

ЗАМЕТКИ И ПИСЬМА

ЛИНЕЙНЫЕ МЕТОДЫ НАИМЕНЬШИХ
АБСОЛЮТНЫХ ВЕЛИЧИН

© 2015 г. Н.М. Мусин

(Москва)

Рассматриваются два метода наименьших абсолютных величин, названные так по аналогии с методом наименьших квадратов. Первый, названный в статье абсолютной ординатной регрессией, был предложен Лапласом, известен как метод наименьших модулей (МНМ) и в наши дни рассчитывается, как правило, методом итераций с использованием компьютеров. В статье рассматривается другой метод расчета. Второй метод назван абсолютной нормальной регрессией и предлагается впервые. Построен алгоритм вычисления коэффициентов уравнений регрессии и оценки их качества. Эти методы необязательно связаны только со случайными величинами. Результаты имеют большое значение для решения таких актуальных экономических задач, как, например, построение систем дорог с наименьшей суммарной длиной, и тому подобных.

Ключевые слова: метод наименьших квадратов (МНК), метод наименьших модулей (МНМ), метод наименьших абсолютных величин, абсолютная ординатная регрессия, абсолютная нормальная регрессия.

Классификация JEL: C10, C13, C51, C60.

ОСЬ ОРДИНАТ КАК ОБРАЗУЮЩАЯ

Пусть на плоскости имеется совокупность точек $M_1, \dots, M_n$, $n \geq 3$. Выберем некоторую декартову систему координат. Возьмем произвольную прямую a (рис. 1), не параллельную оси ординат, и спроецируем эти точки на эту прямую вдоль некоторой прямой d , не параллельной прямой a . Прямую d назовем *образующей*.

Из всех возможных вариантов выбора образующей d выделим два случая:

- 1) в качестве образующей выбрана ось ординат;
- 2) в качестве образующей выбрана нормаль к прямой a .

Как отмечалось ранее, мы выбираем некоторую декартову систему координат. Тогда данные на плоскости точки, на которые мы указывали выше, можно определить через их координаты: $M_1(x_1, y_1), \dots, M_n(x_n, y_n)$.

Данные точки проецируем на прямую a параллельно оси ординат (выбранной в качестве образующей d); обозначим длины отрезков, соединяющих эти точки с их проекциями $N_1, \dots, N_n$, соответственно $t_1, t_2, t_3, \dots, t_i, \dots, t_n$.

Определение. Сумму $T = t_1 + \dots + t_n$ назовем *абсолютным ординатным примыканием* прямой a к данным точкам.

Нужно отыскать прямую, которая наиболее тесно ординатно примыкала бы к данным точкам, т.е. чтобы T приняла наименьшее значение.

Линию, наиболее тесно ординатно примыкающую к данным точкам, назовем *линией абсолютной ординатной регрессии*. Как следует из определения, в зависимости от выбора системы координат эта линия получается не одной и той же.

Если прямую a задавать уравнением $y = kx + b$, то ординатное примыкание вычисляется по формуле

$$T(k, b) = \sum_{i=1}^n t_i = \sum_{i=1}^n |kx_i + b - y_i|.$$

Идея минимизации этой суммы представляется очевидной и адекватной¹, и именно эту идею предлагал в свое время П. Лаплас, но она осталась без реализации. С появлением компьютеров эта идея возродилась, ее реализация известна под названием метода наименьших модулей (МНМ) (Мудров, Кушко, 1971), или метода наименьших отклонений (МНО). Как указано в книге Б.В. Гнеденко (Гнеденко, 1988, с. 419): “Этот прием не получил в ту пору распространения, поскольку вскоре был предложен другой метод, приводящий к более простым результатам. Разработка этого нового метода связана с именами Гаусса (1777–1855), Лежандра (1752–1833) и американского математика Р. Эдрейна (1775–1843)”.

Метод, предложенный указанными тремя учеными, известен как метод наименьших квадратов, и простота его связана с возможностью применить методы дифференциального исчисления. По этой же причине метод минимизации суммы абсолютных величин разностей ординат не получил в это время распространения.

Дело в том, что в приведенной формуле содержатся абсолютные величины, что делает невозможным применение дифференциального исчисления для нахождения наименьшего значения ординатного примыкания T .

В данной статье показано, что “спасать” дифференциальное исчисление вовсе не обязательно. Напомним, что эта попытка предпринимается в широко известном методе наименьших квадратов, идея которого состоит в минимизации суммы

$$S(k, b) = \sum_{i=1}^n (kx_i + b - y_i)^2,$$

т.е. суммы квадратов разностей соответствующих ординат. Однако из неравенства

$$\sum_{i=1}^n t_i > \sqrt{\sum_{i=1}^n t_i^2}$$

следует, что, например, с экономической точки зрения, если истолковать $t_1, \dots, t_n$ как, например,

протяженность дорог, то выгоднее минимизировать $\sum_{i=1}^n t_i$, чем $\sqrt{\sum_{i=1}^n t_i^2}$, – хотя бы потому, что до-

стижение наименьшего значения суммой $\sqrt{\sum_{i=1}^n t_i^2}$ вовсе не означает, что сумма $\sum_{i=1}^n t_i$ также достигает наименьшего значения.

Пример 1. Рассмотрим точки $A(-1; 0)$, $B(0; 1)$ и $C(1; 0)$. Метод наименьших квадратов говорит, что к данным точкам наиболее тесно примыкает – с точки зрения этого метода – прямая $y = 1/3$, в то время как ее абсолютное ординатное примыкание равно $4/3$. Очевидно, что более тесно² ординатно примыкает к данным точкам прямая $y = 0$. Если представить, что прямая $y = 1/3$ описывает прямолинейную трассу, а перпендикуляры к ней из точек A , B и C – подъездные пути к трассе, на каждую единицу пути тратится 300 л бензина, то метод наименьших квадратов дает расход бензина в 400 л, а метод наименьших абсолютных величин – в 300 л, т.е. имеется экономия в 100 л бензина.

А в масштабах страны экономия составила бы миллионы тонн горючего!

¹ Однако еще более очевидной и адекватной является развиваемая далее в статье идея минимизации не суммы модулей разностей ординат или их квадратов, а суммы перпендикуляров на искомую линию так называемой абсолютной нормальной регрессии.

² Далее будет показано, что и наиболее тесно.

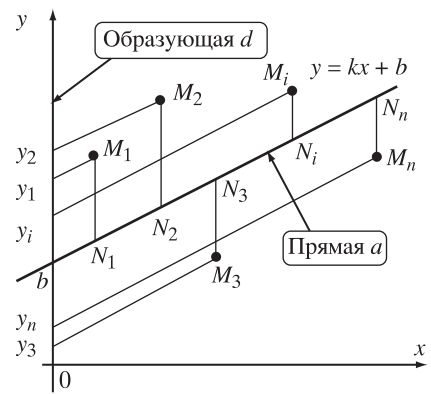


Рис. 1. Проецирование вдоль оси и на ось ординат

Покажем, как можно минимизировать сумму $T = \sum_{i=1}^n t_i = \sum_{i=1}^n |kx_i + b - y_i|$, не пытаясь обратиться к дифференциальному исчислению.

Спроецируем данные точки вдоль прямой $y = kx + b$ на ось ординат. Переиндексируем ординаты получающихся проекций, т.е. перенумеруем их в порядке убывания ординат. Тогда последовательность полученных ординат запишется в виде $y_1, \dots, y_m$, $m \leq n$ (на рис. 1 индексы ординат показаны до переиндексации). Разность между наибольшим и наименьшим из этих чисел, т.е. $y_1 - y_m$ (согласно терминологии статистики, это – размах множества этих ординат) зависит только от параметра k прямой a , которую мы задали уравнением $y = kx + b$. Легко подобрать такое значение этого параметра, чтобы этот размах был наименьшим. Таким образом, параметр k прямой a найден.

Найдем теперь параметр b . Эта задача сводится к тому, чтобы отыскать такую точку b на оси ординат, чтобы сумма расстояний от нее до точек $y_1, \dots, y_m$ была наименьшей. Имеет место очевидная теорема.

Теорема 1. Если $m = 2q + 1$, то $b = y_{q+1}$; если $m = 2q$, то b – любое число из замкнутого отрезка $[y_q, y_{q+1}]$.

Примечание 1. По терминологии математической статистики, число b является медианой упорядоченного множества $y_1, \dots, y_m$.

Примечание 2. В примере 1 то, что прямая $y = 0$ наиболее тесно абсолютно ординатно примыкает к данным точкам, очевидно с геометрической точки зрения. Покажем, что эта же прямая получается и с помощью нашего метода, а в нашем анализе можно будет усмотреть план действий и для большего числа точек. Проецируем точки A, B и C вдоль искомой прямой $y = kx + b$ на ось ординат, получаем три значения: $1, -k$ и k .

Если $|k| \geq 1$, то размах множества проекций равен $2|k|$, наименьший размах достигается при $|k| = 1$, искомая прямая проходит через точку B , при этом ординатное примыкание равно $\sqrt{2}$.

Если $|k| < 1$, то размах равен $1 + |k|$, значит, $k = 0$, искомая прямая проходит через точку A или через точку C (в зависимости от знака k), и ординатное примыкание равно 1. Следовательно, искомой является прямая $y = 0$ с абсолютным ординатным примыканием 1.

НОРМАЛЬ КАК ОБРАЗУЮЩАЯ

Из данных точек проводим к прямой a перпендикуляры; обозначим длины этих перпендикуляров соответственно $h_1, \dots, h_n$ (рис. 2).

Определение. Сумму $H = h_1 + \dots + h_n$ назовем *абсолютным нормальным примыканием* прямой a к данным точкам.

Нужно отыскать прямую, которая наиболее тесно абсолютно нормально примыкала бы к данным точкам, т.е. чтобы H приняла наименьшее значение.

Линию, наиболее тесно нормально примыкающую к данным точкам, назовем *линией абсолютной нормальной регрессии*. Как следует из определения, эта линия не зависит от выбора системы координат.

Пусть в некоторой системе координат известны координаты точек $M_1(x_1, y_1), \dots, M_n(x_n, y_n)$. Если прямую a задавать уравнением $y = kx + b$, то абсолютное нормальное примыкание вычисляется по формуле

$$H(k, b) = \sum_{i=1}^n \frac{|kx_i + b - y_i|}{\sqrt{1 + k^2}}.$$

В формуле содержатся абсолютные величины, что, как отмечалось выше, делает невозможным применение дифференциального исчисления для нахождения наименьшего значения нормального примыкания H .

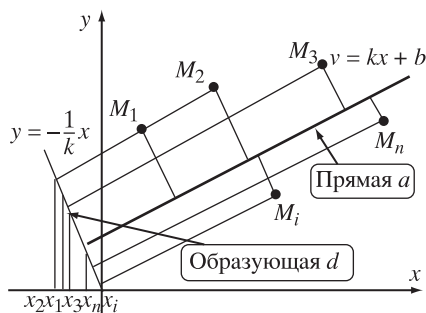


Рис. 2. Проецирование вдоль нормали и на нее

Также можно показать, что попытка исправить метод наименьших квадратов рассмотрением сумм вида $Q(k, b) = \sum_{i=1}^n [(kx_i + b - y_i)^2 / (1 + k^2)]$ не только приводит к весьма сложным формулам, но и, что естественно, дает худший результат. В самом деле, из неравенства

$$\sum_{i=1}^n h_i > \sqrt{\sum_{i=1}^n h_i^2}$$

следует, что минимизация правой части вовсе не означает минимизации левой части.

Как отмечалось раньше, если предположить, что h_i , $i = 1, \dots, n$ – протяженность дорог, проведенных перпендикулярно к искомой прямолинейной магистрали, то экономически выгодно минимизировать левую часть. И в данном случае также можно обойтись без применения дифференциального исчисления.

Выберем произвольно декартову систему координат. Возьмем произвольную прямую $y = kx + b$. Все точки спроецируем на перпендикулярную ей прямую с уравнением $y = -x/k$.

Абсциссы получающихся проекций нумеруем в порядке их возрастания, тогда они будут обозначены в виде $x_1, \dots, x_m$, причем $m \leq n$. Разность $x_m - x_1$ (т.е., согласно терминологии статистики, размах множества этих абсцисс) зависит только от параметра k прямой a . Легко подобрать такое значение этого параметра, чтобы этот размах был наименьшим. Таким образом, параметр k прямой a найден.

Найдем теперь параметр b . Эта задача сводится к тому, чтобы отыскать такую точку x_0 на оси абсцисс, чтобы сумма расстояний от нее до точек $x_1, \dots, x_m$ была наименьшей.

Имеет место очевидная теорема.

Теорема 2. Если $m = 2q + 1$, то $x_0 = x_{q+1}$; если $m = 2q$, то x_0 – любое число из замкнутого отрезка $[x_q, x_{q+1}]$.

То есть, по терминологии математической статистики, число b является медианой упорядоченного множества $x_1, \dots, x_m$.

Следствие. Пусть N – точка пересечения прямых $x = x_0$ и $y = -x/k$. Тогда параметр b подбираем так, чтобы прямая $y = kx + b$ прошла через точку N .

Пример 2. Рассмотрим точки $A(0; 0)$, $B(1; 1)$ и $C(2; 1)$. Метод наименьших квадратов дает прямую $y = \frac{1}{2}x + \frac{1}{6}$, ординатное примыкание этой прямой равно $\frac{1}{6} + \frac{1}{3} + \frac{1}{6} = \frac{2}{3} = 0,666\dots$, а нормальное примыкание равно $4\sqrt{5}/15 = 0,598\dots$. В то же время очевидно, что ординатное приближение прямой AC лучше, так как оно равно 0,5. Вычисление коэффициентов линии абсолютной ординатной регрессии приводит к уравнению $y = 0,5x$, т.е. к этой же прямой AC . У прямой AC абсолютное нормальное примыкание – также лучше, так как оно оказалось равным $\sqrt{5}/5 = 0,447\dots$

Если, как в примере 2, представить, что искомая линия нормальной регрессии описывает прямолинейную трассу, то суммарная длина подъездных путей к прямой AC короче на $0,598 - 0,447 = 0,151$, т.е. на 25,25%. Оказывается, что прямая AC – самый лучший вариант. Проверим это.

Рассчитаем уравнение абсолютной нормальной регрессии по нашей методике в виде $y = kx + b$. Проекции точек A , B и C на прямую $y = -x/k$ имеют абсциссы

$$P = 0, \quad Q = \frac{(k-1)k}{k^2+1}, \quad R = \frac{(2k-1)k}{k^2+1}.$$

Легко убедиться, что всегда $Q \leq R$.

$Q \geq 0$ при $k \leq 0$ или $k \geq 1$. Так как в этом случае $P \leq Q \leq R$, линия регрессии должна пройти через среднюю точку B и поэтому имеет вид $y = kx + 1 - k$. При $k = 0$ или $k = 1$ получаются прямые $y = 1$ и $y = x$ с нормальными примыканиями 1 и $0,5\sqrt{2} = 0,707\dots$, т.е. эти варианты хуже.

$Q \leq P \leq R$ при $0,5 \leq k \leq 1$. Линия регрессии должна пройти через точку A , поэтому имеет вид $y = 0,5x$. При $k = 1$ размах проекций равен $R - Q = 0,5$, при $k = 0,5$ он равен $0,2$, поэтому выбираем второй вариант. Прямая $y = 0,5x$ пройдет также и через точку C .

Аналогично – случай $Q \leq R \leq P$ приводит к этому же результату, т.е. прямая AC – искомая.

СПИСОК ЛИТЕРАТУРЫ

Гнеденко Б.В. (1988). Курс теории вероятностей. М.: Наука.

Мудров В.И., Кушко В.Л. (1971). Метод наименьших модулей. М.: Знание.

Поступила в редакцию
14.11.2014 г.

Linear Least Absolute Values Methods

N.M. Musin

Two of the least absolute value methods are considered, called so by analogy with the method of least squares. In this paper, the first method, which is known as the least absolute deviations method, is called the absolute ordinate regression. It was proposed by Laplace; nowadays calculations, which are usually carried out by iteration, require computers. In this paper, we present another method of calculation. The second method is called the absolute normal regression and is presented here for the first time. An algorithm for calculating the coefficients of regression and an assessment for their quality are developed. These methods are not necessarily associated with random variables only; they are of much wider use than just that. The results are important for addressing the most economic problems of current interest such as building a road system with the least total length and the like.

Keywords: least squares method (LSM), least absolute deviations method (LADM), least absolute values method, absolute ordinate regression, absolute normal regression.

JEL Classification: C10, C13, C51, C60.