Raymond Manna Bangura et al. 2019, 7:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Curve Smoothing in a Local Polynomial: Local Weighted Error Sum of Squares (Lowess)

Raymond Manna Bangura
Biometric Unit/Statistics Unit
Sierra Leone Agricultural Research Institute
Freetown
Sierra Leone
mannahray@yahoo.com

Sahr Milton John Bull
Mathematics Department
Milton Margai College of Education and Technology
Freetown
Sierra Leone
sahrmiltonjohnbull@gmail.com

Abstract- The objective of this paper is to provide a summary approach to curve fitting in a local polynomial; local weighted error sum of squares. We proposed a fit diagnostics for the value Y and also compared quadratic and linear interpolation method in a local polynomial of second order degree. Again, we re-established the fact that curve fits better than line interpolations of a given set of points.

Keywords- Approximation, Bandwidth, Interpolation, Nonparametric and Polynomial.

I. INTRODUCTION

Local weighted error of sum of squares has been used to approximate residuals and expected value from a given set of observation [15]. They are relatively local, in the sense that the position and orientation of the fitted curve in any vicinity of X values is primarily dependent upon the data points in that vicinity [10]. Nonparametric statistical procedures had also been used in order to aid this process of approximation, especially in a local polynomial regression of degree K.

But, for this process of curve fitting, the local weighted error of sum of squares (lowess) has proven to be more accurate, since it does not take into consideration the step by step orientation, which monitors change in each independent variable that gives different levels of output. Because of it nonparametric status, it can measure the relationship that exists between a given set of observation within the same neighborhood. It might be more realistic to partition the range of values into disjoint regions and to approximate the relationship by a sequence of sub-models which are smoothly connected, in some sense, at the boundaries of neighboring regions [4].

The aim of this research is to clearly demonstrate a diagnostic fit of Y values and its predicted values and to briefly compare both direct fit method and the general smoothing method (quadratic) of the same local polynomial of second degree with 95% confidence interval. We obtained an approximation for the expected value of Y, through both linear and quadratic interpolation and a local weighted

error of sum of squares in order to fit a smooth curve from the data points. We further investigate the comparison between the direct fit method and finding an appropriate bandwidth that will best

smooth the curve in consideration. The research also takes into consideration 95% confidence interval to compare both linear and quadratic approximation to deduce which one best fit the data points. Local polynomial could be viewed also, to have the tendency of minimizing variance of the residuals or the prediction error from the data points.

A local weighted error of sum of squares (lowess) is regarded as a nonparametric procedure standard way of approximating, because it does not follow assumptions of normal distribution. It also measures the relationship between the residual term with the change in the independent values by range of linear and quadratic functions. Locally weighted regression is geared towards separately predicting each case and to reduce weight on distinct observations.. One of the greatest advantage of the loess is that they are very flexible about the exact nature of the relationship between the variables [10].The

regression is local in the sense that each one only uses the subset of observations that fall closest to that evaluation point along the horizontal axis of.

II. BODY TEXT

The primary data was taken from a study that was conducted in 2014-2015 of rice varieties. A portion was taken in order to provide some vital information on local polynomial, especially for mathematics or science students and even researchers who may find this topic very interesting. We extracted fifteen observations, which were used to demonstrate some graphics.

In order to compare both linear and quadratic interpolation, the researchers used two software applications. The reason behind this is to provide a sufficient data representation of results which could clearly distinguish one representation from on system of analysis to the other, when using the same observation (data points). A statistical analysis system (SAS) was used to give us a linear interpolation through the direct fit method. Whilst analytic with R is simplified (ARIS) was used to show quadratic interpolation.

III. RESULT AND DISCUSSION

1.General Diagnostics Fit for 'Y'

Figure 1 shows the diagnostics fit for Y. It shows the residual and the predicted value of Y, percent and residual, residual and quantile (σ), proportion less of fit-mean and residual, Y and its predicted value, with a summary of data. The summary data includes; the fifteen (15) observations, 0.9 smooth bandwidth, local boundary points of 13, polynomial degree of 2, residual sum squares of 1260.3, points fit 9 and the type of interpolation within two points.

The residual and the predicted value in figure 1 show that, the residual values is not significantly related to the predicted value of Y and that the there is significantly large information in that data, which has not been captured by the local weighted error sum of squares model. It further shows so many noisy values which may cause distortion in the observation. The diagram showing residual (Y_i-μˆ) versus quantile (σ). (σ), as the critical value of 95% confidence interval best describe the points, as most points lie on the line, with few noisy values.

The diagnostic fit for Y also shows points which can follow a smooth curve for both fit-mean and the residual compared with proportion less respectively. It further shows y values versus the predicted values. The Y and predicted diagram shows many points which are far away from the line (noisy points), which can result to a weak correlation from Spearman's rank coefficient (Nonparametric). We also checked for what Y percentage of all cases which fall within the standard deviation of the mean.

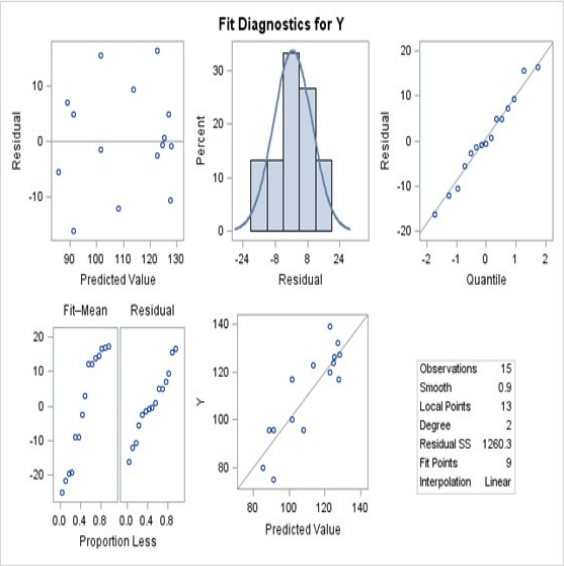

Figure 1: Diagnostics fit for Y value.

2. Scatter Plot Polynomial Regression X And Y

Figure 2 shows a scatter diagram of both X and Y value, with Y following assumptions of normal distribution. A scatter diagram helps to make a few assumptions about the form of relationship that exist between the response and one or more predictor variables. It is seen that

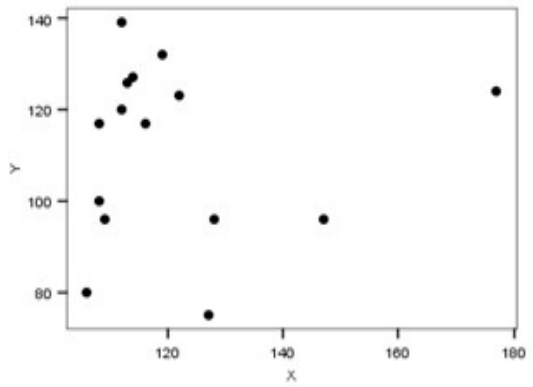

Figure 2: Scatter Plot Polynomial Regression X and Y.

Figure 3 shows a direct fit method through a local weighted error of sum of squares. This method is similar to parametric spline linear function. This is different from spline in the sense that, splines are not generally local and they optimize a global criterion. The values of the bandwidth could be assumed until a smooth curve is been obtained. The sharp connections of points from figure 3 are linearly oriented, because they are directly fitted to each other, but through a local polynomial regression. To acquire a direct fit curve from the given set of data points, a smoothing parameter of 0.9 and a bandwidth of 8 were used with 95% confidence interval of all cases which fall within the standard deviation of the mean. According to Taylor's theorem, any differentiable function can be approximated locally by a straight line, and a twice differentiable function can be approximated by a quadratic polynomial. The straight lines connect one point to the other and attempt to form a curve that best fits the scatter plot of X and Y values. Our next aim is to construct a smooth curve of these connected lines. Locally weighted regression also requires a weight function and a specification of neighborhood size [15]. The tricube function is given equation (6).

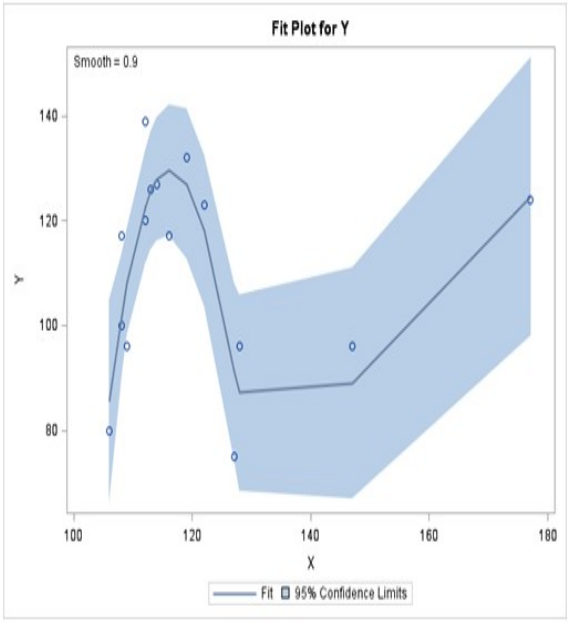

Figure 3: Curve Fitting (Direct fit) from the Scatter Plot X and Y.

Linear approximation

$$\mu(x_i - x) = a_0 + a_1(x_i - x) \quad 1$$

The model for this process of fits is given as

$$Y_i = \mu(x_i) + \varepsilon_i \quad 2$$

$$\varepsilon_i = Y_i - \mu(x_i) \quad 3$$

$$\widehat{\theta}(x) = \operatorname{argmin}_{\theta \in R} \sum_{i=1}^{n} w_i(x)(Y_i - \mu(x))^2 \quad 4$$

w_i is a Kernel function that weights the influence of the ith observation according to the (oriented) distance of x_i from x [13].

$$|x_i - x| < h(x) \quad 5$$

$$\text{weight } (x) = \begin{cases} \left(1 - \left|\frac{x_i - x}{h(x)}\right|^3\right)^3 \\ 0 \end{cases} \quad 6$$

If; $\left|\frac{x_i - x}{h(x)}\right| < 1$ 7

Linear interpolation

$$\widehat{y_0} = \widehat{b_0} + b_0 x_0 \quad 8$$

3. Smooth Curve Fitting

Figure 4 showing the smooth curve fitting of the data points. A second order quadratic polynomial equation was considered in giving a smooth curve among the observation of points.

Quadratic model;

(y_i)ˆ=(b_1)ˆ+(b_1)ˆx_1+(b_2)ˆx_2^2 9

The local quadratic approximation is (Taylor's series)

μ(x_i-x)=a_0+a_1 (x_i-x)+1/2 a_2 (x_i-x)^2 10

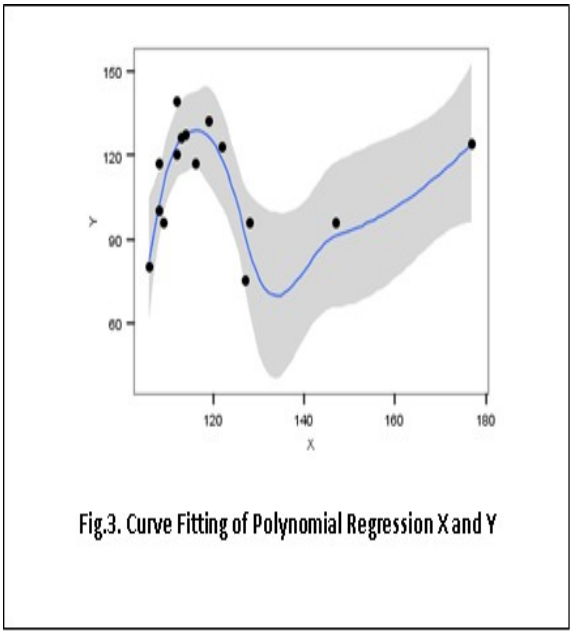

Fig.3. Curve Fitting of Polynomial Regression X and Y

Figure 4: Smooth Curve Fitting with 95% Confidence Interval.

In order to estimate locally for variance, 95% confidence interval was used for both linear and quadratic interpolation. The local weighted error of sum squares function is fitted using weighted least

squares regression with the weight corresponding to each observed Y value calculated using a separate weighting function [12]. A local quadratic estimate can reduce bias, but may face problem of increased variance at boundaries [12].

4. Comparing Linear and Curve Interpolation Methods

Figure 5 shows both linear and quadratic curve of the scatter plot between X and Y values. Each curve showed the line that best fit the scatter plot. Diagram A represents the direct fit (linear approximation) and figure B represents quadratic interpolation among the data points. It could be seen that, figure B is quite similar to figure B.

It may be assumed that diagram B is derived from diagram A. But, this is so, because both curves depend on the smoother parameter and the bandwidth, which is normally determined by the researcher. This should not be misinterpreted as if it were splines when a curve could be obtained from two points in-between segmented or partitioned lines in order to find an approximation to the curve in diagram B. It further shows 95% confidence interval from the original scattered plot in both A and B. These two diagrams were drawn from two different statistical software, namely; Statistical Analysis Systems (SAS) and Analytic with R is Simplified (ARiS). The method used to produce the diagram A was a direct fit with local weighted error of sum of square. Whilst diagram B on the right hand side was done by quadratic interpolation fit method.

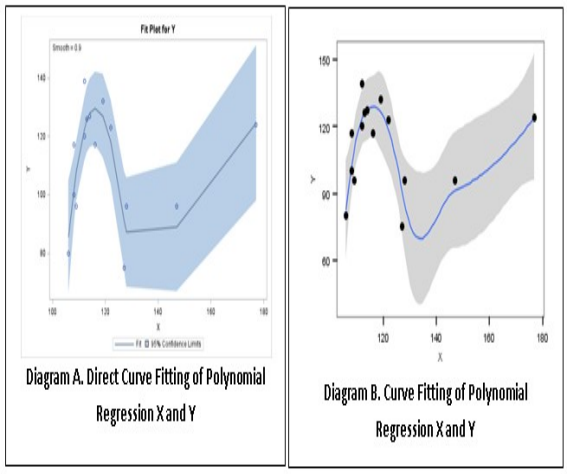

Figure 5: Diagram A and B. Curve Fitting of Polynomial Regression X and Y.

These two diagrams demonstrate that, local polynomial regression of the same degree can have similar shape at some bandwidth and smoother parameter with the same confidence intervals. It could be seen that, the smooth method quadratic function divide the data points better than the direct method of the local polynomial.

IV. CONCLUSION

Scatter plot is normally used to demonstrate a local weighted sum of squares (loess) for smoothing a curve, because of it flexibility potential in the study of empirical results. Local weighted sum of squares (lowess) provides a very flexible approach to the problem of representing structure within a dataset. When lowess is employed merely as a scatter plot smoother, it can be very helpful for a number of important research tasks, including the exploration of vicariate and multivariate data, assessment of functional forms for relationships among variables, examination of model assumptions in regression analysis and representation of complex structures within empirical data [12]. The study was able to find a diagnostics fit for y and a comparative analysis of the two interpolation methods of a local polynomial of second degree order with the same confidence intervals. The research was able to deduce that, the quadratic interpolation of the second degree fits better within the points of the observation at some smoother parameter and bandwidths.

ACKNOWLEDGEMENT

We appreciate the Sierra Leone Agricultural Research Institute (SLARI), for providing data for this short communication. We also appreciate the department of mathematics, Milton Margai College of Education and Technology for providing useful materials for this work. Our profound thanks to families, friends and relatives for their usual supports; this project could not have been completed without their assistance.

REFERENCE

[1]. J. Fan: Local polynomial modeling and its applications: monographs on statistics and applied probability

[2]. W. Cleveland, S. Devlin (1979). Robust locally weighted regression and smoothing scatter plots, J. Amer. Statist. Assoc. 74(368): 829–836.
[3]. W. S. Cleveland and C. Loader (1996). Smoothing by local regression: Principle and method
[4]. J Fan, I. Gijbels, T. C Hu and L. S. Huang (1996). A study of variable bandwidth selection for local polynomial regression
[5]. E. Masry (1996). Multivariate regression estimation of local fitting for time series
[6]. E. Masry (1995). Multivariate local polynomial regression for time series: uniform strong consistency and routes
[7]. P.J.Bickel and B. Li (2007). Local polynomial on unknown manifolds
[8]. Q. Li, X Lu and A. Ullah (2003). Multivariate local polynomial regression for estimating average derivatives
[9]. D. Rupert (1997). Empirical bias bandwidths for local polynomial nonparametric regression and density estimation
[10]. W.G. Jacoby (1999). A nonparametric, graphical tool for depicting relationship between variables
[11]. W. S. Cleveland and S.J Devlin (2007). Locally weighted regression: An approach to regression analysis
[12]. J. Chambers et al, (1999). Local regression likelihood
[13]. D. L. Banks, R.T. Olszewski and R.A. Maxion (1999). Comparing methods of multivariate nonparametric regression
[14]. J.E. Walsh (1962). Handbook of nonparametric statistics
[15]. W. Cleveland and S. Devlin (1988). Robust locally weighted regression and smoothing scatter -plot.

Supplementary Information

Data of X and Y Values

X	Y
147	96
177	124
113	126
119	132
116	117
112	120
108	117
112	139
109	96
114	127
128	96
122	123
106	80
108	100
127	75

Output Statistics

Output Statistics								
Obs.	X	Y	Predicted Y	Estimated Prediction Std Deviation	Residual	t Value	95% Confidence Limits	
147	96	88.90742	9.79118	7.09258	0.72	66.90805	110.9068	
177	124	124.5438	11.86171	-0.54378	-0.05	97.89222	151.1953	
113	126	125.6898	5.03798	0.31016	0.06	114.3702	137.0095	
119	132	127.0311	6.42696	4.96892	0.77	112.5906	141.4715	
116	117	129.668	5.57592	-12.668	-2.27	117.1397	142.1963	
112	120	122.5873	4.78481	-2.58728	-0.54	111.8365	133.3381	
108	117	101.5323	5.30347	15.46771	2.92	89.61615	113.4484	

127	108	106	122	128	114	109	112
75	100	80	123	96	127	96	139
91.25681	101.5323	85.55035	118.109	87.12293	127.942	108.1331	122.5873
7.52465	5.30347	8.64905	6.45548	8.29673	5.22197	4.54656	4.78481
-16.2568	-1.53229	-5.55035	4.89096	8.87707	-0.94202	-12.1331	16.41272
-2.16	-0.29	-0.64	0.76	1.07	-0.18	-2.67	3.43
74.35	89.61615	66.11718	103.6045	68.48137	116.209	97.91767	111.8365
108.1636	113.4484	104.9835	132.6136	105.7645	139.675	118.3486	133.3381