Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/120113

New Simple and Multiple Nonparametric Regression Model Using a Standard Logistic Kernel with Application to Fire Accidents Data

Samah M. Abo-El-Hadid

Department of Mathematics, Insurance and Applied Statistics, Faculty of Commerce and Business Administration, Helwan University Cairo, Egypt

Received: 15 Feb. 2022, Revised: 11 May 2022, Accepted: 13 May 2022
Published online: 1 Jan. 2023

Abstract: In this paper, new simple and multiple Nadaraya-Watson regression models are introduced by using a standard logistic kernel density. The statistical properties of the resulting regression functions are studied, and then the suggested models are applied to Egyptian fire accidents real data during the year 2020. Also, the suggested regression models are compared with other parametric and nonparametric regression models.

Keywords: Nadarya-Watson regression; standard logistic distribution; Kernel density estimation; Nonparametric regression; fire accidents.

1 Introduction

Suppose that Y and X are related by a simple regression model of the form:

$$y = m(x) + \in \tag{1}$$

where m(x) is the regression function and $\in$ is a random error term. If we assume a certain structure on the regression function m(x), then the regression model is called a parametric regression model. In the nonparametric regression model, the function m(x) is unknown and its estimation is the objective of predictive modelling. The regression function m(x) depends on the densities of the explanatory variable f(x), and the joint density function of both the dependent and independent variable f(x,y). Rosenblatt [1] introduce a nonparametric estimator of the density function f(x), which called the kernel density estimator. Rosenblatt's kernel density estimator is extended to the bivariate case by Epanechnikov [2]. Nadaraya [3] and Watson [4] use the kernel estimators of Rosenblatt [1] and Epanechnikov [2] to estimate the regression function m(x). Härdle and Gasser [5] introduce a robust non-parametric estimation for the regression function based on the theory of M-estimation, and studied the statistical properties of the resulting estimator. Michels [6] suggested using asymmetric kernel functions to estimate the regression function. He found that the asymmetric kernel techniques outperform the usual symmetric kernel methods and lead to better predictions according to real data application. Aalen [7] developed the kernel regression model in the case of survival analysis. He also uses the bootstrap replications to judge the accuracy of the cumulative regression plots. Brown and Chen [8] use the beta family of density functions as kernels function in estimating the regression function. The method they used is a generalization of Bernstein polynomials. Peristera and Kostaki [9] use kernel regression estimators in graduating age-specific mortality data for France, Japan, and Sweden. They found that The Gasser-Muller estimator is superior to the other kernels estimators. Park et. Al. [10] investigate the properties of L2 boosting with kernel regression estimates. They found that L2 boosting reduces the bias of the estimate, while it does not deteriorate the order of the variance. Martın-Baosa et. Al. [11] used the kernel estimator to estimate the utility function of the alternatives for the decision-maker whose stochastic part has a multinomial logistic model. In this paper, we introduce a simple nonparametric standard logistic regression model and a

* Corresponding author e-mail: s_aboelhadid@yahoo.com

multiple regression model. The rest of this paper is organised as follows: In section 2 the kernel regression model is overviewed, while in section 3, the new simple and multiple nonparametric regression model are introduced, and its statistical properties are obtained. In section 4 the new model is applied to Egyptian fire accidents real data for the simple and multiple regression models.

2 the nonparametric regression model

Note that the simple regression function can be expressed as a function in both $f(x)$ and $f(x,y)$ as follows:

$$m(x) = E\,(Y\,|X\ ) = \int \mathrm{y\ f}\,(Y\,|X\ ) \tag{2}$$

$$\therefore m(x) = \int y\frac{f(x,y)}{f(x)} \qquad \mathrm{dy} \tag{3}$$

The Rosenblatt's kernel density estimator for the density $f(x)$ takes the following form:

$$\widehat{f}(x) = \frac{1}{\mathrm{nh}}\sum_{i=1}^{n} K(\frac{x-x_i}{h}) \tag{4}$$

where n is the sample size; $K(\blacksquare)$ and h are the kernel function and the bandwidth respectively, where the kernel function $K(\blacksquare)$ is assumed to be a density function.

Epanechnikov [2] extends the Rosenblatt's kernel estimator to the bivariate case, which takes the form:

$$\widehat{f}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\mathrm{hd}}K_1\left(\frac{x-x_i}{h}\right)K_2\left(\frac{y-y_i}{d}\right) \tag{5}$$

Nadaraya [3] and Watson [4] suggested replacing the previous densities in (3) by their kernel density estimators in (4) and (5), then

$$\widehat{m(x)} = \frac{1}{\widehat{f}(x)}\int \frac{1}{\mathrm{nhd}}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)K\left(\frac{y-y_i}{d}\right)y\,dy \tag{6}$$

$$= \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}{\widehat{f}(x)}\int \frac{1}{\mathrm{nhd}}K\left(\frac{y-y_i}{d}\right)y \quad dy \tag{7}$$

Let

$$[u = \frac{y-y_i}{d}, \quad c = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}{\widehat{f}(x)}] \tag{8}$$

substituting from (8) into (7), then:

$$\widehat{m(x)} = c\left[\int \frac{1}{nh}K(u)(ud+y_i)du\right]$$

$$= c\left[\int \frac{d}{nh}uK(u)du + \int \frac{1}{nh}y_iK(u)du\right]$$

Note that, $\int_{-\infty}^{\infty}\mathrm{uK}(u)du = 0$, then

$$[\widehat{m(x)} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{K_h(x-x_i)}{n^{-1}\sum_{i=1}^{n}K_h(x-x_i)}\right\}y_i = \frac{1}{n}\sum_{i=1}^{n}L_i(x)y_i] \tag{9}$$

The above estimator in (9) is biased estimator, to prove that the estimated regression function in (9) can be rewritten as:

$$\widehat{m(x)} = \frac{\frac{1}{n}\sum_{i=1}^{n}K_h(x-x_i)y_i}{\frac{1}{n}\sum_{i=1}^{n}K_h(x-x_i)} = \frac{\widehat{r}(x)}{\widehat{f}(x)} \tag{10}$$

then

$$\widehat{m(x)} - m(x) = \left[\frac{\widehat{r}(x)}{\widehat{f}(x)} - m(x)\right] \cdot \left[\frac{\widehat{f}(x)}{f(x)} + \left\{1 - \frac{\widehat{f}(x)}{f(x)}\right\}\right]$$

$$= \frac{\widehat{r}(x) - m(x)\widehat{f}(x)}{f(x)} + \left\{\widehat{m(x)} - m(x)\right\} \cdot \frac{f(x) - \widehat{f}(x)}{f(x)} \tag{11}$$

The second term in (11) can be ignored, then

$$\widehat{m(x)} \approx \frac{\widehat{r}(x) - m(x)\widehat{f}(x)}{f(x)} + m(x)$$

$$\approx \frac{\widehat{r}(x) - m(x)\left\{\widehat{f}(x) - f(x)\right\}}{f(x)}$$

$$\therefore E(\widehat{m(x)}) = \frac{1}{f(x)}[E(\widehat{r}(x)) - m(x)^{E\{\widehat{f}(x) - f(x)\}}] \tag{12}$$

Note that, $Bias(\widehat{f}(x)) \approx \frac{h^2}{2} f''(x)\sigma_K^2$, where:

$$\sigma_K^2 = \int u^2 K(u) du \tag{13}$$

$$E(\widehat{r}(x)) = \frac{1}{nh} E\left[\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i\right]$$

$$= \frac{1}{h} \int \int K\left(\frac{x - t}{h}\right) \text{y f}(t, y) \text{ dt dy}$$

$$\int y f(y,t) dy = \widehat{m(x)} f(t)$$

then

$$E(\widehat{r}(x)) = \frac{1}{h} \int K\left(\frac{x - t}{h}\right) \widehat{m(x)} f(t) dt$$

Using Taylor's expansion of $\widehat{m(x + uh)}$, $f(x + uh)$, yields:

$$E(\widehat{r}(x)) \approx f(x)m(x) + \frac{h^2}{2} f(x)m''(x)\sigma_K^2 + \cdots + h^2 f'(x)m'(x)\sigma_K^2 + \frac{h^2}{2} m(x)f''(x)\sigma_K^2 \tag{14}$$

then

$$E(\widehat{m(x)}) \approx m(x) + \frac{h^2}{2} m''(x)\sigma_K^2 + h^2 m'(x)\frac{f'(x)}{f(x)}\sigma_K^2$$

Then the bias of the estimated regression function is as follows:

$$\text{Bias}\left(\widehat{m(x)}\right) \approx \frac{h^2}{2}\sigma_K^2 \left\{m''(x) + 2m'(x)\frac{f'(x)}{f(x)}\right\} \tag{15}$$

And the variance of the estimated regression function is as follows:

$$\text{Var}\left(\widehat{m(x)}\right) \approx \frac{\sigma^2(x)}{\text{nhf}(x)} \int K^2(u) \text{du} \tag{16}$$

Where $\sigma^2(x) = Var(Y|X = x)$.
By the same way as in (3), the multiple Nadaray-Watson regression takes the following form:

$$m(x) = \int \frac{f(y,x_1,\ldots,x_p)y}{f(x_1,\ldots,x_p)}dy \tag{17}$$

And using the kernel density estimators, the estimated multiple regression model is:

$$\widehat{m(x)} = \frac{\sum_{i=1}^{n}\left[\prod_{j=1}^{p}K(\frac{x_j-x_{ij}}{h})\right]y_i}{\sum_{i=1}^{n}\left[\prod_{j=1}^{p}K(\frac{x_j-x_{ij}}{h})\right]} \tag{18}$$

In the next section, we introduce the suggested simple and multiple regression models using the univariate and bivariate standard logistic kerel density estimators.

3 The new simple and multiple nonparametric regression models

In this section, the new simple and multiple nonparametric regression models with Logistic kernel density estimators are introduced. Abo-El-Hadid [12] used the standard logistic distribution as a kernel function to introduce the univariate logistic kernel density estimator. This estimator takes the form:

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{e^{-\left(\frac{x-x_i}{h}\right)}}{\left[1+e^{-\left(\frac{x-x_i}{h}\right)}\right]^2} \qquad , -\infty \le x \le \infty \tag{19}$$

with the following statistical properties:

$$E(u) = \int_{-\infty}^{\infty} \mathrm{uK}(u)du = 0 \tag{20}$$

$$\sigma_K^2 = \mathrm{var}(u) = \int_{-\infty}^{\infty} u^2K(u)du = \frac{\pi^2}{3} \tag{21}$$

$$\int_0^{\infty} K^2(u)du = \frac{1}{6} \tag{22}$$

And the optimal bandwidths which minimize the integrated mean squared error (IMSE)

$$h_{\mathrm{opt}} = \left[\frac{n\pi^4}{63}\right]^{-\frac{1}{5}} \tag{23}$$

In 2021 Abo-El-Hadid [13] extended this univariate logistic kernel density estimator in (19) to the bivariate case. The bivariate logistic kernel density estimator introduced by Abo-El-Hadid (2021), and its statistical properties is as follows:

$$\widehat{f}(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n}\left[\frac{e^{-\left(\frac{x-X_i}{h}\right)}}{\left[1+e^{-\left(\frac{x-X_i}{h}\right)}\right]^2}\bullet\frac{e^{-\left(\frac{y-y_i}{h}\right)}}{\left[1+e^{-\left(\frac{y-y_i}{h}\right)}\right]^2}\right] \tag{24}$$

with the following statistical properties:

$$\mathrm{Bias}\left[\widehat{f}(x,y)\right] \approx \frac{\pi^4h^2}{18}\left[\frac{\partial^2}{\partial x_1^2}f(x,y)+\frac{\partial^2}{\partial x_2^2}f(x,y)\right] \tag{25}$$

$$\mathrm{Var}\left[\widehat{f}(x,y)\right] \approx \frac{f(x,y)}{36nh^2} \tag{26}$$

$$IMSE\left[\widehat{f}(x,y)\right] = \frac{1}{36nh^2} + \iint_{-\infty}^{\infty} \frac{\pi^8 h^4}{324}\left[\frac{\partial^2 f(x,y)}{\partial x_1^2} + \frac{\partial^2 f(x,y)}{\partial x_2^2}\right]^2 dx\, dy \tag{27}$$

And the optimal bandwidths which minimize the IMSE

$$h_{\text{opt}} = \left[\frac{32}{14195} n\pi^8\right]^{-\frac{1}{6}} \tag{28}$$

Then the new simple regression function takes the following form:

$$\widehat{m(x)} = \frac{\left(\sum_{i=1}^{n} \frac{e^{-\left(\frac{x-x_i}{h}\right)}}{\left[1+e^{-\left(\frac{x-x_i}{h}\right)}\right]^2} y_i\right)}{\left(\sum_{i=1}^{n} \frac{e^{-\left(\frac{x-x_i}{h}\right)}}{\left[1+e^{-\left(\frac{x-x_i}{h}\right)}\right]^2}\right)} \tag{29}$$

The bias of this new regression function is obtained by substituting (19) into (15). then:

$$\text{Bias}\left(\widehat{m(x)}\right) \approx \frac{h^2\pi^2}{6}\left\{m''(x) + 2m'(x)\frac{f'(x)}{f(x)}\right\} \tag{30}$$

Abo-El-Hadid [12] suggested replacing the unknown term $f'(x)$ in (30) by the standard logistic density as a reference distribution. then

$$f'(x) = \frac{2e^{-2x}}{(1+e^{-x})^3} - \frac{e^{-x}}{(1+e^{-x})^2} \tag{31}$$

Then the estimated biased of the estimated regression function is as follows:

$$\widehat{\text{Bias}}\left(\widehat{m(x)}\right) \approx \frac{h^2\pi^2}{6}\left\{m''(x) + \frac{2m'(x)}{f(x)}\left[\frac{2e^{-2x}}{(1+e^{-x})^3} - \frac{e^{-x}}{(1+e^{-x})^2}\right]\right\} \tag{32}$$

Substituting (22) into (16) yields that the variance of the estimated regression function is:

$$\text{Var}\left(\widehat{m(x)}\right) \approx \frac{\sigma^2(x)}{6nhf(x)} \tag{33}$$

Then the asymptotic mean squared error of the new regression function is as follows:

$$\text{MSE}\left(\widehat{m(x)}\right) \approx \frac{1}{6nh}\frac{\sigma^2(x)}{f(x)} + \left[\frac{h^2\pi^2}{6}\left\{m''(x) + \frac{2m'(x)}{f(x)}\left[\frac{2e^{-2x}}{(1+e^{-x})^3} - \frac{e^{-x}}{(1+e^{-x})^2}\right]\right\}\right]^2$$

By the same way, the estimated multiple standard logistic regression model with two explanatory variables is obtained by substituting (24), into (9), so the suggested multiple regression function takes the following form:

$$\widehat{m(x)} = \frac{\left(\sum_{i=1}^{n} \frac{e^{-\left(\frac{x_1-x_{i1}}{h}\right)}}{\left[1+e^{-\left(\frac{x_1-x_{i1}}{h}\right)}\right]^2} \frac{e^{-\left(\frac{x_2-x_{i2}}{h}\right)}}{\left[1+e^{-\left(\frac{x_2-x_{i2}}{h}\right)}\right]^2} y_i\right)}{\left(\sum_{i=1}^{n} \frac{e^{-\left(\frac{x_1-x_{i1}}{h}\right)}}{\left[1+e^{-\left(\frac{x_1-x_{i1}}{h}\right)}\right]^2} \frac{e^{-\left(\frac{x_2-x_{i2}}{h}\right)}}{\left[1+e^{-\left(\frac{x_2-x_{i2}}{h}\right)}\right]^2}\right)} \tag{34}$$

Then the bias and variance of the multiple regression function are as follows:

$$Bias(\widehat{m(x)}) \approx \frac{1}{2}\sigma_{\kappa}^{2} trH^{T} H_{m} H + \frac{\sigma_{\kappa}^{2}}{f()}\{\nabla_{m}^{T} HH^{T} \nabla_{f}\} \tag{35}$$

$$\mathrm{Var}\left(\widehat{m(x)}\right) = \frac{1}{n \cdot \det(H)} \cdot \frac{\sigma^{2}(x)}{f()} \int \kappa^{2}(u)\mathrm{du} \tag{36}$$

Where (H) is a diagonal matrix of bandwidth;(H_m) is the second derivative matrix of the function $\widehat{m(x)}$; and ∇ is the first derivatives.

4 Real Data Application

In this section the new simple and multiple nonparametric regression models are applied to Egyptian fire accidents real data. The data used in this section is obtained from the annual report of fire accidents in Egypt 2020 (issue April 2021) which published by the Central Agency for Public Mobilization and Statistics (CAPMS).

At the governorate level, Cairo comes first in terms of fire accidents with 13.3% of the total fire accidents during the year 2020, followed by Giza Governorate with 7.4%, and in the last rank North Sinai Governorate with 0.2%. also at the governorate level, Cairo comes first in terms of fire points with 11% of the total fire points, followed by Sharkia Governorate with 7%, and in the last rank red sea and North Sinai Governorates with 1.1%.

4.1 Application of simple regression model

In this subsection, we study the relationship between the percentage fire stations (% fire points) an independent variable, and the percentage of fire accidents (% fire accidents) as a dependent variable.

The following figure illustrate percentage of both fire accidents and fire points in each governorate during 2020.

Figure (2) gives the scatter plot of the percentage of fire accidents and fire points in each Egyptian governorate during 2020.

The parametric linear regression model is estimated and the regression line with the scatter plot is given in figure (3): Then the new regression model using the logistic kernel is applied to the fire data and then compared with the parametric one; and the nonparametric regression model using the uniform kernel. Figure (4) and table (1) compares the three regression models.

From figure (4), it appears that the new regression model (in green) fits the fire data better than both the parametric and nonparametric uniform regression models (in red).

Table 1: Comparison between parametric regression; and nonparametric regression models with logistic kernel and uniform kernel

criteria	**Parametric linear**	**Nonparametric uniform**	**Nonparametric logistic**
Residual standard error	1.569	1.567	1.346749
R-squared	0.7118	0.7125	0.7822888

Table (1) shows that new regression model for fire data outperform the parametric and the uniform nonparametric regression models according to the residuals standard error and R-squared criteria, and the worst model is the parametric one.

4.2 Application of multiple regression model

In this subsection, we study the relationship between percentage of deaths due to fire accidents as a dependent variable and the percentage of fire stations, and percentage of fire accidents as independent variables.

Figure (5) gives a 3D scatter plot, of the dependent and independent variables.

Figure (6) gives the scatter plot matrix and the estimated densities of the dependent and independent variables.

Then the suggested multiple regression model using the logistic kernel is applied to the fire data and then compared with the parametric linear regression model. Figure (7) and table (2) compares the parametric and nonparametric regression models.

Figure (7) and table (2), show that the suggested multiple regression model fits the fire data better than the parametric linear regression model.

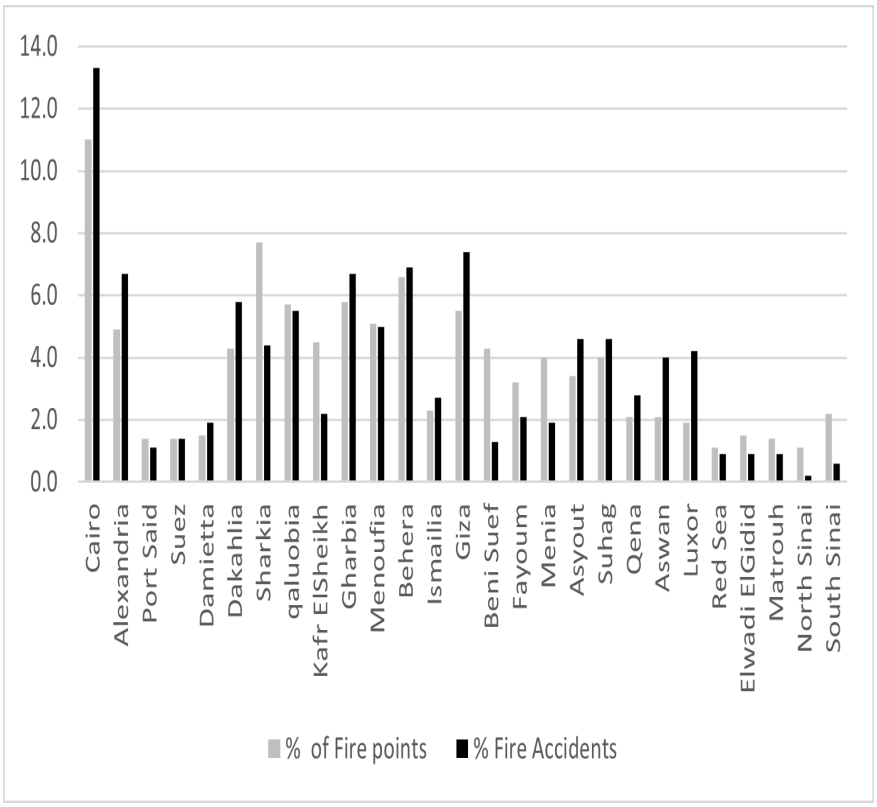

Fig. 1: % of total number fire accidents and % of total fire stations in governorates in 2020

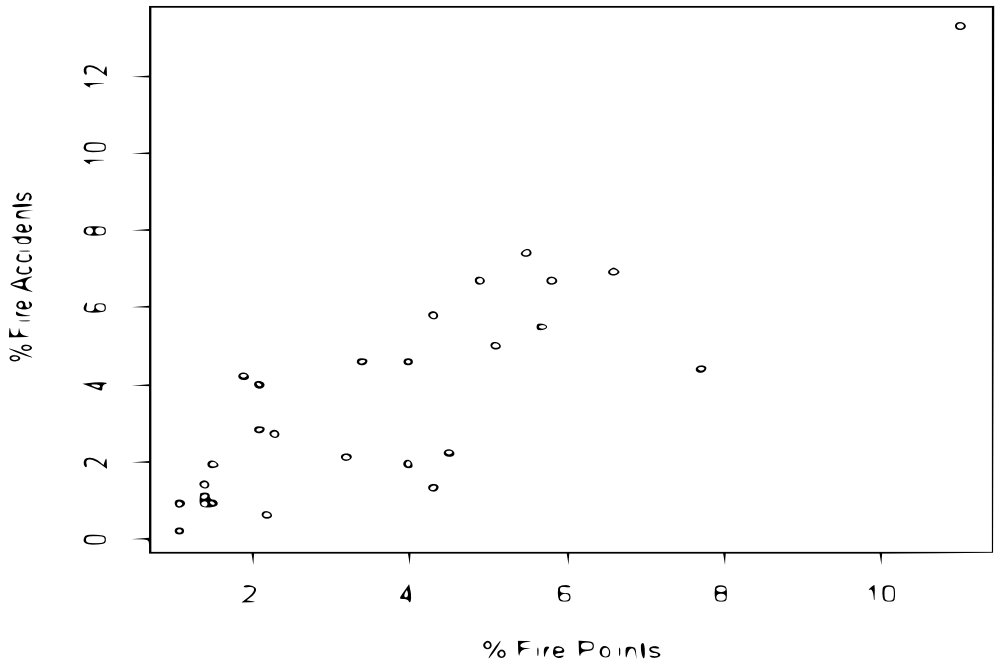

Fig. 2: scatter plot of percentages fire accidents and percentages fire stations in 2020

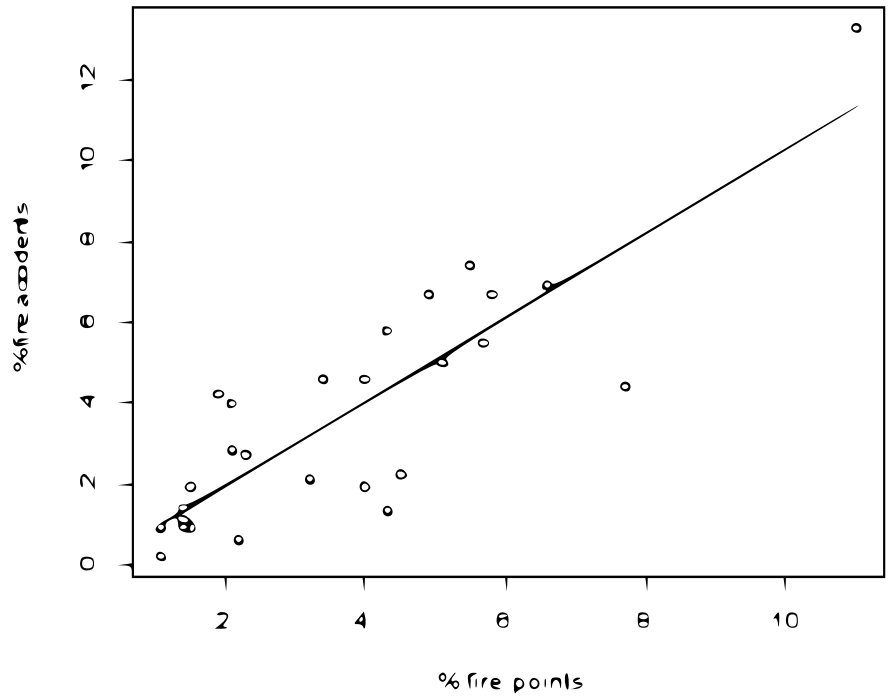

Fig. 3: scatter plot with the parametric regression line for percentage fire accidents and percentage of fire stations in 2020

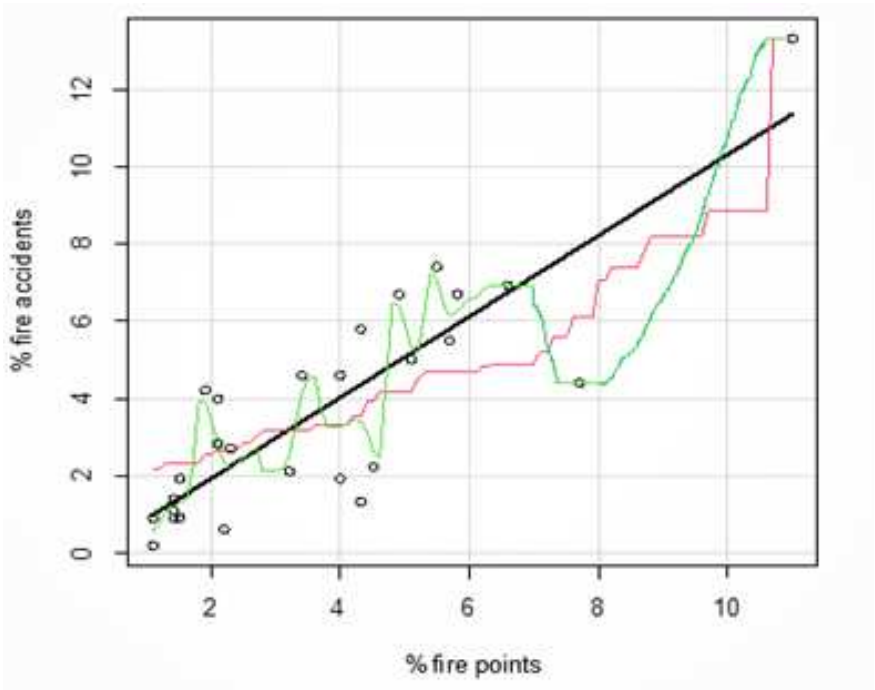

Fig. 4: scatter plot; parametric regression line; and nonparametric regression curves with logistic kernel (in green) and uniform kernel (in red)

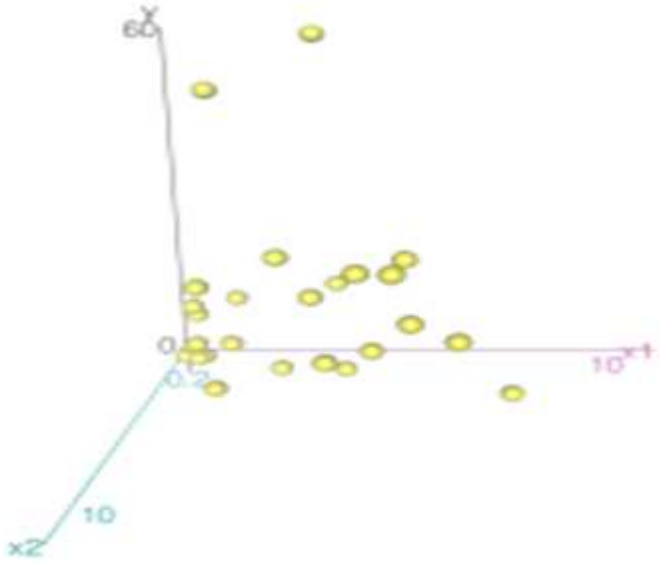

Fig. 5: 3Dscatter plot of% of deaths due to fire accidents and % fire accidents and % fire stations in 2020

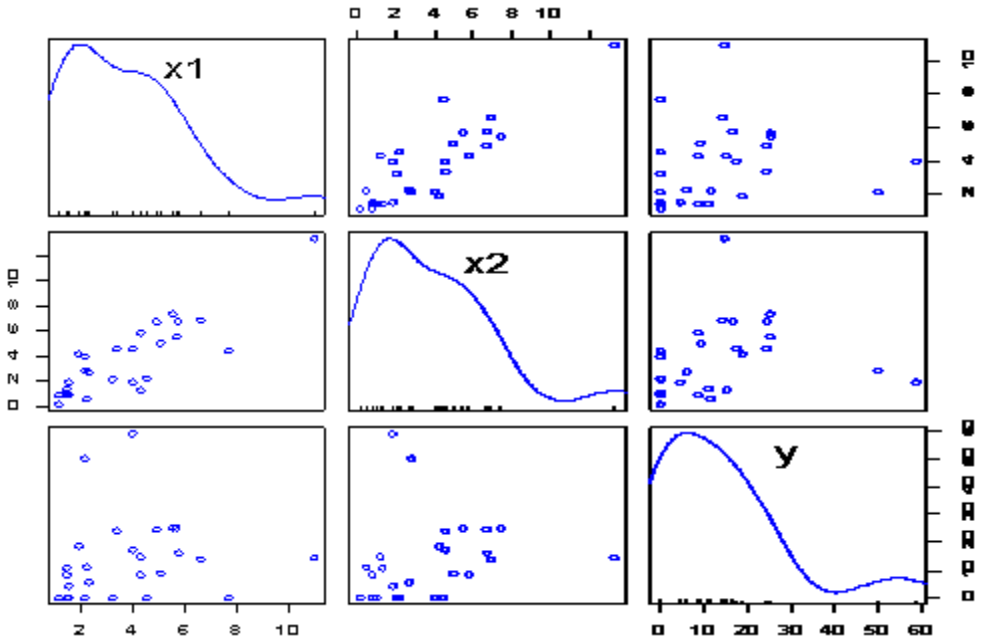

Fig. 6: scatter plot matrix and density estimate for the dependant and independent variables of fire data

5 Conclusion

In this paper, a new regression model was derived using nonparametric logistic density functions presented by Abo-El-Hadid (2018) and (2021). The statistical properties of the presented regression model were also derived, which are bias, variance, and mean squares of errors. Finally, the suggested model was applied to analyze the fire accidents real data in Egyptian governorates during 2020, and we found that the nonparametric regression models overcome the parametric linear regression model.

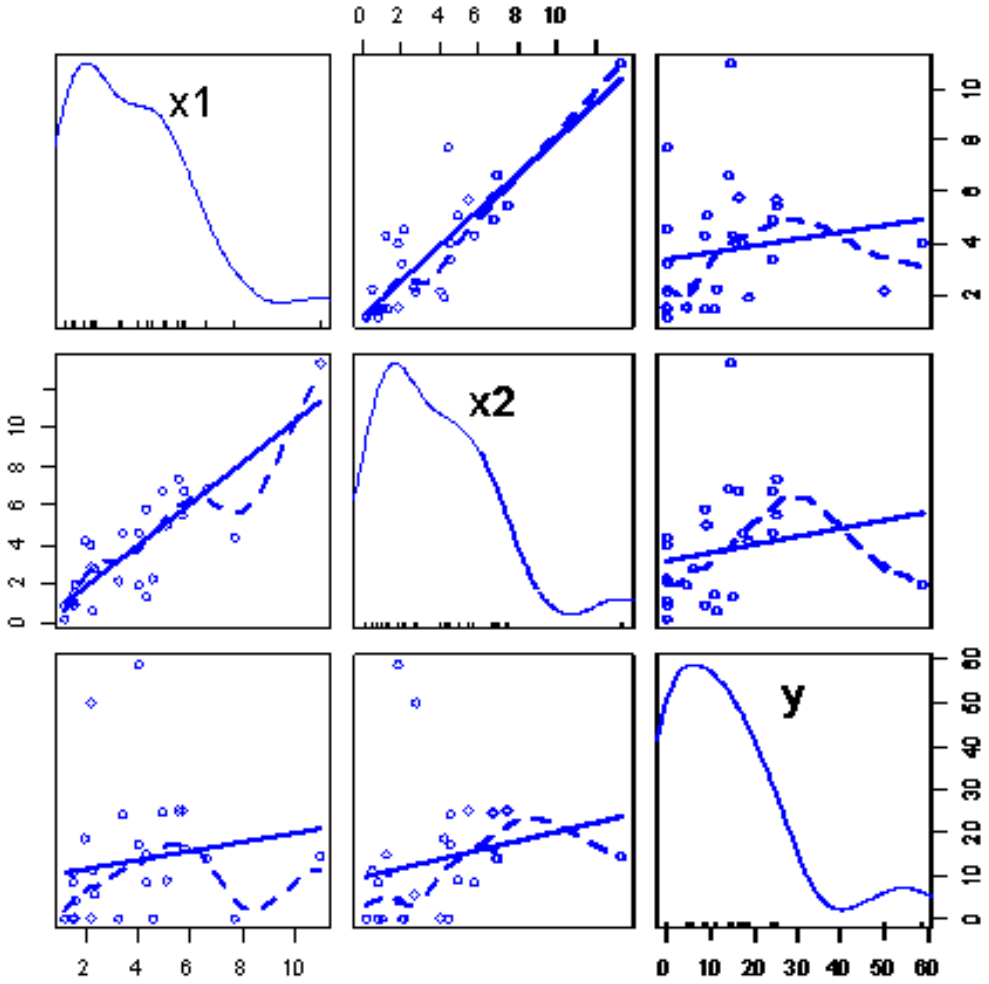

Fig. 7: scatter plot; parametric regression line (solid line); and nonparametric regression curves with logistic kernel (dashed line)

Table 2: Comparison between multiple parametric regression; and nonparametric regression models with logistic kernel

criteria	**Parametric linear**	**Nonparametric logistic**
Residual standard error	14.970	12.035
R-squared	0.04439	0.295

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] M.Rosenblatt, Remarks on Some Nonparametric Estimates of Density Function. The Annals of Mathematical Statistics **27**, 832-837 (1956).
[2] V.A Epanechnikov, Nonparametric Estimation of Multivariate Probability Density. Theory of Probability and its Application **14**, 153-158 (1969).
[3] E. A. Nadaraya, On Estimating Regression. Theory of Probability and its Application **9**, 141-142 (1964).
[4] G. S. Watson, Smooth Regression Analysis, Sankhyã **26**, 359-372 (1964).

[5] W. Härdle, and T. Gasser, Robust Non-Parametric Function Fitting. Journal of the Royal Statistical Society: Series B (Methodological) **46**, 42-51 (1984).
[6] P.Michels, Asymmetric Kernel Functions in Non-Parametric Regression Analysis and Prediction. Journal of the Royal Statistical Society: Series D (The Statistician) **41**, 439-454 (1992).
[7] O.O Aalen, Further results on the non-parametric linear regression model in survival analysis, Statistics in medicine **12**, 1569-1588 (1993).
[8] B.M Brown and S.X Chen, Beta-Bernstein Smoothing for Regression Curves with Compact Support,Scandinavian journal of statistics **26**, 47-59 (1999).
[9] P. Peristera, and A. Kostaki, An Evaluation of The Performance of Kernel Estimators for Graduating Mortality Data. Journal of Population Research **22**, 185-197 (2005).
[10] B.U. Park, Y.K. Lee and S. Ha, L_2 boosting in kernel regression. Bernoulli **5**, 599-613 (2009).
[11] J.A. Martın-Baosa, R. G. Rodenas., M. L. Garcia, iscrete choice modeling using Kernel Logistic Regression **47**, 457–464 (2020).
[12] S. M. Abo-El-Hadid, Logistic Kernel Estimator and Bandwidth Selection for Density Function, International Journal of Contemporary Mathematical Sciences **13**, 279 – 286 (2018).
[13] S. M. Abo-El-Hadid, A Suggested Nonparametric Bivariate Logistic Density Estimator with application on the productivity of Egyptian wheat during 2019/2020,Journal of Mathematics and Statistics, **17**, 44-49 (2021).
[14] T. Gasser and H. G. Mülle, Kernel estimation of Regression Functions, in Gasser and Rosenblatt (eds), Smoothing Techniques for Curve Estimation. Heidelberg, Springer Verlag (1979).
[15] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. . Nonparametric and Semi-parametric Models: An Introduction. New York, Springer-Verlag. (2004).