مجلة العلوم البحثة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 15/11/2017 Revised 16/02/2018 Published online 30/06/2018

A comparison of Several Bandwidth Selection Methods for Local Polynomial Regression

*Alsaidi M. Altaher , Ali S. Ambark , Abdslam k. Suliman
Department of Statistics , faculty of Science,University of Sebha, Libya
*Corresponding author: als.altaher@sebhau.edu.ly

Abstract In local polynomial regression, choosing the smoothing parameter (bandwidth) is a crucial issue. A too large value provide over smoothing. Conversely, a too small value gives a wiggly estimate which result in under smoothing. However, the proper choice of bandwidth can be considered as a careful balance of these principles. In this paper, intensive simulation experiments are carried out using R software to compare the practical performance of several bandwidth selection methods, namely the Cross Validation (CV), Generalized Cross Validation (GCV), and Adaptive (ADP).Within the context of these strategies of selecting the optimal bandwidth(s), four different example-regression models have been used under different sample sizes and kernel functions. Results showed that the (GCV) bandwidth selection criterion appears to give better (smaller) estimates of MSE when the sample sizes (n) are small; with Gaussian kernel function. However, the (Adp) bandwidth selection appears to give better (smaller) estimates of MSE when the sample sizes (n) are large with Triweight l kernel function.
Keywords: local polynomial, bandwidth, kernel function, simulation.

مقارنة بين عدة طرق لاختيار معلمة التنعيم لمقدر الانحدار اللامعلمي

*السعيدي المهدي الطاهر و علي صالح أمبارك و عبدالسلام كامل سليمان
قسم الإحصاء – كلية العلوم –جامعة سبها، ليبيا
*للمراسلة:als.altaher@sebhau.edu.ly

الملخص عند تقدير دالة الانحدار باستخدام local polynomial regression فإن اختيار معلمة التنعيم تلعب دورا حساسا. القيم الكبيرة لمعلمة التنعيم تؤدي إلى (over smoothing) بينما القيم الصغيرة جدا تعطي مقدر متذبذب ((wiggly smoothing بينما الاختيار الأمثل سوف يكون أقرب ما يكون لوصف دالة الانحدار.في هذه الورقة تم إجراء العديد من تجارب المحاكاة باستخدام برنامج Rلأجل مقارنة الأداء العملي لعدة طرق لاختيار معلمة التنعيم وهي: (CV), (GCV), (ADP). باستخدام أربعة نماذج انحدار مختلفة وأحجام عينة مختلفة ودوال وزن kernel functions والنتائج بينت أن طريقة (GCV) تعطي نتائج أفضل في حالة أحجام العينة صغير وعند استخدام دالة وزن (Gaussian) بينما طريقة (ADP) تعطي نتائج أفضل في حالة أحجام العينة كبير وعند استخدام دالة وزن.(triweight)

الكلمات المفتاحية: معلمة التنعيم ، دالة الوزن ، المحاكاة.

Introduction

As one basic form of statistical inference, regression analysis has been usually used in discovering the relationship between one quantity (called dependent variable) and one or more other quantities (called explanatory variables). Non-parametric regression eliminates all parametric assumptions (i.e it comes to signify the absence of the parameters in the regression model).There exist many smoothing methods to obtaining non-parametric function. Some of the most widely used in the literature of the smoothing methods are Kernel based smoothing, K-Nearest Neighbor, Spline smoothing, Orthogonal series estimators and Wavelet. Within the context of the Kernel-based smoothing, there are many well-known approaches namely: (Nadaraya-Watson Estimator [Nadaraya and Watson(1964)], Priestley-Chao Estimator[Priestley and Chao(1972)], Gasser-Muller Estimator[Gasser and Muller(1979)], Locally Weighted Scatter Plot Smoother LOWESS[Cleveland (1979)] and Local Polynomial Kernel Estimator[Fan and Gijbels(1995)]). In local polynomial regression, the choice of bandwidth (h) is considered to be the most sensitive topic. One might ask how wide the local neighborhood should be so that the local approximation is a suitable one. If we take a very small bandwidth, the modeling bias will be small since the number of data points falling in this local neighborhood is also small but the variance will be large. On other hand, If we take a very large bandwidth creates a large modeling bias depending on the underlying function. This means that the bandwidth governs the complexity of the model during the trade-off between quantities of bias and variance (Fan and Gijbels (1996)). An extensive literature addresses this problematic subject, especially in the context of nonparametric mean regression. The classical techniques used for mean kernel smoothing, such as cross-validation, plug in, rule-of-thumbs, and bootstrap, also can be used (after adaptation) to select the bandwidth for quintile regression. (For more details, see Yu and Jones 1998; Zheng and Yang 1998; Leung 2005. In this paper, we shall compare the practical performance of several bandwidth selection methods, namely the Cross

Validation (CV),Generalized Cross Validation (GCV), and Adaptive (ADP).Within the context of these strategies of selecting the optimal bandwidth(s), four different example-regression models have been used under different sample sizes and kernel functions.

Method and Material

Local Polynomial Regression: local polynomial regression fits a weighted least squares polynomial locally rather than a weighted average. Thus the usual regression setup is as follows:

The response variables $y_i's$ are modeled as

$$y_i = g(x_i) + \varepsilon_i \quad i = 1,2,...,n$$

Where ε_i are i.i.d random errors from a unimodal symmetric density Centered about 0 and the $g(x)$ is a continuous mean function with continuous derivative. In the matrix notation, for a particular point x_0 we can write

$$X_{p,x} = \begin{bmatrix} 1 & (x_1 - x_0) & \cdots & (x_1 - x_0)^p \\ 1 & (x_2 - x_0) & \cdots & (x_2 - x_0)^p \\ \vdots & \vdots & & \\ 1 & (x_n - x_0) & \cdots & (x_n - x_0)^p \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$w_x = \begin{bmatrix} w_{1,1} & 0 & \cdots & 0 \\ 0 & w_{2,2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & w_{n,n} \end{bmatrix}$$

Here $w_{ij} = \frac{1}{h} K(\frac{x_j - x_i}{h})$ where K is a kernel function (usually a symmetric and has a bounded support). Table 1 displays some poplar kernel functions.

The symbol h is a positive constant referred to as a bandwidth.

In this regard, the local polynomial regression estimate of the mean function $\hat{g}(x)$ is the first element of the $\hat{\beta}_i$ vector given by

$$\hat{\beta}_i = e_1^T (x_p^T w_X x_{p,x})^{-1} (x_p^T w_x y)$$

$$\hat{\beta} = LY$$

Where *L* is called smoothing matrix

$$e_{p*1}^T = (1,0,0,....,0)$$

Table (1) kernel functions

Kernel	Formula	Support
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$	$-\infty < u < +\infty$
Triweight	$K(u) = \frac{35}{32}(1-u^2)^3$	$\lvert u \rvert \le 1$
Box	$K(u) = \frac{1}{2}$	$\lvert u \rvert \le 1$
Epanechnikov (parabolic)	$K(u) = \frac{3}{4}(1-u^2)$	$\lvert u \rvert \le 1$

Bandwidth Selection

• **Leave-One-Out Cross-Validation:** Cross validation is an important idea in regression. The idea is to estimate the smoothing parameter by minimizing cross validation score $CV(h)$:

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{g}_{h,-i}(x_i) \right)^2$$

Where $\hat{g}_{h,-i}$ means the smooth estimate for smoothing parameter h and data $(x_1,....., x_{i-1}, x_{i+1},.....x_n)$ that "leaves out" x_i. Amazing shortcut formula for cross validation score is

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_{h,i}}{1 - l_{ii}} \right)^2$$

Where l_{ii} is the i^{th} diagonal element in matrix L.

• **Generalized Cross Validation:** A minor variant on cross validation is, so-called generalized cross validation, which, of course, like most things statisticians call "generalized," isn't.

It replaces the l_{ii} in the denominator with their average $tr(L)/n$ giving

$$GCV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_{h,i}}{1 - \frac{tr(L)}{n}} \right)^2$$

• **Adaptive bandwidth:** Adaptive bandwidth is obtained by a similar procedure to the one proposed by Fan and Gijbels (1995). The interval is split into $[1.5 \times n/(10 \times \log(n))]$ intervals, a leave-one-out cross validation is performed in each interval to obtain a local bandwidth. These bandwidths are then smoothed to obtain the bandwidth for each point in X.

Experimental Work

Simulation Setup: The purpose of this simulation is to compare the precision of three bandwidth selection methods for local polynomial regression (Leave-One-Out Cross-Validation ,Generalized Cross Validation and Adaptive bandwidth).

To that end, we have used

Four different test functions,

Test function 1: $g(x) = 20e^{\left(-4(x-1)^2\right)}$

Test function 2: $g(x) = \sin(2x) + 2e^{(-16x)} + 0.3$

Test function 3: $g(x) = \frac{3}{4}\sin(\frac{3\pi x}{2} + 1.25)$

Test function 4: $g(x) = 1 + \sin(x) + 2(\cos(x) + 3\sin(5x)$

Three different sample sizes $n = 50, n = 100, n = 200$

Four kernel functions (epanech, box, triweight, gaussian).

Each simulation study involves 1000 repetitions. The Mean Square Error

$$MSE(\hat{g}) = \frac{1}{n}\sum_{i=1}^{n}(g(x_i) - \hat{g}(x_i))^2$$

All computations have been carried out using the R statistical package.

Discussion: Having conducted the simulation runs, results of MSE have been tabulated in Table 2, Table 3 Table 4, and Table 5 and we have observed the following empirical findings:

In most cases Cross validation method performs better than other to methods for small sample sizes with Gaussian kernel function. However Adaptive method performs better than other to methods for large sample sizes or triweight kernel function. In addition All three methods perform differently with respect to kernel function. However, the best estimate is often obtained with Gaussian kernel function. The worst estimate is observed when box kernel function is used.

Table 2: simulation results of MSE for test function1

n	Method	Epanech	box	triweight	Gaussian
50	CV	0.3885604	1.322681	0.2762529	0.1586460
	Adp	0.3835315	1.313034	0.2766261	0.1585023
	GCV	0.3931859	1.308724	0.2751616	0.1578063
100	CV	0.1113023	15.5983826	1.76432933	0.3409894
	Adp	0.1022484	0.5845154	0.09888788	0.5942468
	GCV	0.1104038	15.3128563	1.76238672	0.3383542
200	CV	4.519204	14.11726	0.51067595	0.2610333
	Adp	2.651003	10.76982	0.09238398	0.3031902
	GCV	4.505492	13.94645	0.50629880	0.2617503

Table 3: simulation results of MSE for test function 2

n	Method	epanech	box	triweight	gaussian
50	CV	0.9215217	1.036236	0.8822267	0.7852551
	Adp	0.9185002	1.046121	0.8820987	0.7855062
	GCV	0.9133377	1.045376	0.8777225	0.7849031
100	CV	1.4032346	1.802942	0.9318802	0.8889426
	Adp	0.9085951	1.334389	0.8020218	0.8567398
	GCV	1.4067769	1.773179	0.9315370	0.8886935
200	CV	0.9472898	1.088610	1.218585	0.9137279
	Adp	0.8860952	0.965193	1.093466	0.8747591
	GCV	0.9473895	1.090380	1.219794	0.9139164

Table 4: simulation results of MSE for test function 3

n	Method	Epanech	box	triweight	Gaussian
50	CV	1.094858	8.271399	1.436837	0.8029657
	Adp	1.101221	**8.247743**	1.440555	0.8029069
	GCV	**1.086157**	8.389363	**1.399306**	**0.8026283**
100	CV	0.9293517	3.906464	1.2342131	0.8893607
	Adp	**0.8629295**	**3.093572**	**0.9647979**	**0.8605289**
	GCV	0.9264548	3.811529	1.2213032	0.8880302
200	CV	0.9748502	1.1220330	0.9425849	0.9184896
	Adp	**0.8974806**	**0.9531895**	**0.8279044**	**0.8814559**
	GCV	0.9750830	1.1219109	0.9428955	0.9184919

Table 5: simulation results of MSE for test function 4

N	Method	epanech	box	triweight	Gaussian
50	CV	0.8978851	1.416605	0.9484048	0.7803267
	Adp	0.8965303	**1.412582**	0.9502768	**0.7797178**
	GCV	**0.8827724**	1.419494	**0.9403477**	0.7833331
100	CV	0.9661152	1.0938644	0.8613612	0.8918565
	Adp	**0.8838907**	**0.9636717**	**0.7760621**	**0.8513759**
	GCV	0.9650178	1.0870621	0.8641446	0.8921334
200	CV	0.9745632	1.584137	0.9537848	0.9397907
	Adp	**0.9011993**	**1.284437**	**0.8406634**	**0.8856389**
	GCV	0.9746647	1.589106	0.9537880	0.9400075

Conclusion: This paper considered the issue of choosing the bandwidth for local polynomial regression. Three methods have been investigated (CV,GCV,Adp).Results showed that the (GCV) bandwidth selection criterion appears to give better (smaller) estimates of MSE when the sample sizes (n) are small; with Gaussian kernel function. However, the (Adp) bandwidth selection appears to give better (smaller) estimates of MSE when the sample sizes (n) are large with Triweight l kernel function.In addition The best estimate is often obtained with Gaussian kernel function. The worst estimate is observed when box kernel function is used.

References

[1]- Abdelkder A. K. (2009). A Comparison Between Global and Adaptive Bandwidths Using(DPI) and (RSC) Selection Methods. Unpublished M.Sc thesis. Dept of Statistics, University of Garyounis ,Libya.

[2]- Fan, J., and Gijbels, I (1995a). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adoption. Journal of Royal Statistical Society (B)57 ,No.2,371-394.

[3]- Fan, J., and Gijbels, I (1995b). Adaptive order Polynomial fitting: bandwidth robustification and bias reduction. Journal of Computational and Graphical Statistics 4,213-277.

[4]- Fan, J., and Gijbels, I. (1996). Local Polynomial Modeling and Its Applications. Chapman& Hall. London.

[5]- Mami, A.(2002). Local Polynomial Regression with Applications To both Independent and Longitudinal Data. Unpublished Ph.D thesis Dept. of Statistics, University of Manchester United Kingdom.

[6]- R Development Core Team, R: A Language and environment for statistical computing, R Foundation for statistical computing Vienna, ISBN 3-900051-07-0, 2005.

[7]- Wand, M. P. and Jones, M. C. (1995). Kernel Smooting .Chapman&Hall London.

[8]- Yang, L. (2006).Have you considered nonparametric regression? Michigan State University 1,1-2.