International Journal of Advanced Culture Technology Vol.10 No.1 236-241 (2022)
DOI https://doi.org/10.17703/IJACT.2022.10.1.236

IJACT 22-3-29

Nonparametric Estimation of Univariate Binary Regression Function

[1]Shin Ae Jung, [2]Kee-Hoon Kang

[1]Manager, Automobile Insurance Team, Hanwha General Insurance, Korea
[2]Professor, Department of Statistics, Hankuk University of Foreign Studies, Korea
khkang@hufs.ac.kr

Abstract

We consider methods of estimating a binary regression function using a nonparametric kernel estimation when there is only one covariate. For this, the Nadaraya-Watson estimation method using single and double bandwidths are used. For choosing a proper smoothing amount, the cross-validation and plug-in methods are compared. In the real data analysis for case study, German credit data and heart disease data are used. We examine whether the nonparametric estimation for binary regression function is successful with the smoothing parameter using the above two approaches, and the performance is compared.

Keywords: *Bandwidth, Cross-validation, Nadaraya-Watson Estimator, Plug-in Bandwidth, Smoothing Parameter*

1. INTRODUCTION

Let $\{(X_i, Y_i) \mid Y_i = 0, 1,\ i = 1, \cdots, n\}$ denote the observed values of the covariate and the binomial response variable for binomial regression, and let $f(x)$ and $g(x)$ be the density function of the covariates X_i when $Y = 1$ and $Y = 0$, respectively. The goal of this paper is to predict the Y_i value given the covariate X_i. That is, estimating the univariate binomial regression function is the same as the estimation problem of the binomial classification function. As an estimation method, a nonparametric Nadaraya-Watson estimator is used with an appropriate single or double smoothing amount.

A smoothing amount should be selected for nonparametric estimation. Two approaches, plug-in and cross-validation are introduced and their effects are examined. As mentioned in [1], the basic idea for estimating a binomial regression function is as follows. The binomial regression function formula is expressed as $\lambda(x) = \Pr(Y = 1|X = x)$, where Y is the binary response variable, $Y = 1$ is "success", $Y = 0$ is "failure", and X is a continuous covariate. Let $h(x)$ denote the density function for all covariates, regardless of success or failure, which is defined by $h(x) = \pi_1 f(x) + \pi_2 g(x),\ \pi_2 = 1 - \pi_1$. Then, the binomial regression function can be expressed as the following Equation (1).

$$\lambda(x) = Pr(Y = 1|X = x) = \frac{\pi_1 f(x)}{\pi_1 f(x) + (1 - \pi_1) g(x)} = \frac{\pi_1 f(x)}{h(x)} \tag{1}$$

Manuscript received: January 25, 2022 / revised: March 1, 2022 / accepted: March 8, 2022
Corresponding Author: khkang@hufs.ac.kr
Tel: +82-31-330-4486, Fax: +82-31-330-4566
Professor, Department of Statistics, Hankuk University of Foreign Studies, Korea

If s out of sample size n is success and $m = n - s$ is failure, then the kernel density function estimators of f and g can be expressed as Equation (2), respectively.

$$\hat{f}(x) = s^{-1} \sum_{i=1}^{n} Y_i K_a(x - X_i), \ \hat{g}(x) = m^{-1} \sum_{i=1}^{n} (1 - Y_i) K_b(x - X_i) \quad (2)$$

Here, a and b are the bandwidths that control the smoothness of the density function estimated based on success and failure data, respectively, and the kernel function K is a symmetric probability density function, meaning $K_c(u) = c^{-1}K(u/c)$. In this study, we examine the selection criteria for smoothing parameters a and b, and check whether there is a difference in the performance of binary regression function according to them.

2. NONPARAMETRIC FUNCTION ESTIMATION

2.1 Kernel Density Estimation

Estimation method of the probability density function can be divided into a parametric and nonparametric one depending on whether the target density function is formalized by a parameter or not. The most basic nonparametric estimation method is a histogram. In the case of continuous data, histogram divides the entire given data into several classes and expresses the relative frequency belonging to each class as a bar. However, histogram cannot find the derivative of the estimator, and the shape of the estimated distribution varies a lot depending on how the class size is determined. Kernel density estimation is a method that compensates for these shortcomings and has good statistical properties. When a random sample obtained from a continuous probability density function f is $\{X_1, X_2, \cdots, X_n\}$, the kernel density estimator is defined as follows.

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

Here, for the kernel function K we normally use a symmetric probability density function, and h is the smoothing parameter (bandwidth), which determines the smoothness of the estimates. Histogram has the same weight regardless of the distance between the data X_i and x within a certain interval, but in the case of the kernel estimation method, the closer the distance between the data X_i and x, the more weight is given. Kernel functions include Uniform, Triangle, Epanechnikov, Biweight, Tricube, Triweight, and Gaussian. For a detailed description of the kernel function, refer to [2]. Parametric and nonparametric estimation of probability density functions has also been discussed in [3]. In this paper, the biweight kernel function defined as follows is used.

$$K(u) = \begin{cases} \frac{15}{16}(1 - u^2)^2, & if \ |u| \le 1 \\ 0, & otherwise \end{cases}$$

2.2 Binary Regression Function Estimation

In order to nonparametrically estimate the binomial regression function $\lambda(x)$, the unknown f and g in (1) are estimated and substituted into the kernel probability density function of (2). Since π_1 and π_2 correspond to prior probabilities from each group, if the sample size is reflected and replaced with s/n and

m/n, the result is as shown in (3).

$$\hat{\lambda}_{a,b,c}(x) = \frac{s\hat{f}_a(x)}{s\hat{f}_b(x) + m\hat{g}_c(x)} \tag{3}$$

Here, a, b and c are the smoothing parameter required when estimating each probability density function using the kernel. Since a and b are smoothing amounts for estimating the same distribution, there is no need to set them differently, so $a = b$ was set in all cases. Therefore, $\hat{\lambda}_a$, a case in which the same smoothing amount is estimated for f and g, and $\hat{\lambda}_{a,c}$, a case in which different smoothing amounts are used for f and g, are compared. Specifically, $\hat{\lambda}_a$ and $\hat{\lambda}_{a,c}$ are expressed as (4) and (5), and it can be seen that they are in the form of Nadaraya-Watson estimators.

$$\hat{\lambda}_a = \frac{\hat{\pi}_1\hat{f}_a(x)}{\hat{\pi}_1\hat{f}_a(x) + \hat{\pi}_2\hat{g}_a(x)} = \frac{\sum_{i=1}^{n} Y_i\, K_a(x - X_i)}{\sum_{i=1}^{n} K_a(x - X_i)} \tag{4}$$

$$\hat{\lambda}_{a,c} = \frac{\hat{\pi}_1\hat{f}_a(x)}{\hat{\pi}_1\hat{f}_a(x) + \hat{\pi}_2\hat{g}_c(x)} = \frac{\sum_{i=1}^{n} Y_i\, K_a(x - X_i)}{\sum_{i=1}^{n} Y_i K_a(x - X_i) + \sum_{i=1}^{n}(1 - Y_i)K_c(x - X_i)} \tag{5}$$

2.3 Bandwidth Selection Method

1) Plug-in Method

To select the smoothing amount for nonparametric estimation of the binomial regression function, divide by grid points in the appropriate range of each smoothing amount, and obtain $\hat{f}$ and $\hat{g}$, finally estimate λ. Then, compare the weighted integrated squared error for each smoothing parameter to select the best one. This is called the optimal grid search method. This is a good method based on the optimal weighted integrated squared error, but another method should be tried because it is difficult to implement in practice. A plug-in method is one of the methods that can automatically select a smoothing quantity that fits the data well in the binomial regression function. This method used to determine the smoothing amount in kernel density function estimation has been discussed in detail in [4] and [5].

A plug-in method is used to select the smoothing amount for estimating each kernel density function in $X|Y = 1 \sim f(x)$ and $X|Y = 0 \sim g(x)$. This is to replace an unknown function value and a related value with an estimator in the formula for asymptotically obtaining the optimal smoothing amount, and the first step uses a value calculated from the normal distribution.

2) Cross-validation Method

One of the methods for selecting a bandwidth is cross-validation, which includes least squares cross-validation, maximum likelihood cross-validation, and biased cross-validation, etc. In this paper, the maximum likelihood cross-validation criterion was used, and [6] presented a method for selecting a smoothing quantity using the concept of the likelihood function. The proposed method is to select the smoothing amount h which maximize the likelihood function presented in (6).

$$MLCV(h) = \frac{1}{n}\sum_{i=1}^{n} log\hat{f}_{-i}(X_i) \tag{6}$$

Here, $\hat{f}_{-i}(X_i)$ is the estimated value of the probability density function from the remaining observation data except for the value X_i, as in Equation (7).

$$\hat{f}_{-i}(X_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x - X_j}{h}\right) \tag{7}$$

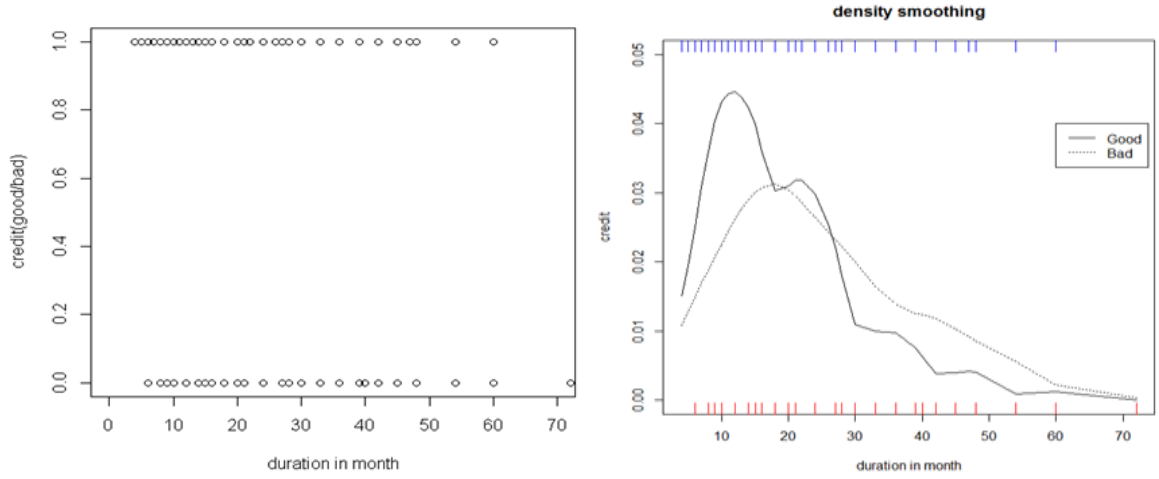
3. DATA ANALYSIS

In this section, we will compare the results of estimating the binomial regression function through the case of two real data whose population distribution is unknown. The first case is German credit data according to several independent variables, and the second case is data on the presence or absence of heart disease according to causes. Both data were obtained from the University of California (UCI) repository and there are no missing values.

3.1 German Credit Data

This is a data set on the credit evaluation of 1000 Germans, and consists of 7 continuous variables such as loan amount, account balance, delinquency period, past credit history, income, and current employment status, and 13 categorical variables, a total of 20 independent variables. The dependent variable is a binary variable consisting of an individual's creditworthiness, that is, "Good" or "Bad". In accordance with the above-mentioned binomial regression analysis, one continuous variable was selected out of a total of 20 independent variables, and the credit was estimated accordingly. Duration in month was selected as a continuous independent variable that affects credit rating, which is a dependent variable. Therefore, the independent variable is the overdue period, and the dependent variable is a binary variable with a value of 0 or 1 in terms of creditworthiness.

Among the given 1000 data, 600 samples selected by sampling without replacement were divided into the training set, and the remaining 400 samples were divided into the test set. After selecting a bandwidth estimate from the training set, the credit is estimated by applying this to the 400 test samples. The performance of estimation is measured by comparing it with the actual value. That is, the smoothing parameter is estimated using the cross-validation method and the plug-in method, and f and g are estimated. If $\hat{f}(x) > \hat{g}(x)$, the credit rating is good, that is, $\hat{\lambda}(x) = 1$, and in the opposite case, the credit rating is bad, that is, classified as $\hat{\lambda}(x) = 0$. By comparing the credit rating $\hat{\lambda}(x)$ classified through estimation and the actual value $\lambda(x)$, the rate of the correctly classified cases was obtained. As a result of the analysis, the correct classification rate was about 65%, and the cross-validation gave better performance than the plug-in approach. However, considering that only one variable was selected out of a total of 20 independent variables and the relationship with the dependent variable was estimated using the binomial regression equation, both methods can be considered as not bad classification results.



(a) (b)

Figure 1. Data plot and density function estimates of credit rating according to the duration in month

The Figure 1(a) shows the data distribution of credit rating according to the overdue period in the given data. Since the independent variable is the number of months overdue, the values often overlap, so the plots are displayed as overlapping. When the value of the credit rating is 1, it is good and the density function of the independent variable is $f(x)$, and when it is 0, which is bad, and the density function is $g(x)$. The Figure 1(b) shows estimated $\hat{f}(x)$ and $\hat{g}(x)$ by using the cross-validation bandwidth.

3.2 Heart Disease Data

This data set is about whether a total of 1341 people were diagnosed with heart disease and various factors related to it. There are a total of 76 independent variables, including age, sex, cholesterol level, blood pressure, and categorical variables, such as smoking or not, and the dependent variable is the diagnosis of heart disease. Among the 76 independent variables, 'cholesterol level (serum cholestoralin mg/dl)', which is continuous and thought to be highly related to the presence or absence of heart disease, was selected as the independent variable.

The distribution of the independent variable in the presence of heart disease is $f(x)$, and the distribution of the independent variable in the absence of heart disease is $g(x)$. As in the case of the previous credit data, among the total 1341 data, 800 observations are taken and into the training set and the remaining 541 observations are classified into the test set. With the estimate of the smoothing amount obtained from the training set, it is applied to the test set to estimate the presence or absence of heart disease, and the accuracy of the estimation is checked by comparing it with the actual value. The estimation results were similar to the previous data, and the accuracy was about 61-64%. The cross-validation method showed slightly better results than the plug-in method, too. Plots of data and estimates are omitted to save space.

4. CONCLUDING REMARKS

In this paper, the kernel estimation method is used to estimate the binomial regression function in the case of single covariate. To analyze the effect, two smoothing parameter selection methods were used and the performance was compared with the weighted integrated squared error. We evaluated the performance of the

estimation method using actual data. The two available data relate to German credit data and heart disease data. The plug-in method and the likelihood cross-validation method were used to select the smoothing parameter for estimating the probability density of each data, and the performance was evaluated by comparing the correct classification rate.

In both cases, considering that the relationship between only one of the many independent variables and the response variable was estimated, the correct classification probability of 0.6 or higher is judged to be a satisfactory result. There was no significant difference in performance between the two bandwidth selection methods. In general, the cross-validation method had slightly better estimation results, but it was difficult to conclude which method was better.

ACKNOWLEDGEMENT

This research was supported by Hankuk University of Foreign Studies Research Fund of 2021.

REFERENCES

[1] D.F. Signorini, D. F and M.C. Jones, "Kernel Estimators for Univariate Binary Regression," *Journal of the American Statistical Association*, Vol. 99, No. 465, pp. 119-126, Mar. 2004. http://doi.org/10.1198/016214504000000115

[2] M.P. Wand and M.C. Jones, *Kernel Smoothing*, Chapman & Hall, London, pp. 175-176, 1995.

[3] S. Woo and K.H. Kang, "Parametric nonparametric methods for estimating extreme value distribution," *The Journal of the Convergence on Culture Technology* (JCCT), Vol. 8, No. 1, pp. 531-536, Jan. 2022. http://dx.doi.org/10.17703/JCCT.2022.8.1.531

[4] S.J. Sheather and M.C. Jones, "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, ser. B, Vol. 53, No. 3, pp. 683-690, Sep. 1991. http://www.jstor.org/stable/2345597

[5] M.P. Wand and M.C. Jones, "Multivariate Plug-in Bandwidth Selection," *Computational Statistics*, Vol. 9, No. 2, pp. 97-116, 1994.

[6] R.P.W. Duin, "On the choice of smoothing parameters for Parzen estimators of probability density functions," *IEEE Transaction on Computers*, Vol. 25, No. 11, pp. 1175-1179, 1976. http://doi.org/10.1109/TC.1976.1674577