СБОРНИК НАУЧНЫХ ТРУДОВ НГТУ. - 1996. - № 2. - 9 - 18

УДК 519.24

К ВОПРОСУ О РОБАСТНОСТИ ОЦЕНОК ПО ГРУППИРОВАННЫМ ДАННЫМ

Б.Ю. ЛЕМЕШКО♣, С.Н. ПОСТОВАЛОВ♥

В статье рассматриваются функции влияния оценок параметров распределений по группированным и негруппированным данным. На основании сравнительного анализа функций влияния для оценок максимального правдоподобия по негруппированным наблюдениям и для оценок по сгруппированным данным делается вывод, ранее базирующийся на опыте применения разработанной программной системы статистического анализа одномерных наблюдений, о неробастности оценок максимального правдоподобия для параметров большинства законов распределений. Это следует из неограниченности функций влияния для этих оценок на области определения случайных величин. В то же время подчеркивается, что предварительная группировка данных в совокупности с асимптотически оптимальным группированием с последующим вычислением оценок максимального правдоподобия или оптимальных оценок параметров сдвига и масштаба по выборочным квантилям позволяет получать робастные оценки. Последний вывод подтверждается, в том числе ограниченностью функций влияния для этих оценок.

1. ВВЕДЕНИЕ

В [1-3] подчеркивается высокая устойчивость оценок максимального правдоподобия (ОМП) по группированным наблюдениям к наличию в выборке аномальных измерений, к отклонению реально наблюдаемого закона от предполагаемого, к засорению выборки данными, принадлежащими другому закону. Всё это подтверждается опытом эксплуатации программной системы [1] и многочисленными результатами модельных экспериментов. В данной работе свойство робастности ОМП исследуется с позиций *функции влияния*, предложенной Хэмпелом [4-5]. Именно анализ функций влияния ОМП параметров различных распределений, в том числе того множества распределений (такие как нормальное, Лапласа, Копи, Вейбулла, логистическое и др.), которое включено в программную систему [1], позволяет утверждать, что ОМП по негруппированным данным, вопреки порой бытующему заблуждению, в большинстве своём

♣ Доцент кафедры прикладной математики, канд. техн. наук

♥ Аспирант кафедры прикладной математики

являются неробастными. В то же время ОМП по группированным данным всегда оказываются робастными.

2. АНАЛИЗ ФУНКЦИЙ ВЛИЯНИЯ

Влияние ещё одного наблюдения на очень большую выборку может характеризоваться функцией (кривой) влияния, которая определяется следующим образом [6]:

$$IF(x;F,T)=\lim_{s\to 0}\frac{T\big((1-s)F+s\delta_x\big)-T(F)}{s},$$

где δ_x – единичная масса в точке x; F – функция распределения, к которому принадлежит выборка; $T(F)$ – вычисляемая статистика.

Функция влияния позволяет оценить относительное влияние отдельного наблюдения на значение статистики критерия или оценку параметров. Если функция влияния неограничена, то резко выделяющиеся наблюдения могут приводить к существенным изменениям оценок или статистик. *Чувствительность к большой ошибке* может характеризоваться величиной

$$\gamma^*=\sup_x\left|IF(x;F,T)\right|.$$

Для асимптотически эффективных оценок, к которым относятся оценки максимального правдоподобия по негруппированным данным, функция влияния удовлетворяет равенству [6]

$$IF(x;F_\theta,T)=J^{-1}(F_\theta)\frac{\partial\ln f_\theta}{\partial\theta}, \qquad (1)$$

где $J(F_\theta)$ - количество информации Фишера.

Для оценок типа максимального правдоподобия (M-оценок), где всякая оценка T_n определяется как решение экстремальной задачи на минимум вида

$$\sum_{i=1}^{n}\rho(x_i;T_n)\to\min$$

или как решение неявного уравнения

$$\sum_{i=1}^{n} \psi(x_i; T_n) = 0,$$

где ρ – произвольная функция, $\psi(x;\theta) = \partial\rho(x;\theta)/\partial\theta$ – функция влияния имеет вид [7]

$$IF(x;F,T) = -\frac{\psi(x;T)}{\lambda'_F(T(F))}, \quad -\infty < x < \infty,$$

где

$$\lambda'_F(T(F)) = \frac{d}{dt}\int \psi(x;t)dF(x), \quad -\infty < t < \infty.$$

В случае ОМП по группированным данным

$$\psi(x;\theta) = \frac{\partial \ln P_i(\theta)}{\partial\theta}, \quad x_{i-1} < x < x_i,$$

$$\lambda'_F(\theta)) = \frac{d}{d\theta}\int \psi(x;\theta)dF(x) = \sum_{j=1}^{k}\left(\frac{\partial^2 \ln P_j(\theta)}{\partial\theta^2}\right)P_j(\theta)$$

и функция влияния будет иметь вид

$$IF(x;F,\theta) = -\frac{\dfrac{\partial \ln P_i(\theta)}{\partial\theta}}{\sum_{j=1}^{k}\left(\dfrac{\partial^2 \ln P_j(\theta)}{\partial\theta^2}\right)P_j(\theta)}, \quad x_{i-1} < x < x_i. \qquad (2)$$

Для оценок, использующих квантили, соответствующие асимптотически оптимальному группированию [8], и являющихся одним из частных случаев L-оценок, функция влияния имеет вид [7]

$$IF(x;F,T) = \sum_{j=1}^{k-1} a_j\left(p_j - c\left(F^{-1}(p_j) - x\right)\right) \Big/ f\left(F^{-1}(p_j)\right), \qquad (3)$$

где a_j - коэффициенты при выборочных квантилях в формуле для вычисления L-оценок,

$$p_j = \sum_{l=1}^{j} P_l(\theta)\,,\ c(u) = \begin{cases} 1, & u \geq 0, \\ 0, & u < 0. \end{cases}$$

Были рассмотрены функции влияния для оценок параметров множества распределений, включенных в программную систему [1].

Приводимые ниже функции влияния построены при конкретных значениях параметров и характеризуют качественную картину их поведения на области определения случайных величин. На рис.1, 2 показаны функции влияния для оценок параметров сдвига и масштаба нормального распределения с функцией плотности

$$f(x) = \frac{1}{\theta_1\sqrt{2\pi}} e^{-\frac{(x-\theta_0)^2}{2\theta_1^2}},$$

определяемых методом максимального правдоподобия по негруппированным и сгруппированным данным. Функция влияния для ОМП параметра сдвига по негруппированным данным имеет вид

$$IF(x; F_\theta, \theta_0) = x - \theta_0\,,$$

для ОМП параметра масштаба -

$$IF(x; F_\theta, \theta_1) = \frac{1}{2}\left[\frac{(x-\theta_0)^2}{\theta_1} - \theta_1\right].$$

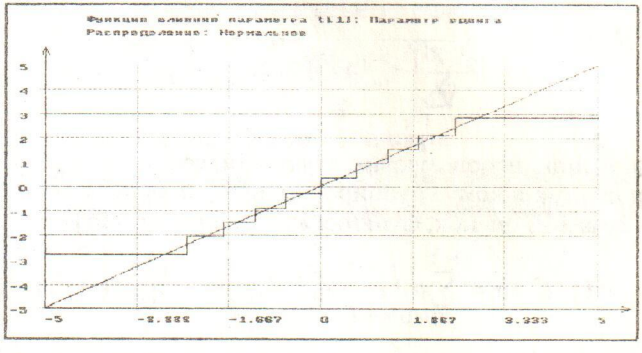

Рис.1. Функции влияния для параметра сдвига нормального распределения по негруппированным (прямая) и сгруппированным данным (ступенчатая линия)

Функции влияния неограничены и этим определяется чувствительность данных оценок к ошибкам измерения и засорению выборки. Напротив, функции влияния оценок параметров нормального распределения по группированным данным ограничены. Это ещё раз подчеркивает высокую устойчивость получаемых по группированным наблюдениям оценок, подтверждаемую практикой. На рис.1, 2 и последующих рис. 3 - 8 функции влияния для ОМП по группированным данным соответствуют случаю использования асимптотически оптимального группирования. Аналогично на рис. 3, 4 даны функции влияния для параметров распределения Вейбулла с функцией плотности

$$f(x) = \frac{\theta_0 (x-\theta_2)^{\theta_0 - 1}}{\theta_1^{\theta_0}} \exp\left\{ -\left(\frac{x-\theta_2}{\theta_1} \right)^{\theta_0} \right\}.$$

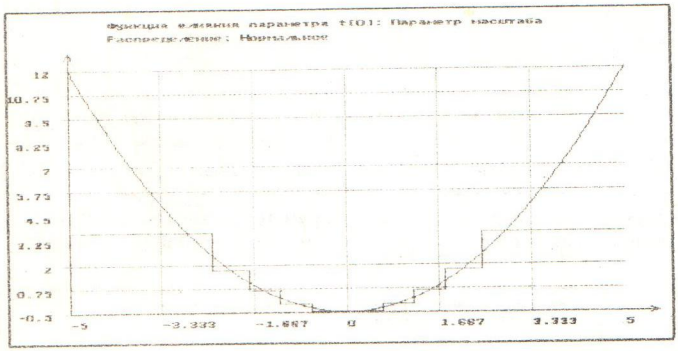

Рис.2. Функции влияния для параметра масштаба нормального распределения по негруппированным и сгруппированным данным (ступенчатая линия)

Функция влияния для ОМП основного параметра по негруппированным данным имеет вид

$$IF(x; F_\theta, \theta_0) = \frac{\theta_0 \left[(1-t)\ln t + 1 \right]}{1 + \pi/6 + C^2 - 2C},$$

где C - постоянная Эйлера и $t = \left(\frac{x-\theta_2}{\theta_1} \right)^{\theta_0}$, для ОМП параметра масштаба -

$$IF(x;F_\theta,\theta_1)=\frac{\theta_1}{\theta_0}(t-1).$$

Для основного параметра функция влияния по негруппированным данным неограничена снизу на левой и правой границе области определения случайной величины, для масштабного параметра - неограничена сверху на правой границе. В то же время для группированных наблюдений функции влияния являются ступенчатыми ограниченными функциями.

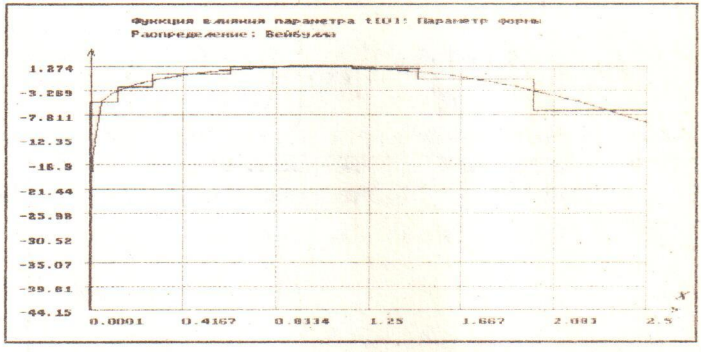

Рис.3. Функции влияния для основного параметра распределения Вейбулла по негруппированным и сгруппированным данным (ступенчатая линия)

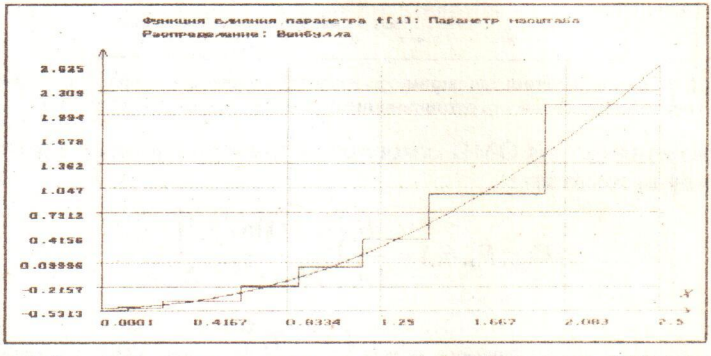

Рис.4. Функции влияния для параметра масштаба распределения Вейбулла по негруппированным и сгруппированным данным (ступенчатая линия)

Совершенно другую картину мы видим для ОМП по негруппированным наблюдениям для параметров распределения Коши с функцией плотности

$$f(x) = \frac{\theta_0}{\pi[\theta_0^2 + (x - \theta_1)^2]},$$

представленную на рис. 5, 6.

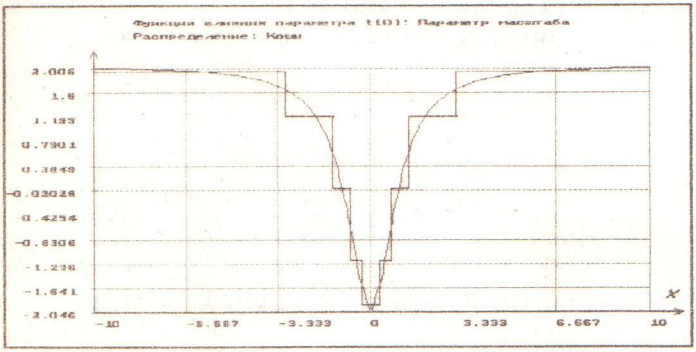

Рис.5. Функции влияния для параметра масштаба распределения Коши по негруппированным (непрерывная) и сгруппированным данным (ступенчатая линия)

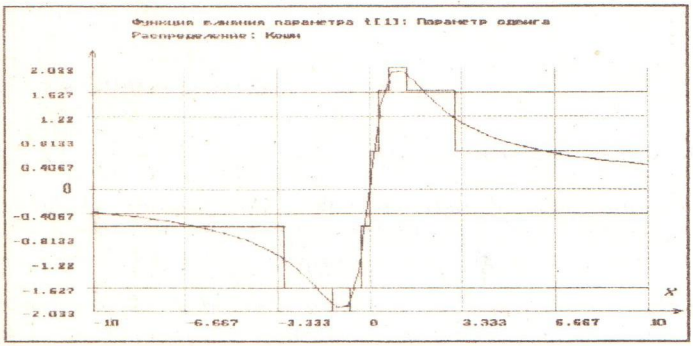

Рис.6. Функции влияния для параметра сдвига распределения Коши по негруппированным (непрерывная) и сгруппированным данным (ступенчатая линия)

Функция влияния для ОМП параметра сдвига по негруппированным данным имеет вид

$$IF(x; F_\theta, \theta_1) = \frac{4\theta_0 t}{1 + t^2}.$$

где $t = (x - \theta_1)/\theta_0$, для ОМП параметра масштаба -

$$IF(x; F_\theta, \theta_0) = 2\theta_0\left(1 - \frac{2}{1+t^2}\right).$$

Их функции влияния ограничены на области определения случайной величины, что говорит о робастности этих оценок, их устойчивости к грубым ошибкам измерений.

Для логистического распределения с плотностью

$$f(x) = \frac{\pi}{\theta_1\sqrt{3}} \frac{\exp\left\{-\dfrac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}}{\left[1+\exp\left\{-\dfrac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\right]^2},$$

функция влияния ОМП параметра масштаба по негруппированным данным имеет вид

$$IF(x; F_\theta, \theta_1) = \frac{9\theta_1 t}{\pi^2+3}\left(t - 1 - 2\frac{te^t}{1-e^t}\right),$$

где $t = \dfrac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}$, (см. рис. 7). Из её неограниченности следует, что соответствующая оценка не робастна.

В то же время функция влияния ОМП параметра сдвига

$$IF(x; F_\theta, \theta_0) = 3\theta_1\left(1 - 2\frac{e^t}{1+e^t}\right)$$

ограничена сверху и снизу, а это свидетельствует о робастности ОМП этого параметра.

Ситуация, которую мы наблюдаем для функций влияния ОМП по негруппированным наблюдениям параметров распределений Коши и логистического (параметр сдвига), оказывается явно нетипичной. Для ОМП параметров большинства законов распределения, включенных в программную систему [1], а это 26 законов и семейств непрерывных распределений, функции влияния неограничены, откуда

следует неробастность этих оценок. С другой стороны, для ОМП по группированным данным фунции влияния всегда представляют собой ограниченные ступенчатые зависимости.

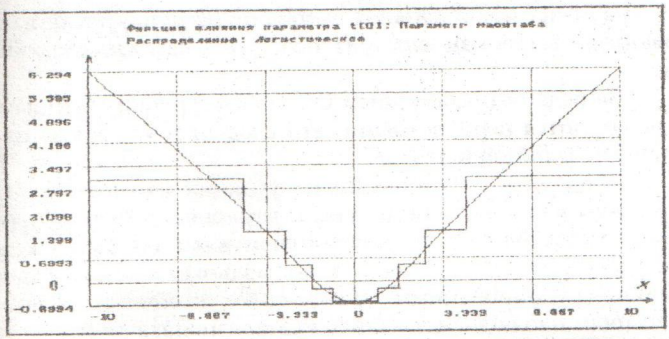

Рис.7. Функции влияния для параметра масштаба логистического распределения по негруппированным и сгруппированным данным (ступенчатая линия)

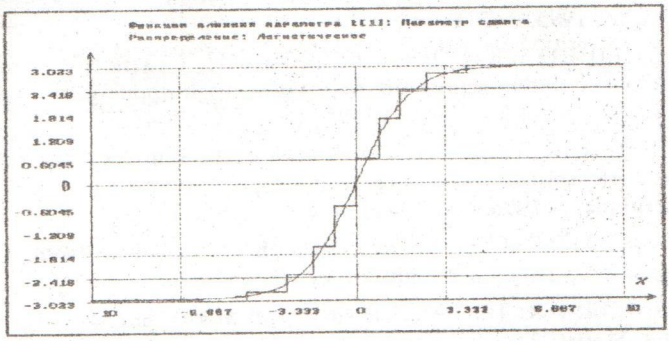

Рис.8 Функции влияния для параметра масштаба логистического распределения по негруппированным и сгруппированным данным (ступенчатая линия)

Функции влияния оптимальных оценок параметров сдвига и масштаба по выборочным квантилям для больших выборок, построенных с использованием решения задач асимптотически оптимального группирования [8], как следует из вида соотношения (3), также представляют собой ступенчатые ограниченные зависимости. Это указывает на робастность таких оценок. Результаты моделирования с засорением выборок аномальными измерениями и наблюдениями,

подчиняющимися другим распределениям, подтверждают устойчивость этих оценок.

ЗАКЛЮЧЕНИЕ

Анализ функций влияния оценок по негруппированным и группированным выборкам ещё раз позволяет сделать следующие выводы.

1. За редким исключением ОМП по негруппированным наблюдениям являются неробастными, что следует в том числе из неограниченности функций влияния.

2. Напротив, ограниченность функций влияния для ОМП по сгруппированным данным и для оптимальных оценок параметров сдвига и масштаба по выборочным квантилям для больших выборок подтверждает их устойчивость к аномальным ошибкам измерений. Оценки по группированным данным более устойчивы и к возможным отклонениям наблюдаемого закона от предполагаемого.

[1] ЛЕМЕШКО Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система. - Новосибирск: Изд-во НГТУ. - 1995. - 125 с.

[2] ЛЕМЕШКО Б.Ю., ПОСТОВАЛОВ С.Н. Статистический анализ одномерных наблюдений по частично группированным данным // Изв. вузов. Физика. - Томск, 1995. - № 9. - С. 39-45.

[3] ЛЕМЕШКО Б.Ю., ПОСТОВАЛОВ С.Н. Вопросы обработки выборок одномерных случайных величин // Научный вестник НГТУ.- Новосибирск. - 1996. - № 2. - С. 3-24.

[4] HAMPEL F.R. Contributions to the theory of robust estimation // Ph. D. Thesis. Berkeley: Univ. California.

[5] HAMPEL F.R. The influence curve and its role in robust estimation // J. Amer. Statist. Ass. - V. 69, № 346. - P. 383-393.

[6] ХЬЮБЕР П. Робастность в статистике. - М.: Мир, 1984. - 303 с.

[7] ШУЛЕНИН В.П. Введение в робастную статистику. - Томск: Изд-во Том. ун-та, 1993. - 227 с.

[8] ЛЕМЕШКО Б.Ю. Оптимальные оценки параметров сдвига и масштаба по выборочным квантилям для больших выборок // Тр. третьей международной НТК "Актуальные проблемы электронного приборостроения" АПЭП-96. - Новосибирск, 1996.