4)\). The initial stage before modelling is to analyze the pattern of relation between response and each predictor. The pattern of relation between gini ratio with each predictor can be seen in figure 1.

![Figure 1. Scatter plot between Gini ratio data with each predictors](image)

Based on Figure 1, it was clear that the relation between Gini ratio with its predictors didn’t follow a certain pattern. Further identification is required to understand if the model follows a certain pattern or not. This study used Ramsey’s Reset Test to perform linearity test with hypothesis:

- \(H_0\): It is sufficient to model predictors and response with linear model
- \(H_1\): It is insufficient to model predictors and response with linear model

The decision criteria are to reject \(H_0\) if \(p\)-value < 0.05. The results of the linearity test by using Ramsey’s Reset Test were presented in table 1.

| Predictor   | \(p\)-value | Relation   |
|-------------|-------------|------------|
| \(x_1\)     | 0.0141      | Nonlinear  |
| \(x_2\)     | 0.01203     | Nonlinear  |
| \(x_3\)     | 0.3034      | Linear     |
| \(x_4\)     | 0.8851      | Linear     |
| \(x_1 + x_2 + x_3 + x_4\) | 0.04887 | Nonlinear  |

From table 1, the relation between \(x_1\) and \(x_2\) predictor and \(y\) response followed a nonlinear relation, while \(x_3\) and \(x_4\) predictor and \(y\) response followed a linear relation. Meanwhile, variables of \(x_1 + x_2 + x_3 + x_4\) simultaneously and \(y\) formed a nonlinear relation. Although the relation between \(x_3\) and \(x_4\) predictor and response was linear, the response and the four predictor simultaneously had a nonlinear pattern. Therefore, the gini ratio data and each predictor can be modeled by using spline truncated nonparametric regression.

After modeling the Gini ratio data and its predictors by using linear spline truncated, the best model is chosen from the results. The best model is a model that has the optimal knot, which is determined by the minimum value of GCV or UBR. The spline truncated linear model with optimal GCV value was presented in table 2, while the spline truncated linear model with optimal UBR value was presented in table 3.

| Table 1. Results of the linearity test with Ramsey’s reset test |
|-------------|------------|
| Predictor   | \(p\)-value | Relation   |
|-------------|-------------|------------|
| \(x_1\)     | 0.0141      | Nonlinear  |
| \(x_2\)     | 0.01203     | Nonlinear  |
| \(x_3\)     | 0.3034      | Linear     |
| \(x_4\)     | 0.8851      | Linear     |
| \(x_1 + x_2 + x_3 + x_4\) | 0.04887 | Nonlinear  |

**Table 2. Spline linear 1, 2, 3 and combination knot with GCV method**
| Optimal Knot at Variable | GCV  |
|--------------------------|------|
| 1 knot                   |      |
| $x_1$                    | 57.507 | 64.809 | 3.613 | 2.914 | 0.00123 |
| 2 knot                   |      |
| $x_1$, $x_2$             | 55.741 | 64.174 | 3.376 | 2.655 | 0.00112 |
| 3 knot                   |      |
| $x_1$, $x_2$, $x_3$      | 57.507 | 64.809 | 3.613 | 2.914 | 0.00103 |
| comb. knot               |      |
| $x_1$, $x_2$, $x_3$, $x_4$ | 57.507 | 64.809 | 3.613 | 2.91445 | 0.00085 |

Table 3. Spline Linear 1, 2, 3 and Combination Knot with UBR Method

| Optimal knot at variable | UBR  |
|--------------------------|------|
| 1 knot                   |      |
| $x_1$                    | 53.975 | 63.538 | 3.140 | 2.395 | 0.09331 |
| 2 knot                   |      |
| $x_1$, $x_2$             | 87.535 | 75.612 | 7.632 | 7.333 | 0.07623 |
| 3 knot                   |      |
| $x_1$, $x_2$, $x_3$      | 89.302 | 76.247 | 7.869 | 7.593 | 0.06106 |
| comb. knot               |      |
| $x_1$, $x_2$, $x_3$, $x_4$ | 89.302 | 76.247 | 7.869 | 7.593 | 0.06106 |

Based on table 2, the best model was generated by combination knot of 3-2-2-3 with GCV minimum of 0.00085, while based on Table 3, the best model was generated by three knot with UBR minimum of 0.06106. The best model was measured by MSE and $R^2$ criteria. The MSE and $R^2$ value of GCV and UBR method were presented on table 4. From table 4, the spline linear model with GCV method generated smaller MSE and greater $R^2$ than spline linear model with UBR model. So, by considering the MSE and $R^2$ criteria, GCV generated a better model than UBR in this modelling.

Table 4. MSE and $R^2$ in spline linear with GCV and UBR method

| Minimum Value | MSE    | $R^2$  |
|---------------|--------|--------|
| GCV           | 0.00085 | 0.00053 | 79.183 |
| UBR           | 0.06106 | 0.00095 | 66.420 |

From the best model obtained by GCV and UBR method, simultaneous test of parameter significance is performed with hypothesis:

$H_0$: $\beta_{11} = \beta_{12} = \cdots = \beta_{prp} = 0$

$H_1$: there is minimal one $\beta_{jl} \neq 0$, $j = 1, 2, \ldots, 4$; $l = 1, 2, \ldots, r$

The ANOVA results for simultaneous hypothesis testing on the best model of GCV were presented in table 5.

Table 5. ANOVA of simultaneous test with GCV

| df | SS   | MS   | F-stat | p-value |
|----|------|------|--------|---------|
| Regression | 14 | 0.038 | 0.003  |
Error 19 0.010 0.001 5.162 0.001
Total 33 0.048

Based on the result in table 5, p-value (0.001) was less than the value of α (0.05), so the decision was to reject H₀. The conclusion was there is at least one significant parameter in the model. The simultaneous hypothesis testing was also performed on the best model of UBR. The results were presented in table 6.

Table 6. ANOVA of simultaneous test with UBR

|            | df | SS    | MS    | F-stat | p-value |
|------------|----|-------|-------|--------|---------|
| Regression | 16 | 0.032 | 0.002 |        |         |
| Error      | 17 | 0.016 | 0.001 | 2.102  | 0.07    |
| Total      | 33 | 0.048 |       |        |         |

Based on the result in table 6, p-value (0.07) was more than the value of α (0.05), so the decision was failed to reject H₀. In conclusion, the best models of GCV method was simultaneously significant, while the best models of UBR methods was not simultaneously significant. After simultaneous test, the significant test of partial parameters was performed on the best model obtained by GCV and UBR method. The hypothesis for significant test of partial parameters was:

H₀ : β jl = 0
H₁ : β jl ≠ 0, j = 1,2, ..., 4; l = 1,2, ..., r

The results of significant test of partial parameters of the best model by using GCV was presented in Table 7.

From the results of ANOVA simultaneous test and significant test of partial parameters, the conclusion is the best model of GCV generates significant model and parameters, while the best model of UBR is simultaneously not significant and partially only generated three significant parameters. By performing residual assumption test on model of GCV method, the models meet the identical, independent and normally distributed assumptions. In conclusion, from the result of MSE, R² and significance of the model, GCV generated a better model than UBR for modelling gini ratio data in Indonesia, 2017.
By estimating $\beta$ parameter, the best model of spline truncated with GCV method for multivariable data was:

$$\hat{y}_i = 0.439 + 0.151x_{i1} -1.138(x_{i2} -55.74) + 0.985(x_{i3} -57.51) + 0.008(x_{i4} -91.07) -0.108x_{i5}$$

$$+0.23(x_{i6} -64.17) -0.12(x_{i7} -64.81) + 0.019x_{i8} -0.268(x_{i9} -3.38) + 0.248(x_{i10} -3.61)$$

$$+0.025(x_{i11} -2.66) + 0.297(x_{i12} -2.91) -0.054(x_{i13} -7.85)$$

The model generated $R^2$ of 79.18. It means the model can explain Gini ratio in Indonesia by
79.18%. Figure 2 shows the comparison between estimation \(\hat{y}\) and actual \(y\) value of Gini ratio of the best model. Blue line is the estimation by the best model. Meanwhile, red line is the actual data. It can be seen from Figure 2 that the estimation follow the pattern of actual data. Thus, it can be concluded that the linear spline truncated method with selection of optimal knot by using GCV method produce estimation value that adjusts to actual data.

\[\hat{y}_i = \begin{cases} 
0.019x_{3i} & , x_{3i} < 3.38 \\
-0.268x_{3i} + 0.9 & , 3.38 \leq x_{3i} < 3.61 \\
-0.001x_{3i} + 0.01 & , x_{3i} \geq 3.61 
\end{cases}\]

If a region has Unemployment Rate less than 3.38%, then an increase of 1% in Unemployment Rate will cause Gini Ratio to increase by 0.019%. Regions with Unemployment Rate between 3.38% to 3.61%, any increase of Unemployment Rate by 1%, the Gini Ratio will decrease by 0.268%. Meanwhile, for the region with Unemployment Rate above 3.61%, if Unemployment Rate increases by 1%, the Gini Ratio tend to decrease by 0.001%. In the same way, the interpretation of the influence of The percentage of Households with PLN Electricity Sources \(x_1\), Human Development Index \(x_2\), and Economic Growth \(x_4\) variables on gini ratio can be determined.

5. Summary
From linear spline truncated modeling on Gini ratio multivariable data, for the selection of optimal knot by using GCV method, it generated MSE value of 0.00053, \(R^2\) a value of 79.183 and optimal model that was simultaneously and partially significant. While for the selection of optimal knot by using UBR method, it generated MSE value of 0.00095, \(R^2\) value of 66.42 and optimal model that was not significant simultaneously and partially only generated three significant parameters. In conclusion, for spline truncated modeling on Gini ratio data, GCV method produced a better model than UBR.
References

[1] Umam K 2018 Interval konfidensi parameter regresi nonparametrik spline truncated untuk data longitudinal, *Tesis*, Institut Teknologi Sepuluh Nopember.

[2] Eubank R L 1999 *Nonparametric Regression and Spline Smoothing* (New York: Marcel Dekker).

[3] Suparti and Tarno 2015 *Media Statistika* 8 103.

[4] Suparti 2013 *Media Statistika* 6 91.

[5] Hardle W 1990 *Applied Nonparametric Regression* (New York: Cambridge University Press).

[6] Kayri M and Zirhhoglu G 2009 *Ozean J. Appl. Sciences* 2 49.

[7] Diana R, Budiantara I N, Purhadi and Darmesto S 2014 *Model Assisted Stat. Appl.* 9 159.

[8] Lestari B, Budiantara I N, Sunaryo S and Mashuri M 2012 *J. Math. Stat.* 8 377.

[9] Wahba G 1990 *Spline Model for Observational Data* (Philadelphia: SIAM).

[10] Astuti E T, Budiantara I N, Sunaryo S and Dokhi M 2013 *Int. J. Appl. Math. Stat.* 33 92.

[11] Chamidah N, Budiantara I N, Sunaryo S and Zain I 2012 *J. Math. Stat.* 8 342.

[12] Asrini L J and Budiantara I N 2014 *ARPN J. Engineering and Appl. Sciences* 9 1501.

[13] Bilodeau M 1992 *The Canadian J. of Stat.* 3 257.

[14] Pane R, Budiantara I N, Zain I and Otok B W 2014 *Applied Math. Sciences* 8 5053.

[15] Budiantara I N, Suryadi F, Otok B W and Guritno S 2006 *J. Teknik Industri* 8 1.

[16] Montoya E, Ulloa N and Miller V 2014 *Int. J. Stat. Prob.* 3 3

[17] Wang Y 2011 *Smoothing Splines Methods and Applications* (California: CRC Press).

[18] Gu C 1992 *J. of Computational and Graphical Stat.* 1-2 169

[19] Wang Y 1998 *Royal Stat. Soc. Series B* (5r0) 341.

[20] Devi A R, Budiantara I B and Ratnasari V 2018 *AIP Conf. Proc.* 2021(1).

[21] Craven P and Wahba G 1979 *Num. Math.* 31 377.