ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

УДК 519.2

НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ КВАНТИЛЕЙ В МОДЕЛИ БИНАРНОЙ РЕГРЕССИИ

Тихов М.С., Шкилева К.Н.
Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, г. Нижний Новгород

Поступила в редакцию 30.10.2019, после переработки 12.12.2019.

В статье предлагается новая оценка квантильной функции. Она основана на непараметрических модифицированных Рида и Менча оценках функции распределения $F(x)$ в модели бинарной регрессии. Приведены условия состоятельности и асимптотической нормальности оценок. Мы сравниваем предложенные оценки с некоторыми существующими методами. К ним относятся двухъядерный метод Yu и Jones (1998), скорректированная версия оценки Stute (1986), оценки, предложенные Бородиной (2019) на основе подхода Надарая-Ватсона. Сравнение производится с помощью асимптотической среднеквадратичной ошибки и асимптотического среднего. Предлагаемый метод имеет практическое применение, например, для оценки квантиля.

Ключевые слова: зависимость доза-эффект, квантильная функция, модифицированный метод Рида и Менча.

Вестник ТвГУ. Серия: Прикладная математика. 2020. № 1. С. 5–19.
https://doi.org/10.26456/vtpmk552

Введение

В статье рассматриваются новые оценки квантилей функции распределения на базе модифицированных Рида и Менча оценок функции распределения в модели доза-эффект. которая описывается следующим образом.

Пусть случайная величина X имеет распределение с неизвестной функцией распределения $F(x)$ и плотностью $f(x) > 0$. Пусть имеется еще одна случайная величина U, плотность распределения которой равна $g(x) > 0$ и которая может быть как известна, так и неизвестна. Наблюдению доступны пары (u_i, w_i), $1 \le i \le n$, (повторная выборка), где w_i есть значение случайной величины W_i − индикатора события $(X_i < U_i)$. Требуется по наблюдениям $\mathcal{U}^{(n)} = \{(u_i, w_i)\}_{i=1}^{n}$ оценить неизвестную функцию распределения $F(x)$ или её квантили.

В работе [1] была предложена следующая модифицированная Рида и Менча оценка функции распределения.

Пусть

$$S_{1n}(x)=\frac{1}{nh}\sum_{i=1}^{n}W_iH\left(\frac{U_i-x}{h}\right),\ S_{2n}(x)=\frac{1}{nh}\sum_{i=1}^{n}(1-W_i)\left(1-H\left(\frac{U_i-x}{h}\right)\right),$$

$$S_{3n}(x)=S_{1n}(x)+S_{2n}(x).$$

Определим $H(x)=\int\limits_{-1}^{x}K(t)\,dt$, для $-1\le x\le 1$, и $H(x)=0$, если $x\notin[-1,1]$.

Оценку функции распределения $F(x)$ в точке x будем находить по формуле

$$\hat{F}_n(x)=\frac{S_{1n}(x)}{S_{3n}(x)}, \tag{1}$$

считая, что $\hat{F}_n(x)=0$, если $S_{3n}(x)=0$.

В работе [1] было показано, что оценка $\hat{F}_n(x)$ является состоятельной и асимптотически нормальной: при некоторых условиях регулярности

$$n^{2/5}(\hat{F}_n(x)-\mathbf{E}(\hat{F}_n(x)))\xrightarrow[n\to\infty]{d}N(0,\|H\|^2F(x)(1-F(x))/g(x)),$$

где $\|H\|^2=\int\limits_{-\infty}^{\infty}H^2(x)\,dx$.

Если $u_i, 1\le i\le n$ – неслучайны, соответствующий план испытаний $(u_i,W_i), 1\le i\le n$ будем называть *фиксированным*. В дальнейшем мы будем рассматривать *фиксированные* планы испытаний, более того, будем предполагать, что $u_i=i/n, 1\le i\le n$.

Определим оценку $\hat{\xi}_{n,\lambda}$ квантиля ξ_λ порядка $0<\lambda<1$ как

$$\hat{\xi}_{n\lambda}=\inf\{x\in R,\ F_n^*(x)\ge\lambda\}. \tag{2}$$

Здесь мы докажем, что оценка $\hat{\xi}_{n,\lambda}$, определенная формулой (2), является состоятельной и асимптотически нормальной оценкой ξ_λ с предельной дисперсией $\sigma^2=\lambda(1-\lambda)\|H\|^2/f^2(\xi_\lambda)$.

1. Вспомогательные результаты

Для исследования оценок по фиксированным планам эксперимента нам понадобится понятие *вариации* функции ([2], с.234).

Пусть на отрезке $[a,b]$ задана функция $f(x)$. Разобьем $[a,b]$ на части точками $x_0=a<x_1<x_2<...<x_n=b$ и составим сумму

$$\boldsymbol{VS}(f)=\sum_{k=0}^{n-1}|f(x_{k+1})-f(x_k)|.$$

Определение 1. *Точная верхняя грань множества всевозможных сумм $\boldsymbol{VS}(f)$ называется вариацией функции $f(x)$ на отрезке $[a,b]$ и обозначается $\boldsymbol{V}_a^b(f)$, т.е.*

$$\boldsymbol{V}_a^b(f)=\sup_{\mathcal{P}_n}\boldsymbol{VS}(f)=\sup_{\mathcal{P}_n}\sum_{k=0}^{n-1}|f(x_{k+1})-f(x_k)|,\ \textit{где}\quad \mathcal{P}_n=\{x_0,x_1,...,x_n,x_{n+1}\}.$$

Если $\boldsymbol{V}_a^b(f) < \infty$, то говорят, что $f(x)$ есть *функция конечной* (*ограниченной*) *вариации* на отрезке $[a, b]$. Если функция $f : [a, b] \to \mathbf{R}$ принадлежит классу $C^1[a, b]$, то есть имеет непрерывную производную первого порядка, то f – функция ограниченной вариации на этом отрезке, а ее вариацию можно вычислить как

$$\boldsymbol{V}_a^b(f) = \int\limits_a^b |\, f'(x)\,|\, dx.$$

Далее, пусть P есть множество точек $x_1, x_2, ..., x_N$ из $[0, 1]$ и $\mathcal{B}$ – лебегова $\sigma-$алгебра на $[0, 1]$, а μ – лебегова мера на $\mathcal{B}$. Определим

$$A(B, P) = \sum_{i=1}^{N} I_B(x_i), \quad D_N(B, P) = \left| \frac{A(B, P)}{N} - \mu(B) \right|,$$

где $I_B(x)$ – индикатор множества B. Положим $D_N^*(P) = D_N(J_c^*, P)$, где J_c^* есть семейство подинтервалов на $[0, 1]$ вида $[0, u_i)$.

Теорема 1 (3, с.18, Koksma, Hlawka)**.** *Если функция* $f(u)$ $(0 \le u \le 1)$ *имеет ограниченную вариацию* $\boldsymbol{V}_a^b(f)$, *то для любых* $0 < x_1 < x_2 < ... < x_N < 1$ *мы имеем*

$$\left| \, \frac{1}{N} \sum_{i=1}^{N} f(x_i) - \int_0^1 f(x)\, du \, \right| \le \boldsymbol{V}_a^b(f) D_N^*(x_1, ..., x_N)\,.$$

Приведем также еще две леммы из [3], которые проясняют суть этого неравенства.

Лемма 1. *Если* $x_1, ..., x_N$, $y_1, ..., y_N$, $x_i, y_i \in [0, 1]$, *удовлетворяют неравенствам* $|\, x_i - y_i| \le \varepsilon$ *для* $\varepsilon > 0$, $1 \le i \le N$, *то* $|\, D_N^*(x_1, ..., x_N) - D_N^*(y_1, ..., y_N)\,| \le \varepsilon$.

Из леммы 1 следует, что $D_N^*(x_1, ..., x_N)$ есть непрерывная функция переменных $x_1, ..., x_N$.

Лемма 2. *Если* $0 < x_1 < ... < x_N < 1$, *то*

$$D_N^*(x_1, ..., x_N) = \frac{1}{2N} + \max_{1 \le i \le N} \left| \, x_i - \frac{2i - 1}{2N} \, \right|.$$

Если $x_i = \dfrac{i}{N}$, то $\dfrac{i}{N} - \dfrac{2i - 1}{2N} = \dfrac{1}{2N}$ и $D_N^*(x_1, ..., x_N) = \dfrac{1}{N}$.

2. Предположения

Пусть $K(x)$ – ядерная функция, где $x \in \mathbf{R}$.

Предположения (K).

$(\mathbf{K}_1)$ $K(x) \ge 0$, *причем* $K(x) = 0, \ x \notin [-1, 1]$. $\qquad (\mathbf{K}_2)$ $\int\limits_{-\infty}^{\infty} K(x)\, dx = 1.$

($\mathbf{K}_3$) $K(x) = K(-x),\ x \in \mathbf{R}$. ($\mathbf{K}_4$) $\|K\|_\infty = \sup\limits_{x\in\mathbf{R}} |\,K(x)\,| = \kappa < \infty$.

($\mathbf{K}_5$) *Существуют непрерывные ограниченные производные функции* $K(x)$ *до третьего порядка включительно в интервале* $(-1, 1)$.

Так как при приведенных условиях $\int\limits_{-\infty}^{\infty} K^2(x)\,dx < \infty$ и $\int\limits_{-\infty}^{\infty} H^2(x)\,dx < \infty$, то положим

$$\|K\|^2 = \int\limits_{-\infty}^{\infty} K^2(x)\,dx, \quad \|H\|^2 = \int\limits_{-\infty}^{\infty} H^2(x)\,dx.$$

При условиях **(K)** преднорма $\|K\|^2$ ядра K конечна и существуют четвертые моменты для распределений с плотностями $K(x)$:

$$\nu^2 = \int_{-\infty}^{\infty} x^2 K(x)\,dx, \quad \mu^4 = \int_{-\infty}^{\infty} x^4 K(x)\,dx.$$

Кроме того,

$$\int_{-1}^{1} H(t)\,dt = \int_{-1}^{0} H(t)\,dt + \int_{0}^{1} (H(t) - 1)\,dt + 1 = 1 - \mu_1(K) = 1,$$

поскольку $\mu_1(K) = 0$ и

$$\int_{-1}^{1} t\,H(t)\,dt = \frac{1}{2}\,(1 - \mu_2(K)),\ \mu_3(K) = \int_{-1}^{1} t^2 H(t)\,dt = \frac{1}{3},$$

где $\mu_2 = \int\limits_{-1}^{1} t^2 K(t)\,dt,\ K(t) = K(-t)$.

Примером функции, удовлетворяющей условиям ($\mathbf{K}$), служит:

1. $K(x) = (3/4)(1 - x^2)I(|x| \le 1)$ (ядро Епанечникова),

$$\|K\|^2 = 3/5 = 0.6, \quad \nu^2 = 1/5 = 0.2,$$

2. $K(x) = (15/16)(1 - x^2)^2 I(|x| \le 1)$ (квартическое ядро),

$$\|K\|^2 = 5/7 = 0.714, \quad \nu^2 = 1/7 = 0.143.$$

Мы будем рассматривать ядра $K(x)$ *такие, что* $\mu_2(K) = 1$.

Пример 1. Пусть $K_0(x) = (1/\sqrt{5})K_1(x/\sqrt{5}),\ -\sqrt{5} < x < \sqrt{5}$, и ноль в остальных случаях, тогда $\mu_2(K_0) = 1$ и $1 - \mu_2(K_0) = 0$.

Пусть $h = h(n)$ – ширина окна просмотра данных.

Предположение (H).

($\mathbf{H}_1$) *При* $n \to \infty, \quad h \to 0,\ nh \to \infty$.

Условие ($\mathbf{H}_1$) означает, что по мере получения большего количества информации из выборки, т.е. при $n \to \infty$, усреднение данных происходит по более узкой области ($h \to 0$), но в то же время количество «локальной информации» (nh) должно увеличиваться.

Примером числовой последовательности, удовлетворяющей условию $(\mathbf{H}_1)$, является $h = cn^{-1/5}$, c – некоторая положительная константа.

Предположение (F).

$(\mathbf{F}_1)$ *Плотность* $f(x)$ *определена на отрезке* $[0,1]$, *отделима от нуля, т.е* $f(x) \geq c_0 > 0$ *для* $0 \leq x \leq 1$, *и существует третья непрерывная ограниченная производная плотности на интервале* $(0,1)$.

Пусть $\mathcal{P}_n = \{u_0, u_1, ..., u_n, u_{n+1}\}$ – упорядоченное разбиение отрезка $[0,1]$, где $u_0 = 0 < u_1 < ... < u_n < 1 = u_{n+1}$.

Предположение (P).

$(\mathbf{P}_1)$ *При* $n \to \infty$, $\max\limits_{k\in\{0,1,\dots,n\}} \max\left\{\left|u_k - \dfrac{k}{n}\right|, \left|u_{k+1} - \dfrac{k}{n}\right|\right\} = O(n^{-1})$.

Из условия $(\mathbf{P})$ следует, что $u_k = \dfrac{k}{n} + O(n^{-1})$, причем последовательность $n\left(u_k - \dfrac{k}{n}\right)$ равномерно по $0 \leq k \leq n$ ограничена некоторой константой C.

Лемма 3 ([4]). *Если выполнены предположения **(K)**, **(H)** , **(F)** , **(P)**, то*

$$\sup_{\mathcal{P}_n} \sum_{j=1}^{l} |\ \tilde{q}(u_j) - \tilde{q}(u_{j-1})\ | = O(h_d^{-1}),$$

где sup *берется по всем возможным упорядоченным разбиениям* $\mathcal{P}_n$.

3. Результаты

Сейчас мы рассмотрим и исследуем асимптотическое поведение оценок квантилей в зависимости доза-эффект по фиксированным планам эксперимента, а именно, считая, что величина U неслучайна и принимает значения $u_i = i/n$, $i = 1, 2, ..., n$. Определим следующую оценку функции распределения $F(x)$ по фиксированным планам:

$$F_n^*(x) = \frac{S_{1n}^*(x)}{S_{3n}^*(x)}\,, \tag{3}$$

где

$$S_{1n}^*(x) = \frac{1}{nh}\sum_{i=1}^{n} w_i H\left(\frac{i/n - x}{h}\right), S_{2n}^*(x) = \frac{1}{nh}\sum_{i=1}^{n}(1 - w_i)\left(1 - H\left(\frac{i/n - x}{h}\right)\right),$$

$$S_{3n}^*(x) = S_{1n}^*(x) + S_{2n}^*(x).$$

Имеем:

$$\mathbf{E}(S_{1n}(x)^*) = \frac{1}{nh}\mathbf{E}\left(\sum_{i=1}^{n} W_i H\left(\frac{i/n - x}{h}\right)\right) = \frac{1}{nh}\mathbf{E}\left(\sum_{i=1}^{n} F\left(\frac{i}{n}\right) H\left(\frac{i/n - x}{h}\right)\right)$$

$$= \frac{1}{h}\int_{-\infty}^{+\infty} H\left(\frac{u-x}{h}\right) F(u)\ du + o\left(\frac{1}{\sqrt{nh}}\right) = \int_{-\infty}^{+\infty} H(t)\,F(x+ht)\ dt + o\left(\frac{1}{\sqrt{nh}}\right) =$$

$$= F(x)\int_{-1}^{1} H(t)\,dt + \frac{1}{2}f'(x)h^2\int_{-1}^{1} t^2 H(t)\,dt + o(h^2) = F(x) + \frac{1}{6}f'(x)h^2 + o(h^2), \quad (4)$$

а

$$\mathbf{D}(S_{1n}(x)^*) = \frac{\sigma^2}{nh}(1+o(1), \quad n\to\infty.$$

Поэтому

$$S^*_{1n}(x) \xrightarrow[n\to\infty]{p} F(x).$$

Аналогично показывается, что последовательность статистик $S^*_{3n}(x)$ сходится к 1, поэтому вместо оценки $F^*_n(x)$ можно использовать оценку $S^*_{1n}(x)$.

В теореме 2 мы докажем состоятельность, а в теореме 3 асимптотическую нормальность оценок $\hat{\xi}_{n\lambda}$. Установим предварительно следующий результат.

Лемма 4. *(i) Пусть c_n и y_n – последовательности констант таких, что*

$$\underline{\lim}_{n\to\infty}(F(c_n) - y_n) = l > 0.$$

Тогда

$$\mathbf{P}\left(F^*_n(c_n) < y_n\right) \to 0(n\to\infty).$$

(ii) Если последовательности d_n и z_n удовлетворяют условию

$$\overline{\lim}_{n\to\infty}(F(d_n) - z_n) = -L < 0,$$

то

$$\mathbf{P}\left(F^*_n(d_n) > z_n\right) \to 0(n\to\infty).$$

Доказательство. Заметим, что для всех достаточно больших n справедливы неравенства $F(c_n) - y_n > l/2$, поэтому

$$\mathbf{P}(\hat{F}_n(c_n) < y_n) = \mathbf{P}(F^*_n(c_n) - F(c_n) < y_n - F(c_n)) \le \mathbf{P}(F^*_n(c_n) - F(c_n) < -l/2),$$

откуда, пользуясь неравенством Чебышева, получим

$$\mathbf{P}(F^*_n(c_n) < y_n) \le \mathbf{P}(\,|F^*_n(c_n) - F(c_n)|\ > l/2) \le \frac{4\,\mathbf{E}((F^*_n(c_n) - F(c_n))^2)}{l^2}\,.$$

Заметим, что

$$\mathbf{E}((F^*_n(c_n) - F(c_n))^2) = \mathbf{D}(F^*_n(c_n)) + \mathbf{E}^2((F^*_n(c_n) - F(c_n)).$$

Используя результаты работы [5], согласно которым

$$\mathbf{E}(F^*_n(c_n)) = F(c_n) + \frac{||\,H\,||^2h^2}{2}f'(c_n)(1+o(1)), \quad \text{где} \quad \nu^2 = \int_{-\infty}^{\infty} x^2K(x)\,dx,$$

можно показать, что при $n\to\infty$,

$$\mathbf{D}(F^*_n(c_n)) = \frac{F(c_n)(1-F(c_n))\,||H||^2}{nh}(1+o(1)).$$

Значит,

$$\mathbf{E}((F_n^*(c_n) - F(c_n))^2) = \left(\frac{F(c_n)(1 - F(c_n))||H||^2}{nh} + \frac{||H||^4 h^4}{4} (f'(c_n))^2 \right) (1 + o(1)).$$

Поскольку $|\, f'(x)| \ \le M_1$, то для $n \ge n_0$

$$\mathbf{E}((F_n^*(c_n) - F(c_n))^2) \le \frac{||H||^2}{nh} + \frac{||H||^4 h^4 M_1^2}{4}.$$

Следовательно,

$$\mathbf{P}(F_n^*(c_n) < y_n) \le \frac{4\,||K||^2}{l^2 nh} + \frac{||H||^4 h^4 M_1^2}{l^2} \underset{n\to\infty}{\longrightarrow} 0.$$

Аналогично показывается, что при $n \to \infty$,

$$\mathbf{P}\left(F_n^*(d_n) > z_n\right) \to 0.$$

Лемма доказана. □

Установим теперь состоятельность оценок (2).

Теорема 2 (Слабая состоятельность). *Пусть выполнены предположения (**H**), (**K**), (**F**), (**P**). Тогда*

$$\hat{\xi}_{n\lambda} \xrightarrow[n\to\infty]{p} \xi_\lambda.$$

Доказательство. Для произвольного $\varepsilon > 0$ имеем:

$$\mathbf{P}(\hat{\xi}_{n\lambda} > \xi_\lambda + \varepsilon) = \mathbf{P}(F_n^*(\xi_\lambda + \varepsilon) \le \lambda).$$

Поскольку $F(\xi_\lambda + \varepsilon) - \lambda = l > 0$, то

$$\lim_{n \to \infty} \mathbf{P}(\hat{\xi}_{n\lambda} > \xi_\lambda + \varepsilon) = \lim_{n \to \infty} \mathbf{P}(F_n^*(\xi_\lambda + \varepsilon) \le \lambda) = 0.$$

Точно так же показывается, что $\lim\limits_{n \to \infty} \mathbf{P}(\hat{\xi}_{n\lambda} < \xi_\lambda - \varepsilon) = 0$. Из этих соотношений получаем, что

$$\lim_{n \to \infty} \mathbf{P}(|\hat{\xi}_{n\lambda} - \xi_\lambda| \ > \varepsilon) = 0, \quad \Leftrightarrow \quad \hat{\xi}_{n\lambda} \xrightarrow[n\to\infty]{p} \xi_\lambda, \quad \text{т.е. результат теоремы 2.}$$

□

В следующей теореме утверждается асимптотическая нормальность оценок $\hat{\xi}_{n\lambda}$.

Теорема 3. *Пусть $\hat{\xi}_{n\lambda}-$ оценка квантиля порядка $0 < \lambda < 1$, определенная формулой (2), выполнены предположения (**H**), (**K**), (**F**), (**P**), $h = n^{-1/5}$ и*
$a = \dfrac{f'(\xi_\lambda)\nu^2}{2\sqrt{\lambda(1-\lambda)}}$. *Тогда*

$$\sqrt{nh}(\hat{\xi}_{n\lambda} - \xi_\lambda - ah^2) \xrightarrow[n\to\infty]{d} N\left(0, \frac{\lambda(1-\lambda)||H||^2}{f^2(\xi_\lambda)}\right).$$

Доказательство. Обозначим $\sigma^2 = \lambda\,(1-\lambda)$, $H_h(x) = \dfrac{1}{h} H\left(\dfrac{x}{h}\right)$,

$$\mu_i = \mathbf{E}\left(W_i H_h\left(\frac{i}{n} - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right)\right) = F\left(\frac{i}{n}\right)\, H_h\left(\frac{i}{n} - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right).$$

Тогда

$$\mathbf{P}\left(\frac{\sqrt{nh} f(\xi_\lambda)(\hat{\xi}_{n\lambda} - \xi_\lambda)}{\sigma} \le x\right) = \mathbf{P}\left(\hat{\xi}_{n\lambda} \le \xi_\lambda + \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right) =$$

$$= \mathbf{P}\left(F_n^*\left(\xi_\lambda + \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right) \ge \lambda\right) = \mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n W_i H_h\left(\frac{i}{n} - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right) \ge \lambda\right) =$$

$$= \mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n \left(\mu_i - W_i H_h\left(\frac{i}{n} - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right)\right) \le \frac{1}{n}\sum_{i=1}^n (\mu_i - \lambda)\right) =$$

$$= \mathbf{P}\left(\frac{\sqrt{nh}}{\sigma}\cdot\frac{1}{n}\sum_{i=1}^n \left(\mu_i - W_i H_h\left(\frac{i}{n} - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right)\right) \le \frac{\sqrt{nh}}{\sigma}\cdot\frac{1}{n}\sum_{i=1}^n (\mu_i - \lambda)\right).$$

Отсюда и из теоремы 2 получаем

$$\frac{1}{n}\sum_{i=1}^n \mu_i = \frac{1}{n}\sum_{i=1}^n F\left(\frac{i}{n}\right)\, H_h\left(\frac{i}{n} - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right) =$$

$$= \int_0^1 F(u) H_h\left(u - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right)\, du + o\left(\frac{1}{\sqrt{nh}}\right).$$

Делая в последнем интеграле замену переменной $z = \dfrac{1}{h}\left(u - \xi_\lambda - \dfrac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right)$, заключаем, что

$$\alpha_n = \int_0^1 F(u) H_h\left(u - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right)\, du =$$

$$= \int_{-(\xi_\lambda + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}})/h}^{(1-\xi_\lambda - \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}})/h} F\left(\xi_\lambda + zh + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}}\right) H(z)\, dz,$$

поскольку функция $H_h\left(u - \xi_\lambda - \dfrac{x\sigma}{\sqrt{nh} f(\xi_\lambda)}\right)$ равна нулю вне интервала $\left(-\dfrac{\xi_\lambda + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}}}{h}, \dfrac{1 - \xi_\lambda - \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}}}{h}\right)$. Для достаточно большого n $(n \ge n_1)$

$$\int_{-(\xi_\lambda + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}})/h}^{(1-\xi_\lambda - \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}})/h} F\left(\xi_\lambda + zh + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}}\right) H(z)\, dz =$$

$$= \int_{-1}^1 F\left(\xi_\lambda + zh + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}}\right) H(z)\, dz.$$

Так как $\mu_2(K) = 1$, то

$$\alpha_n = \int_{-1}^{1} \left(F(\xi_\lambda) + f(\xi_\lambda) \left(zh + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}} \right) \right) H(z)\, dz +$$

$$+ \int_{-1}^{1} \left(\frac{f'(\xi_\lambda)}{2} \left(zh + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}} \right)^2 + \right.$$

$$\left. + \frac{f''(\theta_n)}{6} \left(zh + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}} \right)^3 \right) H(z)\, dz = F(\xi_\lambda) +$$

$$+ \left(\frac{x\sigma}{\sqrt{nh}} + \frac{f'(\xi_\lambda)\nu^2 h^2}{2} + \frac{x^2\sigma^2 f'(\xi_\lambda)}{2f^2(\xi_\lambda)nh} \right) + \int_{-1}^{1} \left(\frac{f''(\theta_n)}{6} \left(zh + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}} \right)^3 \right) H(z)\, dz,$$

где $-1 \le \theta_n \le 1$ и поскольку $\int_{-1}^{1} z\, K(z)\, dz = 0$ (функция $K(z)$ –четная).

Обозначим

$$\rho_1 = \int_{-1}^{1} |z|\, H(z)\, dz \le 1, \quad \rho_2 = \int_{-1}^{1} |z|^3\, K(z)\, dz \le 1.$$

Используя ограниченность второй производной плотности ($|f''(\theta_n)| \le M_2$), имеем:

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mu_i - \lambda - \left(\frac{x\sigma}{\sqrt{nh}} + \frac{f'(\xi_\lambda)h^2}{2} + \frac{x^2\sigma^2 f'(\xi_\lambda)}{2f^2(\xi_\lambda)nh} \right) \right| \le$$

$$\le \frac{M_2}{6} \left(h^3 \rho_2 + \frac{3\,\nu^2 |x|\sigma}{f(\xi_\lambda)} n^{-1/2} h^{3/2} + \frac{3\rho_1\, x^2\, \sigma^2}{n\, f^2(\xi_\lambda)} + \frac{|x|^3 \sigma^3}{f^3(\xi_\lambda)} (n\, h)^{-3/2} \right).$$

(Так как $h = n^{-1/5}$, то $\sqrt{nh} = n^{2/5}$, $h^2 = n^{-2/5}$, $n^{-1/2}h^{3/2} = n^{-4/5}$, $(nh)^{-3/2} = n^{-6/5}$). Умножая обе части на $\dfrac{\sqrt{nh}}{\sigma}$ и учитывая, что $nh^5 = 1$, получим

$$\left| \frac{\sqrt{nh}}{\sigma} \frac{1}{n} \sum_{i=1}^{n} (\mu_i - \lambda) - \left(x + \frac{f'(\xi_\lambda)}{2\sigma} + \frac{x^2 \sigma f'(\xi_\lambda)}{2f^2(\xi_\lambda)\sqrt{nh}} \right) \right| \le$$

$$\le \frac{M_2}{6} \left(\frac{h\rho_2}{\sigma} + \frac{3\nu^2 x}{f(\xi_\lambda)} h^2 + \frac{3\rho_1\, x^2 \sigma}{n^{1/2} f^2(\xi_\lambda)} h^{1/2} + \frac{|x|^3 \sigma^2}{f^3(\xi_\lambda) nh} \right) \equiv A_n.$$

Отсюда следует, что величина $\dfrac{\sqrt{nh}}{\sigma} \dfrac{1}{n} \sum_{i=1}^{n} (\mu_i - \lambda) - x - \dfrac{f'(\xi_\lambda)\nu^2}{2\sigma}$ равномерно сходится к нулю по $|x| \le T$ при $n \to \infty$.

Пусть

$$\Sigma_n(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\mu_i - W_i H_h \left(\frac{i}{n} - \xi_\lambda - \frac{x\sigma}{\sqrt{nh} f(\xi_\lambda)} \right) \right).$$

Так как

$$\left| \int_{-1}^{1} \left(\frac{f''(\theta_n)}{6} \left(zh + \frac{x\sigma}{f(\xi_\lambda)\sqrt{nh}} \right)^3 \right) H(z)\, dz \right| \le A_n \underset{n\to\infty}{\longrightarrow} 0,$$

то для доказательства теоремы 3 достаточно показать, что

$$\frac{\sqrt{nh}}{\sigma}\cdot\Sigma_n(x)\xrightarrow[n\to\infty]{d}N(0,||H||^2).$$

Для этого рассмотрим дисперсию величины $\Sigma_n(x)$. Имеем:

$$\mathbf{D}(\Sigma_n(x))=\frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbf{D}\left(W_iH\left(\frac{1}{h}\left(\frac{i}{n}-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right)\right)\right)=$$

$$=\frac{1}{n^2h^2}\sum_{i=1}^{n}H^2\left(\frac{1}{h}\left(\frac{i}{n}-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right)\right)\mathbf{D}\left(W_i\right)=$$

$$=\frac{1}{n^2h^2}\sum_{i=1}^{n}H^2\left(\frac{1}{h}\left(\frac{i}{n}-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right)\right)F\left(\frac{i}{n}\right)\left(1-F\left(\frac{i}{n}\right)\right)=$$

$$=\frac{1}{nh^2}\int_0^1F(u)(1-F(u))H^2\left(\frac{u-\xi_\lambda}{h}\right)\,du\,(1+o(1))$$

равномерно по x в замкнутом (и ограниченном) интервале.

Действительно, разобьем отрезок $[\,0,1\,]$ точками $\frac{i}{n}$. Тогда

$$\int_0^1F(u)(1-F(u))H^2\left(\frac{u-\xi_\lambda}{h}\right)\,du\,=\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n}F(u)(1-F(u))H^2\left(\frac{u-\xi_\lambda}{h}\right)\,du\,.$$

Положим $a_i\,=\,F(u)(1-F(u))$ на интервале $\left[\frac{i-1}{n},\frac{i}{n}\right]$, $b_i\,=\,H^2\left(\frac{u-\xi_\lambda}{h}\right)$ и $\tilde{a}_i=F\left(\frac{i}{n}\right)\left(1-F\left(\frac{i}{n}\right)\right)$, $\tilde{b}_i=H^2\left(\frac{1}{h}\left(\frac{i}{n}-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right)\right)$.

В силу неравенств

$$|\,a_ib_i-\tilde{a}_j\tilde{b}_j\,|\ \le\ |a_i|\cdot|\,b_i-\tilde{b}_j\,|\,+\,|\tilde{b}_j|\cdot|a_i-\tilde{a}_j|,F(x)(1-F(x))\le\frac{1}{4},H(u)\le1,$$

и того, что плотность $f(x)$ обладает ограниченными производными, а $H(u)$ имеет ограниченную производную, выводим, что разность

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}H^2\left(\frac{1}{h}\left(\frac{i}{n}-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right)\right)F\left(\frac{i}{n}\right)\left(1-F\left(\frac{i}{n}\right)\right)-$$

$$-\frac{1}{nh^2}\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n}F(u)(1-F(u))H^2\left(\frac{u-\xi_\lambda}{h}\right)\,du$$

будет ограничена величиной $(C_2+C_3|\,x\,|)h^2$, поэтому для $|\,x\,|\ \le\ T$ эта разность равномерно сходится к нулю. Делая замену переменных $z=\frac{u-\xi_\lambda}{h}$ в последнем интеграле, получаем

$$\mathbf{D}(\Sigma_n(x))=\frac{1}{nh}\int_{-1}^{1}F(\xi_\lambda+zh)(1-F(\xi_\lambda+zh))H^2(z)\,dz\ \,(1+o(1)).$$

Так как

$$F(\xi_\lambda+zh)=F(\xi_\lambda)+f(\xi_\lambda)zh+\frac{f'(\xi_\lambda)}{2}z^2h^2+o(h^2)=\lambda+f(\xi_\lambda)zh+\frac{f'(\xi_\lambda)}{2}z^2h^2+o(h^2),$$

то равномерно по x, $\mathbf{D}(\Sigma_n(x))=\dfrac{\lambda\,(1-\lambda)\,||H||^2}{nh}\,(\,1+o(1))$. Следовательно,

$$\mathbf{D}\left(\frac{\sqrt{nh}}{\sigma}\Sigma_n(x)\right)\underset{n\to\infty}{\longrightarrow}\|H\|^2.$$

Проверим теперь условия Ляпунова. Для этого рассмотрим

$$C_n=\frac{1}{n^4}\sum_{i=1}^{n}\mathbf{E}\left(\left(W_i-F\left(\frac{i}{n}\right)\right)^4H_h^4\left(\frac{i}{n}-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right)\right)=$$

$$=\frac{1}{n^4}\sum_{i=1}^{n}\left(F\left(\frac{i}{n}\right)-4F^2\left(\frac{i}{n}\right)+6F^3\left(\frac{i}{n}\right)-3F^4\left(\frac{i}{n}\right)\right)\cdot H_h^4\left(\frac{i}{n}-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right).$$

Определим функцию $A(t)=t-4t^2+6t^3-3t^4$, $0\le t\le 1$, которая достигает своего наибольшего значения в точках $t_{1,2}=\dfrac{1}{2}\pm\dfrac{\sqrt{3}}{6}$ (при этом $A(t_{1,2})=\dfrac{1}{12}$). Из неравенства Koksma-Hlawka и ограниченности функции $A(t)$ следует

$$|\,C_n\,|\le\frac{1}{12n^4}\sum_{i=1}^{n}H_h^4\left(\frac{i}{n}-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right)=$$

$$=\frac{1}{12n^3}\int_{-1}^{1}H_h^4\left(u-\xi_\lambda-\frac{x\sigma}{\sqrt{nh}f(\xi_\lambda)}\right)du+$$

$$+O\left(\frac{1}{n^4h}\right)=\frac{1}{12n^3h^3}\int_{-1}^{1}H^4(z)\,dz\,(1+o(1)).$$

Из предположений ($\mathbf{K}$) на ядро следует, что $\rho_4=\int_{-1}^{1}H^4(z)\,dz<2$. Значит, для дроби Ляпунова имеет место соотношение:

$$L_n=\frac{C_n}{\mathbf{D}^2(\Sigma_n(x))}\sim\frac{n^2h^2\rho_4}{12n^3h^3\lambda^2(1-\lambda)^2||K||^4}=\frac{\rho_4}{12nh\lambda^2(1-\lambda)^2||K||^4}\underset{n\to\infty}{\longrightarrow}0.$$

Таким образом, выполнены условия центральной предельной теоремы Ляпунова [20], откуда $\sqrt{nh}\Sigma_n(x)\xrightarrow[n\to\infty]{d}N\left(0,\lambda\,(1-\lambda)||H||^2\right)$. Теорема 3 доказана. □

Заключение и выводы

В работе рассмотрена задача статистического оценивания квантиля ξ_λ порядка $0<\lambda<1$ неизвестной функции распределения $F(x)$ на основе оценок модифицированного метода Рида и Менча (см. [1]) по выборке фиксированного объема. Они состоятельны и асимптотически нормальны. Это позволяет при больших объемах выборки рассчитывать доверительные интервалы для значений функции распределения. Наша оценка квантиля (эффективных доз) в основном сохраняет свойства смещения и дисперсии при сравнении с оценками, построенными Yu и Jones [6], Stute [7], Бородиной [8]. Однако предложенная здесь оценка более устойчива при оценивании экстремальных квантилей.

Список литературы

[1] Тихов М.С., Шкилева К.Н. Модифицированный метод оценивания Рида и Менча в зависимости доза-эффект // Вестник ТвГУ. Серия: Прикладная математика. 2019. № 4. С. 5–26. https://doi.org/10.26456/vtpmk543

[2] Натансон И.П. Теория функций вещественной переменной. М.: Наука, 1974. 480 с.

[3] Niederreiter H. Random Number Generation and Quasi-Monte Carlo-Method. Pennsylvania: SIAM Philadelphia, 1992. 241 p.

[4] Тихов М.С. Непараметрическое оценивание эффективных доз по данным бинарных откликов // Уфимский математический журнал. 2013. Т. 5, № 2. С. 94–108.

[5] Смирнов Н.В. Предельные законы распределения для членов вариационного ряда // Труды Математического института им. В.А. Стеклова Академии наук СССР. 1949. Т. 49. С. 5–60.

[6] Yu K., Jones M. Local Linear Quantile Regression // Journal of the American Statistical Association. 1998. Vol. 93, № 441. Pp. 228–237.

[7] Stute W. Conditional Empirical Processes // Annals of Statistics. 1986. Vol. 14, № 2. Pp. 638–647.

[8] Бородина Т.С. Построение и исследование математической модели совместного действия нескольких веществ в зависимости доза-эффект: дис. ... канд. физ.-мат. наук. Нижний Новгород, 2019. 166 с.

[9] Hayes R.L., Mantel N. Procedures for computing the mean age of eruption of human teeth // Journal of Dental Research. 1958. Vol. 35, № 5. Pp. 938–947.

Образец цитирования

Тихов М.С., Шкилева К.Н. Непараметрическое оценивание квантилей в модели бинарной регрессии // Вестник ТвГУ. Серия: Прикладная математика. 2020. № 1. С. 5–19. https://doi.org/10.26456/vtpmk552

Сведения об авторах

1. **Тихов Михаил Семенович**
профессор лаборатории прикладной теории вероятностей кафедры программной инженерии института информационных технологий, математики и механики Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского.

Россия, 603950, г. Нижний Новгород, пр. Гагарина, д. 23, ННГУ им. Н.И. Лобачевского. E-mail: tikhovm@mail.ru

2. **Шкилева Ксения Николаевна**
студентка института информационных технологий, математики и механики Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского.

Россия, 603950, г. Нижний Новгород, пр. Гагарина, д. 23, ННГУ им. Н.И. Лобачевского. E-mail: shkileva98@mail.ru

NONPARAMETRIC ESTIMATION FOR QUANTILE IN BINARY REGRESSION MODELS

Tikhov Mikhail Semenovich
Professor of Applied Probability Theory laboratory at Software Engineering department, Institute of Information Technology, Mathematics and Mechanics, Lobachevsky State University of Nizhniy Novgorod
Russia, 603950, Nyzhniy Novgorod, 23 Gagarin av., UNN.
E-mail: tikhovm@mail.ru

Shkileva Ksenia Nikolaevna
Student of the Institute of Information Technology, Mathematics and Mechanics, Lobachevsky State University of Nizhniy Novgorod
Russia, 603950, Nyzhniy Novgorod, 23 Gagarin av., UNN.
E-mail: shkileva98@mail.ru

Received 30.10.2019, revised 12.12.2019.

In this article we propose a new estimator of the quantile function. It is based on nonparametric modified Reed-Muench estimators of a distribution function $F(x)$ in the binary regression models. Conditions for weak consistency and asymptotic normality are given. We compare the new proposal with some existing methods. Those include the double-kernel technique of Yu and Jones (1998), the adjusted version of the Stute (1986), estimator suggested by Borodina (2019) based on the Nadaraya-Watson type estimators. The Comparison is done by asymptotic mean squared error and asymptotic mean. Our methods also have the practical application, for example to quantile estimation to the work Hayes and Mantel (1958). Calculations on these data are made in Tikhov and Shkileva (2019).

Keywords: dose-effect relationship, quantile function, modified Reed-Muench estimators.

Citation

Tikhov M.S., Shkileva K.N., “Nonparametric estimation for quantile in binary regression models”, *Vestnik TvGU. Seriya: Prikladnaya Matematika [Herald of Tver State University. Series: Applied Mathematics]*, 2020, № 1, 5–19 (in Russian). https://doi.org/10.26456/vtpmk552

References

[1] Tikhov M.S., Shkileva K.N., “A modified Reed-Muench method of estimation in dose-effect relationship”, *Vestnik TvGU. Seriya: Prikladnaya Matematika [Herald of Tver State University. Series: Applied Mathematics]*, 2019, № 4, 5–26 (in Russian), https://doi.org/10.26456/vtpmk543.

[2] Natanson I.P., *Theory of Functions of a Real Variable*, Dover Publications, New York, 2016, 544 pp.

[3] Niederreiter H., *Random Number Generation and Quasi-Monte Carlo-Method*, SIAM Philadelphia, Pennsylvania, 1992, 241 pp.

[4] Tikhov M.S., "Nonparametric estimation of effective doses at quantal response", *Ufimskij matematicheskij zhurnal [Ufa Mathematical Journal]*, **5**:2 (2013), 94–108 (in Russian).

[5] Smirnov N.V., "Limit distribution laws for members of a variation series", *Trudy Matematicheskogo instituta im. V.A. Steklova Akademii nauk SSSR [Proceedings of the Steklov Institute of Mathematics]*, **49** (1949), 5–60 (in Russian).

[6] Yu K., Jones M., "Local Linear Quantile Regression", *Journal of the American Statistical Association*, **93**:441 (1998), 228–237.

[7] Stute W., "Conditional Empirical Processes", *Annals of Statistics*, **14**:2 (1986), 638–647.

[8] Borodina T.S., *Construction and studying of the mathematical model of a drug combination in dose-effect relation*, dis. ... kand. fiz.-mat. nauk, Nizhnij Novgorod, 2019 (in Russian), 166 pp.

[9] Hayes R.L., Mantel N., "Procedures for computing the mean age of eruption of human teeth", *Journal of Dental Research*, **35**:5 (1958), 938–947.