УДК 514.172

И.В. Пономарев, В.В. Славский

О геометрической интерпретации метода наименьших квадратов*

I.V. Ponomarev, V.V. Slavsky

About Geometrical Interpretation of the Least Squares Method

В данной статье рассматриваются два метода построения линейной регрессионной модели. Приводится геометрическая интерпретация функционала качества. Доказывается неравенство, связывающее эти функционалы.

Ключевые слова: линейная регрессия, метод наименьших квадратов, объем симплекса.

In this article two methods of construction linear regression models are considered. Geometrical interpretation functional qualities is resulted. The inequality connecting these functionals is proved.

Key words: linear regression, method of the least squares, simplex volume.

Пусть R^{k+1} – $k+1$-мерное арифметическое евклидово пространство. Пусть Ω – конечное подмножество точек:

$$\Omega = \{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, \ldots, N\},$$

которое можно рассматривать как результат N экспериментов. В приложениях часто возникает вопрос о существовании функциональной зависимости между переменными y и $x_1, x_2, \ldots, x_k$.

Наиболее простая зависимость – линейная, которая в классическом случае имеет вид

$$y_i = a_1 x_{i1} + \ldots + a_k x_{ik} + \varepsilon_i,$$

где y_i – значение зависимой переменной; x_{ij} – значение j-й независимой переменной; $a_j \in R$ – параметры модели; ε_i – случайная ошибка; $j = 1,\ldots,k$, $i = 1, \ldots, N$.

Обозначим

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Nk} \end{pmatrix},$$

$$x_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{Ni} \end{pmatrix}, a = \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$

Тогда модель линейной регрессии будет иметь вид

$$y = Xa + \varepsilon.$$

В статистике разработаны мощные методы для анализа множества Ω на линейную зависимость основанные на Евклидовой норме.

*Работа выполнена при финансовой поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 гг. (гос. контракт №02.740.11.0457).

Классическим подходом к оценке параметров модели является метод наименьших квадратов, суть которого заключается в минимизации функционала

$$\alpha_2^2 = \min_a (y - Xa)^T (y - Xa). \qquad (1)$$

Теорема 1 (теорема Гаусса-Маркова). *Предположим, что*

1. $y = Xa + \varepsilon$;
2. X – *детерминированная* $N \times k$ *матрица, имеющая максимальный ранг* k;
3. $M(\varepsilon) = 0$, $D(\varepsilon) = \sigma^2 E_N$.

Тогда оценка метода наименьших квадратов наиболее эффективна (в смысле наименьшей дисперсии) в классе линейных (по y*) несмещенных оценок.*

Уравнение гиперплоскости, на котором достигается (1), назовем уравнением L_2 регрессии:

$$\widehat{y}_i = \widehat{a}_1 x_{1i} + \ldots + \widehat{a}_k x_{ki}, \qquad (2)$$

где $\widehat{a}_j$ — оценка метода наименьших квадратов для коэффициента a_i; $\widehat{y}_i$ — прогнозные значения зависимой переменной.

В векторной форме равенство (2) будет иметь вид

$$\widehat{y} = X\widehat{a}, \qquad (3)$$

где $\widehat{a}$ — оценка метода наименьших квадратов векторов параметров; $\widehat{y}$ — прогнозные значения вектора зависимых переменных.

Рассмотрим геометрическую интерпретацию метода наименьших квадратов. Представим $y, x_1, \ldots, x_k$ как векторы в R^N. Эти векторы линейно независимы (в противном случае нет смысла ставить задачу об оценке параметров),

т.е. образуют $(k+1)$-мерное пространство П. По предположению теоремы Гаусса-Маркова, векторы $x_1, \ldots, x_k$ также линейно независимы и порождают в пространстве П k-мерное подпространство π. Вектор $\widehat{y} = X\widehat{a}$ – ортогональная проекция вектора y в подпространство π. Соответственно, $e = y - \widehat{y}$ – вектор, ортогональный подпространству π. Следовательно, функционал $\alpha_2^2 = e^T e$ равен квадрату расстояния между y и π.

Квадрат этого расстояния может быть вычислен с использованием определителя Грама [1]

$$\alpha_2^2 = \frac{G(x_1, x_2, \ldots, x_k, y)}{G(x_1, x_2, \ldots, x_k)}, \tag{4}$$

где $G(x_1, x_2, \ldots, x_k)$ — определитель Грама системы векторов $x_1, x_2, \ldots, x_k$.

Теорема 2 [1]. *Определитель Грама может быть вычислен по формуле*

$$G(x_1, x_2, \ldots, x_k) = \\ = \frac{1}{k!} \sum_{i_1, \ldots, i_k} \begin{vmatrix} x_{1i_1} & x_{2i_1} & \ldots & x_{ki_1} \\ x_{1i_2} & x_{2i_2} & \ldots & x_{ki_2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1i_k} & x_{2i_k} & \ldots & x_{ki_k} \end{vmatrix}^2, \tag{5}$$

где $i_1, \ldots, i_k$ независимо изменяются от 1 до N.

Следствие 1.

$$G(x_1, \ldots, x_k) = \frac{((k-1)!)^2}{k!} \sum_{i_1, \ldots, i_k} V^2_{i_1, \ldots, i_k},$$

$$G(x_1, \ldots, x_k, y) = \frac{(k!)^2}{(k+1)!} \sum_{t_1, \ldots, t_{k+1}} U^2_{t_1, \ldots, t_{k+1}},$$

где $V_{i_1, \ldots, i_k}$ и $U_{t_1, \ldots, t_{k+1}}$ — объемы симплексов с вершинами $\{A_{i_s}(x_{i_s 1}, \ldots, x_{i_s k})\}_{s=1,\ldots,k}$ и соответственно $\{B_{t_s}(x_{t_s 1}, \ldots, x_{t_s k}, y_{t_s})\}_{s=1,\ldots,k+1}$.

Доказательство непосредственно следует из теоремы 2 и формулы ориентированного объема симплекса [2, 3].

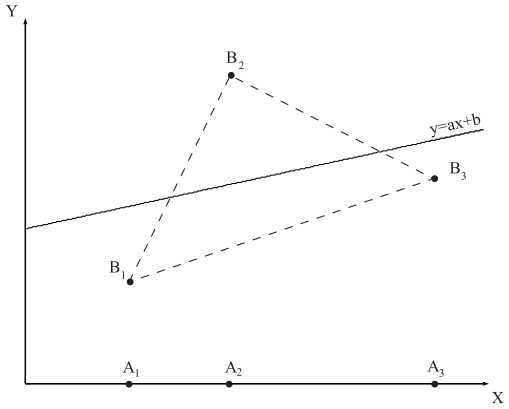
Теорема 3. *Функционал метода наименьших квадратов может быть вычислен по формуле*

$$\alpha_2^2 = k^2 \frac{\frac{1}{(k+1)!} \sum\limits_{t_1, \ldots, t_{k+1}} U^2_{t_1, \ldots, t_{k+1}}}{\frac{1}{k!} \sum\limits_{i_1, \ldots, i_k} V^2_{i_1, \ldots, i_k}}. \tag{6}$$

Доказательство. Справедливость теоремы следует из равенства (4) и следствия 1.

Проиллюстрируем результат теоремы 3. Пусть на плоскости XY даны три точки $B_1(x_1, y_1)$, $B_2(x_2, y_2)$, $B_3(x_3, y_3)$. Функционал качества уравнения для регрессии вида $y = ax + b$, построенного методом наименьших квадратов, пропорционален частному квадрата площади треугольника и суммы квадратов длин проекций его сторон на ось X (рис.), т.е.

$$\alpha_2^2 = 4 \cdot \frac{S^2_{B_1 B_2 B_3}}{(x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_1)^2}.$$



Пример парной регрессии

Таким образом, геометрический смысл величины α_2^2 сводится к отношению суммы квадратов объемов $k+1$-мерных симплексов (k – число регрессоров) и суммы квадратов проекций гипергранней этих сиплексов на гиперплоскость, образованную регрессорами.

В работе [4] в качестве основы берется Чебышевская норма равномерного отклонения.

Определение 1. *Минимальной шириной множества Ω вдоль переменной y назовем число*

$$\alpha_\infty = 2 \cdot \min_{a_s, s \neq j; b} \left\{ \max_{i=1,\ldots,N} |x_{ij} - \sum_{s \neq j}^{k} a_s x_{is} - b| \right\}. \tag{7}$$

С геометрической точки зрения величина α_∞ равна минимуму ширины «полосы», ограниченной двумя параллельными гиперплоскостями и содержащей множество Ω, ширина берется вдоль оси Y в R^{k+1} (т.е. длина пересечения полосы с осью Y).

Уравнение гиперплоскости, на котором достигается (7), назовем уравнением L_∞ регрессии:

$$y = \sum_{s=1}^{k} a_s^0 x_s - b^0, \tag{8}$$

или уравнением регрессии относительно Чебышевской нормы.

Теорема 4. *Справедливо неравенство, связывающее α_2 и α_∞*

$$\frac{\alpha_2}{\alpha_\infty} \leq \sqrt{\frac{(k+1)! \cdot N}{4k^2}}.$$

Доказательство. Очевидно, что для произвольных $i_1, \ldots, i_{k+1}$ справедливо неравенство:

$$U_{i_1,\ldots,i_{k+1}} \le \frac{1}{k} \cdot \alpha_\infty \cdot \frac{1}{2} \sum_{j_1,\ldots,j_k} V_{j_1,\ldots,j_k},$$

где $j_1, \ldots, j_k$ – всевозможные сочетания из номеров $i_1, \ldots, i_{k+1}$.

Возведем последнее равенство в квадрат

$$U^2_{i_1,\ldots,i_{k+1}} \le \frac{\alpha^2_\infty}{4k^2} \cdot \left(\sum_{j_1,\ldots,j_k} V_{j_1,\ldots,j_k} \right)^2.$$

Согласно неравенству Коши-Буняковского $\left(\sum\limits_{i=1}^{n} z_i \right)^2 \le n \sum\limits_{i=1}^{n} z_i^2$ имеем:

$$U^2_{i_1,\ldots,i_{k+1}} \le \frac{\alpha^2_\infty}{4k^2} \cdot (k+1) \cdot \sum_{j_1,\ldots,j_k} V^2_{j_1,\ldots,j_k}.$$

Суммируя, получим:

$$\sum_{i_1,\ldots,i_{k+1}} U^2_{i_1,\ldots,i_{k+1}} \le \frac{\alpha^2_\infty}{4k^2}(k+1)\frac{(k+1)N}{k!} \sum_{i_1,\ldots,i_k} V^2_{i_1,\ldots,i_k}.$$

$$\frac{(k!)^2}{(k+1)!} \sum_{i_1,\ldots,i_{k+1}} U^2_{i_1,\ldots,i_{k+1}} \le$$

$$\le \frac{\alpha^2_\infty}{4k^2}(k+1)\frac{(k+1)N}{k!}\frac{(k!)^2}{(k+1)!} \sum_{i_1,\ldots,i_k} V^2_{i_1,\ldots,i_k}.$$

Воспользовавшись результатами теоремы 3, получаем искомое неравенство:

$$\frac{\alpha^2_2}{\alpha^2_\infty} \le \frac{(k+1)! \cdot N}{4k^2};$$

$$\frac{\alpha_2}{\alpha_\infty} \le \sqrt{\frac{(k+1)! \cdot N}{4k^2}}.$$

Библиографический список

1. Шилов Г.Е. Математический анализ (конечномерные линейные пространства). — М., 1969.

2. Берже М. Геометрия: пер. с франц. — М., 1984. – Т. 1.

3. Берже М. Геометрия: пер. с франц. — М., 1984. – Т. 2.

4. Пономарев И.В., Славский В.В. Равномерно нечеткая модель линейной регрессии // Вестник Новосибирского государственного университета. Сер.: Математика, механика, информатика. — 2010. — Т. 10, №2.