საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 14, №1, 2020
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 14, no. 1, 2020

Mathematics

On the Estimating the Bernoulli Regression Function Using Bernstein Polynomials

Petre Babilua[*] and Elizbar Nadaraya[**]

[*]*Department of Mathematics, Faculty of Exact and Natural Sciences, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia*
[**] *Academy Member, Department of Mathematics, Faculty of Exact and Natural Sciences, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia*

The estimate for the Bernoulli regression function is constructed using the Bernstein polynomial. The question of its consistency and asymptotic normality is studied. Testing hypothesis is constructed on the form of the Bernoulli regression function. Also, the test is constructed for the hypothesis on the equality Bernoulli functions. The question of consistency of the constructed tests is studied. © *2020 Bull. Georg. Natl. Acad. Sci.*

Bernstein polynomial, Bernoulli regression function, consistency, power of test, one-sided alternatives

Let a random variable Y take two values 1 and 0 with probabilities p ("success") and $1-p$ ("failure"). Assume that the probability of "success" p is the function of an independent variable $x\in[0,1]$, i.e. $p=p(x)=\mathbb{P}(Y=1|\,x)$ [1-4]. Assume that $x_k=\frac{k}{n}$, $k=0,1,\ldots,n$, are the points of division of the interval $[0,1]$ and we have Y_i, $i=0,1,\ldots,n$, which are independent Bernoulli random variables with

$$\mathbb{P}(Y_i=1|\,x_i)=p(x_i),\quad \mathbb{P}(Y_i=0|\,x_i)=1-p(x_i).$$

The problem consists in estimating a function $p(x)$, $x\in[0,1]$, by means of the sampling $Y_0,Y_1,\ldots,Y_n$. A problem like this one arises, for instance, in biology [1, 3, 4], also when studying corrosion processes [5], and so on.

1. As an estimate for $p(x)$ let us consider the following statistic

$$\hat{p}_n(x)=\sum_{k=0}^{n}Y_k b_k(n,x), \tag{1}$$

where $b_k(n,x)=\binom{n}{k}x^k(1-x)^{n-k}$, $k=0,1,\ldots,n$ is a binomial distribution with probability of "success", $x\in(0,1)$.

Note that $E\hat{p}_n(x) = B_n(x) = \sum_{k=0}^{n} p\left(\frac{k}{n}\right) b_k(n,x)$, where $B_n(x)$ is a Bernstein polynomial of order *n* of the function $p(x)$. It is well known that if $p(x)$ is continuous on $[0,1]$, then $\lim_{n\to+\infty} B_n(x) = p(x)$ uniformly with respect to $x \in [0,1]$. Moreover, the order of bias $E\hat{p}_n(x) - p(x)$ is established from the result given in Lorentz' monograph [6, Section 1.6]. Namely, the following assertion is true.

Lemma. Let $p(x)$, $x \in [0,1]$ have a bounded derivative of the second order (we denote the class of such functions by $W[0,1]$). Then

(a) $E\hat{p}_n(x) - p(x) = O\left(\frac{1}{n}\right)$ uniformly with respect to $x \in [0,1]$;

(b) let $p''(x)$ satisfy the Lipschitz condition, there exist $c > 0$, such that

$$|p''(x) - p''(y)| \le c|x-y| \quad \text{for all} \quad x, y \in [0,1],$$

then

$$E\hat{p}_n(x) = p(x) + n^{-1}x(1-x)\,p''(x)2^{-1} + O\left(n^{-\frac{3}{2}}\right)$$

uniformly with respect to $x \in [0,1]$.

Moreover, the estimate $\hat{p}_n(x)$ is free from the boundary effect, which motivates us to study estimates like (1) for the Bernoulli regression function $p(x)$. Note that the kernel-type estimates of the function $p(x)$, which we have considered in [7], do not have such a good property.

Theorem 1. Let $p(x) \in W[0,1]$. Then

1^0. $\hat{p}_n(x)$ is a consistent estimate of $p(x)$ at all points $x \in (0,1)$;

2^0. $\sqrt{n}\left(\hat{p}_n(x) - p(x)\right)\sigma^{-1}(x) \xrightarrow{d} N(0,1)$, $x \in (0,1)$,

$$\sigma^2(x) = p(x)(1-p(x))\left[4\pi x(1-x)\right]^{-\frac{1}{2}}$$

where $\xrightarrow{d}$ denotes converges in distribution and $N(0,1)$ is a random variable that has a standard normal distribution $\Phi(x)$.

Corollary 1.

$$\sqrt{n}\left(\hat{p}_n(x) - p(x)\right)\sigma_n^{-1}(x) \xrightarrow{d} N(0,1), \quad x \in (0,1),$$

where

$$\sigma_n^2(x) = \hat{p}_n(x)(1-\hat{p}_n(x))\left[4\pi x(1-x)\right]^{-\frac{1}{2}}.$$

This makes it possible to construct the confidence interval for $p(x)$:

$$p_n^{\pm}(x) = \hat{p}_n(x) \pm \frac{\sigma_n(x)}{\sqrt{n}}\lambda_\alpha, \quad \lambda_\alpha = \Phi^{-1}\left(\frac{1+\alpha}{2}\right).$$

Theorem 2.

(a) Let $p(x)$ have the bounded derivative. Then

$$\sqrt{n}\,\overline{T}_n \xrightarrow{d} N\left(0,\sigma^2(p)\right),$$

$$\overline{T}_n = \int_0^1 \left[\hat{p}_n(x) - E\hat{p}_n(x)\right] dx;$$

(b) Let $p(x) \in W[0,1]$. Then

$$\sqrt{n}\,T_n \xrightarrow{d} N\left(0,\sigma^2(p)\right),$$

$$T_n = \int_0^1 \left[\hat{p}_n(x) - p(x)\right] dx,$$

$$\sigma^2(p) = \int_0^1 p(x)\left(1-p(x)\right) dx.$$

We give several comments on the application of T_n as a test statistic for the testing hypothesis $H_0: p(x) = p_0(x)$ ($p_0(x)$, $x \in [0,1]$, is the well-known dose-response curve). This statistic is informative because the sign of T_n may carry information on the character of an alternative when the hypothesis H_0 is not true, i.e. the sign of the test statistic indicates the direction of deviation of the alternative from H_0. Can be shown that

$$E\int_0^1 \left(\hat{p}_n(x) - p_0(x)\right) dx \sim \int_0^1 \left(p(x) - p_0(x)\right) dx.$$

Thus, for an alternative of the form $H_1^+: p(x) > p_0(x)$ the statistic T_n will have the tendency to deviate to the right from zero, while for the alternative $H_1^-: p(x) < p_0(x)$ it deviates to the left. Hence it is natural to use the statistic T_n in problems of testing the hypothesis H_0 against the one-sided alternatives H_1^+ and H_1^-.

The assertion (b) of Theorem 2 enables us to construct the test of the asymptotic level α, $0 < \alpha < 1$, for testing the hypothesis H_0, according to which $p(x) = p_0(x)$, $x \in [0,1]$:

Test I. Reject the hypothesis H_0 against the right-side alternative $H_1^+: p(x) > p_0(x)$, $x \in [0,1]$ when $T_n \geq \frac{\lambda_\alpha}{\sqrt{n}}\sigma(p_0)$, where $\lambda_\alpha = \Phi^{-1}(1-\alpha)$.

Test II. Reject the hypothesis H_0 against the left-side alternative $H_1^-: p(x) < p_0(x)$, $x \in [0,1]$ when $T_n \leq \frac{\lambda_\alpha}{\sqrt{n}}\sigma(p_0)$, where $\lambda_\alpha = \Phi^{-1}(\alpha)$.

Corollary 2. Tests I and II are consistent against the one-sided alternatives H_1^+ and H_1^-, respectively.

Corollary 3. Let us consider a sequence of Pitman-type alternatives that are close to the hypothesis H_0:

$$H_n^+: \; p_1^{(n)}(x) = p_0(x) + n^{-\frac{1}{2}}u(x),$$

where $p_0(x) \in W[0,1]$, $0 \leq p_0(x) \leq p_1$, $0 < p_1 < 1$, for each $x \in [0,1]$, $u(x) > 0$, $x \in [0,1]$ and $u(x) \in W[0,1]$. Then

$$P_{H_n^+}\left(T_n \geq \frac{\lambda_\alpha}{\sqrt{n}}\sigma\left(p_0\right)\right) \longrightarrow 1-\Phi\left(\lambda_\alpha - \frac{c}{\sigma\left(p_0\right)}\right),$$

$$c = \int_0^1 u\left(x\right) dx > 0.$$

Hence it follows that for alternative H_n^+ Test I for testing the hypothesis H_0 is asymptotically strictly unbiased since $c > 0$, and is equal to 0 if and only if $u\left(x\right) = 0$ (For Test II, the argumentation is analogous).

2. There often arises a necessity to verify whether the Bernoulli regression functions are equal by using mutually independent samplings ([8, 9]).

Let $Y_i^{(k)}$, $i = 0,1,\ldots,n_k$, $k = 1,2$, be mutually independent Bernoulli random variables with

$$P\left(Y_i^{(k)} = 1 \middle| x_i^{(k)}\right) = p_k\left(x_i^{(k)}\right), \quad P\left(Y_i^{(k)} = 0 \middle| x_i^{(k)}\right) = 1 - p_k\left(x_i^{(k)}\right),$$

$$x_i^{(k)} = \frac{i}{n_k}, \quad i = 0,1,\ldots,n_k, \ k=1,2 \ .$$

Based on the samplings $Y_1^{(1)}, Y_2^{(1)},\ldots,Y_{n_1}^{(1)}$ and $Y_1^{(2)}, Y_2^{(2)},\ldots,Y_{n_2}^{(2)}$, it is required to test the hypothesis $H_0:\ p_1\left(x\right) = p_2\left(x\right) = p\left(x\right)$ against the alternative $H_1:\ p_1\left(x\right) > p_2\left(x\right)$, $x \in \left[0,1\right]$.

Denote

$$\hat{p}_{n_k}\left(x\right) = \sum_{j=0}^{n_k} Y_j^{(k)} b_j\left(n_k, x\right), \quad k = 1,2,$$

$$T_{n_1 n_2} = \int_0^1 \left[\hat{p}_{n_1}\left(x\right) - \hat{p}_{n_2}\left(x\right)\right] dx \, .$$

Theorem 3. Let $p\left(x\right) \in W\left[0,1\right]$ and $\frac{n_1}{n_2} \to \tau > 0$. Then for the hypothesis H_0

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, T_{n_1 n_2} \xrightarrow{d} N\left(0, \sigma^2\left(p\right)\right),$$

where

$$\sigma^2\left(p\right) = \int_0^1 p\left(x\right)\left(1 - p\left(x\right)\right) dx \, .$$

Note that the statistic $T_{n_1 n_2}$ is normed by the value $\sigma^2\left(p\right)$, which depends on an unknown $p\left(x\right) \in W\left[0,1\right]$. However, if $p\left(x\right)$ is not defined by the hypothesis, then $\sigma^2\left(p\right)$ should be replaced by its estimate. For this, we consider two consistent estimates:

$$S_{n_1 n_2}^{(1)} = \int_0^1 \lambda_{n_1 n_2}\left(x\right) dx,$$

$$\lambda_{n_1 n_2}\left(x\right) = \frac{n_2}{n_1 + n_2}\hat{p}_{n_1}\left(x\right)\left(1 - \hat{p}_{n_1}\left(x\right)\right) + \frac{n_1}{n_1 + n_2}\hat{p}_{n_2}\left(x\right)\left(1 - \hat{p}_{n_2}\left(x\right)\right)$$

and

$$S_{n_1 n_2}^{(2)} = \frac{n_2}{n_1 + n_2} S_{n_1} + \frac{n_1}{n_1 + n_2} S_{n_2},$$

where

$$S_{n_k} = \frac{1}{2(n_k - 1)} \sum_{j=1}^{n_k - 1} \left(Y_{j+1}^{(k)} - Y_j^{(k)} \right)^2, \quad k = 1, 2.$$

The statistic S_{n_k} was originally introduced by E. Abbe [10] (see also [11, 12]).

Theorem 4. Let $p(x) \in W[0,1]$ and $\frac{n_1}{n_2} \to \tau$, $0 < \tau < \infty$. Then for the hypothesis H_0 random variables

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(S_{n_1 n_2}^{(k)} \right)^{-\frac{1}{2}} T_{n_1 n_2}, \quad k = 1, 2$$

are asymptotically normal with mean 0 and variance 1.

Using Theorem 4 we can construct the test for the testing hypothesis $H_0: p_1(x) = p_2(x)$, $x \in [0,1]$. The critical domain is established by the inequality

$$T_{n_1 n_2} \geq d_{n_1 n_2}^{(k)}(\alpha) = \lambda_\alpha \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \left(S_{n_1 n_2}^{(k)} \right)^{\frac{1}{2}}, \quad k = 1, 2,$$

where $\Phi(\lambda_\alpha) = 1 - \alpha$, $0 < \alpha < 1$.

Now let us consider the question whether the test based on $T_{n_1 n_2}$ is consistent. The following assertion is true.

Theorem 5. Let $p_1(x), p_2(x) \in W[0,1]$ and $\frac{n_1}{n_2} \to \tau$, $0 < \tau < \infty$. Then for $n_1, n_2 \to \infty$

$$P_{H_1}\left(T_{n_1 n_2} \geq d_{n_1 n_2}^{(k)}(\alpha) \right) \longrightarrow 1, \quad k = 1, 2$$

i.e. the test constructed is consistent. In that case, an alternative hypothesis H_1 here is any pair (p_1, p_2) such that $p_1(x) > p_2(x)$, $x \in [0,1]$.

მათემატიკა

ბერნულის რეგრესიის ფუნქციის ბერნშტეინის პოლინომებით შეფასების შესახებ

პ. ბაბილუა* და ე. ნადარაია*

** ივანე ჯავახიშვილის თბილისის სახელმწიფო უნივერსიტეტი, ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი, მათემატიკის დეპარტამენტი, თბილისი, საქართველო*
*** აკადემიის წევრი, ივანე ჯავახიშვილის თბილისის სახელმწიფო უნივერსიტეტი, ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი, მათემატიკის დეპარტამენტი, თბილისი, საქართველო*

ნაშრომში ბერნშტეინის პოლინომების საშუალებით აგებულია ბერნულის რეგრესიის ფუნქციის შეფასება. შესწავლილია შეფასების ძალდებულება და ასიმპტოტური ნორმალობა. ბერნულის რეგრესიის ფუნქციის სახის ჰიპოთეზის შემოწმებისთვის აგებულია კრიტერიუმი. აგებულია აგრეთვე ბერნულის რეგრესიის ორი ფუნქციის ტოლობის ჰიპოთეზის შემოწმების კრიტერიუმი. შესწავლილია აგებული კრიტერიუმების ასიმპტოტური ყოფაქცევა.

REFERENCES

1. Efromovich S. (1999) Nonparametric curve estimation. Methods, theory, and applications. Springer Series in Statistics. Springer-Verlag, New York.
2. Copas J. B. (1983) Plotting *p* against *x*. *Appl. Statist.,* **32**, 1: 25-31.
3. Okumura H., Naito K. (2004) Weighted kernel estimators in nonparametric binomial regression. *J. Nonparametr. Stat.,* **16**, 1-2: 39-62.
4. Aerts M., Veraverbeke N. (1995) Bootstrapping a nonparametric polytomous regression model. *Math. Methods Statist.,* **4**, 2: 189-200.
5. Mandzhgaladze K. V. (1986) On some estimate of the distribution function and its moments. *Soobshch. Akad. Nauk Gruz. SSR*, **124**, 261-263 (in Russian).
6. Lorentz G. G. (1986) Bernstein polynomials. Second edition. Chelsea Publishing, New York.
7. Nadaraya E., Babilua P., Sokhadze G. (2013) About the nonparametric estimation of the Bernoulli regression. *Comm. Statist. Theory Methods,* **42**, 22: 3989-4002.
8. Bhattacharya P. K., Gastwirth J. L. (1999) Estimation of the odds-ratio in an observational study using bandwidth-matching. *J. Nonparametr. Statist.* **11**, 1-3: 1-12.
9. Babilua P. K., Nadaraya E. A., Sokhadze G. A. (2015) On the square-integrable measure of the divergence of two kernel estimations of the Bernoulli regression functions. *Ukraïn. Mat. Zh.,* **67**, 1: 3-18 (in Russian); translation in Ukrainian *Math. J.* **67**, 1: 1-18.
10. Abbe E. (1906) Über die Gesetrzmässigkeit in der Vorteilung der Fehler bei Beobachtungsreiben. *Worke,* Bd. 2: 55-81. Jena.
11. Linnik Yu. V. (1958) The method of least squares and the foundations of the mathematical-statistical theory of reduction of observations. M. (in Russian).
12. Absava R., Nadaraya E. (1999) Limit distribution of the mean square deviation of the Gasser-Müller nonparametric estimate of the regression function. *Georgian Math. J.* **6**, 6: 501-516.

Received December, 2019