International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 4, Issue 2, Mar 2015, 15-26

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

NONPARAMETRIC MODELS BY USING SMOOTHING KERNEL FUNCTION WITH APPLICATION

MONEM A. M & SAMIRA M. S

Al Sulamania University, Collage of Administration and Economic, Sulaymaniyah, Iraq

ABSTRACT

The important issue in the main applications of statistical represented by the distribution and the assumptions for the parent population (from which the sample is drawn) has a specific distribution characteristics to be represented community representation, but in many cases does not know the form of the basis distribution so we needs **statistical techniques** do not depend on the distribution or assumptions about the phenomenon required study (depend on the free distribution for the data), and these methods are the **nonparametric regression methods** that depend directly on the data when estimating equation.

In this research was review some methods in nonparametric regression methods, like the method, (Local polynomial regression) the some methods for estimating the smoothing parameters (with one of these methods have been proposed to find an initial value for the smoothing parameter with Kernel functions) , and then compared the results of the methods mentioned above, among themselves using tests and statistical standards following: (MSE, RMSE , R^2 , R^2_{adj} , and F-statistic). That by application for the real data is the (elements of climate), daily average temperatures for the period from (1/1/2011 to 31/12/2013) registered with the Directorate of Meteorological and Seismology in Sulaimani with different sample sizes

(365, 730,1095), to show which the sample size with climate data and geography is more efficient with simple nonparametric regression model (Local Linear) and multiple regression model (Additive model).To achieve the objective of this study we will using statistical programs through the program (SYSTA- 12).

KEYWORDS: Nonparametric Models by Using Smoothing Kernel Function with Application

1. INTRODUCTION

The important issue in statistical applications is knowing distribution for population we need study and knowledge of the properties of that community to be represented community representation intact ,through statistical methods commonly so over the past years were distributions parametric debated topic of many researchers as assume modalities parametric that the sample comes from the population his known family of distributions, including normal family (Gaussian) or family (Gamma) and then work to estimate the parameters unknown to those families using methods (MLE, Moments, Bays) and other methods, or work tests for confidence limits or interval of confidence for the parameters unknown, depending on the sample, but those assumptions mentioned above are often strict because distribution teachers are supposed to not be necessarily the actual distribution.

The philosophy of statistics in terms of the mechanism of the application is to try different modeling phenomena models are what can be closer to reality, that these examples measured the degree of strength depending on the degree of

convergence with statistical evidentiary properties. Then regression is the most widely used of all analyses, modeling linear broadest sense well understood in both developed and in addition there is a variety of techniques useful to verify the assumptions involved. However nonparametric regression aims to provide a means of modeling and even when appropriate linear models not yet been brought into question and smoothing techniques are still useful by enhancing scatter plots to display the data infrastructure.

Today regression models are one of the most important varieties statistical theory and so for her to researchers in various fields of science and humanity of scientific solutions to their problems and because of the diverse areas of work had to be variations on the other. On the basis of this divided regression models into two classes essential to the nature of the data, namely (parametric regression models and nonparametric regression models), and that's where the nonparametric regression need to restrictions or conditions less than parametric models and this is precisely who made the tool of nonparametric regression models very desirable to researchers to the fact that the actual data is not always have ideal specifications. So these models evolved in many areas and are used by scientists in the fields such as (Computer Engineering to distinguish the picture and sound, Geography, Physics, Economics, Medicine, Environment, etc…).

The parametric regression and nonparametric regression represent two different ways in the regression analysis, but this does not mean that the method prevents the use of other. For example, the study of multiple regressions in general will produce the problems dimensional suffered by most researchers comply and prevent their progress toward a general state of univariate to the binary variables and then multivariate based on the developing modalities nonparametric analysis.

2. AIM OF THE PAPER

The aims of this paper to analysis nonparametric models (Local polynomial regression model) comparing with the linear regression model with different sample sizes and different bandwidth (h).

3. THEORETICAL PART

Nonparametric Regression Models

The traditional nonlinear regression model (described in the Appendix on nonlinear regression) fits the model

$$y_i = f(\beta, \acute{x}_i) + \varepsilon_i$$

Where $\beta = (\beta_1, ..., \beta_p)$ is a vector of parameters to be estimated, and

$\acute{x}_i = (x_1, \ldots, x_k)$ is a vector of predictors for the (i^{th}) of (n) observations; the errors (ε_i) are assumed to be normally and independently distributed with

Mean (0) and constant variance (σ^2). The function $f(\cdot)$, relating the average value of the response (y) to the predictors, is specified in advance, as it is in a linear regression model.

The general nonparametric regression model is written in a similar way, but the function (f) is unspecified

$$y_i = f(\acute{x}_i) + \varepsilon_i \quad (1)$$

$$= f(x_{i1}, x_{i2}, \ldots, x_{ik}) + \varepsilon_i$$

With $E(\varepsilon)=0$, $V(\varepsilon_i)=\sigma_\varepsilon^2$, $Cov(\varepsilon_i,\varepsilon_j)=0$ for $i \neq j$

The object of nonparametric regression is to estimate the regression function f (·) directly, rather than to estimate parameters. Most methods of nonparametric regression to assume that f (·) is a smooth, continuous function as in nonlinear regression. An important special case of the general model is nonparametric simple regression, where there is only one predictor:

$$y_i = f(x_i) + \varepsilon_i \tag{2}$$

Nonparametric simple regression is often called 'scatter plot smoothing' because an important application is to **tracing a smooth curve through a scatter plot of (y) against (x)**. Because it is difficult to fit the general nonparametric regression model when there are many predictors, and ***because it is difficult to display the fitted model when there are more than two or three predictors. One such model is the additive regression model***,

$$E(Y_i / x_{i1},\cdots,x_{id}) = \sum_{j=1}^{d} f_j(x_{ij})$$

$$Y_i = \alpha_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_d(x_{id}) + \varepsilon_i \tag{3}$$

Where the partial-regression functions fj(·) are assumed to be smooth, and are to be estimated from the data, $X_i = (X_{i1},\cdots,X_{id})$ are random design, (ε_i) are unobserved error variables, Where ($f_1, \ldots, f_d$) are smooth functions and (α_0) is a constant.

In this model we assume that all of the partial-regression functions are linear.

4. KERNEL ESTIMATOR

By replace the weight function for the naive estimator above by a kernel function we denoted a new function; it is the kernel estimator. The behavior of the estimator is dependent on the choice of width of the intervals used, and also to some extent on the starting position of the grid of intervals.

- A smooth kernel function rather than a box is used as the basic building block.
- These smooth functions are centered directly over each observation.

Because the kernel estimated is bona fide estimator and a symmetric, then probability density function for kernel estimator with exists the conditions defined by:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{h}\right) \quad \text{Or}$$

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x_0 - X_i}{h}\right) \tag{4}$$

Then the kernel estimators or (local average) is equal to

$$= \sum_{i=1}^{n} w_{hi}(x) y_i$$

Where (w_i) denoted the weight function:

$$w_{hi}(x) = \frac{k\left(\frac{x_0 - X_i}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x_0 - X_i}{h}\right)} \quad (5)$$

Notice: $\sum_{i=1}^{n} w_{hi}(x_i) = 1$

Now the (MSE) of the kernel density estimate can be expressed as:

$$MSE \equiv \mathrm{E}\left(\hat{f}_h(x) - f(x)\right)^2$$

5. ESTIMATION

There are several approaches to estimating nonparametric regression models, one of them is "**The local polynomial simple regression".**

Local polynomial It was studied by (Stone-1977, 1980, 1982) and (Cleveland -1979), which forwardly from simple to multiple regression, local polynomial regression generalizes easily to binary and other non-normal data, The aim of local polynomial simple regression is to estimate the regression function (μ|x) at a focal predictor value

(x = x_0).

Local polynomial estimates are computed by weighted least squares regression. Let (x) be some fixed value at which we want to estimate *f*(x). We can approximate a smooth regression function *f*(x) in a neighborhood of the target value (x_0) by the polynomial:

$$y_i = \alpha_0 + \alpha_1(x_i - x_0) + \alpha_2(x_i - x_0)^2 + \ldots + \alpha_p(x_i - x_0)^p + e_i \quad (6)$$

Suppose that locally the regression function (f) can be approximated. For (x) is a neighborhood of (x_0), by using Taylor's expansion we get:

$$f(x) \approx \sum_{j=0}^{p} \frac{r^{(j)}(x)}{j!}(x - x_0)^j \equiv \sum_{j=0}^{p} \alpha_j (x - x_0)^j \quad (7)$$

From equation (3) model *f*(x) locally by a simple polynomial model. This suggests using a locally weighted

polynomial regression to minimize with respect to ($\alpha_0, \cdots \alpha_p$), $\sum_{i=1}^{n}\left\{Y_i - \sum_{j=0}^{P}\alpha_j (X_i - x_0)^j\right\}^2 K_h(X_i - x_0)$

Or;

$$\sum_{i=1}^{n}\left\{Y_i - \alpha_0 - \alpha_1(X_i - x_0) - \cdots \alpha_p (X_i - x_0)^p\right\}^2 K_h\left(\frac{X_i - x_0}{h}\right) \quad (8)$$

Where K (.) denotes a kernel function and (h) is a bandwidth.

Most commonly, the order of the local polynomial is taken as

(p = 0, 1 ,2 ,….) .

If setting p = 0, we get the kernel estimator or the local average estimator which is a special case of local polynomial. If (p = 1), that is, a local linear (fit) estimator. And if (p = 2) that is a local quadratic polynomial or local quadratic regression . . .

6. APPLICATION PART

In this part, we through the application of real data of Sulaimani city between latitude (34 - 36) degrees longitude (45 - 46) the degree of the globe and the rise of sea level (882 m) is located (355) km northeast of the capital Baghdad. The Information of this research gated from Meteorological department in the city as the average of (temperature, humidity, sun shine hour, wind speed and Station pressure) for period from (1/1/2011 to 31/12/2013), with different samples sizes (n1= 365, n2 = 730 and n3 = 1095). The daily average information that is including one response variable

(y_i) and four explanatory variables (X_{ij}) which data descript as follows:

y_i = Average of temperature

X_{i1} = Average of humidity

X_{i2} = Average of sun shine hour

X_{i3} = Average of wind

X_{i4} = Average of Station pressure

The results of methods with statistical analysis as follows:

Sample (1):

First: Simple Regression

Sample 1: The regression equation is:

Temp. = 8.96 + 0.0514 (hum.)

Predictor	Coef.	SE Coef.	T	P
Constant	8.9594	0.4365	20.53	0.000
hum	0.05141	0.05559	0.92	0.356

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	11.14	11.14	0.86	0.356
Residual Error	363	4728.12	13.03		
Total	**364**	**4739.26**			

Second: The Polynomial Multiple Regression Result:

Temp. = - 0.149 - 0.698 (hum.) + 0.329 (press.) + 0.0486 (sun.) + 0.0325(wind.)

With MSE= 0.227, R^2=0.77 and Analysis of Variance:

Source	DF	SS	MS	F	P
Regression	4	278.056	69.514	305.70	0.000
Residual Error	360	81.862	0.227		
Total	**364**	**359.918**			

Sample (2):

First: The regression equation is: temp. = 0.0008 - 0.837 (hum.)

Predictor	Coef.	SE Coef.	T	P
Constant	0.00076	0.02026	0.04	0.970
humz	-0.83748	0.02027	-41.31	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	512.05	512.05	1706.57	0.000
Residual Error	729	218.73	0.30		
Total	**730**	**730.78**			

Second: The Polynomial Multiple Regression Equation is:

Temp. = - 0.0008 - 0.702 (hum.) + 0.399 (press.) + 0.0120 (sun.) + 0.0181 (win.)

With MSE= 0.048, R^2=0.84 and Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	616.17	154.04	975.73	0.000
Residual Error	726	114.62	0.16		
Total	**730**	**730.78**			

SAMPLE (3):

First: The simple regression equation is: Temp. = 0.0005 - 0.628 (hum.)

Predictor	Coef.	SE Coef.	T	P
Constant	0.00046	0.01636	0.03	0.978
humz	-0.62817	0.01222	-51.39	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	774.80	774.80	2640.83	0.000
Residual Error	1094	320.97	0.29		
Total	**1095**	**1095.77**			

Second: The regression equation is:

Temp. = - 0.0003 - 0.537 (hum.) - 0.0016 (sun.) + 0.0265 (wind.) + 0.416 (vap.)

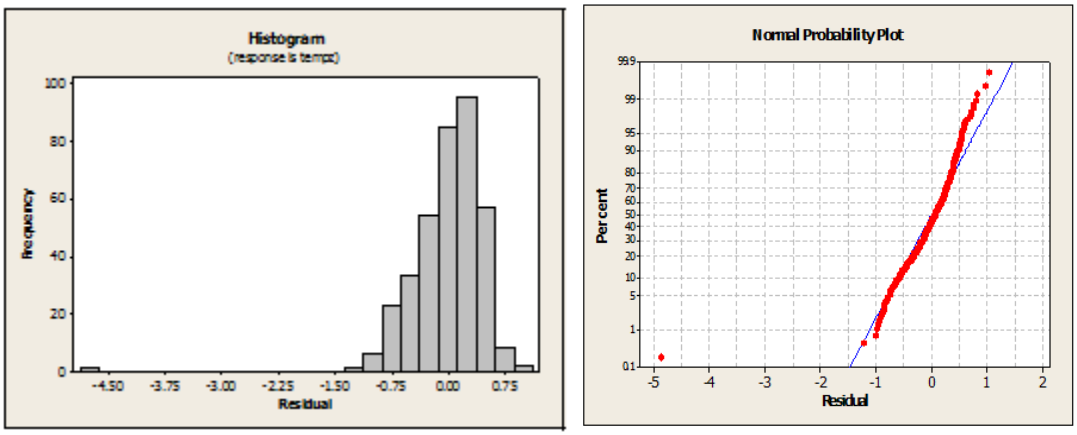
With MSE=0.13, R^2=0.86 and Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	4	950.81	237.70	1789.06	0.000
Residual Error	1091	144.96	0.13		
Total	**1095**	**1095.77**			

Table 1: Comparing the Results between Linear and Nonparametric Regressions with (N1= 365 and H = 0.08, 0.1, 0.02)

	Simple Linear Regression		**Multiple Linear Regression**		**Nonparametric With Simple Local Polynomial Regression**				**Nonparametric Multiple Regression**	
	R^2	MSE	R^2	MSE	Kernel fu.	h=	R^2	MSE	R^2	MSE
	0.67	0.32	0.773	0.95	Epanechikov	0.08	0.75	0.245	0.99	0.088
					Gaussian	0.08	0.75	0.238	0.99	43.252
					Quadratic	0.08	0.74	0.248	0.99	0.072
					N.W	0.08	0.75	0.242	0.99	0.012
					Epanechikov	0.1	0.75	0.240	0.985	0.014
$n_1 =$ 365					Gaussian	0.1	0.75	0.237	0.97	42.35
					Quadratic	0.1	0.74	0.248	0.988	0.012
					N.W	0.1	0.76	0.230	0.98	0.084
					Epanechikov	0.02	0.70	0.288	0.99	0.001
					Gaussian	0.02	0.73	0.262	0.99	53.66
					Quadratic	0.02	0.71	0.288	0.975	0.001
					N.W	0.02	0.71	0.289	0.99	0.001

From above table we can see, (R^2 and MSE) appear that (N.W. and Gaussian, with bandwidth (h= 0.1) models are the best choose with simple local polynomial regression and also, appear that (N.W. and Epanechikov, with bandwidth (h= 0.02) models are the best choose with multiple local polynomial regression.



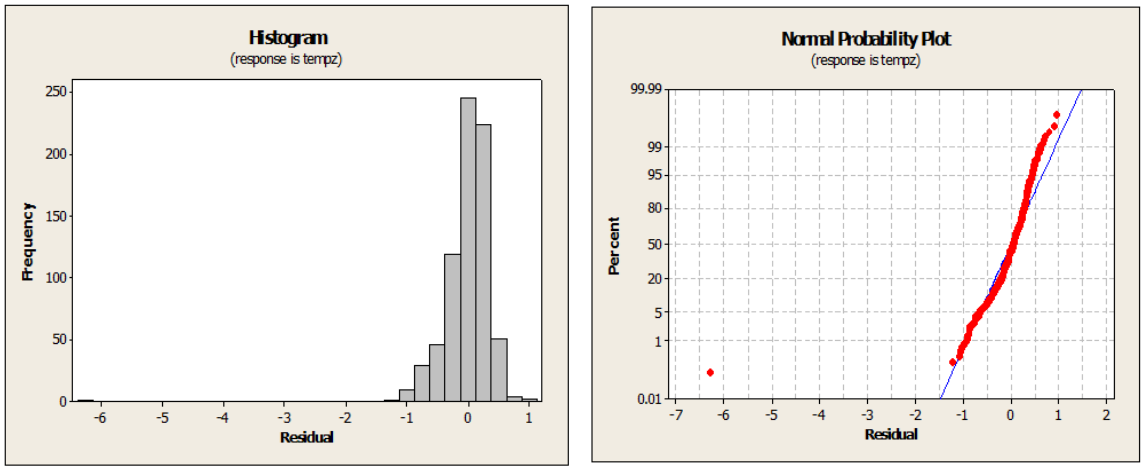
(A) **(B)**

Figure 1: A and B Represent the Test of Residual for Normality with (n1= 365)

Table 2: Comparing the Results between Linear and Nonparametric Regressions with (N2= 730 and H= 0.08, 0.1, 0.02)

	Simple Linear Regression		Multiple Linear Regression		Nonparametric Simple Regression				Nonparametric Multiple Regression	
	R^2	MSE	R^2	MSE	Kernel fu.	h=	R^2	MSE	R^2	MSE
$n_2 = 730$	0.70	0.30	0.84	0.72	Epanechikov	0.08	0.77	0.227	0.99	0.004
					Gaussian	0.08	0.78	0.222	0.99	20.29
					Quadratic	0.08	0.77	0.230	0.98	0.004
					N.W	0.08	0.79	0.221	0.97	0.001
					Epanechikov	0.1	0.77	0.224	0.98	0.020
					Gaussian	0.1	0.78	0.222	0.99	22.07
					Quadratic	0.1	0.77	0.226	0.99	1.69
					N.W	0.1	0.79	0.222	0.996	0.004
					Epanechikov	0.02	0.73	0.271	0.99	0.001
					Gaussian	0.02	0.76	0.237	0.99	20.75
					Quadratic	0.02	0.73	0.271	0.99	0.001
					N.W	0.02	0.73	0.271	0.98	0.002

From above table we can see, (R^2 and MSE) appear that (N.W with (h= 0.1), (Epanechikov and Quadratic) with (h= 0.02) models are the best choose with simple local polynomial regression and also, appear that (N.W.) with (h= 0.1), Epanechikov and Quadratic with bandwidth (h= 0.02) models are the best choose with multiple polynomial regression.



(A) **(B)**

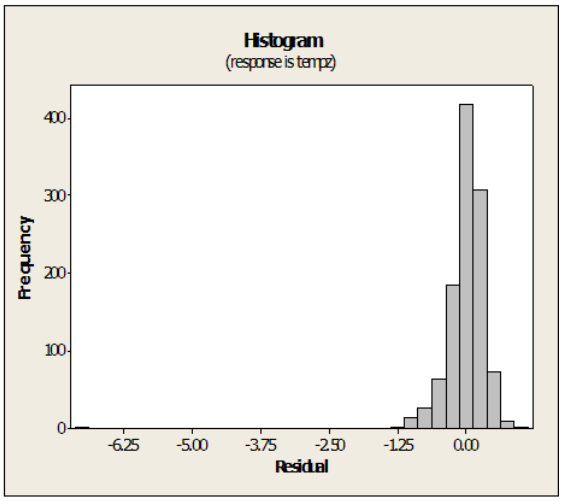
Figure : 2 A and B Represent the Test of Residual for Normality With (n2= 730)

Table 3: Comparing the Results between Linear and Nonparametric Regressions with (n3= 1095 and h=0.08, 0.1 and 0.02)

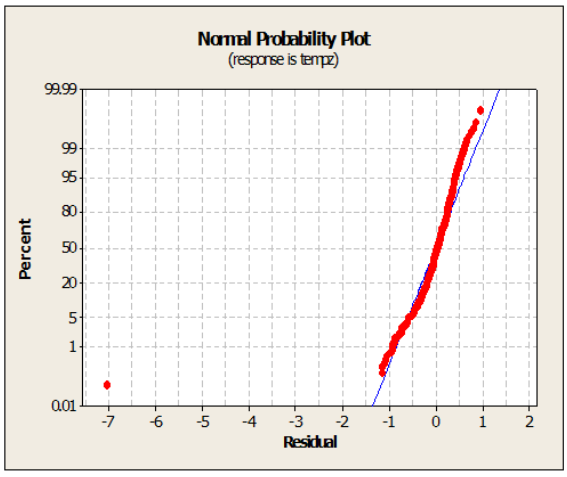
	Simple Linear Regression		Multiple Linear Regression		Nonparametric Simple Regression				Nonparametric Multiple Regression	
	R^2	MSE	R^2	MSE	Kernel fu.	h=	R^2	MSE	R^2	MSE
$n_2 = 1095$	0.70	0.29	0.868	0. 65	Epanechikov	0.08	0.77	0.228	0.998	0.002
					Gaussian	0.08	0.77	0.224	0.997	13.191
					Quadratic	0.08	0.77	0.230	0.998	0.001

					N.W	0.08	0.78	0.226	0.99	0.003
					Epanechikov	0.1	0.77	0.227	0.99	2.071
					Gaussian	0.1	0.77	0.223	0.99	2.144
					Quadratic	0.1	0.77	0.228	0.99	3.029
					N.W	0.1	0.75	0.226	0.99	0.003
					Epanechikov	0.02	0.74	0.255	0.99	0.02
					Gaussian	0.02	0.76	0.237	0.99	0.02
					Quadratic	0.02	0.75	0.255	0.99	0.02
					N.W	0.02	0.74	0.255	0.99	0.001

From above table we can see, (R^2 and MSE) appear that (N.W with h= 0.08) and (Gaussian with h= 0.1) models are the best choose with simple local polynomial regression and also, appear that (Quadratic and Epanechikov) with (h= 0.08) and (N.W., Epanechikov, Quadratic) with bandwidth (h= 0.02) models are the best choose with multiple polynomial regression.



(A)



(B)

Figure 3: A and B Represent the Test of Residual for Normality with (n3= 1095)

7. CONCLUSIONS

1. **Simple Regression**

I. The (N.W.) Nonparametric with simple local polynomial regression with bandwidth (h =0.1) is better than kernel functions Gaussian, Epanechikov, Quadratic and simple linear regression models with (n1= 365).

II. By compare between (simple local polynomial estimators) by (N.W., Epanechikov, Gaussian and Quadratic) Kernel functions for same bandwidth value and simple linear regression, we see that nonparametric regression estimators is the better.

III. (N.W.) simple local polynomial with bandwidth (h =0.1) with the small sample size and (N.W.) with bandwidth (h =0.08) with the large sample size is the best choose

2. Multiple (Local Polynomial) Regression

I. II. The (N.W.) Nonparametric with multiple polynomial regression with bandwidth (h =0.02) is better than kernel functions Epanechikov, Gaussian, Quadratic and multiple linear regression models with (n1= 365).

II. The (multiple polynomial estimators) with (N.W., Epanechikov, Gaussian and Quadratic) Kernel functions for same bandwidth value is better than multiple linear regression.

III. In higher degree (order) of local polynomial (p= 4-dimenssions) we find (N.W.), Epanechikov and Quadratic) are the best estimators with minimum value of (MSE= 0.001) with all samples.

IV. By compare between the Epanechikov and Gaussian function with Increasing bandwidth value the (MSE) increasing in both functions,

V. The multiple polynomial model in any types of kernel functions with nonparametric regression is better than parametric regression by the values of (MSE andR^2).

REFERENCES

1. Adrian W. Bowman & Adelchi Azzalini -2004, (Applied Smoothing Techniques for Data Analysis), Oxford University press Inc. New York.
2. Bernard. W. Silverman-1986, (Density Estimation for Statistics and Data Analysis) - Published in Monographs on Statistics and Applied Probability, London.
3. John Fox-2000, (Multiple and Generalized nonparametric Regression), SAGE University paper (131).
4. J. Fan & I. Gijbels-2003, (Local polynomial Modelling and its Applications) Chapman &Hall/ CRC /Taylor & Francis Group / Boca Raton Landon – New York.
5. Larry Wasserman-2006, (All of Nonparametric Statistics), Springer Science +business Media. Inc //in the United States of America (MVY).
6. Luke Keele -2008, (Semi parametric Regression for the Social Sciences) Ohio State University, U.S.A.(John Wiley & Sons, Lty The Atrium, Southern Gate, Chichester, West Sussex PO198SQ, England.
7. P. J. Green & B.W. Silverman-2000, (Nonparametric Regression and Generalized Linear Models, Aroughness penalty approach .Boca Raton Landon New York Washington, D.C.

8. Wolfgang Hardle-1994, (Applied Nonparametric Regression) Humboldt-University at Berlin at Institute. Spandauer Str. 1-D {10178 Berlin.
9. Andreas Buja, Trevor Hastie, Robert Tibshirani –(1989), (linear Smoothers and additive models) Bellcore , AT& T Bell Laboratories and University of Toronto- The Annuls of Stahstics-1989,Vol. 17, No. 2, 453-555.
10. Anders Stenman- (1999) (Model on Demand: Algorithms, Analysis and Applications) Linköping Studies in Science and Technology. Dissertations No. 571 Department of Electrical Engineering Linköping University Sweden.
11. Armando Hoare-(2008) (Parametric, non-parametric and statisticalmodeling of stony coral reef data) - Theses and Dissertations USF Graduate School Graduate School, 6-1-2008 , University of South Florida.
12. Balaji Rajagopalan & Upmanu Lall. (Locally Weighted Polynomial Estimation of Spatial Precipitation)-Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 44-51, 1998.
13. Dirk Ormoneit & Trevor Hastie – (1999), (Optimal Kernel Shapes for Local Linear Regression) grant DMS of Health grant.
14. David Ruppert-(1996), (Local polynomial Regression and its applications in Environmental statistics), this research was supported by NSA Grant MDA 904-95-H-1025and NMS-9306196.
15. Gery Geenens-(2011), (Curse of dimensionality and related issues in nonparametric functional regression) Statistics Surveys Vol. 5 (2011) 30–43 ISSN: 1935-7516 DOI: 10.1214/09-SS049 e-mail: ggeenens@unsw.edu.
16. Jean D. Opsomer, David Ruppert –(1997) (Fitting a Bivariate Additive Model by Local polynomial Regression), The Annals of Statistics, Vol. 25,No. 1, 186-211.
17. John Fox-(2002), (Nonparametric Regression)- Appendix to An R and SPLUS Companion to Applied Regression. January 2002.
18. John Fox-(2004), (Nonparametric Regression) Department of Sociology McMaster University - February 2004.