УДК 514.172

Геометрические преобразования модели линейной регрессии

Пономарев И.В.
Алтайский государственный университет, г. Барнаул
igorpon@mail.ru

Аннотация

В статье рассматриваются изменения значений оценок параметров модели парной линейной регрессии при преобразованиях исходных данных. Получены формулы связывающие значения функционалов качества при геометрическом преобразовании исходных данных.

Ключевые слова: регрессия, метод наименьших квадратов, геометрическое преобразование плоскости, функционал качества.

1. Введение

В общей постановке задача, приводящая к модели парной линейной регрессии, формулируется следующим образом: имеется множество $\Omega = \{(x_i; y_i) | x_i, y_i \in \mathbb{R}, i = 1, \ldots, N\}$ необходимо составить уравнение

$$y_i = a_0 + a_1 \cdot x_i + \varepsilon_i$$

аппроксимирующее множество Ω.

Наиболее изученным подходом к решению этой задачи является метод наименьших квадратов. Основная идея которого заключается в минимизации функционала

$$\alpha_2^2(y) = \min \sum_{i=1}^{N} \left(y_i - a_0 - a_1 \cdot x_i\right)^2,$$

аргументом y будем подчеркивать тот факт, что зависимой переменной является y.

Стоит также отметить статистический подход к задаче линейной регрессии

$$y_i = a_0 + a_1 \cdot x_i + \varepsilon_i,$$

т.е. моделируется процесс формирования зависимой (результирующей) переменной y_i, под влиянием независимой переменной x_i, $i = 1, \ldots, N$ и случайных ошибок ε_i; a_j – параметры модели; N – количество наблюдений.

Классические подходы к решению этой задачи основываются на аналитическом методе [1, 2] или с привлечением статистических методов [3, 4].

В данной статье поставим задачу найти связь между показателями парной регрессии после преобразования исходного множества Ω.

2. Геометрическая интерпретация регрессии

С геометрической точки зрения возможны следующие интерпретации задачи парной линейной регрессии:

1. В $\mathbb{R}^N$ имеются три линейно независимых вектора $X(x_1,\ldots,x_N)$, $Y(y_1,\ldots,y_N)$ и $C(1,\ldots,1)$. Необходимо определить вектор $\widehat{Y}(\widehat{y}_1,\ldots,\widehat{y}_N)$, являющийся линейной комбинацией X и C, так, что вектор $Y-\widehat{Y}$ имел наименьшую длину. Длина этого вектора и есть значение функционала $\alpha_2^2(y)$ (см. рисунок 1).

2. Функционал качества $\alpha_2^2(y)$ модели парной регрессии равен отношению учетверенной суммы квадратов площадей всевозможных треугольников с вершинами в точках множества Ω к сумме квадратов длин всех отрезков с концами в проекциях на ось x точек множества Ω [5,6] (см. рисунок 2).

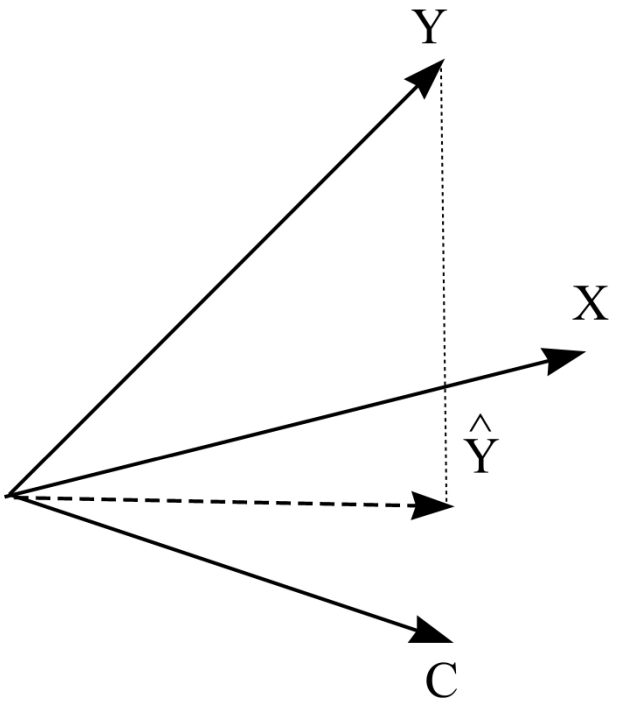

Рисунок 1. Геометрическая интерпретация с помощью N-мерных векторов

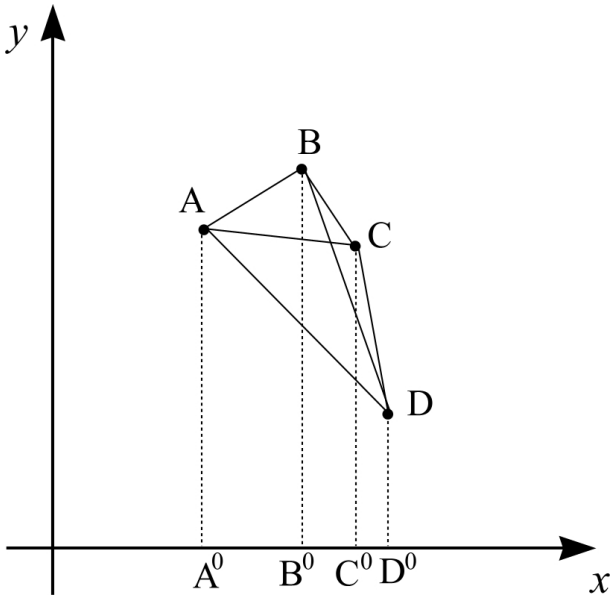

Рисунок 2. Геометрическая интерпретация с помощью проекций

3. Центрированное множество

Пусть центр масс множества точек Ω совпадает с началом координат. Заметим, что в этом случае вектор $Y-\widehat{Y}$ перпендикулярен вектору X и в двух возможных регрессионных моделях

$$\begin{aligned} y_i &= a\cdot x_i+\varepsilon_i,\\ x_i &= b\cdot y_i+\nu_i, \end{aligned} \tag{1}$$

отсутствуют свободные коэффициенты. Оценки параметров данных моделей a и b должны минимизировать функционалы $\alpha_2^2(y)$ и $\alpha_2^2(x)$ соответственно.

Справедливы следующие теоремы

Теорема 1. *Справедливо равенство*

$$(X,Y)=a|X|^2=b|Y|^2,$$

где a,b – коэффициенты регрессий из (1); (X,Y) – скалярное произведение векторов; $|X|,|Y|$ – длины векторов.

Доказательство следует непосредственно из геометрической интерпретации и определения скалярного произведения.

Теорема 2. *Справедливо равенство*

$$|X|^2\alpha_2^2(y) = |Y|^2\alpha_2^2(x).$$

Доказательство основывается на формуле площади треугольника.

Подвергнем множество Ω аффинному преобразованию

$$\begin{cases} x'_i &= k_1x_i + k_2y_i, \\ y'_i &= s_1x_i + s_2y_i. \end{cases}$$

Относительно полученного множества $\Omega' = \{(x'_i, y'_i), i = 1\ldots, N\}$ возникают новые регрессии

$$\begin{aligned} y'_i &= a' \cdot x'_i + \varepsilon'_i, \\ x'_i &= b' \cdot y'_i + \nu'_i, \end{aligned} \tag{2}$$

с функционалами $\alpha_2^2(y')$ и $\alpha_2^2(x')$ соответственно.

Теорема 3. *Значения функционалов качества регрессий (1) и (2) связаны следующим равенством*

$$|X'|^2\alpha_2^2(y') = d^2|X|^2\alpha_2^2(x),$$

где d – определитель матрицы $\left(\begin{smallmatrix} k_1 & k_2 \\ s_1 & s_2 \end{smallmatrix}\right)$.

В частности, если преобразование будет являться поворотом плоскости, т.е.

$$\begin{pmatrix} k_1 & k_2 \\ s_1 & s_2 \end{pmatrix} = \begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix},$$

то формула из теоремы 3 может выть записана в виде

$$|X'|^2\alpha_2^2(y') = |X|^2\alpha_2^2(x).$$

4. Нецентрированное множество

Пусть Ω произвольное множество из $\mathbb{R}^{\not\vDash}$. В этом случае в регрессионных моделях необходимо наличие свободного коэффициента. Рассмотрим множество Ω' полученное из Ω поворотом на угол β относительно начала координат. Тогда для справедлива следующая теорема

Теорема 4. *Функционалы качества парных линейных регрессий на множествах Ω и Ω' связаны равенствами*

$$\begin{aligned} \frac{1}{\alpha_2^2(y')} &= \frac{\cos^2\beta}{\alpha_2^2(y)} + \frac{\sin^2\beta}{\alpha_2^2(x)} - \frac{\sin 2\beta \cdot r(X,Y)}{\alpha_2(y)\cdot\alpha_2(x)}, \\ \frac{1}{\alpha_2^2(x')} &= \frac{\sin^2\beta}{\alpha_2^2(y)} + \frac{\cos^2\beta}{\alpha_2^2(x)} + \frac{\sin 2\beta \cdot r(X,Y)}{\alpha_2(y)\cdot\alpha_2(x)}, \end{aligned}$$

где $r(X,Y)$ – коэффициент корреляции между векторами X И Y.

Доказательство. Воспользуемся второй геометрической интерпретацией. Заметим, что при повороте Ω площади треугольников меняться не будут. Координаты проекций будут меняться по формуле $x'_i = x_i \cos\beta + y_i \sin\beta$. Тогда сумма квадратов длин отрезков равна

$$\sum_{i<j}(x'_i - x'_j)^2 = \cos^2\beta \sum_{i<j}(x_i - x_j)^2 + \sin^2\beta \sum_{i<j}(y_i - y_j)^2 - 2\sin\beta \sum_{i<j}(x_i - x_j)(y_i - y_j).$$

Также исходя из геометрической интерпретации получаем, что

$$4\sum S_\Delta^2 = \alpha_2(y) \cdot \alpha_2(x) \cdot \sqrt{\sum_{i<j}(x_i - x_j)^2 \cdot \sum_{i<j}(y_i - y_j)^2}$$

и

$$\frac{\sum\limits_{i<j}(x_i - x_j)(y_i - y_j)}{\sqrt{\sum\limits_{i<j}(x_i - x_j)^2 \cdot \sum\limits_{i<j}(y_i - y_j)^2}} = r(X, Y).$$

Следовательно, доказываемые формулы верны. □

Список литературы

1. Weisberg S. Applied linear regression. — 3rd edition. — Jonh Wiley & Sans, Inc., 2005.
2. Айвазян С.А. Прикладная статистика. Основы эконометрики. — М. : Юнити-Дана, 2001. — Т. 2.
3. Draper N.R., Smith H. Applied Regression Analysis. — 2nd edition. — Wiley, 1981.
4. Гмурман В.Е. Теория вероятностей и математическая статистика. — 4-е, доп. изд. — М. : Высшая школа, 1972.
5. Пономарев И.В. Геометрический подход к задаче линейной регрессии // Сборник научных статей международной школы-семинара: в 4 частях / Под ред. Е.Д. Родионова. — Барнаул : АлтГПА, 2012. — С. 333–335.
6. Пономарев И.В., Славский В.В. О геометрической интерпретации метода наименьших квадратов // Известия Алтайского гос. ун-та. — 2012. — № 1-1(73). — С. 119–121.