А.Г. Белов, Б.М. Щедрин

ОТНОСИТЕЛЬНАЯ ЭФФЕКТИВНОСТЬ МНК И МНМ ОЦЕНОК

Введение. В «вечном» споре методов обработки данных по методу наименьших квадратов (МНК) и методу наименьших модулей (МНМ) [1] предлагается использовать новый подход к их сравнению через оценку относительной эффективности выборочных среднего и медианы. Полученные ниже результаты подтверждают выводы об относительной эффективности этих методов в оценке параметра сдвига, сделанные ранее в работах других авторов [2] с более трудоемким анализом. Хотя описанный ниже подход применен для нормального и Лапласа распределений ошибок, он может быть распространен и на другие случаи.

Постановка задачи. Предположим, что исследователя интересует значение некоторого неслучайного количественного показателя θ. Величина его неизвестна и должна быть получена из наблюдения (измерения) y. Если бы измерения величины y производились без погрешностей, то для определения θ было бы достаточно одно измерение, $y = \theta$. Однако точные измерения на практике чаще всего невозможны, поэтому наблюдаемые значения y известны с какими-то случайными ошибками измерения (погрешностями) e_i, т.е. реально вместо точного значения θ мы имеем случайный результат

$$y_i = y + e_i = \theta + e_i \quad . \tag{1}$$

Чтобы исключить влияние погрешностей измерений нужна дополнительная информация. Для этого проводят большое число $n > 1$ повторных независимых измерений. Величины $e_1,...,e_n$ можно интерпретировать как значения одной и той же с.в. ε, описываемой некоторым заданным вероятностным распределением с функцией плотности $f_\varepsilon(z)$, а полученные в результате независимых воспроизведений эксперимента величины $y_1,..., y_n$ – как значения с.в. η с функцией плотности $f_\eta(z)$. Но это то же самое, что $e_1,...,e_n$ есть значения независимых и одинаково распределенных, как и с.в. ε, с.в. $\varepsilon_1,...,\varepsilon_n$, а $y_1,..., y_n$ есть наблюдаемые значения независимых с.в. $\eta_1,...,\eta_n$ - копий с.в. η. Тогда в силу (1) с.в. η_i и с.в. $\theta + \varepsilon_i$ эквивалентны

$\eta_i \sim \theta + \varepsilon_i, i = 1,...,n$. Отсюда следует взаимосвязь их функций плотности $f_\eta(z) = f_\varepsilon(z-\theta)$ и равенство их математических ожиданий (м.о.)

$$E\eta_i = E(\theta + \varepsilon_i) = \theta + E\varepsilon_i. \qquad (1')$$

Будем предполагать, что условия эксперимента обеспечивают отсутствие систематической ошибки $(E\varepsilon_i = 0)$, равноточность $\left(D\varepsilon_i = \sigma^2 < \infty\right)$, некоррелированность $\left(\mathrm{cov}\left(\varepsilon_i, \varepsilon_j\right) = 0,\ i \neq j\right)$ результатов измерений. Полученные в ходе эксперимента реализации $(y_1,..., y_n)$ выборки $(\eta_1,...,\eta_n)$ обрабатывают различными статистическими методами. В результате находят не точное значение параметра θ, а определяют его оценку $\theta_n = \theta(y_1,..., y_n)$. Общим методом оценивания параметра θ является метод максимального правдоподобия (ММП), заключающийся в максимизации по θ функции правдоподобия (ФП) $L(\theta) = \prod_{i=1}^{n} f_\varepsilon(y_i - \theta)$ или чаще ее логарифма. В зависимости от того или иного закона распределения $f_\varepsilon(z)$ получаются разного вида оценки θ_n. Иногда удается получить для них формулу, но чаще они находятся численно как решение задачи максимизации ФП. В случае, когда ошибка ε_i имеет двойное экспоненциальное распределение (первый закон ошибок, Лапласа), оценка ММП эквивалентна оценке МНМ, получаемой при минимизации по θ функционала $Q^{МНМ}(\theta) = \sum_{i=1}^{n} |y_i - \theta|$. В случае нормального распределения (второй закон ошибок), оценка ММП эквивалентна оценке МНК, получаемой при минимизации по θ функционала $Q^{МНК}(\theta) = \sum_{i=1}^{n} (y_i - \theta)^2$. Возникает задача: какому методу, МНМ или МНК, отдать предпочтение при оценке параметра сдвига в случае, когда теоретически предполагаемое нормальное распределение ошибки на практике фактически является распределением Лапласа, и наоборот? Для ответа на поставленный вопрос необходимо вспомнить свойства некоторых характеристик «центра» распределения вероятностей.

«Центр» распределения. Как следует из (1), искомому значению θ соответствуют множество значений с.в. η_i. Поскольку $E\eta_i = \theta$, то естественно в качестве оценки параметра выбрать математическое ожидание (м.о.) $E\eta_i$ или медиану $Med\eta_i$, которые совпадают в случае симметричного распределения. Такой выбор связан еще и с тем, что, как

известно [3,с.41,с.43], медиана $Med\eta$ и м.о. $E\eta$ обладают свойством оптимальности характеристики «центра» распределения. Так для любого числа θ справедливы равенства, соответствующие принципу наименьших модулей и наименьших квадратов,

$$E\left|\eta - Med\eta\right| = \min_{\theta} E\left|\eta - \theta\right|,\ E\left(\eta - E\eta\right)^2 = \min_{\theta} E\left(\eta - \theta\right)^2 \quad .$$

То есть, из всех возможных способов приближения значений с.в. η наилучшим в смысле среднего абсолютного и среднего квадратичного является медиана и м.о. соответственно. Отсюда следует, что для параметра θ наилучшими в среднем абсолютном и среднем квадратичном являются значения равные соответственно медиане и м.о.

$$\begin{aligned} \theta^{MHM} &= \arg\min_{\theta} E\left|\eta - \theta\right| = Med\eta, \\ \theta^{MHK} &= \arg\min_{\theta} E\left(\eta - \theta\right)^2 = E\eta \end{aligned} \qquad (2)$$

Различие значений θ^{MHM} и θ^{MHK} соответствует различию значений медианы и м.о. с.в. η. В случае симметричных распределений с.в. $\eta\left(\varepsilon\right)$ значения (2) параметра сдвига θ, гипотетически равны $\theta = \theta^{MHM} = \theta^{MHK}$. А как соотносятся оценки (2) на практике?

Эффективность оценок. Предположим, что в выборке $\left(\eta_1,...,\eta_n\right)$ все элементы независимы и имеют одну и ту же непрерывную плотность распределения $f_\eta(z)$ и $f_\eta(Med\eta_i) > 0$. Обозначим через $\overline{m}_n$ и $\overline{e}_n$ выборочные медиану и среднее, построенные по реализации $\left(y_1,...,y_n\right)$ и вариационному ряду $\left(y_{(1)},...,y_{(n)}\right)$ выборки $\left(\eta_1,...,\eta_n\right)$:

$$\overline{m}_n = \underset{1\le n\le n}{median}\left\{y_i\right\} = \begin{cases} y_{(k+1)},\ n = 2k+1, \\ \left(y_{(k)} + y_{(k+1)}\right)\big/2,\ n = 2k, \end{cases} \qquad \overline{e}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \ .$$

Пусть θ_n^{MHM} и θ_n^{MHK} статистические аналоги значений (2) параметра θ, для которых выполняются соответствующие выборочные равенства,

$$\begin{aligned} \theta_n^{MHM} &= \arg\min_{\theta} \sum_{i=1}^{n} \left|y_i - \theta\right| = \overline{m}_n, \\ \theta_n^{MHK} &= \arg\min_{\theta} \sum_{i=1}^{n} \left(y_i - \theta\right)^2 = \overline{e}_n. \end{aligned} \qquad (2')$$

Очевидно, что выборочные оценки θ_n^{MHM} и θ_n^{MHK}, полученные по выборочным данным, будут отличаться от соответствующих значений θ^{MHM} и θ^{MHK}, полученных по генеральной совокупности. Тогда ответить на вопрос о том, какая из оценок θ_n^{MHM} и θ_n^{MHK} лучше приближает θ^{MHM} и θ^{MHK}, можно оценив их относительную эффективность, которая равносильна относительной эффективности выборочных медианы и среднего, а именно,

$$\frac{E\left(\theta_n^{MHK}-\theta^{MHK}\right)^2}{E\left(\theta_n^{MHM}-\theta^{MHM}\right)^2}=\frac{E\left(\overline{e}_n-E\eta_1\right)^2}{E\left(\overline{m}_n-Med\eta_1\right)^2}.$$

Но для последнего соотношения справедлива оценка [3, с.126]

$$\frac{E\left(\overline{e}_n-E\eta\right)^2}{E\left(\overline{m}_n-Med\eta\right)^2}\approx 4f_\eta^2\left(Med\eta\right)D\eta,$$

которая будет верна и для относительной эффективности оценок θ_n^{MHM} и θ_n^{MHK}. Если рассмотреть класс симметричных распределений $f_\varepsilon(z)$ ошибок ε_i, то $Med\eta=E\eta=\theta$, и получаем соотношение

$$\frac{E\left(\theta_n^{MHK}-\theta\right)^2}{E\left(\theta_n^{MHM}-\theta\right)^2}\approx 4f_\eta^2\left(Med\eta\right)D\eta.$$

Примеры. Рассмотрим некоторые примеры применимости полученных выше результатов.

Пример 1. Пусть с.в. ε_i имеет распределение Лапласа с параметром $\lambda=\frac{1}{\sigma}$ и функцией плотности вероятностей

$$f_\varepsilon(z)=\frac{\lambda}{2}e^{-\lambda|z|}=\frac{1}{2\sigma}e^{-\frac{|z|}{\sigma}},$$

при этом $E\varepsilon_i=0,\ D\varepsilon_i=\frac{2}{\lambda^2}=2\sigma^2$. Тогда с.в. $\eta_i\sim\theta+\varepsilon_i$ будет подчинена распределению Лапласа с плотностью

$$f_\eta(t)=\frac{1}{2\sigma}e^{-\frac{|t-\theta|}{\sigma}}.$$

При этом $E\eta_i = Med\eta_i = \theta,\ D\eta_i = 2\sigma^2$ и $f_\eta(Med\eta) = f_\eta(\theta) = \dfrac{1}{2\sigma}$. Отсюда справедлива оценка относительной эффективности

$$\frac{E\left(\theta_n^{МНК} - \theta\right)^2}{E\left(\theta_n^{МНМ} - \theta\right)^2} \approx 4\frac{1}{4\sigma^2}2\sigma^2 = 2.$$

Отметим, что логарифм функции правдоподобия $LL(\theta) \propto -\sum_{i=1}^{n}\left|y_i - \theta\right|$ дает оценку $\theta_n^{ММП} = \underset{1\le n\le n}{median}\{y_i\} = \theta_n^{МНМ}$.

Таким образом, в предположении распределения Лапласа ошибок измерений МНК дает оценку параметра сдвига с эффективностью в два раза хуже, чем МНМ.

Пример 2. Пусть теперь с.в. ε_i имеет нормальное распределение с функцией плотности вероятностей

$$f_\varepsilon(z) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{z^2}{2\sigma^2}},$$

при этом $E\varepsilon_i = 0,\ D\varepsilon_i = \sigma^2$. Тогда с.в. $\eta_i \sim \theta + \varepsilon_i$ будет подчинена нормальному распределению с плотностью

$$f_\eta(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\theta)^2}{2\sigma^2}}.$$

При этом $E\eta_i = Med\eta_i = \theta,\ D\eta_i = \sigma^2$ и $f_\eta(Med\eta) = f_\eta(\theta) = \dfrac{1}{\sigma\sqrt{2\pi}}$. Отсюда справедлива оценка относительной эффективности

$$\frac{E\left(\theta_n^{МНК} - \theta\right)^2}{E\left(\theta_n^{МНМ} - \theta\right)^2} \approx 4\frac{1}{2\pi\sigma^2}\sigma^2 = \frac{2}{\pi}.$$

Отметим, что в этом случае логарифм функции правдоподобия $LL(\theta) \propto -\sum_{i=1}^{n}\left(y_i - \theta\right)^2$ дает оценку $\theta_n^{ММП} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \theta_n^{МНК}$.

Таким образом, в предположении нормального распределения ошибок МНК дает оценку параметра сдвига с эффективностью в $\frac{\pi}{2} \approx 1{,}57$ раза лучше, чем МНМ.

Результаты примеров 1 и 2 также получены в [2, с.167] с помощью более трудоемкого аппарата оценивания дисперсии оценок.

Предложенный подход, основанный на использовании вероятностных и статистических свойств характеристик с.в., позволяет более оперативно проводить анализ МНК и МНМ оценок параметра сдвига и может быть распространен на более сложные параметрические модели изучаемых явлений.

ЛИТЕРАТУРА

1. Мудров В.И., Кушко В.Л. Метод наименьших модулей. М., Знание, 1971.
2. Мудров В.И., Кушко В.Л. Методы обработки измерений: Квазиправдоподобные оценки. – Изд. 2-е, перераб. и доп. – М.: Радио и связь, 1983.
3. Королев В.Ю. Теория вероятностей и математическая статистика: учеб. – М.: ТК Велби, Изд-во Проспект, 2006.