A Comparative Study of Nonparametric Kernel estimators with Gaussian Weight Function

Mohammed A. Dakhil 1, Jassim N. Hussain2
1,2 Department of Statistics Administration and Economics College University of Karbala.

Corresponding author: muhammed.a@s.uokerbala.edu.iq  
jasim.nasir@uokerbala.edu.iq,  muhammed.a@s.uokerbala.edu.iq

Abstract. Nowadays, Parametric methods become unfavorable by researchers because of the restrictions on using them and losing the flexibility in estimating and analysis the data. Therefore, the researchers preferred the nonparametric method which proved their efficiency and capable to analysis the data without of predetermined assumptions. Consequently, the data and their included information are becoming who determine the functional shape for the studied population and there are no parameters instead of the observations. The objective of estimating the nonparametric regression function is to approximate the regression function to the real regression function. On the other hand, COVID-19 pandemic nowadays speared in all the countries one of them is Iraq. The function of infection speared have been studied in different countries but not in Iraq. Therefore, the aim of our research is to apply three nonparametric Kernel estimators with Gaussian weighted function to model and forecast the number of infections of COVID-19 in Iraq. R software and the data represent the daily number of COVID-19 infections for the period 23/2/2020 to 21/6/2020 are used to apply many models and choose the appropriate one. The results of applying three nonparametric Kernel model that the Priestley-Chao model is the appropriate one in all the sample sizes and other conditions

Key Word: COVID-19 pandemic, Nonparametric Kernel models, Parametric methods, Priestley-Chao model.

1. Introduction
As it was Known that regression analysis is a statistical method have been used to build a statistical casual model to predicate the mean of random variable based on values of another one or more random variables. Also, it was looking for the association between the dependent variable and one or more of explanatory variables. These models may be parametric or nonparametric each one has its characteristics. The parametric models assume many assumptions (e.g., the sample was coming from a population has a known family of probability distribution, the variance was constant, the observations are independent,...) and they used classical methods
to estimate the parameters of these models. Often, these assumptions of parametric models are not available then using these models lead to unreliable results (Yu & Jones, 1997). Consequently, the researchers looking for the nonparametric regression model because they request less assumptions and the data will give the shape of the model (Sheather, 2004). Nonparametric regression was used for the first time by [(Rosenblatt, 1956) and (Parzen, 1962)] which were they referred to estimate of the probability density function. Several nonparametric regression models may be used to model the phenomenon, one of them is the nonparametric Kernel regression. Kernel regression is used to build a casual mathematical model represent the association between a dependent variable and one or more independent variables (AydinD, 1999) Fitting of Kernel regression model was depending on three factors; the first one is choosing the weight function from the available functions (e.g. Gaussian, uniform, etc.). The second factor is choosing the bandwidth (h) by using one method of estimation to estimate the appropriate value. The last factor is choosing the appropriate method of estimate from available methods. Then in this paper we focus on three methods of estimation which are Local Constant Regression Estimator or it was called Nadaraya-Watson (N-W) (Nadaraya, (Watson, 1964) estimator, Local Linear Smoother (LLS), and Priestly-Chao (P-Ch) estimator (Zhang et al., 2007).

On the other hand, COVID-19 pandemic in this year have a huge effect on different sides of life for all the world. The data of this pandemic are driven analysis because there is no previous information about this data. Therefore, the nonparametric is a appropriate models may be used to model and forecast the number of infections of COVID-19 in Iraq.

Several researchers were studied this pandemic in different countries, (literature review for about five to six papers) The problem of this paper is Covid-19 pandemic affected all the sectors of the life such as economics losses, job losses, life losses, health sector losses and others. This pandemic was studied in different countries to know the behavior of this virus. Meanwhile there are few studies about this pandemic in Iraq. Therefore, the aim of this study is to study the spred of this virus in Iraq using the nonparametric models to give a vision of behavior of Covid-19 spreaed.

The plane of paper
2. Methods

Nonparametric regression used to estimate the regression function \( \mu(x_i) \) directly without any specific function and without estimate the parameters such as the parametric regression. This function is a continuous function and smoothing may be used to represent the relationship between the response variable and the explanatory variables in the sample of \( n \) observations \( \{y_i, x_i\}_{i=1}^n \) as the following (Hoti, 2001)

\[
y_i = \mu(x_i) + \varepsilon_i \quad i = 1, 2, \ldots, n
\]

where \( y_i \) is an \( i \) observation of a response variable, \( x_i \) is an \( i \) observation of explanatory variable, \( \mu(x_i) \) is the regression function of \( Y \) on \( X \) which is equal \( \mu(x_i) = E(Y|X = x) \), it does not has parameters and it was increasing in \( (x_i) \) if \( \mu(x_i) \leq \mu(x'_i) \) when \( x_i \leq x'_i \). \( \varepsilon_i \) is a random error distributed as anormal with mean equal zero and c constant variance equal \((\sigma^2)\).

The estimation of this model may be done with one of two methods; the first one is the parametric methods such as ordinary least square (OLS), maximum likelihood estimation (MLE), etc. when we have a prior information about the data. But when we do not have a prior information about the data then the nonparametric methods are used such as Kernel estimation. The estimation of Kernel regression based on three factors (Hansen, 2009). The first factor is the Kernel weight function. It is a real function named with many names such as window functions, shape functions and basic functions and it was defined as \( K(u) \) with the following characteristics, it is positive \( K(u) \geq 0 \); the integral of function is equal \( 1 \int_{-\infty}^{\infty} K(u) \, du = 1 \), it is symmetric \( K(u) = K(-u) \) \( \forall u \), it is continuous \(-1 < K(u) < 1\) (Shirahata & Chu, 1992). There are many weight functions such as Gaussian, Uniform, Epanechnikov and others (Epanechnikov, 1969). These functions may be classified into two types the first one is working to decrease the nearest neighbor variance, the second type is the optimal kernel functions which is working to decrease the mean integrated squared error (MISE). Then choosing the function based on the data and the objective of the analysis. The second factor is Bandwidth or the smoothing parameter \( h \). The important point in the local polynomial regression is how to choose the smoothing parameter. This parameter is considered as threshold between the variance and the bias. The small value of this parameter gives a high variance and small bias, meanwhile the large value gives a high bias with small variance. Therefore, choosing the appropriate value of smoothing parameter has a main role in the Kernel estimators. There are many methods to choose this parameter such as normal scale (optimal bandwidth) method, cross-validation method, and direct plug-in method. According to normal scale method we can find the smoothing parameter as follow (Silverman, 1986)

\[
AMISE(h) = (nh)^{-1}R(K) + h^4 \left( \frac{M_2(K)}{2} \right)^2 R(f^2) = \frac{R(K)}{nh} + \frac{h^4 d^2}{4} R(f''') \tag{2}
\]

Where \( M_2(K) \) and \( R(K) \) are Kernel constants as follow: \( M_2(K) = \int u^2 K(u)^2 \, du \) and \( R(K) = \int K(u)^2 \, du \), then the optimal value of smoothing parameter is the value which is decreasing AMISE become as:

\[
h_{AMISE} = \left( \frac{R(K)}{M_2(K) R(f''')} \right)^{1/5} n^{-1/5} \tag{3}
\]

Since the assumed function is normal then:

\[
R(f''') = \int_{-\infty}^{\infty} (f'''(x))^2 \, dx = \frac{3\sigma^{-5}}{8\pi^{1/2}} \tag{4}
\]

Then substitute the Eq.4 in Eq. 3 gives:
\[ h_{AMISE} = \left( \frac{6\pi^{1/2}R(K)}{3M^2(K)n} \right)^{1/5} \ast \sigma \]  
\[ (5) \]

After substituting the value of Gaussian kernel constants in Eq. 5 (Jones et al., 1996) the smoothing parameter according to Normal scale method become as:

\[ h = \delta(1.06)n^{-1/5} \]  
\[ (6) \]

The last factor is the Nonparametric Estimation methods these methods have high care because they are flexible to apply the function regardless if it is linear or nonlinear. Also, these methods did not need for restricted assumptions about the shape of the unknown functions. Local Polynomial Kernel method (LPK) is considered as one of statistical estimation method of nonparametric regression. It is the best method of smoothing method even that it is preferred than other Kernel methods (Wand & Ripley, 2006). This smoother has a capability of smoothing parameter which is used constant smoothing parameter. The basic idea of this method is estimating the function (\( \mu(x_i) \)) locally instead of using global polynomial which is using all the \( n \) data to estimate this function by estimating \( p + 1 \) parameters as in parametric regression (Altaher & Ambar, 2018) To estimate the function (\( \mu(x_i) \)) locally at the point \( x \) we need to determined neighborhood contain the point \( x \) such as \( (x-h, x+h) \) where \( h \) is the bandwidth or smoothing parameter which is determined the neighborhood around \( x \). Then only the observation of \( (y_i) \) which its data \( (x_i) \) are included in this interval used to estimate the function (\( \mu(x_i) \)). Applying LPK to estimate this function based on two factors are the degree of polynomial and the bandwidth \( h \) (Zhang et al., 2007) There are many methods to estimate Kernel nonparametric regression we choose three of them as follow (Goldenshluger & Lepski, 2011):

2.1 Local Constant Regression (Nadaraya- Watson) Estimator

It is one of oldest nonparametric estimator and most famous. Its name due to the work of two researcher Nadaraya and Watson at 1964. It is one of the famous estimation methods used to estimate the nonparametric regression (Raid B. Salha & Hazem I. El Shekh Ahmed, n.d.). It is a determinant and continuous function with positive values an its integration equal one. The estimator of N-W for the function (\( \mu(x_i) \)) using the weighted least square (WLS) method is as follow (Raid B. Salha & Hazem I. El Shekh Ahmed, n.d.)

\[ \hat{\mu}_{N-W}(x_i) = \frac{\sum_{i=1}^{n} K_h(x_i-x) y_i}{\sum_{i=1}^{n} K_h(x_i-x)} \]  
\[ (7) \]

where \( K(u) \) is a Kernel function, \( h \) is the bandwidth which is control the smoothing quantity in the Kernel regression. Therefore, if we choose a small value of \( h \) then we will use only the point \( x \) to find the estimation as:

\[ K \left( \frac{x_i-x}{h} \right) = K \left( \frac{0}{h} \right) = K(0) \]  
\[ (8) \]

Then the estimated curve will pass through all the data points, with high variance and low bias and it is under smooth. Meanwhile, if it is chosen with a big value approximate to () then the N-W estimator becomes as:

\[ \hat{\mu}_{N-W}(x_i) = \lim_{h \to \infty} \left\{ \frac{\sum_i K \left( \frac{x_i-x}{h} \right) y_i}{\sum_i K \left( \frac{x_i-x}{h} \right)} \right\} \]
\[ \bar{y} = \frac{\sum_i K(0)y_i}{\sum_i K(0)} = \frac{K(0)\sum_i y_i}{nK(0)} = \frac{\sum_i y_i}{n} \]  

(9)

This means that the curve became as straight line and named over smoothing which with small variance and high bias. Therefore, the value of bandwidth must be chosen to balance these two cases (Araveeporn, 2011)

2.2 Local Linear Regression Estimator (L.L)

Stone (1984) and Fan (1992) (Fan, J., 1992) proposed this estimator. It was one of the best estimator methods of Kernel regression. It was used with constant and inconstant models. It has high efficiency if we choose an appropriate Kernel weight function and bandwidth. If the degree of smoothing equal one (p=1) then the result estimator is the local liner regression (Reference). To illustrate the work and the formula of this estimator we assume that the second derivative of the regression function (\( \mu''(x_i) \)) in the small space of neighborhood of the point (x) and the vector of the parameters is:

\[ \hat{\beta} = \left( \frac{\hat{\beta}_0}{\hat{\beta}_1} \right) \]

which was results from weighted least square when (p=1), the L.L estimator is as follow (Lu, 1996):

\[ \hat{\beta}_{(x_i,h)} = n^{-1} \left( \frac{\sum_{i=1}^n K_n(x-x_i)y_i(\hat{s}_2(x,h) - \hat{s}_1(x,h)(x-x_i))}{\hat{s}_0(x,h)\hat{s}_2(x,h) - \hat{s}_1(x,h)^2} \right) \]  

(10)

Often choosing the appropriate smoothing degree (p) but it is not important such as choosing the bandwidth. Where, the local constant regression estimator when (p = 0) and the local linear smoother (L.L) when (p = 1) have enough efficiency for most application if we choose the appropriate \( K(u) \) and \( (h) \) (Kocaguneli et al., 2013)

2.3 Priestly-Chao Estimator (P-Ch)

It is an estimator proposed by (Priestley and Chao) at 1972 (Priestley & Chao, 1972). The estimated function is defined as:

\[ \hat{\beta}_{P,Ch}(x_i) = h^{-1} \sum_{i=2}^n (x_i - x_{i-1}) K(u) \left( \frac{x-x_i}{h} \right) y_i \]  

(11)

Often (P-Ch) estimator is equivalent to local constant regression estimator and has same asymptotic formula of variance and bias as follow (Pinelis, 2020):

\[ E[\hat{\mu}(x_i)] = \frac{1}{nh} \sum_{i=2}^n K(u) \left( \frac{x-x_i}{h} \right) \mu(x_i) \]  

(12)

\[ Var[\hat{\mu}(x_i)] = \frac{\sigma^2}{n^2h^2} \sum_{i=1}^n K(u) \left( \frac{x-x_i}{h} \right)^2 \]  

(13)

2.4 Assessing Criteria

The values of different assessing criteria are representing the equivalent of the estimated function (\( \hat{\mu}(x_i) \)) and the real function (\( \mu(x_i) \)); where if the values of assessing criteria approximated to zero then the estimated function approximated to the real function. There are many assessing criteria we will choose three of them to assess the performance of the estimator in applications (Shirahata & Chu, 1992):

1. Average Mean Square Error (AMSE) as follow:

\[ AMSE = \frac{\sum_{i=1}^N MSE(l)}{N}, \text{ where } N \text{ is the number of replication} \]  

(14)

2. Average Mean Absolute Error (AMAE) as follow:
3. Mean Integrated Square Error (MISE) as follow:

\[
MISE = \frac{1}{100} \sum_{i=1}^{100} \left( \mu(x_i) - \hat{\mu}(x_i) \right)^2
\]  

(16)

Usually solving the integration faced some problems therefore we use approximate method by dividing the interval of the variable (x) into 100 parts, this number may be increased or decreased depend on the convenient of the statistical analyst in the standard results (Marron & Wand, 1992). We apply and compare these estimation methods by using a sample of Covid-19 infection cases in Iraq as in the following section.

3. Results and Discussion

The COVID-19 pandemic is the crucial global health crisis in this time because the virus has spread to every country. The first infection case of this virus appeared in Iraq at the last of Feb. 2020. Therefore, we consider the series of the confirmed infection cases in Iraq from 23\textsuperscript{rd} of Feb. 2020 to 22\textsuperscript{nd} of Jun. 2020 which is divided into three samples the small sample with size (n=30) days, the medium sample with size (n=60) days and the large sample with size (n=120) days. The resource of this data is the official site of World Health Organization (WHO) (Who(2020), n.d.)and the official site of the Iraqi health ministry on the Facebook / Daily infection of Covid-19 in Iraq. Daily number of infection cases as a dependent variables Y and X represent the index of the date. The first step is to identify if the data has a linear or nonlinear trend, the shape of infection cases time series is as in Fig.1:

Fig1, the sector plot of the daily infection of Covid-19 cases from 23/2/2020 to 21/6/2020

Fig.1 shows that the data has nonlinear trend, then to apply the nonparametric regression analysis and the estimation methods need to transform the data to the standardized form, after that we used the transformed data with three sample sizes (30, 60, and 120) and the three methods of Kernel estimators with gaussian weight function (N.W estimator, L.L estimator, and P.Ch estimator). To assess and compare these estimators we used three criteria (AMSE, AMAE, and MISE). Consequently, the results of these criteria for three sample sizes and for three estimators as in Table1.
Table 1: the values of assessing criteria for three samples and for three estimators.

| Sample size | Estimator | AMSE   | AMAE   | MISE   |
|-------------|-----------|--------|--------|--------|
| 30          | N.W       | 0.5473 | 0.5859 | 0.0048 |
|             | L.L       | 0.5329 | 0.5347 | 0.0043 |
|             | P.Ch      | 0.5284 | 0.5230 | 0.0042 |
| 60          | N.W       | 0.3506 | 0.3708 | 0.0035 |
|             | L.L       | 0.4695 | 0.3826 | 0.0046 |
|             | P.Ch      | 0.3127 | 0.3654 | 0.0031 |
| 120         | N.W       | 0.0230 | 0.0825 | 0.0002 |
|             | L.L       | 0.0234 | 0.0834 | 0.0002 |
|             | P.Ch      | 0.0227 | 0.0820 | 0.0002 |

The results in Table 1 show that P.Ch estimator gives less value of all criteria at all sample sizes. Also these results show that when the sample size is increasing the values of criteria are decreasing. When the sample size equal to 120 the values of MISE are equal for all the estimators. Therefore the estimated values by using different estimators to compare their results with the real data according to the sample size as shown in Fig. 2 for sample size equal 30, Fig.3 for sample size equal 60, and Fig.4 for sample size equal 120.

Fig 2. the estimated values using three estimators and sample size 30
Fig 3. the estimated values using three estimators and sample size 60.

Fig 4. the estimated values using three estimators and sample size 120.

These figures show that these three estimators give approximately equal estimated values for these data. This behavior may be due to use the standardized values of the data but when we use the actual data to forecast the infection case of Covid-19 in Iraq for the last 11 days from 15/6/2020 to 25/6/2020 they gave different estimated values as shown in Table 2

Table 2. the estimated values using different Estimators an actual data with sample size 60.

| Equal values | 15/6 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
|--------------|------|----|----|----|----|----|----|----|----|----|----|
| N. W         | 1106 | 1385 | 1554 | 1463 | 1635 | 1870 | 1646 | 1646 | 1808 | 1826 | 2200 |
| L. L         | 1306 | 1367 | 1429 | 1490 | 1545 | 1593 | 1635 | 1689 | 1782 | 1901 | 2167 |
| P. Ch        | 1314 | 1386 | 1471 | 1568 | 1677 | 1797 | 1931 | 1647 | 1780 | 1901 | 2162 |

N. W: N. Weibull
L. L: L. Lognormal
P. Ch: P. Chi
4. Conclusions
We use three estimation methods of nonparametric Kernel regression with Gaussian weight function to compare the performance of these estimation methods. Therefore, we use a data set represent the Covid-19 infection cases in Iraq to apply these estimation methods and three criteria are used to evaluate their performance. The results show that P.Ch estimator was outperformed of the other Estimators because it was giving less values of all criteria for three samples. Also, these results show that the increasing in sample size gives a decreased value of these criteria. Because of using the standardized values of the data; the figures show an approximately same behavior of these estimators. When we use the actual data to apply these estimators to estimate the infection cases for the last 11 days; the results in Table 2 show that

For the future works may be examine other estimators of Kernel regression or using another weight function such as Epanchinkov function or uniform function.

5. References

[1]. Altaher, A. M., & Ambark, A. S. (2018). A comparison of Several Bandwidth Selection Methods for Local Polynomial Regression. 191–194.
[2]. Araveepon, A. (2011). The Comparison of Bandwidth Selection Methods using Kernel Function. CURRENT APPLIED SCIENCE AND TECHNOLOGY, 11(2), 64–78.
[3]. AydinD. (1999). An comparison of the nonparametric regression model using smoothing spline and kernel regression “international journal of mathematical, physical and engineering scinces – vol. 2, NO.2, pp.75–79.
[4]. Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, 14(1), 153–158.
[5]. Fan, J. (1992). Design-adaptive nonparametric regression. Journal of the American Statistical Association, 87(420), 998–1004.
[6]. Goldenshluger, A., & Lepski, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. Annals of Statistics, 39(3), 1608–1632. https://doi.org/10.1214/11-AOS883
[7]. Hansen, B. E. (2009). Lecture Notes on Nonparametrics 2: Kernel Density Estimation. Lecture Notes. http://www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf
[8]. Hoti, F. (2001). Kernel regression via binned data. Research Reports C38.
[9]. Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association, 91(433), 401–407. https://doi.org/10.1080/01621459.1996.10476701
[10]. Kocaguneli, E., Menzies, T., & Keung, J. W. (2013). Kernel methods for software effort estimation: Effects of different kernel functions and bandwidths on estimation accuracy. Empirical Software Engineering, 18(1), 1–24. https://doi.org/10.1007/s10664-011-9189-1
[11]. Lu, Z. Q. (1996). Multivariate locally weighted polynomial fitting and partial derivative estimation. Journal of Multivariate Analysis, 59(2), 187–205. https://doi.org/10.1006/jmva.1996.0060
[12]. Marron, J. S., & Wand, M. P. (1992). Exact mean integrated squared error. The Annals of Statistics, 712–736.
[13]. Nadaraya, E. A. (1964). Some new estimates for distribution functions. Theory of Probability & Its Applications, 9(3), 497–500.
[14]. Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics, 33(3), 1065–1076.
[15]. Pinelis, I. (2020). Monotonicity preservation properties of kernel regression estimators. 1–13. http://arxiv.org/abs/2007.01757
[16]. Priestley, M. B., & Chao, M. T. (1972). Non-parametric function fitting. Journal of the Royal Statistical Society: Series B (Methodological), 34(3), 385–392.
[17]. Raid B. Salha & Hazem I. El Shekh Ahmed. (n.d.). Reweighted Nadaraya-Watson Estimator of the Regression Mean.pdf.
[18]. Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. In The Annals of Mathematical Statistics (Vol. 27, Issue 3, pp. 832–837). https://doi.org/10.1214/aoms/1177728190
[19]. Sheather, S. J. (2004). Density estimation. Statistical Science, 19(4), 588–597. https://doi.org/10.1214/088342304000000297
[20]. Shirahata, S., & Chu, I.-S. (1992). Integrated squared error of kernel-type estimator of distribution function. Annals of the Institute of Statistical Mathematics, 44(3), 579–591.
[21]. Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press.
[22]. Wand, M., & Ripley, B. (2006). KernSmooth: Functions for kernel smoothing for Wand & Jones (1995). R Package Version, 2, 19–22.
[23]. Watson, G. S. (1964). Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A, 359–372.
[24]. who(2020). (n.d.). https://www.who.int/ar/emergencies/diseases/novel-coronavirus-2019
[25]. Yu, K., & Jones, M. C. (1997). A comparison of local constant and local linear regression quantile estimators. Computational Statistics and Data Analysis, 25(2), 159–166. https://doi.org/10.1016/S0167-9473(97)00006-6
[26]. Zhang, X., Brooks, R. D., & King, M. L. (2007). Department of Econometrics and Business Statistics A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Regression with an Application to State- Price Density Estimation A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Regr. Victoria, August.