AJAS / SPAS

African Journal of Applied Statistics
Vol. 7 (2), 2020, pages 993 - 1000.
DOI: http://dx.doi.org/10.16929/ajas/2020.993.252

ISSN 2316-0861

A note on Nonparametric Regression Modeling using a Density Function

Jakperik Dioggban [1,*]

[1]Address Department of Statistics, Faculty of Mathematical Sciences, CK Tedam University of Technology and Applied Sciences, Box 24, Navrongo, Ghana

Received on February 11, 2020. Accepted on December 01, 2020

Abstract. The nonparametric regression offers alternative to classical regression analysis when the data are not well behaved or when the classical regression model throws significant lack of fit. It has in recent times been estimated using Kernel estimators and the smoothing splines, but these methods wields some bias of estimation. In this study, a semi-parametric multiplicative bias reduction density function was used to develop a non parametric regression model. Simulation studies conducted showed that the proposed estimator performs better than both the Kernel and the smoothing splines estimators especially with large samples.

Résumé (Abstract in French) La régression nonparamétrique offre une alternative à la regression classique lorsque les données ne sont pas conformes ou lorsqu'il y a defaut d'ajustage. Récemment, cette méthode a été appliquée avec des noyaux et des lissage de splines, avec apparition de biais importants. Ici, nous présentons une méthode semi-paramétrique multiplicatif conduisant à une réduction de biais. Une étude de simulation a été menée pour supporter la méthode, spécialement avec de grands échantillons.

Key words: nonparametric regression; multiplicative semi-parametric bias reduction density function; density estimator; risk analysis;
AMS 2020 Mathematics Subject Classification Objects : 62G07; 62G08.

Corresponding Author: jdioggban@cktutas.edu.gh

The author.

Jakperik Dioggban, Ph.D., is a lecturer at the Department of Statistics, Faculty of Mathematical Sciences, CK Tedam University of Technology and Applied Sciences, Box 24,Navrongo, Ghana.

1. Introduction

Nonparametric regression models the dependent variable to the independent variable(s) without explicitly specifying the functional form in advance. It has a general form

$$E\left(y_i\right) = f\left(x_{1i}, ..., x_{pi}\right)$$

with similar restrictions on the parameter in regression, however, these restrictions are relaxed in applications Silverman (1986), Green and Silverman (1993).

Unlike the parametric regression analysis which emphasis on the estimation of the parameters of the regression equation, non parametric regression estimates the equation directly. Most studies have commonly used Kernel estimation, local polynomial regression, and smoothing splines for non parametric regression Fox (2000). The efficiency of the non parametric function being dependent on the rate of convergence of these density estimation methods; hence reducing its bias and improving precision. Smoothing splines regression estimators are better than those of the kernel regression Aydn (2007).

In classical regression analysis, the bias stems from estimates of coefficients, largely emanating from violation of the least squares estimation assumptions. The non parametric regression however, uses estimate of an unknown smooth function to determine the smooth equation. Several approaches have been used in estimating non parametric smooth function albeit with bias. This could be as a result of boundary bias, or biases accruing from slow convergence of the density functions used in the estimation process. This study seeks to address this inherent bias.

The outline of the paper is as follows: in section 2 the various approaches to nonparametric regression are discussed, the proposed method used in this study and derivation of its theoretical properties are accomplished in section 3, simulation study is conducted in section 5 which is preceded by the performance criteria of the models in section 4. Conclusions are in section 6.

2. Approaches to non Parametric Modeling

Assuming a non parametric regression model of the form

$$y_i = f\left(x_i\right) + \epsilon_i, a < x_i < ... < x_n < b$$

with $f \in C^2[a,b]$ an unknown smooth function, $(y_i)_{i=1}^n$ being observed values of the response variable, y, $(x_i)_{i=1}^n$ represents observed values of the explanatory variable, x whilst $(\epsilon_i)_{i=1}^n$ are normally and independently distributed random errors with mean zero and constant variance, σ^2 Aydn (2007).

The interest in non parametric regression is to estimate unknown function $f \in C^2[a,b]$ of all functions, f that has continuous first and second derivatives. Since this f is unknown, a data-driven technique is used to determine the best form of f. Two most common methods for used in these estimations are the smoothing splines and kernel regressions. The next subsections discuss these methods.

2.1. Smoothing Splines Regression

In smoothing splines regression, the estimate of function f is obtained as a solution to a minimization problem

$$\hat{\mathbf{f}}_\lambda = \mathrm{s}_\lambda \mathbf{y} \tag{1}$$

f_λ is a natural cubic splines with knots $x_1, ..., x_n$ for a fixed smoothing parameter, $\lambda > 0$, s_λ is a well-known positive-definite symmetrical smoother matrix that depends on λ and the knots $x_1, ..., x_n$, but not $\mathbf{y}$ Wahba (1990), Green and Silverman (1993), Eubank (1999).

2.2. Kernel Regression

In this approach, estimate of the regression function, f is obtained as a function of a weighted average of the raw data, with weights being a decreasing function of relative distance. This weighting scheme is the so-called Nadaraya-Watson estimator Nadaraya(1965), Watson (1964) associating the observations y_i for prediction at x_i given by

$$\begin{aligned} w_{ij} &= \frac{K\left(\frac{x_i - x_j}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_j}{h}\right)} \\ &= \frac{K(u)}{\sum_{j=1}^n K(u)} \end{aligned} \tag{2}$$

$K(u)$, the kernel function is a decreasing function of u, $h > 0$ is the bandwidth. Usually the kernel function, $K(u)$ is a probability density function. Commonly used densities are the normal, Epanechnikov, biweight, triweight among others Yatchew (2003), Wand and Jones (1994). The kernel estimate of the function for the regression model at distinct x_i points is

$$\begin{aligned} \hat{y}_i &= \hat{f}(x_i) \\ &= \sum_{j=1}^n w_{ij} y_j \\ &= w_i^{'} \mathbf{y}, i = 1, 2, ..., n \end{aligned} \tag{3}$$

where each data point has a distinct weight, $w_{ij}, j = 1, 2..., n$ for all points of $f(x_i)$. It can be represented in matrix form as

$$\hat{\mathbf{f}} = \mathbf{WY}$$

where $\mathbf{W}$ is a *kernel hat matrix or kernel smoother matrix*. The disadvantage of this estimator is that, it suffers from poor bias at the boundaries of inputs domain. It is largely due to the poor asymmetry of the kernel weights in the boundary regions.

3. Semi-parametric Multiplicative Bias Reduction Density Function

In this study, a semi-parametric multiplicative bias reduction density estimator is used in the estimation of the non-parametric smooth equation. The general form of the density is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{f\left(x, \hat{\theta}\right)}{f\left(X_i, \hat{\theta}\right)} \tag{4}$$

where $K_h(X_i - x)$ is the kernel, n is the sample size, $f\left(x, \hat{\theta}\right)$, and $f\left(X_i, \hat{\theta}\right)$ are non-parametric and parametric density estimators respectively.Details and properties of this estimator can be found in Jakperik *et al.*(2018).

3.1. Risk Analysis

This examines the accuracy of the density function estimation. The accuracy of the density function determines the precision of the estimator. The following definitions are stated.

Definition 1 (Holder class). Suppose $X_i \in \chi \subset R$ where χ is a compact. Let β and L be positive numbers. Given a vector $s = (s_1, ..., s_d)$, define $|s| = s_1 + ... + s_d, s! = s_1!...s_d!, x_1^s...x_d^{s_d}$ and

$$D^s = \frac{\partial^{s_1+\cdots+s_d}}{\partial x_1^{s_1}\ldots\partial x_d^{s_d}}$$

Suppose β is a positive integer. Define the Holder class

$$\Sigma(\beta, L) = \{g : |D^s g(x) - D^s g(y)| \leq L\|x - y\|, \forall s \quad such \quad that |s| = \beta - 1, and \quad \forall \quad x, y\}$$

If $g \in \Sigma(\beta, L)$ then $g(x)$ is close to its Taylor series approximation:

$$|g(u) - g_{x,\beta}(u)| \leq L\|u - x\|^{\beta}$$

where

$$g_{x,\beta}(u) = \Sigma_{|s|\leq\lfloor\beta\rfloor}\frac{(u - x)^s}{s!}D^s g(x)$$

Demnati and Rao(2004)

Proposition 1 (Bias Risk Bound). *The bias of $\hat{f}\left(x\right)$ satisfies:*

$$sup_{f\in\Sigma(4,L)} \mid \hat{f}\left(x\right)-f\left(x\right)\mid\leq ch^4 \tag{5}$$

for some c.

Proof. The bias is defined as

$$\begin{aligned}\mid \hat{f}\left(x\right)-f\left(x\right)\mid &= \int\frac{1}{h}K\left(\|X_i-x\|/h\right)dx-f\left(x\right)\\ &= \mid\int K\left(\|z\|\right)\left(f\left(x+hz\right)-f\left(x\right)\right)dz\mid\\ &\leq \mid\int K\left(\|z\|\right)\left(f\left(x+hz\right)-f_{x,4}\left(x+hz\right)\right)dz\mid+\mid\int K\left(\|z\|\right)\left(f_{x,4}\left(x+hz\right)-f\left(x\right)\right)dz\mid\end{aligned}$$

The first term is bounded by $Lh^4\int K\left(s\right)\mid s\mid^4$ since $f\in\Sigma\left(4,L\right)$ as stated in definition above. The second term is 0 from the properties on K since $f_{x,4}\left(x+hz\right)-f\left(x\right)$ is a polynomial of degree less than 4 with no constant term.

Proposition 2 (Variance Risk Bound). *The variance of $\hat{f}\left(x\right)$ satisfies*

$$sup_{f\in\Sigma(4,L)}Var\left(\hat{f}\left(x\right)\right)\leq\frac{c}{nh} \tag{6}$$

for some $c>0$.

Proof. The density can be written as

$$\hat{f}\left(x\right)=\frac{1}{n}\Sigma_{i=1}^{n}Z_i$$

where

$$Z_i=\frac{1}{h}K\left(\frac{\|X_i-x\|}{h}\right)$$

Thus,

$$\begin{aligned}Var\left(Z_i\right) &\leq E\left(Z_i^2\right)\\ &= \frac{1}{h^2}\int K^2\left(\frac{\|X_i-x\|}{h}\right)f\left(x\right)dx\\ &= \frac{h}{h^2}\int K^2\left(\|z\|\right)f\left(x+hz\right)dz\\ &\leq \frac{sup_x f\left(x\right)}{h}\int K^2\left(\|z\|\right)dz\leq\frac{c}{h}\end{aligned}$$

for some c since the densities in $\Sigma\left(4,L\right)$ as stated above in definition above are uniformly bounded. Then

$$sup_{f\in\Sigma(4,L)}Var\left(\hat{f}\left(x\right)\right)\leq\frac{c}{nh}$$

Since c is chosen to optimize convergence and is usually small and positive, the ratio $\frac{c}{nh}$ for n large and h close to 0 ensures the variance is always kept minimum.

Proposition 3 (Risk bound for Mean Square Error). *The MSE has a risk bound of* $\left(\frac{1}{n}\right)^{-\frac{8}{9}}$

Proof. Since the $MSE = bias^2 + variance$ implies $MSE \preceq h^8 + \frac{1}{nh}$ since $h \asymp n^{-\frac{1}{9}}$ then

$$sup_{f\in\Sigma(4,L)} E \int \left(\hat{f}(x) - f(x)\right)^2 dx \preceq \left(\frac{1}{n}\right)^{-\frac{8}{9}}$$

The bounded condition talks about how the bias of a function is affected by its smoothness. This therefore increases the precision of the proposed non-parametric regression estimator and makes it suitable for application under varying conditions and different sampling designs.

4. Performance criteria of the models

This evaluates the closeness of prediction value to the observed value. The prediction consistency criteria used in this study are Mean Square Error (MSE) (or Root Mean Square Error$(RMSE)$), Mean Absolute Error (MAE), and Mean Absolute Percentage Error $(MAPE)$ as defined below

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \\ or \;\; RMSE &= \sqrt{MSE} \end{aligned}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{|y_t|} \times 100$$

Aydn (2007)

5. Simulations

The study used data from the Ghana Living Standards Survey Round 6 Service (2014) for the simulation study. It relied on two samples of sizes 250 and 500 drawn from the population and fitted to non-parametric regression using the above mentioned smoothing approaches over 1000 iterations. In each case, the accuracy measures shown in table 1 and table 2 were computed for sample sizes 250 and 500 respectively. The results of these simulations are given below.

Consistently, the proposed estimator performed better than both the Nadaraya-Watson and Smoothing Splines estimators especially as the sample size increased from 250 to 500. The proposed estimator does not offer extreme efficiency for small sample sizes, thus its performance was notably close to that of the smoothing

Table 1. ACCURACY MEASURES FOR THE THREE ESTIMATORS

Accuracy Measure	Nadaraya-Watson	Splines	Propose Estimator
MSE	579.4247	497.3547	490.8527
MAE	40.40317	34.0067	34.0571
MAPE	1550.1630	1547.3250	1554.9681

Table 2. ACCURACY MEASURES FOR THE THREE ESTIMATORS

Accuracy Measure	Nadaraya-Watson	Splines	Propose Estimator
MSE	453.3581	389.7261	350.4792
MAE	31.5483	28.3921	25.7640
MAPE	1497.0398	1458.8765	1437.9376

splines estimator. In fact, the smoothing splines estimator was superior to the proposed estimator under the sample size 250 for MAE. This comes as no surprise as the efficiency of the proposed estimator is expected to improve with increasing sample size since it is a large sample estimator. Notably also, is the efficiency of the smoothing splines estimator which closely matches that of the proposed estimator and hence offers a better alternative to the Nadaraya-Watson estimator Aydn (2007). Clearly, the Nadaraya-Watson estimator is seemingly inferior to its competitors in this study, especially as the sample size increases. Therefore, the proposed estimator is preferred for non-parametric regression estimation in cases where there are large samples with high heterogeneity.

6. Conclusions

A new estimator for non-parametric regression based multiplicative semi-parametric density function is proposed. The proposed estimator was compared to the Non-parametric regression based on the kernel and the smoothing splines estimators. The multiplicative semi-parametric density function produced estimates with high precision than its competitors. It however produces sparing results for small samples. Its accuracy measures improves with increasing sample sizes and thus is ideal for large sample estimations.

References

Aydn, D.(2007) A comparison of the nonparametric regression models using smoothing splines and kernel regression. World Academy of Science, Engineering and Technology 36.

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*. Vol 30, 17-34

Eubank, R. L. (1999) *Nonparametric regression and spline smoothing*. CRC press

Fox, J. (2000) *Nonparametric simple regression: Smoothing scatterplots*. Number 130. Sage.

Green, P. J. and Silverman, B. W. (1993) *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall/CRC

Jakperik D., Odhiambo, R. O., Orwa G. O., 2018. A semi-parametric multiplicative bias reduction density with a parametric start. *Advances in Statistics and Applications*, Vol 53(6):715-729.

Nadaraya, E., 1965. On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, Vol(1):186-190.

Service G. S. (2014). Ghana Living Standards Survey Round 6 $(GLSS \quad 6)$:Poverty profile in Ghana (2005 -2013). *Ghana Statistical Service*.

Wahba, G., 1990. Spline models for observational data. *Siam*, Vol 59.

Wand, M. P. and Jones, M. C., 1994. *Kernel smoothing*. Chapman and Hall/CRC.

Watson G. S., 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359-372.

Yatchew, A., 2003. *Semiparametric regression for the applied econometrician*. Cambridge University Press.

Silverman B.W., 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.