Integral Estimator with Kernel Approach for Estimating Nonparametric Regression Functions

Rahmat Hidayat*¹, Ma’rufi¹, and Yuliani²
¹Master of Mathematics Education Program, Cokroaminoto Palopo University, Palopo, Indonesia.
²Department of Mathematics, Cokroaminoto Palopo University, Palopo, Indonesia.

*Email: dayatmath@gmail.com

Abstract. Derivatives are measurements of how a function change as the input value changes, or in general a derivative shows how one quantity changes due to a change in another quantity. The concept of universal or comprehensive function derivatives is widely used in various scientific fields. For example, in economics, people are interested in studying the condition of the derivative of an objective function as the result of an optimization problem. In this study, nonparametric procedures are used to estimate a function where the form of the function does not lead to a particular function model. Suppose we are given a nonparametric regression model where \( f \) is an unknown function. The main problem of regression analysis is to determine the form of estimation \( f \). To determine the estimation of \( f \), one approach that can be used is the integral estimator with the Gaussian Kernel approach. Furthermore, as an application, the Labour Force Participation Rate \((y)\) data is used with the predictor variable, namely the Average Length of Schooling \((x)\). By using the GCV (Generalized Cross Validation) method, the optimal bandwidth is obtained at \( h = 80 \) with a GCV value of 0.243 with an MSE value of 32.1864.

Keywords: nonparametric regression; integral estimator; Kernel; Labour Force Participation Rate

1. Introduction

In most cases, one is only interested in observing the application of the derivative of a function. For example, in the field of economics, people are only interested in studying the condition of the derivative of an objective function as the result of an optimization problem. For example, the problem of the cost function of water pollution depreciation in the French industry, with constraints namely gross pollution, abatement, and net pollution. The company will maximize profit against the above constraints. So, the condition of the first derivative is an optimization problem [1].

In that case, the condition resulting from this optimization behaviour is the derivative condition while the function of interest is the integral. Usually, economists postulate a parametric form to represent the boundary of a condition, then use statistical methods to estimate the unknown parameters from the observed data. However, since there is neither prior data nor information on the objective function of a water agency that monitors water pollution in France, there is no reason to assign a parametric form to the first-derived conditions to be estimated. Therefore, in this case it is not suitable to use the derivative conditions to estimate an economic function, but to use a nonparametric method with integration techniques in estimating the function [2]. The real purpose of nonparametric methods is to estimate functions without specifying functional forms.
The main problem of nonparametric regression analysis is determining the form of estimation $f$. To determine the estimation of $f$, one approach that can be used is the Kernel approach. In this study, a Kernel method is proposed which is the most popular method, mainly because its properties are well understood since it was defined by Nadaraya and Watson [1,3].

Another important problem related to the integral utility problem, is the estimation of the derivative of a function as well as the function itself. The first derivative is often used, for example in regression analysis, to measure the response coefficient of the dependent variable with respect to the predictor. The second derivative indicates the convexity of a function and is useful for testing certain form hypotheses. In 1989, Mack and Muller proposed a simple Kernel estimator that allows an easy derivation and developed an integral estimator as well as the derivative estimator. The problem of selecting bandwidth using the Cross Validation (CV) and Generalized Cross Validation (GCV) methods has also received great attention in research [4]. In this paper, we propose an integral estimator with a Kernel approach for estimating regression functions, which was developed on the Mack Muller Kernel integral estimator.

2. Nonparametric regression

Suppose there are as many as $q = p + 1$ random variable $(X, Y)$ where $Y$ is the dependent variable and $X$ is the vector $p \times 1$ is the predictor variable. If $E(|Y|)$ is finite, then the nonparametric regression model can be expressed as follows:

$$Y = f(x) + \varepsilon$$

where $f(x) = E[Y|X = x]$ and $\varepsilon$ is an error. If $\varphi(x, y) = \varphi(x_1, x_2, ..., x_p, y)$ is the unknown joint density function of $(X, Y)$ at the point $(x, y)$, and $\varphi(x)$ is the marginal density function of $X$ at $x$, then the function $f(x)$ can be written as:

$$f(x) = E[Y|X = x] = \int y\varphi(y|X = x)dy = \int y\frac{\varphi(x, y)}{\varphi(x)}dy = \frac{g(x)}{\varphi(x)}.$$  \hfill (2)

Next, take a random sample of size $n$ from the predictor variable $X$ and the response variable $Y$ so that $(X_i, Y_i)$ will be i.i.d and the relationship between the two variables is stated as follows:

$$Y_i = f(X_i) + \varepsilon_i$$

where $f$ is the unknown function. There are two approaches to estimating $f$ namely the parametric approach and the nonparametric approach. The parametric approach is used if there is previous information about the form of $f$ whether it is linear or nonlinear. The information is obtained based on theory or past experience. While the nonparametric approach is used if there is no information about the shape of the regression curve [5].

3. Kernel estimator

Kernel estimators were introduced by [6] and [7]. Rosenblatt gives weight to each observation. According to [1,8] the Kernel density estimator has several advantages, namely:

a. The Kernel density estimator has a flexible form and is relatively easy to work with mathematically.

b. The Kernel density estimator has a relatively fast average convergence.

In general, the Kernel function for one-dimensional (one-dimensional) according to [1] and [7], is defined as follows.

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right); \quad -\infty < u < \infty$$ \hfill (4)

The Kernel $K(\bullet)$ is a continuous, finite, and symmetric real function with an integral equal to one, $\int K(u)du = 1$. From this definition, if $K$ is a non-negative function, then $K$ is also interpreted as a function of solid chance (density function) [9].

In Kernel estimation, Kernel shape is affected by Kernel function $K(\bullet)$ and bandwidth $h$. The most important issue associated with using kernel density estimates is choosing the optimal bandwidth. The
optimal value of \( h \) depends on the criteria used to measure the overall accuracy of \( f \). The criterion commonly used is Mean Square Error (MSE) [10].

4. Integral estimator
The Kernel approach introduced by [11,12,13] to estimate the density function \( f(x) \) is a multivariate Kernel estimator, namely:

\[
f_\lambda(x) = \frac{1}{n} \sum_{i=1}^{n} K_i(x - X_i) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)
\]

(5)

where \( K(.) \) is the Kernel function and \( h \) is the bandwidth.

Most classes of nonparametric estimators have been defined as a weighted sum of the dependent variables as well as the known Nadaraya-Watson Kernel estimators, namely:

\[
f(x) = \sum_{i=1}^{n} Y_i W_{xy}(x, X_i, h_n) = \frac{1}{n h_n^p} \sum_{i=1}^{n} Y_i \frac{1}{n h_n^p} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)
\]

(6)

where \( p \) is the order bandwidth \( h_n \), and \( K(\ast) \) is the Kernel function and \( h_n \) is the bandwidth, such that \( h_n \to 0 \) and \( n h_n \to \infty \) for \( n \to \infty \). Therefore, depending on the value of the bandwidth \( h_n \) where for very small \( h_n \) the estimator will go to the data and for large \( h_n \) the estimator will lead to the average of the response variables.

A modified version of the Kernel estimator above where the denominator of (6) is an estimate of the marginal density at a fixed point \( x \). Furthermore, if evaluated on the sample value \( X_i \), for \( i = 1, 2, \ldots, n \), it is an estimate of the marginal density function. The estimation of the marginal density function is as follows:

\[
\phi(X_i) = \frac{1}{n h_n^p} \sum_{i=1}^{n} K\left(\frac{X_j - X_i}{h_n}\right)
\]

(7)

So the Mack Muller Kernel estimator can be written as:

\[
f_{MM}(x) = \frac{1}{n h_n^p} \sum_{i=1}^{n} Y_i \frac{1}{n h_n^p} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) = \frac{1}{n h_n^p} \sum_{i=1}^{n} Y_i \frac{1}{n h_n^p} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)
\]

(8)

Then, Mack and Muller found an integral estimator as well as its derivative estimator for nonparametric regression functions known as the Mack Muller Kernel approximation integral estimator. Suppose an estimator function is an integral of the unknown function \( f \) is an integral of the estimate defined as follows:

\[
F(v) = \int_{-\infty}^{v} f(x) \, dx
\]

(9)

using the Mack and Muller estimator it becomes:

\[
F(v) = \frac{1}{n h_n^p} \sum_{i=1}^{n} Y_i \int_{-\infty}^{v} K\left(\frac{x - X_i}{h_n}\right) \, dx
\]

\[
= \frac{1}{n h_n^p} \sum_{i=1}^{n} Y_i \int_{-\infty}^{v} K\left(\frac{x - X_i}{h_n}\right) \, dx
\]

(10)

The integral of the estimate in (9) above includes all of its components, for \( s = 1, 2, \ldots, p \), is an integral of the estimate \( f \) which is defined as follows:
\[ F(x_1, \ldots, v, \ldots, x_p) = \int_{-\infty}^{v} f(x) \, dx \]  

or

\[ F(x_1, \ldots, v, \ldots, x_p) = \frac{1}{n h_n^p} \sum_{i=1}^{n} \frac{Y_i - \bar{X}_i}{\varphi(X_i)} K \left( \frac{x - \bar{X}_i}{h_n} \right) \, dx_i \]

\[ = \frac{1}{n h_n^p} \sum_{i=1}^{n} \frac{Y_i}{\varphi(X_i)} \int_{-\infty}^{v} K \left( \frac{x - \bar{X}_i}{h_n} \right) \, dx_i \]  

(12)

5. Bandwidth selection

The most important issue associated with using Kernel density estimates is choosing the optimal bandwidth [14, 15]. The optimal value of \( h \) depends on the criteria used to measure the overall accuracy of. One of the criteria commonly used is the Mean Square Error (MSE). Optimal bandwidth is obtained by minimizing MSE. MSE is defined as follows:

\[ \text{MSE}(h) = n^{-1} \sum_{i=1}^{n} \left( f_n(x_i) - f(x_i) \right)^2 \]  

(13)

The GCV method is defined as follows:

\[ \text{GCV}(h) = n^{-1} \sum_{i=1}^{n} \left( \frac{1 - \text{trace} H(h)}{1 - \frac{1}{n} \text{trace} H(h)} \right)^2 \left( \frac{y_i - f_n(x_i)}{1 - H(h)} \right)^2 \]  

(14)

\[ \text{GCV}(h) = \frac{n^{-1} \sum_{i=1}^{n} \left( y_i - f_n(x) \right)^2}{\left( 1 - \frac{1}{n} \text{trace} H(h) \right)^2} \]  

(15)

Optimal bandwidth is obtained from the bandwidth that produces the smallest GCV value.

\[ h_{opt} = \text{GCV}(h). \]

6. Result and discussion

Suppose that a random sample of size \( n \) is taken from the predictor variable \( X \) and the response variable \( Y \) the relationship between the two variables is stated as follows:

\[ Y_i = f \left( X_i \right) + \epsilon_i \]  

(16)

where \( f \) is the unknown function.

To determine the estimation of \( f \), one approach that can be used to estimate the density function \( f(x) \) with Kernel \( K \) and bandwidth \( h \) is the Mack Muller Kernel estimator, namely:

\[ f_{MM}(x) = \frac{1}{n h_n^p} \sum_{i=1}^{n} Y_i K \left( \frac{x - \bar{X}_i}{h_n} \right) \]  

(17)

where the Kernel function \( K(\cdot) \) is polynomial, continuous, finite, and symmetric.

In this paper, as explained in the literature review, the form of the proposed Kernel estimator [10] to estimate the regression curve \( f \) is:

\[ F(v) = \frac{1}{n h_n^p} \sum_{i=1}^{n} Y_i \int_{-\infty}^{v} K \left( \frac{x - \bar{X}_i}{h_n} \right) \, dx \]  

(18)

Where \( \varphi(X_i) \) is the estimator of the marginal density function of the vector \( x \) evaluated at the sample value \( X_i \) for \( i = 1, 2, \ldots, n \). Hereinafter \( F(v) \) known as the Kernel approximation integral estimator.
6.1. Integral estimator

Theoretically the regression function is defined as follows:

\[
f(x) = E[Y|X = x] = \frac{\int y \varphi(x, y)dy}{\varphi(x)} \tag{19}
\]

for \( \varphi(x) \neq 0 \). The denominator is estimated by the marginal density of \( X \) at the value of \( x \). If evaluated on the sample value \( X_i, i = 1, 2, ..., n \), then it is an estimate of the marginal density function \( \varphi(X_i) \). Meanwhile, the numerator can be estimated by estimating the density function together \( \varphi(x, y) \) using Kernel multiplication.

For example, given \( n \) samples of observations \((X_i, Y_i)\) and \( X_i = (X_i^{(1)}, X_i^{(2)}) \), if \( p = 2 \) then a bivariate Kernel will be used, namely \( K\left(\frac{x^{(1)} - X_i^{(1)}}{h_1}, \frac{x^{(2)} - X_i^{(2)}}{h_2}\right) \) which is defined as the result of 2 univariate Kernels with two different bandwidths \( h \) namely \( h_1 \) and \( h_2 \). The bivariate density \( \varphi(x, y) \) at (19) is unknown but can be estimated by multiplying the bivariate Kernel estimator with the Kernel function \( K(\cdot) \) and the bandwidth \( h \) as follows:

\[
\varphi(x, y) = \frac{1}{n h_1 h_2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_1}\right) K\left(\frac{y - Y_i}{h_2}\right) = \frac{1}{n} \sum_{i=1}^{n} K_{h_1}(x - X_i) K_{h_2}(y - Y_i) \tag{20}
\]

Therefore, the numerator of (19) becomes,

\[
\int y \varphi(x, y)dy = \frac{1}{n} \sum_{i=1}^{n} K_{h_1}(x^{(1)} - X_i^{(1)}) K_{h_2}(x^{(2)} - X_i^{(2)}) K_{h_2}(y - Y_i) \ dy
\]

\[
f(x^{(1)}, x^{(2)}) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{\varphi(X_i)} K_{h_1}(x^{(1)} - X_i^{(1)}) K_{h_2}(x^{(2)} - X_i^{(2)}) \tag{21}
\]

next obtained

\[
\varphi(X_i) = \varphi(X_i^{(1)}) \varphi(X_i^{(2)}) = \frac{1}{n} \sum_{j=1}^{n} K_{h_1}(X_j^{(1)} - X_i^{(1)}) K_{h_2}(X_j^{(2)} - X_i^{(2)})
\]

So that it is obtained,

\[
f(x^{(1)}, x^{(2)}) = -\frac{1}{n h_1 h_2} \sum_{i=1}^{n} Y_i \frac{1}{\varphi(X_i^{(1)}) \varphi(X_i^{(2)})} K\left(\frac{x^{(1)} - X_i^{(1)}}{h_1}\right) K\left(\frac{x^{(2)} - X_i^{(2)}}{h_2}\right) \tag{22}
\]

6.2. Determining \( \hat{f} \) with the gaussian kernel approach.

The form of the given Gaussian Kernel is:

\[
K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right), -\infty < u < \infty. \tag{23}
\]

So that,

\[
K_{h_1}(x^{(1)} - X_i^{(1)}) = \frac{1}{h_1 \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x^{(1)} - X_i^{(1)}}{h_1}\right)^2\right), K_{h_2}(x^{(2)} - X_i^{(2)}) = \frac{1}{h_2 \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x^{(2)} - X_i^{(2)}}{h_2}\right)^2\right)
\]

Therefore, based on (22) the equation becomes,

\[
f(x^{(1)}, x^{(2)}) = \frac{1}{2\pi n h_1 h_2} \sum_{i=1}^{n} \frac{Y_i}{\varphi(X_i^{(1)}) \varphi(X_i^{(2)})} \exp\left(\frac{1}{2} \left(\frac{x^{(1)} - X_i^{(1)}}{h_1}\right)^2 + \left(\frac{x^{(2)} - X_i^{(2)}}{h_2}\right)^2\right) \tag{24}
\]
Thus obtained as follows:

\[
f(x^{(1)}, x^{(2)}) = \frac{1}{2\pi n h_1 h_2} \sum_{i=1}^{n} \frac{Y_i}{\phi(X_i^{(1)}) \phi(X_i^{(2)})} \exp \left\{ -\frac{1}{2} \left( \frac{x^{(1)} - X_i^{(1)}}{h_1} \right)^2 + \frac{1}{2} \left( \frac{x^{(2)} - X_i^{(2)}}{h_2} \right)^2 \right\}
\]  

(25)

6.3. Getting the integral estimator \( \hat{F} \)

In this section we will show an integral estimator using the Gaussian Kernel approach. As discussed in the literature review, the general form of an integral estimator is,

\[
F(x_1,...,x_s,...,x_p) = \int_{-\infty}^{\infty} f(x)dx,
\]

(26)

Using (26) \( F \) is obtained as follows:

\[
F(u^{(1)}, v^{(2)}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi n h_1 h_2} \sum_{i=1}^{n} \frac{Y_i}{\phi(X_i^{(1)}) \phi(X_i^{(2)})} \exp \left\{ -\frac{1}{2} \left( \frac{x^{(1)} - X_i^{(1)}}{h_1} \right)^2 + \frac{1}{2} \left( \frac{x^{(2)} - X_i^{(2)}}{h_2} \right)^2 \right\} dx^{(2)} dx^{(1)}
\]

\[
= \frac{1}{2\pi n h_1 h_2} \sum_{i=1}^{n} \frac{Y_i}{\phi(X_i^{(1)}) \phi(X_i^{(2)})} \left[ \int_{-\infty}^{u^{(1)}} \exp \left\{ -\frac{1}{2} \left( \frac{x^{(1)} - X_i^{(1)}}{h_1} \right)^2 \right\} dx^{(1)} \right] \left[ \int_{-\infty}^{v^{(2)}} \exp \left\{ -\frac{1}{2} \left( \frac{x^{(2)} - X_i^{(2)}}{h_2} \right)^2 \right\} dx^{(2)} \right]
\]

(27)

7. Modelling the labour force participation rate

Before estimating the model, we will first investigate the pattern of the relationship between the response variable (labour force participation rate) and the predictor variable (mean length of schooling) through a scatter plot.

![Figure 1. The relationship between the average length of schooling and the labour force participation rate.](image-url)
participation rate

The pattern of the relationship between the Average Length of Schooling and the Labor Force Participation Rate (LFPR) is shown in Figure 1, where a low average length of schooling tends to result in a low LFPR, but a high average length of schooling also tends to produce a low LFPR. So, in general it can be said that the pattern of the relationship between the average length of schooling and the LFPR tends to be irregular. For this reason, the use of the average length of schooling and LFPR is modelled nonparametrically.

The most important issue associated with using Kernel density estimates is choosing the optimal bandwidth. Furthermore, the optimal bandwidth $h$ is needed which is obtained by using the GCV method. From the results of running the GCV program, the following results are obtained:

| $h$  | GCV  |
|------|------|
| 10   | 0.397|
| 20   | 0.353|
| 30   | 0.332|
| 40   | 0.298|
| 50   | 0.285|
| 60   | 0.279|
| 70   | 0.250|
| 80   | 0.243|
| 90   | 0.299|
| 100  | 0.311|
| 110  | 0.452|

From the table above, it is obtained that a small GCV of 0.243 at $h = 80$ is the optimal bandwidth. The plot between $h$ and GCV is

![Figure 2. Curve shape estimation](image)

The $R^2$ and MSE values of this model are 84.71 and 32.1864, respectively. This means that 84.71 percent of the variation in LFPR can be explained by the average length of schooling.
8. Conclusion
Based on the results and discussions that have been described in the previous chapter, several conclusions can be drawn, including:

1. The form of the estimation of \( f \) obtained through the Gaussian Kernel approach to the nonparametric regression function is:

\[
f(x^{(1)}, x^{(2)}) = \frac{1}{2\pi n h_1 h_2} \sum_{i=1}^{n} \frac{Y_i}{\phi(X_i^{(1)}) \phi(X_i^{(2)})} \exp \left( -\frac{1}{2} \left( \frac{x^{(1)} - X_i^{(1)}}{h_1} \right)^2 + \left( \frac{x^{(2)} - X_i^{(2)}}{h_2} \right)^2 \right)
\]

While the integral estimator \( f \) is

\[
F(u^{(1)}, v^{(2)}) = \int_{-\infty}^{u^{(1)}} \int_{-\infty}^{v^{(2)}} \frac{1}{2\pi n h_1 h_2} \sum_{i=1}^{n} \frac{Y_i}{\phi(X_i^{(1)}) \phi(X_i^{(2)})} \exp \left( -\frac{1}{2} \left( \frac{x^{(1)} - X_i^{(1)}}{h_1} \right)^2 + \left( \frac{x^{(2)} - X_i^{(2)}}{h_2} \right)^2 \right) dx^{(2)} dx^{(1)}
\]

2. Based on a case study with labor force participation rate data as a response variable, to the average length of schooling as a predictor variable, it can be seen that by using the R program, the optimal bandwidth value obtained is at \( h = 80 \) with a GCV value of 0.243. With \( R^2 \) and MSE values of 84.71 and 32.1864, respectively.

References
[1] Härdle W 1990 Applied Nonparametric regression (New York: Cambridge University)
[2] Bontemps C 2000 Nonparametric Integration of Regression Functions (mimeo LEERNA-INRA Toulouse)
[3] Chaubey Y P, Laib N and Li J 2012 J. of Stat. Plan. Inf. 142 2.
[4] Rice J A 1984 Ann. of Stat. 12 1215.
[5] Hidayat R, Budiantara I N, Otok B W and Ratnasari V 2019 App. Math. and Inf. Sci. 13 1
[6] Rosenblatt M 1956 Ann. of Math. Stat. 27 642
[7] Parzen 1962 Ann. of Math. Stat. 35 1065
[8] Lin X, Wang N, Welsh A, and Carroll R J 2004 Biometrika 91 1.
[9] Kayri M and Zirhhoglu G 2009 Ozean J. of App. Sci. 2 1.
[10] Hidayat R, Budiantara I N, Otok B W and Ratnasari V 2020 Comm. in Stat. - Theory and Methods 50 17
[11] Silverman B W 1986 Density estimation for statistics and data analysis (London: Chapman and Hall, New York)
[12] Scott D W 1992 Multivariate density estimation theory, practice, and visualization (New York: John Wiley and Sons, Inc)
[13] Manzana S and Zerom D 2005 Stat. and Prob. Letters. 72 1
[14] Okumura H and Naito K 2006 J. of Mult. Anal. 97 1
[15] Hidayat R, Budiantara I N, Otok B W and Ratnasari V 2020 J. of Theo. and App. Inf. Tech. 97 2