Tadulako Social Humaniora Journal
Vol 3, No. 1, September 2022, pp. 13-22

Comparison Of Gaussian And Epancehnikov Kernels

Nur Fadillah[1], Priliany Audina Dariah[2], Anisa Anggraeni[3], Nur Cahyani[4], Lilies Handayani [5]

Faculty of Mathematics, Statistics Study Program,
Tadulako University
nrfdllhaaaa@gmail.com; Dinaapril264@gmail.com; anisaanggraeni700@gmail.com;
Cahyaninur298@gmail.com; lilies.stath@gmail.com

Abstract

Kernel regression is a nonparametric analysis with a smoothing method. Smoothing has become synonymous with nonparametric methods used to estimate functions. The purpose of smoothing is to remove variability from data that has no effect so that the characteristics of the data will appear clear. Kernel regression has a flexible form, and the mathematical calculations are easy to adjust. In kernel regression, an estimator is known, which is usually used to estimate the regression function, namely the Nadaraya-Watson estimator. This study aims to show how to estimate data using nonparametric regression Gaussian and Eponocvh kernels with the Nadaraya-Watson estimator, and the bandwidth selection methods are "Rule of Thumb" bandwidth, Unbiased Cross Validation, Biased Cross Validation, and Complete Cross-Validation. The results of this study indicate that the MSE value generated by the Epanechnikov kernel function and the Gaussian kernel uses the optimal bandwidth. Statistically, the MSE value generated by the Epanechnikov kernel is almost close to the value in the Gaussian kernel, so it can be said that the MSE value produced by the two kernel functions is almost the same. Based on the plot of estimation results for the Eponocvh kernel function and the Gaussian kernel using the optimal bandwidth, it is very close, so it can be said that the use of a different kernel function with the optimal bandwidth for each of the kernel functions will produce the same estimated regression curve. The results of this study support the opinion expressed by Hastie and Tibshirani, which states that in kernel regression, the selection of the smoothing parameter (bandwidth) is much more important than choosing the kernel function.

Keywords: Kernel Regression, Epanechnikov, Gaussian

1. INTRODUCTION

Regression analysis is a data analysis method that describes the relationship between the response variable and one or more predictor variables. Suppose X is the predictor variable and Y is the response variable to observe an {x,y} relationship, then the linear relationship between the predictor variables and the response variable can be expressed as follows:

$$Y_i = (X_i) + \varepsilon_i, \quad i=1, 2, 3 ,\ldots,n , \ldots\ldots\ldots(1)$$

Where (i) is the assumed independent residual with zero mean and variance _(i) and m(X_i) is the regression function or regression curve. Hardle (1990:4) revealed that to estimate m(X_(i,)) there are two approaches that can be used in determining the regression curve, namely the parametric regression approach and the nonparametric regression approach.

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

The parametric regression approach assumes the form of the relationship between the response variable and the predictor variable is known or estimated from the regression curve. Hardle (1990:6) revealed that the nonparametric regression approach used to estimate the regression curve has several main objectives, namely providing a method for relating two variables in general, producing predictions from observations even though they are made without references, and being a flexible method for substituting values. -missing values between adjacent predictor variables. The nonparametric regression approach is the appropriate regression approach for data patterns whose shape is unknown or there is no past information about data patterns (Budiantara, 2010).

In the journal Halim and Bisono (2006) entitled Kernel Functions in Nonparametric Regression Methods and Its Application to Priest River Experimental Forest's Data, it is concluded that if the assumption of a parametric model is justified, then the regression function can be estimated in a more efficient way. When compared to using a nonparametric method, if the assumption of the parametric model is wrong, then the results will give the wrong conclusion to the regression function. Sukarsa (2012), in his journal entitled kernel regression in nonparametric regression models, revealed that kernel regression is a nonparametric statistical technique for estimating the value of E(YIX)=m(X) or y=m(X) in a variable. The purpose of kernel regression is to obtain a nonlinear relationship between X and Y.

In nonparametric regression, the data will look for its own estimation form without being influenced by the subjectivity of the researcher, so the nonparametric regression approach has high flexibility, Eubank (1988). Budiantara (2010) revealed that there are several techniques for estimating the regression curve in nonparametric regression, namely kernel, histogram, spline, Fourier series, wavelets, and orthogonal. One of the nonparametric regression approaches used in this research is kernel regression.

Kernel regression is a nonparametric analysis with a smoothing method. Smoothing has become synonymous with nonparametric methods used to estimate functions. The purpose of smoothing is to remove variability from data that has no effect so that the characteristics of the data will appear clear. Kernel regression has a flexible form, and the mathematical calculations are easy to adjust. In kernel regression, an estimator is known, which is usually used to estimate the regression function, namely the Nadaraya-Watson estimator.

Estimation with kernel approach depends on two parameters, namely bandwidth and kernel function. There are seven kernel functions, namely Uniform, Triangle, Epanechnicov, Quartic, Triweight, Gaussian, and Cosinics. Among the seven kernel functions in this study, the Gaussian kernel function was chosen. Meanwhile, bandwidth is a smoothing parameter that functions to control the smoothness of the estimated curve. Bandwidth that is too small will cause the estimated function to be very rough so that the variance relationship is high and has a low potential for bias. On the other hand, if the bandwidth is too large, the estimated function becomes very smooth so that the variance relationship is low and has a large potential for bias. Therefore, it is necessary to choose the optimal bandwidth.

Optimal bandwidth selection is made by minimizing the error rate. The smaller the error rate, the better the estimate. To determine size of the error rate of an estimator can be seen from the Mean Squared Error (MSE). The bandwidth used in Guidom (2015) is the Bandwidth Rule of Thumb, Unbiased Cross Validation, Biased Cross Validation, and Complete Cross-Validation. From the four bandwidths, the bandwidth that has the smallest MSE value will be chosen.

The Gaussian kernel nonparametric regression with the Nadaraya-Watson estimator in time series data can use data on life expectancy, average length of schooling, and human development index.

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

By looking at the conditions above, the author will discuss how to estimate the data using Gaussian kernel nonparametric regression with the Nadaraya-Watson estimator, and the bandwidth selection methods are "Rule of Thumb" bandwidth, Unbiased Cross Validation, Biased Cross Validation, and Complete Cross-Validation.

Based on the background above, the problem can be formulated as follows:

1. How does the Nadaraya-Watson estimator analyze the Gaussian kernel type?
2. How is the selection of bandwidth on the Rule of Thumb, Unbiased Cross Validation, Biased Cross Validation, and Complete Cross Validation?
3. What is the estimation result after selecting the bandwidth?

2. LITERATURE REVIEW

1. Regression analysis

Regression analysis is one of the most widely used data analysis techniques in statistics to determine the pattern of relationships between response variables to one or several predictor variables. Regression was first introduced in 1886 by an expert named Francis Galton. According to Galton, regression analysis is concerned with the study of the dependence of a variable called the dependent variable on one or the explaining variable with the aim of estimating or predicting the values of the dependent variable if the value of the explaining variable is known. Variables that explain are often called independent variables. The relationship between the two variables can be described by a regression curve with a certain form of function. There are two types of approaches that can be used to estimate the regression curve, namely: **(**a) Parametric Regression Approach and (b) Nonparametric Regression Approach. Parametric regression is used to estimate the form of the relationship between the response variable and one or more predictor variables where the shape of the regression curve is known. In parametric regression, there are assumptions that must be met. Nonparametric regression is a statistical method used.

2. Kernel Regression

Kernel regression is an estimation technique according to the available data. Given a data set, we want to find a regression function, such as the function that best fits the data held at the data points. You may also want to interpolate and estimate a bit beyond the data. The idea of kernel regression is to assign a set of identical weighted functions called local kernels to each observation data point. The kernel will assign a weight to each location based on the distance from the data point.

3. Kernel function

The kernel function denoted K(u) is a continuous, symmetrical, finite function, and

$$\int_{-\infty}^{\infty} \boldsymbol{K(u)du = 1} \text{.....(2)}$$

According to Halim and Disono (2006:75), there are three kinds of kernel estimators, namely: the Nadaraya Watson Estimate.

According to Hardle (1991), if there are n observational data $\{(Xi, Yi)\}^{n}{}_{i=1}$, which satisfies equation (2.1) where $X_i \epsilon R$ dan $Y_i \epsilon R$, then the estimator of m(x) is:

$$\hat{m}(x) = E(Y|X = x) = \int \frac{yf(X,Y)}{} dy \text{ .....(3)}$$

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

$(X = x)$..... (4)

The denominator is estimated using the kernel density estimator.

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_n(x - X_i) \quad \text{.....(5)}$$

So, the numerator of the Nadaraya-Watson estimator becomes:

$$\int y\hat{f}_{h1,h2}(x,y)dy = n^{-1}\sum_{i=1}^{n} K_{h1}(x - X_i)\int yK_{h2}(y - Y_i)dy$$

$$= n^{-1}\sum_{i=1}^{n} K_{h1}(x - X_i)\int \frac{y}{h_2}K\left(\frac{y - Y_i}{h_2}\right)dy$$

$$= n^{-1}\sum_{i=1}^{n} K_{h1}(x - X_i)\int (sh_2 + Y_i)K(s)\,ds$$

$$= n^{-1}\sum_{i=1}^{n} K_{h1}(x - X_i)\,Y_i$$

The form of the Nadaraya-Watson estimator can be written:

$$\hat{m}(x_i) = \frac{\frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j)y_j}{\frac{1}{n}\sum_{k=1}^{n} K_h(x - x_k)}$$

$$\hat{m}(x_i) = \frac{\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x_i - X_j}{h}\right)y_j}{\frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x_i - X_k}{h}\right)}$$

$$\hat{m}(x_i) = \frac{\sum_{j=1}^{n} K\left(\frac{x_i - X_j}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x_i - X_k}{h}\right)}Yj$$

$$\hat{m}(x) = \sum_{j=1}^{n} w_{ij}(x_i)yj$$

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

So, $\hat{Y} = WY$, Where

$$w_{hi}(x) = \frac{K\left(\frac{x - x_j}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_k}{h}\right)}$$

The W matrix is also called the Hat Matrix of the m(x) estimator. Equation (2.9) was found by Nadaraya and Watson (1964), so it is called the Nadaraya-Watson estimator.

4. Various kernel functions

- Kernel Uniform: $K^{(x)} = \frac{1}{2} I(|x| \leq 1)$
- Kernel Segitiga (Triangel) : $K(x) = (1 - |x|) I(|x| \leq 1)$
- Kernel Epanechnikov : $K^{(x)} = \frac{3}{4}(1 - x^2) I(|x| \leq 1)$
- Kernel Kuadrat (Quartik) : $K^{(x)} = \frac{15}{16}(1 - x^2)^2 I(|x| \leq 1)$
- Kernel Triweight: $K^{(x)} = \frac{35}{32}(1 - x^2)^3 I(|x| \leq 1)$
- Kernel Gaussian: $K^{(x)} = \frac{1}{\sqrt{2\pi}} exp\left(\frac{1}{2}(-x^2)\right) - \infty < x < \infty$
- Kernel Cosinus: $K^{(x)} = \frac{\pi}{4} cos\left(\frac{\pi}{2} x\right) I(|x| \leq 1)$
- Kernel Tricube: $K^{(x)} = \frac{70}{81}(1 - |x|^3)^3 I(|x| \leq 1)$

5. Life Expectancy

Health is one of the factors that cause poverty. Various health indicators in low- and middle-income countries compared to high-income countries show that morbidity and mortality rates are strongly correlated. Several reasons for the increasing burden of disease on the poor are. First, the poor are more susceptible to disease due to limited access to clean water and sanitation, as well as adequate nutrition. Second, the population is increasingly inclined not to seek treatment even though it is in dire need due to the large gap with health workers, limited resources to meet basic needs, and limited knowledge to deal with disease attacks in the thesis (Tessa, 2017). An attack of a non-fatal disease in early life will have a detrimental effect during the next life cycle. Education is widely recognized as the key to development, but it is not yet appreciated how important health is in achieving educational outcomes. Poor health directly reduces cognitive potential and indirectly reduces school ability. Illness can impoverish families through decreased income, decreased life expectancy, and decreased psychological well-being in the thesis (Tessa, 2017). Life expectancy is the estimated average number of years a person can live during a lifetime.

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

6. Average Length of School

Todaro (2000) states that education is a fundamental development goal. Where education plays a key role in shaping a country's ability to absorb modern technology and to develop capacity for sustainable growth and development. The average length of schooling indicates the higher the level of formal education achieved by the people of an area. The higher the average length of the school, the higher the level of education undertaken. The average length of schooling is the average number of years spent by the population aged 25 years and over at all levels of formal education followed in the thesis (Widiatma, 2012). According to Todaro (2000), this level of income is strongly influenced by the length of time a person receives education. The average length of schooling is an indicator of the level of education in an area. Education is a form of human capital that shows the quality of Human Resources (HR). To be able to maximize the difference between the expected benefits and the estimated costs, the optimal strategy for a person is to try to complete education as high as possible. Investment in human capital will be seen as having higher benefits if we compare the total cost of education spent during their education to the income that will be obtained when they are ready to work. People with higher education will start full-time work at an older age, but their income will rise faster than people who work earlier (Todaro, 2000).

7. Human Development Index (PMI)

According to Susianti (2012), the Human Development Index is a measure of the success of development in a country in the development process by looking at the level of income, health, and education. To achieve the success of a country in development, the country must improve the quality of human development along with several human development factors, namely:

a) Life Expectancy is the average estimated age of a person during life. Life expectancy is calculated using an indirect approach (Indirect Estimation).

b) To measure knowledge, two indicators are used, namely, the average years of schooling (mean years of schooling) taken and the literacy rate. The average length of schooling describes the number of years used by the population in undergoing formal education.

c) A decent standard of living is a standard of living that describes the level of welfare of the population as a result of the improving economy.

d) The diagram for calculating the Human Development Index (HDI) is as follows: the new growth theory emphasizes the importance of the government's role, especially in increasing human capital development and encouraging research and development to shorten human productivity. In fact, it can be seen that investing in education will be able to increase human resources, which is shown by increasing one's knowledge and skills. The higher a person's education level, the more knowledge and skills will also increase, which will encourage an increase in work productivity. The quality of work inputs or human resources is the most important factor for economic success. Almost all other factors of production, namely capital goods, raw materials, and technology, can be purchased or borrowed from other countries. But the application of high productivity techniques to local conditions almost invariably demands the availability of management, productivity skills, and expertise that can only be acquired through an educated, skilled workforce.

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

3. RESEARCH METHOD

1) Sources of data and research variables

The source of data in this case study is secondary data. Secondary data is data obtained or collected from sources that already exist and are reliable (Hassan, 2002). This study uses data from the Central Sulawesi Statistics Agency. The data taken are life expectancy, the average length of schooling, and the human development index. The variables in this study are independent variables and dependent variables. The independent variable is a variable that affects other variables, and in this case, the independent variables are life expectancy (X1) and the average length of schooling (X2). While the dependent variable is a variable that is influenced by other variables, and in this case, the dependent variable is the human development index (Y).

2) Sampling Technique

The sampling technique used was the cluster sampling technique (regional sampling technique). The regional sampling technique is used to determine the sample if the object to be studied or the data source is very broad, for example, the population of a country, province, or district. To determine which population will be used as a data source, the sampling is based on a predetermined population area (Sugiyono, 2002).

3) Data Analysis

This test uses R software. The steps for data analysis are as follows:

- Describing the data
- Selection of optimal bandwidth of gaussian and epanechnikov kernel functions
- Comparing the gaussian and epanechnikov functions.

4. RESULTS AND DISCUSSION

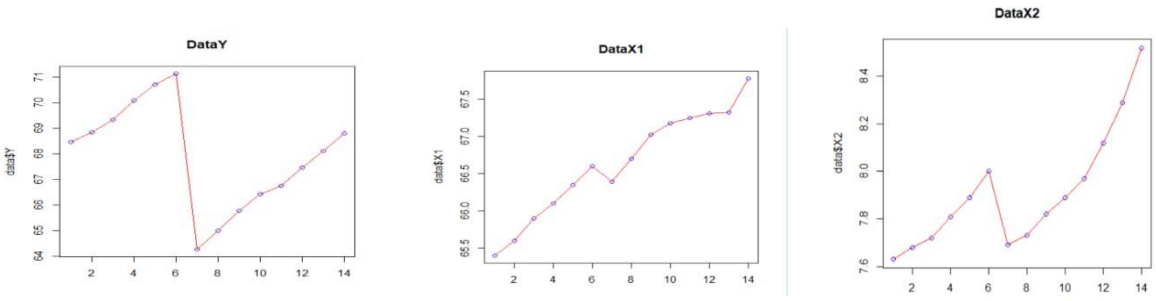

Fig.1. scatter plot Data Y, X1, and X2 Interpretation:

- Variable Y (HDI)

  The Human Development Index data in Central Sulawesi Province from 2005-2018 experienced a significant decline from 2010 to 2011. This occurred because of the change in the calculation of the Human Development Index data from the old method to the new method.

- Variable X1(Life Expectancy)

  From the picture above, it can be seen that life expectancy from 2005-2018 experienced an increasing trend, but from 2010 to 2011, it experienced an insignificant decrease.

- Variable X2(average Length of School)

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

From the picture above, it can be seen that the average length of schooling from 2005-2018 experienced an increasing trend but from 2010 to 2011 experienced an insignificant decrease from 8 to 7.69

4.1 Optimum Bandwidth Selection

Table 1. Optimal Bandwidth

Variable	CV Method Type	Bandwidth	R-Square	MSE
X1	CV.AIC	4783847	0.034079	2.45
	CV.LS	0.1250295	0.999909	1.91
X2	CV.AIC	1699106	0.034079	2.34
	CV.LS	0.03748174	0.999909	2.08

See the optimal bandwidth, it can be done by looking at the smallest MSE value. From the table, it can be seen that the smallest MSE value is the CV.LS method, so the best model in this study is CV.LS with a bandwidth of 0.1250295.

4.2 Prediction Results for Each Bandwidth

Prediction results with the CV.A method

```
> yhat_1=predict(model.np)
> yhat_1
 [1] 67.9439 67.9439 67.9439 67.9439 67.9439 67.9439 67.9439 67.9439 67.9439
[10] 67.9439 67.9439 67.9439 67.9439 67.9439
```

Prediction results with the CV.A method

```
> yhat_2=predict(model.np)
> yhat_2
 [1] 68.50529 68.82796 69.34197 70.07831 70.69250 71.13931 64.29232 64.98117
 [9] 65.83036 66.42037 66.72669 67.46978 68.10998 68.80000
```

4.3 Comparison of Gaussian and Epanechnikov . functions

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

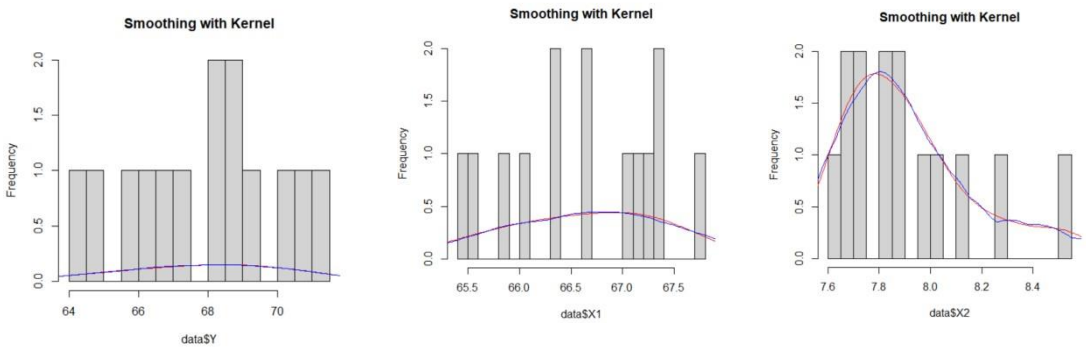

Fig.2. Smoothing with Kernel

Table 2. Bandwidth Value Kernel

Variable	Kernel Functions	Bandwidth	MSE
Y	Gaussian	1.419764	-
	Epanechnikov	3.143079	-
X1	Gaussian	0.454393	2.49
	Epanechnikov	1.005937	2.49
X2	Gaussian	0.117839	2.3
	Epanechnikov	0.2608726	2.3

The table above shows the MSE value generated by the Epanechnikov kernel function and the Gaussian kernel using the optimal bandwidth. Statistically, the MSE value generated by the Epanechnikov kernel is almost close to the value in the Gaussian kernel, so it can be said that the MSE value produced by the two kernel functions is almost the same.

Based on the plot of estimation results for the Eponocvh kernel function and the Gaussian kernel using the optimal bandwidth, it is very close, so it can be said that the use of a different kernel function with the optimal bandwidth for each of the kernel functions will produce the same estimated regression curve. The results of this study support the opinion expressed by Hastie and Tibshirani, which states that in kernel regression, the selection of the smoothing parameter (bandwidth) is much more important than choosing the kernel function.

5. CONCLUSION

Based on the results and discussion, it can be concluded that in kernel regression, the most important thing is the selection of the optimal bandwidth value, not the selection of the kernel function, because the use of a different kernel function with the optimal bandwidth value produces a regression curve estimate that is almost the same. This is in accordance with the opinion expressed by Hastie and Tibshirani 4, namely, in kernel regression, the selection of the smoothing parameter (bandwidth) is much more important than choosing the kernel function.

E-ISSN : 2776-4885 || P-ISSN : 2776-4877

6. REFERENCES

Buxton, A. and P. Greenhalg. (1989). Ramie, Short Live Curiosity or

Fiber. The Future Textile Outlook International, May 1989. The Economist Intelligence Unit, London.

https://jateng.bps.go.id/subject/26/index-pembangun-human.html

BPS, 2011. North Sumatra in Figures 2011, Medan: Central Statistics Agency North Sumatra.

Sugiyono, Administrative Research Methods, (Bandung: CV Alfabeta, 2002), p1m.57.

M. Iqbal Hasan, 2002. Main Materials of Research Methodology and Its Applications. Jakarta, Indonesia Ghalia Publisher

E-ISSN : 2776-4885 || P-ISSN : 2776-4877