Comparison of Kernel regression model with a polynomial regression model on financial data

Journal of Physics: Conference Series **1763** (2021) 012017 doi:10.1088/1742-6596/1763/1/012017

Comparison of Kernel regression model with a polynomial regression model on financial data

Nur'eni [1*], M Fajri[2] and S Astuti[3]

[1]Statistics Study Program, Tadulako University, Indonesia
[2]Statistics Study Program, Tadulako University, Indonesia
[3]BPS, Statistics of Sulawesi Tengah Province, Indonesia

*Email: nureniuntad@gmail.com

Abstract: Regression analysis is constructed for determining the influence of independent variables on the dependent variable. It can be done by looking at the relationship between those variables. This task of approximating the mean function can be done essentially in two approaches, parametric and nonparametric approach. Kernel regression is one of the models with a nonparametric approach, and polynomial quadratic regression is one of the models with the parametric approach. This research aims to find the best model regression with compare to the model of kernel regression and model of polynomial quadratic regression in financial data using *RMSE* criterion. Share data that be used is Mastercard Incorporated (MA) with data periods 02 Januari 2019 until 31st December 2019. Research's result indicated that for MA data, best model regression is kernel regression with *RMSE* value = 16,00147 and *Bandwidth (h)* = 25,64.

1. Introduction

Regression analysis is a statistical analysis used to see the effect of independent variables related to variables, first looks at the pattern of these relationships [1]. In some financial cases, there are many problems with the relationship between the dependent variable and the independent variable where the forms of the relationship do not have a certain pattern, so that they are not resolved, a regression approach is used with a nonlinear form, where the data pattern is assumed to be known [5] or nonparametric form, where the data pattern is not assumed to follow any certain pattern [8]. Not all cases of parametric patterned data in regression analysis follow a certain form like linear, quadratic or cubic, so the other approaches needed like semiparametric or nonparametric approaches [3]. One method that can be used in non-linear regression is polynomial quadratic regression, while the method that can be used in nonparametric regression is kernel regression. Quadratic regression analysis is the development of linear regression, where the data modelled in quadratic regression has or forms a quadratic pattern when visualized in a graph or diagram. Meanwhile, the kernel regression equation is carried out using the smoothing technique, which is based on the kernel function used [7].

This research aims to get the best regression model between the kernel regression model and quadratic polynomial model regression in financial data based on the RSME value and bandwidth obtained from the two methods.

2. Materials and methods

The data used in this study are financial publications data from yahoo financial, from January 1 to December 31, 2019 [10]. In reality, not all data can be estimated with the parametric regression approach because there is no complete information about the shape of the regression curve. In this approach, a nonparametric regression approach can be used [5]. For a sample of size n observational data, the relationship between these variables can be expressed by the regression model as follows:

$$Y_i = m(X_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1,2, \dots, n \tag{1}$$

Nonparametric estimators that are widely used are smoothing estimators; one example of smoothing is Kernel Regression. The purpose of smoothing is to remove data variability that does not affect so that the characteristics of the data will appear clearer so that the resulting curve will be smooth.

2.1 Kernel functions

The kernel function is denoted by K(x) which is a function whose use is applied to each data point. Following are the characteristics of a kernel function [4]

a. $\int_{-\infty}^{\infty} K(x)dx = 1$
b. $\int_{-\infty}^{\infty} xK(x)dx = 0$
c. $\int_{-\infty}^{\infty} x^2K(x)dx = \mu_2(K) \neq 0$
d. $\int_{-\infty}^{\infty} [K(x)]^2 dx = \int_{-\infty}^{\infty} K^2(x)dx = \|K\|_2^2$

In the kernel method, the density estimator for a value of x is denoted by $\hat{f}_h(x)$f, which is expressed in the following formula:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_I}{h}\right) \tag{2}$$

Here is a general form of the *K* (*x*) kernel in terms of *bandwidth* usage:

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right) \tag{3}$$

Kernel density estimator theorem:
If the kernel function is a density function $\int_{-\infty}^{\infty} K(u)du = 1$, then the function estimator using the kernel function is also a probability density function.

2.2 Kernel density estimator

In the kernel density estimator, there are two kinds of parameters [4], namely:

a. *Bandwidth h,* dan
b. Density function of kernel *K*

To select *h* from the kernel density function *K* it is necessary to check the unbiased asymptote of $f_h(x)$ as follows:

$$E[K_h(x)] = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)du \tag{4}$$

Based on the nature of the kernel function:

$$Bias\left|\hat{f}_h(x)\right| = \frac{h^2}{2} f''_h(x)\mu_2(K) + o(h^2), h \to 0 \tag{5}$$

As for the variance:

$$Var\left|\hat{f}_h(x)\right| = n^{-1}h^{-1}\|K\|_2^2 f(x) + o((nh)^{-1}), nh \to \infty \quad (6)$$

After the bias and variance are obtained, then analyze the *MSE* which is a combination of variance and bias squared from $\hat{f}_h(x)$ as follows:

$$MSE\left|\hat{f}_h(x)\right| = \frac{1}{nh} f(x)\|K\|_2^2 + f(x) + \frac{h^4}{4}\left(f''_h(x)\mu_2(K)^2\right) + o((nh)^{-1}) + o(h^4) \quad (7)$$

Using the *MSE* formula is very difficult to use because there is an unknown density function *f (x)*. For this reason the *MISE* (*Mean Integrated Squared Error*) is defined as follows:

$$MSE\left|\hat{f}_h(x)\right| = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\mu_2(K)^2)\|f''\|_2^2 + o((nh)^{-1}) + o(h^4) \quad (8)$$

2.3 Kernel regression

Kernel regression is a nonparametric statistical technique used to estimate the value of the Conditional Expectation of a random variable. Typical expected value is denoted by E(Y|X). Mathematically, for any *x* value, the smoothing estimator for *m(x)* can be expressed as follows:

$$\widehat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} \omega_{hi}(x) Y_i \quad (9)$$

Where $\omega_{hi}(x)$ can be defined as a weighted function:

$$\omega_{hi} = \frac{K_h(x - X_i)}{\hat{f}_h(x)} \quad (10)$$

Where:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - X_I) \quad (11)$$

By substituting (11) into (10), the Nadaraya-Watson kernel estimator $\widehat{m}_h(x)$ from *m(x)* is obtained as follows [2]:

$$\widehat{m}_h^{NW} = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{x - X_i}{h}\right)} \quad (12)$$

The choice of kernel function to be used in this study is the *Gaussian* kernel, namely:

$$K_G(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{x^2}{2}\right) \quad (13)$$

2.4 Optimum bandwidth selection

Given a bandwidth value that is too small will produce a rough estimation curve, on the other hand a bandwidth that is too large will produce a very smooth estimation curve. There are several ways to approach the optimum *bandwidth* selection, one of which is by using the *plug-in* method. The approach using the *plug-in method* is more based on an extension of the *Mean Integrated Square Error* (*MISE)* for kernel smoothing which can be seen in equation (8) so that the assymptotic *MISE (A-MISE)* is obtained by ignoring the use of $o((nh)^{-1}) + o(h^4)$, so:

$$A - MISE = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\mu_2(K)^2)\|f''\|_2^2 \quad (14)$$

So that the optimum *bandwidth* size is:

$$h_{opt} = \left(\frac{\|K\|_2^2}{(\|f''\|_2^2)(\mu_2(K))^2 n}\right)^{1/5}$$
$$h_{opt} = \frac{1{,}06}{n^{1/5}}\sigma \quad (15)$$

2.5 Optimum selection of Kernel regression model with optimal h

In accordance with the objective of the nonparametric regression approach, which is to obtain a smooth curve that has an optimum *h* using *n* data, it is necessary to measure the performance of the estimator that is universally accepted. The performance measure for this estimator is to calculate the average number of squares of residues (*Mean Square Error-MSE*). The performance measure of the simple estimator is the square of the remainder, which is averaged by the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(m_h^{NW}(x) - Y_i)^2 \quad (16)$$

For $i = 1,2,\dots,n$.

This criterion is expected to have a minimum value so that the kernel regression model can be said to have an optimal *h*.

2.6 Quadratic Polynomial Regression

The polynomial regression model is a relationship between two variables consisting of the *dependent* variable (Y) and the *independent* variable (X) so that a curve that forms a curved line will be obtained. Here is a mathematical model of the quadratic polynomial regression equation:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2 + e_i \quad (17)$$

Where variables Y and X show statistical variables: $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\beta}_2$ are estimators for β_0, β_1,, and β_2 which are called simple regression coefficients, *e* states the error component of the regression form [9].

2.7 Approach to Analysis of Variance in Quadratic Polynomial Nonlinear Regression

The ANOVA approach is based on breaking down the sum of the squares (*sum square*), and degrees of freedom associated with the dependent variable Y. decomposing the sum of the total squares and degrees of freedom is usually arranged in the form of an ANOVA table as follows [6]:

Table 1. analysis of variance (ANOVA)

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (DF)	Average Square (MS)
Regression	SSR	*K*	$MSR = \frac{SSR}{k}$
Error	SSE	*n-k-1*	$MSE = \frac{SSE}{n-k-1}$
Total	SST	*n-1*	

2.8 Parametric assumption test

2.8.1. Normality test Normality testing is a test of the normal distribution of data. This test can be done using the Kolmogorov Smirnov test where decision making can be done by looking at the value of the probability.

2.8.2. Linearity test Linearity test can be done using a plot. If the final conclusion obtained is that the data is linear then the data cannot be used but if the data is nonlinear then we can continue the test to the next stage, namely the quadratic polynomial regression coefficient test.

2.8.3. Quadratic polynomial regression coefficient test To determine whether the quadratic polynomial regression is significant, we need to test the significance of the regression parameters. Simultaneous test with ANOVA table for testing regression parameters together and partial test with t-test for testing regression parameters separately [1].

3. Result and discussion

The kernel function used is the Gaussian kernel, where the boundaries of this kernel are between $(-\infty, \infty)$. For the kernel nonparametric regression method with Nadaraya-Watson estimation, the optimum bandwidth is obtained which can provide a curve shape in the estimated regression function. The data used is data on daily closing price of shares (taken in the last trading period each day). The Mastercard Incorporated share value data can be expressed in the following chart plot:

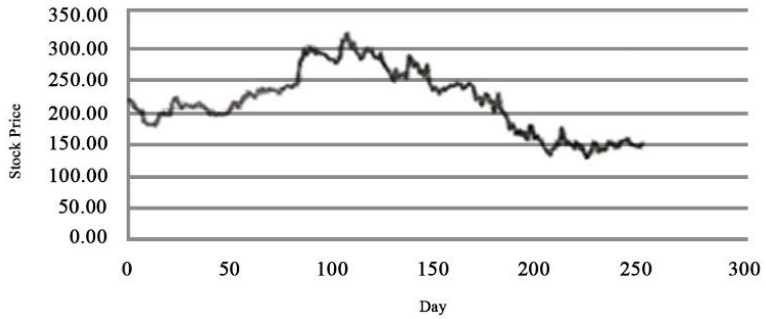

Figure 1. Mastercard Incorporated's value stock

3.1. Kernel regression

From the calculation using the R program, the *RMSE* value = 16.00147 for the *bandwidth* size = 25.64. The following is a graph of the kernel regression results with the Nadaraya-Watson estimate

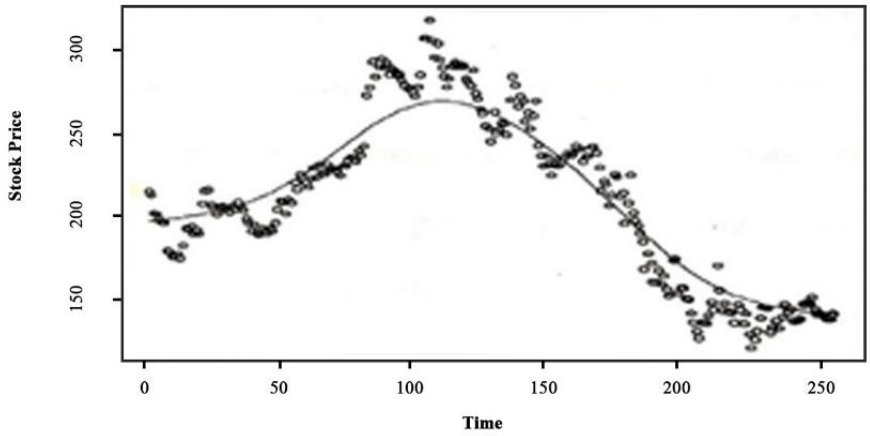

Figure 2. Kernel Regression Graph with Nadaraya-Watson Estimation

In the picture above, it can be seen that the movement of the values generated by the Nadaraya-Watson kernel estimator will follow the shape of each existing observation point so that it will produce an optimum kernel nonparametric regression curve.

3.2. Quadratic Polynomial Regression
The following is the result of the quadratic polynomial regression graph

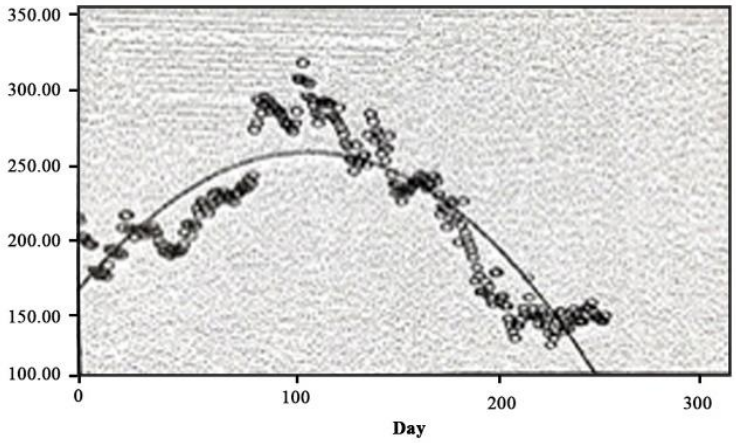

Figure 3. Quadratic Polynomial Regression Graph

From the picture above, it can be seen that the graph of the quadratic polynomial regression is parabolic with the following regression equation:

$$\hat{Y} = 164{,}818 + 1{,}743X - 0{,}008X^2$$

However, the regression above cannot be drawn as a conclusion because there has not been a parametric assumption test and a quadratic polynomial regression coefficient test. For that, the next step is to test the parametric assumptions.

3.3. Parametric assumption test
3.3.1. Normality test Normality was tested using the Kolmogorov Smirnov test.

Table 2. Kolmogorov Smirnov Test for One Sample

Normality Test	Stock Price
Kolmogorov-Smirnov Z	1.314
Asymp. Sig. (2-tailed)	0.063

Based on the table, the probability value for the stock price variable is 0.063. This probability value can be seen from the Asymp value. The significance (P)> 0.05, the data used is data that is normally distributed.

3.3.2. Linearity test The linearity test shows that for the linear regression equation, the relationship between the independent and dependent variables must be linear. This assumption will determine the type of estimation equation used.

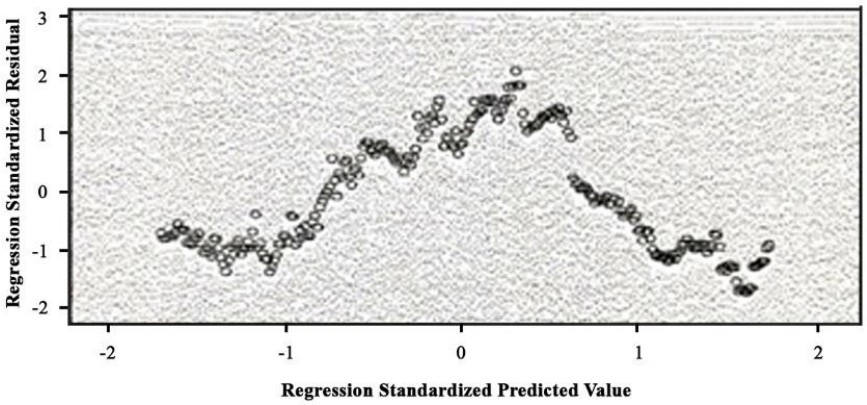

Figure 4. Scatterplot between standardized residual with standardized predicted

Based on the scatterplot above, it can be explained that the linear assumptions are not fulfilled. This can be seen from the scatterplot graph above which forms a certain pattern so that there is a nonlinear relationship between variables.

3.4. Quadratic polynomial regression coefficient test

This test is performed using the F-test through the Anova table. The hypothesis is:
$H_0: \beta_1 = \beta_2 = 0$, means that the quadratic polynomial regression model is not significant
$H_0: \beta_1 = \beta_2 \neq 0$, means that the quadratic polynomial regression model is significant
Based on the results of data processing, the Anova table is obtained as follows:

Table 3. ANOVA results for the quadratic polynomial regression model test

Source of Variation	Sum of Squares	Df	Mean Square	F	Sig.
Regression	501437.947	2	250718.973	408.129	.000
Residual	153578.172	250	614.313		
Total	655016.119	252			

Testing criteria:
H_0 is rejected if $F_{count} \geq F_{table}\ (1-a; k; n-k-1)$
H_1 is accepted if $F_{count} < F_{table}\ (1-a; k; n-k-1)$
Obtained the value of F_{table} is:

$$F_{table}(1-a; k; n-k-1) = F_{table}\ (1-0{,}05; 2; 250) = 3{,}035$$

Based on the Anova table, the mean square column in the residual row, the *MSE* value = 614,313 is obtained, so the *RMSE* (*Root of Mean Squared of Error*) value is:

$$\begin{aligned} \text{RMSE} &= \sqrt{\text{MSE}} \\ &= \sqrt{614{,}313} \\ &= 24{,}785 \end{aligned}$$

Table 4. Summary of the quadratic polynomial regression model

r	R Square	Adjusted R Square	Std. Error of the Estimate
.875	.766	.764	24.785

3.5. Comparison of Kernel regression model with a quadratic polynomial regression model
From the two regression models that have been carried out, namely the kernel regression model with Nadaraya-Watson estimation and the quadratic polynomial regression model, comparisons will be made to determine which regression model is better. The comparison measure used is based on the RMSE value of each model:

Table 5. Model comparison based on *RMSE* value

Regression	Model	RMSE
Nadaraya-Watson Kernel Regression (h_{opt} = 25,64)	$\hat{Y}_i = \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{\left((x-X_i)/25{,}64\right)^2}{2}\right) Y_i}{\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{\left((x-X_i)/25{,}64\right)^2}{2}\right)}$	16.00147
Quadratic Polynomial Regression Model	$\hat{Y} = 164{,}818 + 1{,}743X - 0{,}008X^2$	24,785

From the Table 5, it can be seen that the Nadaraya-Watson kernel nonparametric regression model with a *bandwidth* value of = 25.64 provides a better estimate than the quadratic polynomial regression model because it produces a smaller *RMSE* value than the quadratic polynomial regression so that the best regression model is the kernel regression model with Nadaraya-Watson's estimate.

4. Conclusion
By using the quadratic polynomial regression model, the *RMSE (root of Mean Squared Error)* value is 24.785, while using the Nadaraya-Watson kernel regression model with the Gaussian kernel type, the *RMSE* value is 16.00147 and the bandwidth size used is 25.64. By comparing the two regression models, it is found that the Nadaraya-Watson kernel regression model with a *bandwidth* value of = 25.64 in financial data is better than quadratic polynomial regression model because, in addition to providing a lower error rate, the kernel regression model also provides a better estimate than the quadratic polynomial regression model.

Acknowledgement
Thanks to Statistics Laboratory who assisted this research in the data processing. Thanks also to many academic parts for their contribution to this research.

References
[1] Draper N R & Smith H 1998 *Applied regression analysis* **326** (New York: John Wiley & Sons)
[2] Eubank R L 1988 *Spline smoothing and nonparametric regression* **90** (New York: M Dekker).
[3] Fernandes A A R, Budiantara I N and Otok B W 2015 *Journal of Mathematics and Statistics* **11**(2) 61
[4] Halim S and Bisono I 2006 *Jurnal Teknik Industri* **8**(1) 73
[5] Härdle W 1990 *Applied nonparametric regression* **19** (Cambridge: Cambridge University Press)
[6] Nurgiyantoro B 2004 *Statistik Terapan: Untuk Penelitian Ilmu-Ilmu Sosial* (Yogyakarta: Gajah Mada University Press)
[7] Puspitasari I, Suparti S and Wilandari Y 2012 *Jurnal Gaussian* **1**(1) 93-102
[8] Wahyuni S A, Ratnawati R, Indriyani I and Fajri M 2020 *Natural Science: Journal of Science and Technology* **9**(2) 34-39
[9] Tiro M A 2008 *Dasar-dasar statistika* (Makassar: Andira Publisher)
[10] Yahoo Finance 2019 *Data Saham Mastercard Incorporated* (http://www.yahoofinance.com)