Исследование эффективности некоторых непараметрических критериев проверки гипотез о разрывности функции регрессии и интенсивности пуассоновского процесса

Григорьев Дмитрий Артемович, гр.18.Б04-мм

Санкт-Петербургский государственный университет
Кафедра статистического моделирования

Научный руководитель: д. ф.-м. н., проф. Ермаков М. С.
Рецензент: к. ф.-м. н., с.н.с. Солев В. Н.

8 июня 2022 г.

Постановка задачи

Рассматривается пара случайных величин $(\mathbf{X}, \mathbf{Y}) \sim P$:

- $\mathbf{X}$ — регрессор (например, величина стипендии),
- $\mathbf{Y}$ — зависимая переменная (например, результаты теста).

Есть функция регрессии $m(x)$:

- условное математическое ожидание (УМО),
- условная медиана.

По выборке $(\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ из P нужно проверить гипотезу

$$\mathbb{H}_0 : m(0^+) = m(0^-).$$

Для УМО критерий предложен в [Bertanha, Chung, 2021].

Задача:

- Построить непараметрический критерий для случая условной медианы,
- Сопоставить его с критерием для УМО по устойчивости.

Задачи

Для решения сформулированных задач применены следующие методы:

- Ядерное оценивание условной медианы и плотности интенсивности пуассоновского процесса;
- Перестановочные критерии;
- Сравнение критериев в постановке робастности к шуму.

Полученные результаты:

- Аналитически изучены свойства статистик критериев для условной медианы и пуассоновского процесса, и они провалидированы моделированием;
- Построенный критерий для функции условной медианы устойчив к зашумлённым данным.

Для $(\mathbf{X}, \mathbf{Y}) \sim P$ рассматривается условная медиана в $x = 0$

$$m(x) = \mathbb{M}[\mathbf{Y} \mid \mathbf{X} = x] = \arg\min_a \mathbb{E}[|\mathbf{Y} - a| \mid \mathbf{X} = x].$$

Выборка $(\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ из P разбивается на две:

$(\mathbf{X}_{1i}, \mathbf{Y}_{1i})_{i=1}^{n_1}$ — из P_1,
$\mathbf{X}_{1i} = \mathbf{X}_i \geq 0$,
$\theta(P_1) = m(0^+)$,

$(\mathbf{X}_{2i}, \mathbf{Y}_{2i})_{i=1}^{n_2}$ — из P_2,
$-\mathbf{X}_{2i} = \mathbf{X}_i < 0$,
$\theta(P_2) = m(0^-)$.

Строится статистика T_n:

$$T_n = \sqrt{nh_n}(\hat{\theta}_1 - \hat{\theta}_2),$$

$$\widehat{\theta}_k = \arg\min_a \sum_{i=1}^{n_k} |\mathbf{Y}_{ki} - a| K\left(\frac{\mathbf{X}_{ki}}{h_n}\right),$$

$K : \mathbb{R} \to \mathbb{R}$ — ядро (неотрицательная, симметричная функция, интеграл которой равен 1), $h_n \to 0$, $nh_n \to \infty$

При $h_n = O(n^{-1/3})$, $\frac{n_1}{n} \xrightarrow{p} \lambda \in (0,1)$ было доказано утверждение:

Утверждение

Статистика T_n с оценкой условной медианы асимптотически нормальна с нулевым средним и предельной дисперсией

$$\sigma^2 = \xi^2(P_1)/\lambda + \xi^2(P_2)/(1-\lambda),$$

$$\xi^2(P_k) = \int_0^\infty K^2(u)\,du \frac{1}{f^2_{Y|X}(\theta(P_k) \mid 0^+; P_k) f_X(0^+; P_k)},$$

где плотности $f_X(x; P_k)$ и $f_{Y|X}(y \mid x; P_k)$ дважды дифференцируемы по x всюду, кроме $x = 0$ и отделены от нуля, а их производные ограничены.

Критерий

$$\mathbb{H}_0 \text{ отвергается на уровне } \alpha, \text{ если } |T_n| > \Phi^{-1}_{\sigma^2}(1-\alpha/2),$$

где $\Phi_{\sigma^2}(x)$ — функция распределения $N(0, \sigma^2)$.

Рассматривается модель, где $\mathbf{X} \sim N(0,1)$, $\mathbf{Y} = g(\mathbf{X}) + \varepsilon$, $g(x) = x$, $\varepsilon \sim N(0,1)$; $K(x) = 0.75(1-x^2)_+$, $h_n = n^{-1/3}$.

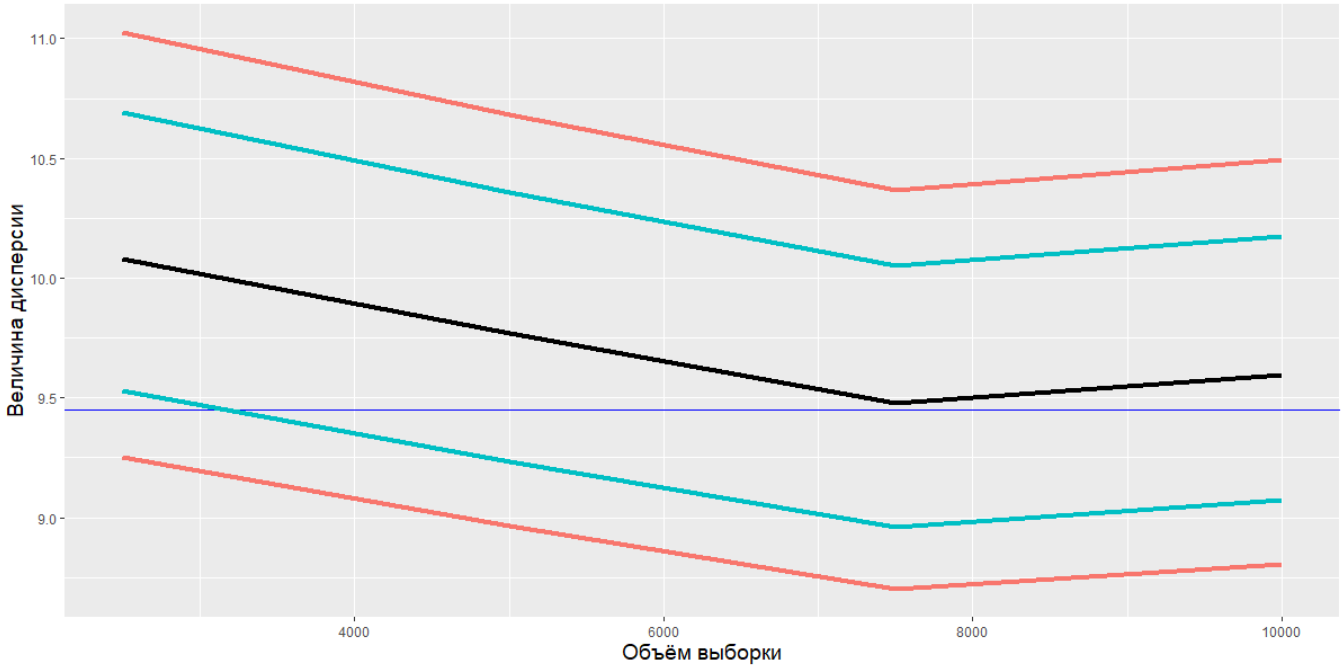

Рис.: Синяя линия демонстрирует теоретическое значение дисперсии. Бирюзовые и красные линии показывают соответственно 80% и 95% асимптотические доверительные интервалы для дисперсии. 1000 повторов эксперимента.

Для меньших объёмов выборки с использованием подхода перестановок [Lehmann, Romano, 2005] построены двухсторонние перестановочные критерии на основе T_n и S_n:

$$S_n = T_n/\widehat{\sigma}_n,\ \widehat{\sigma}_n^2 = \frac{n}{n_1}\widehat{\xi_1^2} + \frac{n}{n_2}\widehat{\xi_2^2},$$

где $\widehat{\xi_k^2}$ — состоятельная оценка ξ_k^2.

Если $P_1 \neq P_2$, то первый некорректен (отсутствует контроль ошибки первого рода).

Моделированием они сравнены друг с другом и с t-критерием: $|S_n| > \Phi^{-1}(1-\alpha/2)$, α — уровень значимости.

Пусть $\mathbf{X} \sim \lambda U(-1, 0) + (1 - \lambda)U(0, 1)$, $\lambda \in (0, 1)$ — смесь равномерных распределений; $\mathbf{Y} = g(\mathbf{X}) + \varepsilon$, где $g(x) = x$ и $\varepsilon \sim N(0, \sigma^2)$ с $\sigma^2 = 1$, если $\mathbf{X} \geq 0$, иначе $\sigma^2 = s^2$, $\lambda \in \{0.5, 0.3\}$, $s^2 \in \{1, 5\}$, $n = 200$.

Уровень значимости $\alpha = 0.05$. Рассматривается контроль ошибки первого рода.

s^2	λ	$\widehat{\alpha}_I(T_n)$	$\widehat{\alpha}_I(S_n)$	$\widehat{\alpha}_I(t)$
1	0.5	0.062	0.053	0.030
5	0.5	0.110	0.039	0.018
1	0.3	0.058	0.032	0.016
5	0.3	0.189	0.072	0.032

Таблица: Значения оценок вероятности ошибки первого рода для перестановочных критериев со статистиками T_n, S_n и t-критерия при различных конфигурациях модели. $K(x) = 0.75(1 - x^2)_+$, $h = 0.1$, 500 повторов эксперимента в 500 случайных перестановках.

Два статистических критерия — на основе УМО и медианы — сравниваются в постановке задачи о робастности их статистик к шуму в данных [Huber, 1981].

Пусть $\beta > 0$, $\mathbf{X} \sim N(0,1)$, $\mathbf{Y} = g(\mathbf{X}) + \varepsilon$, с вероятностью $(1-\beta)$ $\varepsilon \sim N(0,1)$ и с вероятностью β ε имеет распределение Коши.

β	$\widehat{a}_{n,\,mean}$	$\widehat{a}_{n,\,med}$	$\widehat{\sigma^2}_{n,\,mean}$	$\widehat{\sigma^2}_{n,\,med}$
0	0.032	0.075	6.010	8.962
0.01	0.106	0.109	36.577	9.582
0.05	0.160	0.153	1861.970	9.736
0.1	3.348	0.154	10388.807	10.423
0.2	−3.992	0.345	26640.35	10.415
0.5	30.980	0.281	307110.19	11.139

Таблица: Значения оценок среднего и дисперсии двух статистик при шуме Коши, $n = 10000$. 1000 повторов эксперимента.

Рассматривается неоднородный пуассоновский процесс ξ_t с плотностью интенсивности $\lambda(t)$ на отрезке времени $[-1, 1]$. Пусть $X_1, \cdots, X_n$, $X_i = \sum_j \delta_{T_{ji}}$ — выборка из процесса, T_{ji} — время скачков процесса.

$$\mathbb{H}_0 : \lambda(0^+) = \lambda(0^-)$$

Выборка разбивается на две выборки объёмов n_1 и n_2, по каждой оценивается половина разрыва:

$$\widehat{\theta}_m = \frac{1}{2}\left(\widehat{\lambda}(0^+) - \widehat{\lambda}(0^-)\right) = \frac{1}{mh}\sum_{i=1}^{m}\int_{-1}^{1} K\left(\frac{y}{h}\right)\operatorname{sign}(y)\, X_i(dy),$$

где для $B \subset [-1, 1]$ $X_i(B) = \#\{T_{ji} \in B\}$.

Глава 2: Основное утверждения для пуассоновского процесса

Утверждение

Пусть $\lambda(t)$ ограничена, дважды дифференцируема на $[-1,1]$ за исключением точки 0, её производные ограничены, $\frac{n_1}{n} \to 0.5$, $nh_n \to \infty$ и $h_n = O(n^{-1/3})$. Тогда статистика критерия проверки гипотезы о непрерывности плотности интенсивности $T_n = \sqrt{nh_n}(\widehat{\theta}_1 - \widehat{\theta}_2)$ асимптотически нормальна с предельной дисперсией

$$\sigma^2 = 4 \int_0^1 K^2(u)\, du \left(\lambda(0^+) + \lambda(0^-)\right).$$

Критерий

$\mathbb{H}_0$ отвергается на уровне α, если $|T_n| > \Phi^{-1}_{\sigma^2}(1-\alpha/2)$,

где $\Phi_{\sigma^2}(x)$ — функция распределения $N(0,\sigma^2)$.

Рассматривается процесс с плотностью интенсивности $\lambda(t) = \frac{t+1}{2},\ t \in [-1, 1]$. Выбраны $K(x) = 0.75(1 - x^2)_+$, $h_n = n^{-1/3}$.

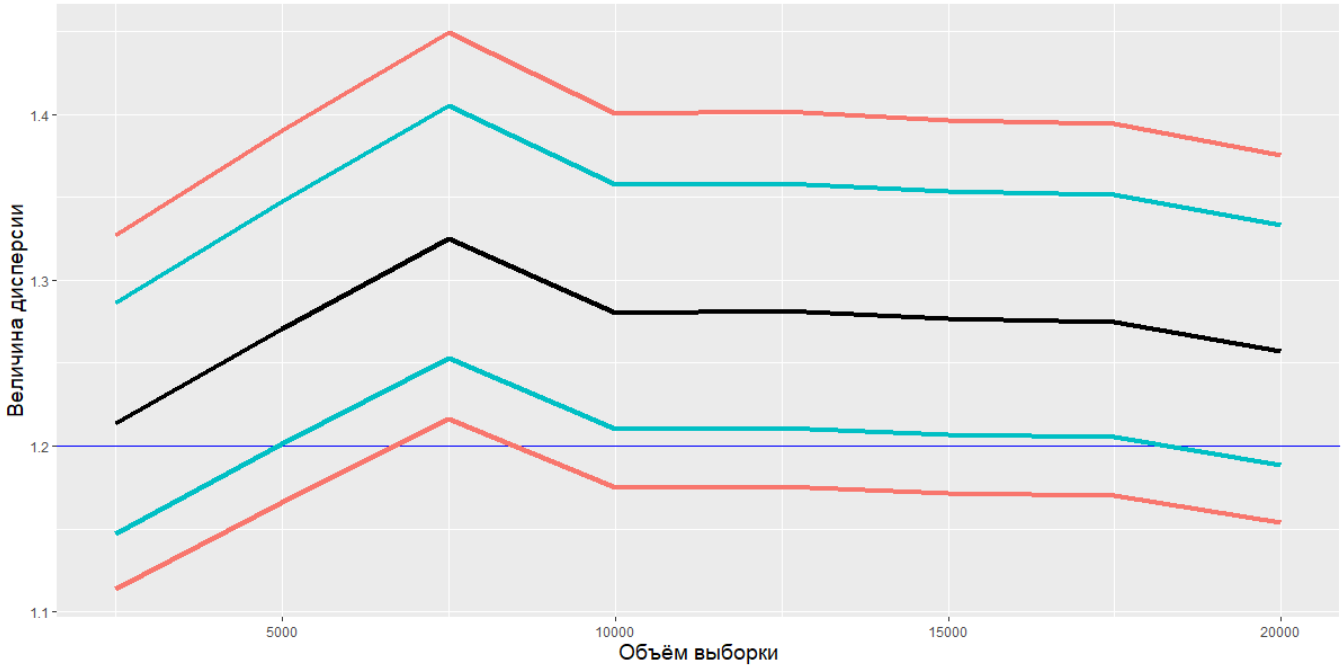

Рис.: Синяя горизонтальная прямая соответствует предельной дисперсии $\sigma^2 = 1.2$. Бирюзовые и красные линии показывают соответственно 80% и 95% асимптотические доверительные интервалы для дисперсии. 1000 повторов эксперимента.

Заключение

- Построен непараметрический критерий проверки непрерывности функции условной медианы;
- Его свойства изучены аналитически и проверены моделированием;
- Представлена перестановочная модификация критерия в двух формах, сходства и различия которых продемонстрированы моделированием;
- Моделированием показано, что предложенный критерий проверки разрывности функции регрессии на основе условной медианы устойчив к шуму;
- Построен непараметрический критерий проверки непрерывности плотности интенсивности пуассоновского процесса, свойства которого изучены аналитически и проверены моделированием.

В дальнейшем для задачи проверки гипотезы разрывности плотности пуассоновского процесса предлагается построить перестановочный критерий, а также сравнить результаты с известными методами обнаружения разладок пуассоновского процесса.