Australian Journal of Basic and Applied Sciences, 9(28) Special 2015, Pages: 111-116

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Kernel Density Estimation on Online Sales Data

[1]Yundari, [2]U.S.Pasaribu, [3]U.Mukhaiyar

[1,2,3]*Institute Technology of Bandung, Mathematics Department, Math &Science Faculty, Bandung City. Indonesia*

ARTICLE INFO

Article history:
Received 23 June 2015
Accepted 25 July 2015
Available online 30 August 2015

Keywords:
kernel density, regression, sales online, nonparametric

ABSTRACT

In this paper, we do some investigation of kernel function characteristics which be applied in estimating the probability density function, especially in nonparametric case. For study case, we use online sales data in order to observe the causal effect of committed fraud number to the amount of transactions. For that purpose, we use and compare ordinary linear regression model and kernel regression. We obtain that kernel regression approach give a better estimation than the ordinary one.

To Cite This Article: Yundari, U.S.Pasaribu, U.Mukhaiyar., Kernel Density Estimation on Online Sales Data. ***Aust. J. Basic & Appl. Sci.,*** *9(28): 111-116, 2015*

INTRODUCTION

Nowadays many providers of services and goods are done online. However, not all online providers of goods and services can be trusted. There may be some that are frauds. Beside that. the frauder is not only come from provider's side but also from consumers. In online sales, there are so many product that are offered, from the cheapest goods to the highest price can be found there. For simplicity and efficiency, the transactions were conducted using credit cards, but this method opens up the oppurtunities for fraud. This paper examined the relationship between number of committed fraud (based on rule hits) and amount of the transaction. We presumed that there is causal effect between both variables, so that regression analysis be favorable to be used here. Commonly, when regression analysis be applied, normal distribution is a must to be met as assumptions. When we face the real data, this assumption is difficult to be satisfied, then nonparametric analysis be a best alternative.

Before doing regression method, it is important to analysis data descriptively in order to find its distribution. The distribution of data is usually described by a probability function, but if the data is not assumed to be normal and independent, as mentioned before, it will be executed using nonparametric analysis.

An important aspect of the statistics, frequently overlooked today is the data presentation in order to provide explanations and illustrations of the obtained conclusions using particular methods (Silverman, B.W., 1986). The plot of probability density function is required for this purpose, since it is quite easy to understand even by non-mathematician. For example, people will no doubt explain that the normal distribution with bell drawing curve of probability density function is more understandable than cumulative distribution function or the formula of probability function. In this paper, the probability density function will be estimated using kernel density estimation.

Kernel Density Estimation:

Consider random variable X with cummulative distribution function and probability density function at *x* consecutively are $F(x)$ and $f(x)=\frac{d}{dx}F(x)$ so that the main goal in this problem is to estimate $f(x)$ of a random sample $\{X_1, X_2, ..., X_n\}$. Cummulative distribution function $F(x)$ is usually estimated by (Hardle, W., 1994)

$$\widehat{F}(x)=\frac{1}{n}\sum_{i=1}^{n}1(\mathrm{X}_i \le x) \qquad (1)$$

Generally, the estimation of probability density function $f(x)$ can be obtained from the derivative of $\widehat{F}(x)$ i.e. $\hat{f}(x)=\frac{d}{dx}\widehat{F}(x)$, but from Equation (1), it is possible to obtain a probe of the irrelevant set point in order to estimate $f(x)$.

By using the definition of derivative to a small number $h>0$, then

Corresponding Author: Yundari, Institute Technology of Bandung, Mathematics Department, Math &Science Faculty, Bandung City. Indonesia

$$\hat{f}(x) = \frac{F(x+h) - F(x-h)}{2h}$$

$$= \frac{1}{2nh} \sum_{i=1}^{n} 1(x-h < X_i \le x+h)$$

$$= \frac{1}{2nh} \sum_{i=1}^{n} 1\left(\frac{x - X_i}{h} < 1\right)$$

$$= \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)$$

with $k(u) = \begin{cases} 1/2, |u| \le 1 \\ 0, |u| > 1 \end{cases}$, which is a function of uniform density on (Hardle, W., 1994).

Estimator $\hat{f}(x)$ may calculate the percentage of observations at point x. If many observations are close to the point x then $\hat{f}(x)$ will be a high value. Conversely, if only a small X_i that close to the point x then $\hat{f}(x)$ is small either. In other hands, smoothing parameter (bandwidth) h has a function to control the level of a curve smoothness (Michalack, M., 2011). For h is a real number, kernel function K can be expressed as

$$K_h(x) = \frac{1}{n} k\left(\frac{1}{h}\right)$$

In a typical case, the estimator $\hat{f}(x)$ is known as kernel estimator (Wand, M., and M.M.C. Jones, 1995). Generally, it is written as

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)$$

with

$k(u)$: kernel function

h : width of the window (Bandwidth)

X_i : the value of the independent variable

x : the value of the independent variable to be estimated

n : the size of data

Definition:

[4]. Function of $k : \mathbb{R}^n \to \mathbb{R}$ will be said as **kernel function** if it met the conditions:

1. $\int_{\mathbb{R}} k(x)\,dx = 1$
2. For each $x \in \mathbb{R}, k(x) = k(-x)$
3.
4. For each $x \in \mathbb{R}, k(0) \ge k(x)$
5. $\int_{\mathbb{R}} x^2 k(x)\,dx < \infty$

The condition of number 3 and 5 respectively state the mean and variance, so in general the *j*-th moment of the kernel is expressed by

$$\kappa_j(K) = \int_{\mathbb{R}} x^j K(x)\,dx$$

The next step in assuming a kernel after kernel function selection is the selection of smoothing parameter (smoothing) *h*. Selection of smoothing parameter (bandwidth) *h* using Thumb rule (Turlach, B.A., 1993) (Rule of Thumb Bandwidth):

$$h = \hat{\sigma} C(k) n^{-1/5}$$

$$C(k) = 2\left(\frac{\pi^{1/2}(2!)^3 \int_{-\infty}^{\infty} k(u)^2\,du}{4.4! \kappa_2^2(k)}\right)^{1/5}$$

Common second-order kernels are listed in the table 1.

Kernel Regression:

Kernel regression is a non-parametric statistical technique to assess the value of the conditional expectation of a random variable. Non-parametric regression model is:

$$Y_i = f(x_i) + \varepsilon_i,\ i = 1, 2, \ldots, n$$

with $E[\varepsilon_i] = 0$ for every i and $f(x_i)$ is the regression function. To estimate the regression function of $f(x_i)$, Nadaraya and Watson (1964) defined a kernel regression estimator known as Nadaraya-Watson estimator (Nadaraya, E.A., 1964) as follows:

Table 1: Some Examples of Kernel Functions

Name of Kernel	Function Form	Advantage	Disadvantage
Uniform	$K(x) = \frac{1}{2} I(-1 < x < 1)$	Is the simplest kernel function and has a degree of smoothness (degree of smoothness 0)	Not continuous, and the resulting large bias
Triangular	$K(x) = (1 - \lvert x \rvert) I(-1 < x < 1)$	Smoothness level 1	The first derivative is not continuous, not symmetric kernel function*
Epanechnikov	$K(x) = \frac{3}{4}(1 - x^2) I(-1 < x < 1)$	The optimal kernel because it has the smallest MSE	The first derivative is not continuous

Name of Kernel	Function Form	Advantage	Disadvantage
Biweight	$K(x)=\frac{15}{16}(1-x^2)I(-1<x<1)$	Effective for higher order*	Should have many data
Triweight	$K(x)=\frac{35}{32}(1-x^2)^3 I(-1<x<1)$	Effective for higher order*	Should have many data
Quartic	$K(x)=\frac{15}{16}(1-x^2)^2 I(-1<x<1)$	Effective for higher order*	Should have many data
Gaussian	$K(x)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}$	Generates a good level of smoothness, can be used for small data*	Generated MSE value is still large

*) it is obtained from some simulation study

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

with

$\hat{f}(x_i)$: estimator for the regression function

n : number of observations

K : Kernel function

h : smoothing parameter (Smoothing)

Research data:

For case study, we use online sales transaction data in February-March 2014 which had been categorized as consumer fraud who are using credit cards when doing transaction. In other words, this data is the amount of the online company loss in February-March 2014.

The Table 2 and Figure 1 is the numerical summary of the online companies loss data.

Table 2: Numerical summary of company loss data(data 1)

Descriptive Statistics	
Mean	2299102.293
Standard Error	114652.076
Median	1488870
Mode	7263711
Standard Deviation	2461684.472
Sample Variance	6.060 E+12
Kurtosis	3.475
Skewness	1.892
Range	12532800
Minimum	16200
Maximum	12549000
Sum	1059886157
Count	461

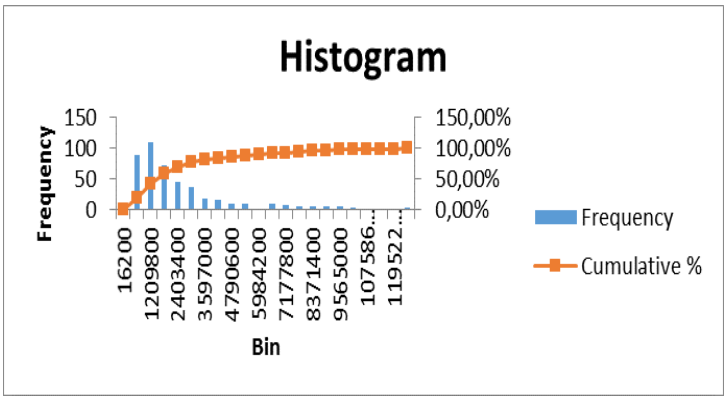

Fig. 1: Histogram of company loss data

The next step to estimate the regression function using kernel approach using secondary data online sales with descriptive statistics as Table 3 follows.

Table 3: Descriptive Statistics data 2

Descriptive Statistics	
Mean	4052723,386
Standard Error	611401,4046
Median	2839016,491
Mode	#N/A
Standard Deviation	2287654.582
Sample Variance	5.233E+12
Kurtosis	-0.383
Skewness	0.949
Range	6842850.188
Minimum	1564005.312
Maximum	8406855.5
Sum	56738127.4
Count	14

RESULT AND DISCUSSION

In this paper density function estimated using a kernel approach. Before that, see figure. 2 the result of a scatter plot of the data.

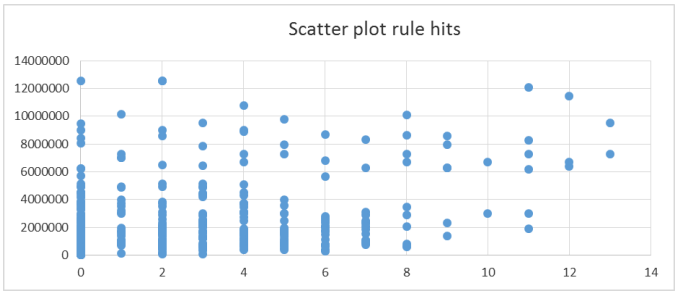

Fig. 2: the result of scatter plot of data

To determine the type of distribution of the data is carried out fitting data. The fitting distribution of research data will obtain the exponential distribution. The curves shown in figure 3.

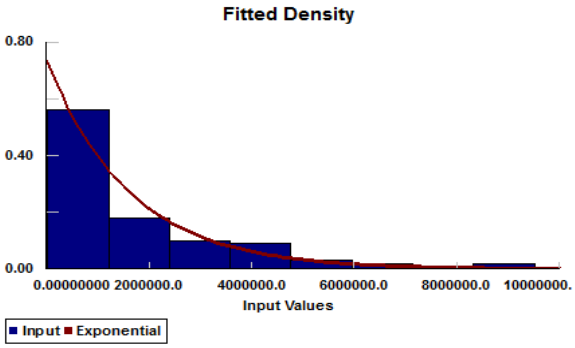

Fig. 3: The fitting distribution

With use Gaussian kernel density estimate for each number of violations, the result on any number of violations of the predictor variable result in different density probability function. The result is given in Figure 4.

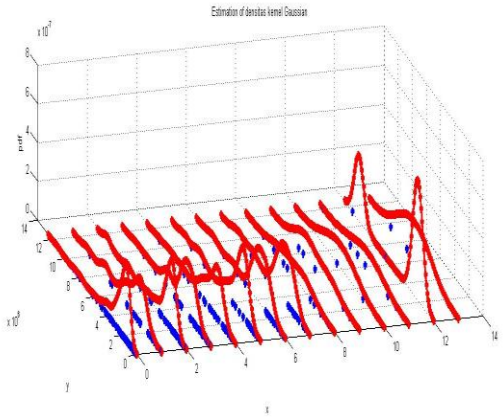

Fig. 4: Density estimation of each committed violations

While the result of using some of kernel density estimation such as Gaussian, Epanechnikov, uniform and triangular kernel shown in the following figure 5.

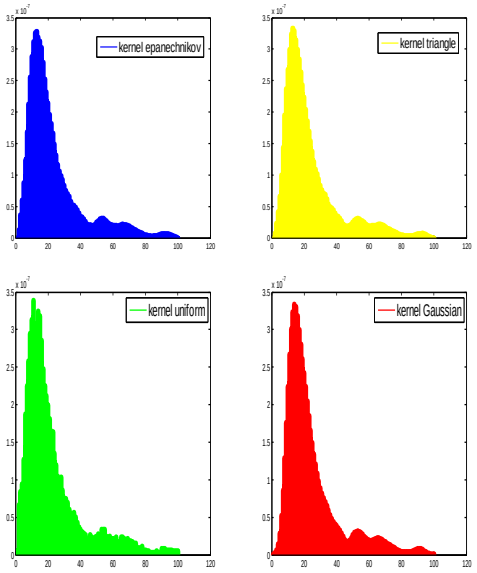

Fig. 5: kernel density estimation of some kernel

From the histogram Figure 1, the approaching distribution is the distribution built by Gaussian, Epanechnikov, and Triangular kernel. As for the distribution built by the uniform kernel is still not smooth especially when compared with results of fitting distribution that approaching exponential distribution.

Next we discusses about the influence of a number of rule violations committed by "fraudster" consumers to the amount of fraudulent transactions. The initial step is done using regression analysis with ordinary regression using least square method. Whereas when compared with Gaussian kernel regression obtained shown in Figure 6.

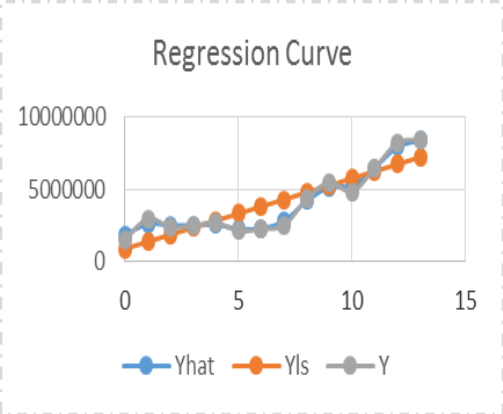

Fig. 6: The comparing least square methods and Gaussian kernel regression

The estimation results are obtained as Table 4.

Table 4: The comparison of the estimation result least square and Gaussian kernel methods

x	Y	Yhat	Yleast squared
0	1564005.312	1873263.447	912474.196
1	2972911.115	2617612.602	1395589.456
2	2397542.392	2514465.089	1878704.716
3	2492454.045	2513862.524	2361819.976
4	2705121.867	2565662.278	2844935.236
5	2148092.355	2269472.255	3328050.496
6	2259534.714	2295926.783	3811165.756
7	2521895.800	2804243.770	4294281.016
8	4319690.300	4200148.164	4777396.276
9	5477072.167	5150543.742	5260511.535
10	4837000	5247976.778	5743626.795
11	6451951.833	6476467.235	6226742.055
12	8184000	7902296.565	6709857.315
13	8406855.500	8349388.912	7192972.575
RMSE		227455.956	1032858.190

From the value of RMSE can be seen that by using kernel regression estimation, obtained the data that almost close to the original data, that is RMSE = 227455.960, while the ordinary linear regression obtained RMSE = 1032858.190.

Conclusion:

An important aspect of the statistics, frequently overlooked today is the presentation of data in order to provide explanations and illustrations of the conclusions that may have been obtained by other ways. The density estimation is required for this purpose, for the simple reason that they are quite easy to understand for the non-mathematician. Density estimation is a first important thing to describe conclusions. From the discussion showed that the estimation of kernel density produces smooth curves and close to the results of the original data histogram. And Also for kernel regresion, from the results of RMSE can be seen that by using Gaussian kernel regression estimation, obtained the data that almost close to the original data.

ACKNOWLEDMENT

This paper are sponsored by Innovation research grant of ITB (Riset KK & Inovasi) 2015 and BOPTN funds for international seminar traveling (DIKTI) 2015.

REFERENCES

Hardle, W., 1994. *Applied Nonparametric Regression*, NewYork: Cambridge University Press.

Michalack, M., 2011. Adaptive Kernel Approach to the time series prediction. *Pattern Analitic Application*, 14: 283-293.

Nadaraya, E.A., 1964. On estimating regression. Theory Probab Appl, 9: 141-142.

Scott, D.W., 1992. *Multivariate density estimation*. Theory Pract Vis. Wiley & Sons

Silverman, B.W., 1986. *Density estimation for statistics and data analysis*. Chapman & Hall.

Turlach, B.A., 1993. *Bandwidth selection in kernel density estimation*: a review. C.O.R.E. and Institut de Statistique, Universite Catholique de Louvain

Wand, M., and M.M.C. Jones, 1995. *Kernel Smoothing*, Chapman & Hall, London.