One Nonparametric Estimation of the Bernoulli Regression

Elizbar Nadaraya[1], Petre Babilua[2], Grigol Sokhadze[3]
I. Javakhishvili Tbilisi State University, Tbilisi, Georgia
[1]*elizbar.nadaraya@tsu.ge*, [2]*petre.babilua@tsu.ge*, [3]*grigol.sokhadze@tsu.ge*

***Abstract*— The nonparametric estimation of the Bernoulli regression function is studied. The uniform consistency conditions are established and the limit theorems are proved for continuous functionals on $C[a,1-a]$, $0<a<1/2$.**

***Keywords*— Bernoulli regression; kernel estimation; consistency; uniform convergence; Wiener process**

I. INTRODUCTION

Let a random value Y takes two values 1 and 0 with probabilities p ("success") and $1-p$ ("failure"). Assume that the probability of "success" p is a function of an independent variable $x\in[0,1]$, i.e. $p=p(x)=P\{Y=1\mid x\}$ ([1]-[3]). Let x_i, $i=\overline{1,n}$, be the division points of the interval $[0,1]$ which are chosen from the relation

$$\int_0^{x_i} h(x)\,dx=\frac{2i-1}{2n},\quad i=\overline{1,n},$$

where $h(x)$ is the known positive bounded distribution density on $[0,1]$. Let further Y_i, $i=\overline{1,n}$, be independent Bernoulli random variables with $P\{Y_i=1\mid x_i\}=p(x_i)$, $P\{Y_i=0\mid x_i\}=1-p(x_i)$, $i=\overline{1,n}$. The problem consists in estimating the function $p(x)$, $x\in[0,1]$, by the sample $Y_1,Y_2,\dots,Y_n$. Such problems arise in particular in biology ([1], [3]), in corrosion studies [4] and so on.

As an estimate for $p(x)$ we consider a statistic ([5], [6]) of the form

$$\hat p_n(x)=p_{1n}(x)\cdot p_{2n}^{-1}(x)=\frac{\sum_{i=1}^{n}h^{-1}(x_i)K\left(\frac{x-x_i}{b_n}\right)Y_i}{\sum_{i=1}^{n}h^{-1}(x_i)K\left(\frac{x-x_i}{b_n}\right)},\qquad (1)$$

$$p_{\nu n}(x)=\frac{1}{nb_n}\sum_{i=1}^{n}h^{-1}(x_i)K\left(\frac{x-x_i}{b_n}\right)Y_i^{2-\nu},\quad \nu=1,2,$$

where $K(x)\ge 0$ is some distribution density (kernel) and also $K(x)=K(-x)$, $x\in(-\infty,\infty)$, $\{b_n\}$ is a sequence of positive numbers converging to zero and $nb_n\to\infty$.

II. STATEMENT OF THE MAIN RESULTS

Theorem 1. Assume that

$K(x)$ is a function with bounded variation.

$p(x)$ and $h(x)$ are also functions with bounded variation on $[0,1]$, and $h(x)\ge\mu>0$, $x\in[0,1]$.

Then the estimate (1) is asymptotically unbiased and consistent at all points $x\in[0,1]$ where $p(x)$ is a continuous function. Moreover, it has an asymptotically normal distribution, namely:

$$\sqrt{nb_n}\left(\hat p_n(x)-E\,\hat p_n(x)\right)\sigma^{-1}(x)\xrightarrow{d}N(0,1),$$

$$\sigma^2(x)=\frac{p(x)(1-p(x))}{h(x)}\int K^2(u)\,du.$$

Theorem 2. Let $K(x)$, $p(x)$ and $h(x)$ satisfy the conditions of Theorem 1, and also $p(x)$ is continuous function on $[0,1]$. Let further $\varphi(t)\in L_1(-\infty,\infty)$,

$$\varphi(t)=\int_{-\infty}^{\infty}e^{itx}K(x)\,dx.$$

(a$_1$) Let $nb_n^2\to\infty$, then

$$D_n=\sup_{x\in\Omega_n}\left|\hat p_n(x)-p(x)\right|\xrightarrow{P}0,$$

$$\Omega_n=\left[b_n^{\alpha},1-b_n^{\alpha}\right],\quad 0<\alpha<1.$$

(b$_1$) If $\sum_{n=1}^{\infty}n^{-s/2}b_n^{-s}<\infty$ for some $s>2$, then $D_n\to 0$ a.s.

Corollary. Under the conditions of Theorem 2,

$$\sup_{x\in[a,b]}\left|\hat{p}_n(x)-p(x)\right|\to 0$$

almost surely for any fixed interval $[a,b]\subset(0,1)$.

Assume that $b_n=n^{-\gamma}$, $\gamma>0$. The conditions of Theorem 2 are fulfilled:

$$n^{1/2}b_n\to\infty \quad \text{if} \quad 0<\gamma<\frac{1}{2},$$

and

$$\sum_{n=1}^{\infty}n^{-s/2}b_n^{-s}<\infty \quad \text{if} \quad 0<\gamma<\frac{s-2}{2s},\ s>2.$$

Let us introduce the following random processes:

$$\bar{T}_n(t)=\sqrt{n}\int_a^t\left(\hat{p}_n(u)-E\,\hat{p}_n(u)\right)\psi(u)\,du,$$

$$T_n(t)=\sqrt{n}\int_a^t\left(\hat{p}_n(u)-p(u)\right)\psi(u)\,du,$$

where

$$\psi(u)=\left(\frac{h(u)}{p(u)\left(1-p(u)\right)}\right)^{1/2}.$$

Theorem 3. Let $K(x)\geq 0$ satisfy the condition (a) of Theorem 1 and, besides, $K(x)=0$ for $|x|\geq 1$, $K(-x)=K(x)$, $\int_{-\infty}^{\infty}K(x)\,dx=1$. Let further $p(x)$ and $h(x)$ satisfy the condition (b) of Theorem 1 and $0<\inf p(x)\leq\sup p(x)<1$, $x\in[0,1]$.

If $p(x)$ is continuous on $[0,1]$ and $nb_n^2\to\infty$ as $n\to\infty$, then for all the continuous functionals $f(\cdot)$ on $C[a,1-a]$, $0<a<1/2$, the distribution of $f\left(\bar{T}_n(\cdot)\right)$ converges to the distribution of $f\left(W(\cdot)\right)$, where $W(t-a)$, $a\leq t\leq 1-a$, is the Wiener process.

If $nb_n^2\to\infty$, $nb_n^4\to 0$ and $p(x)$ has bounded derivatives up to second order, then the distribution of $f\left(T_n(\cdot)\right)$ converges to the distribution of $f\left(W(\cdot)\right)$.

Corollary. By virtue of Theorem 3 and Theorem 1 from [7, p. 371] we can write

$$P\left\{\max_{a\leq t\leq 1-a}T_n(t)>\lambda\right\}\to$$
$$\to G(\lambda)=\frac{2}{\sqrt{2\pi(1-2a)}}\int_{\lambda}^{\infty}\exp\left\{-\frac{x^2}{2(1-2a)}dx\right\},\quad 0<a<\frac{1}{2}.$$

This result enables us to construct the test of the level α, $0<\alpha<1$, for checking Hypothesis H_0, by which

$$H_0:\ p(x)=p_0(x),\quad a\leq x\leq 1-a$$

when the alternative hypothesis is

$$H_1:\ p(x)=p_1(x),\quad \int_a^{1-a}\psi_0(x)\left(p_1(x)-p_0(x)\right)dx>0,$$

$$\psi_0(x)=\sqrt{h(x)}\left(p_0(x)\left(1-p_0(x)\right)\right)^{-1/2}.$$

Further note that the functionals

$$f_1\left(x(\cdot)\right)=\sup_{a\leq t\leq 1-a}\left|x(t)\right|,$$

and

$$f_2\left(x(\cdot)\right)=\int_a^{1-a}x^2(t)\,dt$$

are continuous on $C[a,1-a]$. Therefore Theorem 3 also implies

1. $f_1\left(T_n(\cdot)\right)\xrightarrow{d}f_1\left(W(\cdot)\right)$

and

$$f_2\left(T_n(\cdot)\right)\xrightarrow{d}f_2\left(W(\cdot)\right).$$

REFERENCES

[1] S. Efromovich, Nonparametric curve estimation. Methods, theory, and applications. Springer Series in Statistics. New York : Springer-Verlag, 1999.

[2] J. B. Copas, Plotting p against x. Appl. Statist., 1983, Vol. 32, No. 1, pp. 25-31.

[3] H. Okumura and K. Naito, Weighted kernel estimators in nonparametric binomial regression. J. Nonparametr. Stat., 2004, Vol. 16, No. 1-2, pp. 39-62.

[4] K. B. Manjgaladze, On some estimate of the distribution function and its moments. Soobshch. Akad. Nauk Gruz. SSR, 1986, Vol. 124, pp. 261-263 (in Russian).

[5] E. A. Nadaraya, On estimating regression. Teor. Veroyatn. Primen., 1964, Vo. 9, pp. 157-159 (in Russian).

[6] G. S. Watson, Smooth regression analysis. Sankhyā Ser. A, 1964, Vol. 26, pp. 359-372.

[7] I. I. Gikhman and A. B. Skorokhod, Introduction to the theory of random processes. Moscow: Izdat. Nauka, 1965 (in Russian).