УДК 514.172; 519.654

Действие группы преобразований на показатель качества регрессионной модели

И.В. Пономарев

Алтайский государственный университет (Барнаул, Россия)

The Effect of a Transformation Group on the Quality Indicator of a Linear Regression Model

I.V. Ponomarev

Altai State University (Barnaul, Russia)

Построение функциональных зависимостей между наблюдаемыми явлениями представляет собой важное направление современной прикладной математики. Основой таких построений зачастую является статистический массив данных. От качества этих данных напрямую зависит адекватность получаемых моделей. В общем случае приходится выбирать одну из возможных моделей, основываясь на некотором показателе. Однако полученные выборки могут быть и тождественными, но построенные модели будут отличаться.

Рассматривается один из методов построения линейной регрессии — метод наименьших квадратов. Изучается задача об изменении функционала качества регрессионной модели при ортогональном преобразовании исходного множества данных. Дается геометрическая интерпретация самой регрессионной модели и ее функционала качества, а также статистического показателя связи между переменными — коэффициента корреляции. В явном виде представлены формулы, показывающие зависимость между функционалами качества при вращении множества относительно одной из осей координат в дву- и трехмерном пространствах. Основываясь на полученных формулах, приводится алгоритм, позволяющий получать значение функционала качества при любом собственном движении n-мерного пространства.

Ключевые слова: линейная регрессия, метод наименьших квадратов, группа преобразований, выпуклый анализ.

The construction of functional dependencies between the observed phenomena is an important area of modern applied mathematics. The basis of such constructions is often a statistical data array. The adequacy of the models obtained directly depends on the quality of these data. In general, one has to choose one of the possible models, based on a certain indicator. However, the resulting samples may be in some sense identical, but the models constructed will be different.

This paper discusses one of the methods for constructing linear regression — the method of least squares. The problem of changing the quality functional of a regression model under the orthogonal transformation of the initial data set is studied. A geometric interpretation of the regression model itself and its functional quality, as well as the statistical indicator of the relationship between variables — the correlation coefficient, is given. Formulas are shown explicitly showing the relationship between the functionals of quality during rotation of a set relative to one of the axes of coordinates in two- and three-dimensional spaces. Based on the formulas obtained, an algorithm is presented that allows one to obtain the value of the quality functional with any proper movement of n-dimensional space.

Key words: linear regression, ordinary least squares, transformation group, convex analysis.

DOI 10.14258/izvasu(2019)4-16

1. Введение, постановка задачи. В настоящее время одним из самых распространенных методов изучения закономерностей, по статистическим данным, является регрессионное моделирование. Наиболее востребованным способом оценки линейных регрессионных зависимостей является метод наименьших квадратов [1–4]. В литературе оцениваемые этим методом регрессии получили обозначение L_2. Суть метода наименьших квадратов состоит в нахождении минимального значения квадратов отклонения α_2. Эта величина в дальнейшем применяется при оценке различ-

ных статистических показателей, например, дисперсии ошибок регрессии, коэффициента детерминации и т.п. Соответственно изменение значения α_2 ведет к пересмотру результатов регрессионного моделирования.

Пусть $\mathbb{R}^k$ — k-мерное евклидово пространство. Пусть Ω — конечное подмножество точек:

$$\Omega = \left\{ A_i \left(x_i^1, \ldots, x_i^k \right) : i = 1, \ldots, N \right\},$$

которое можно рассматривать как результат N экспериментов.

Задача линейной регрессии заключается в составлении уравнения

$$x_i^1 = a_0 + a_2 \cdot x_i^2 + \ldots + a_k \cdot x_i^k, \qquad (1)$$

наилучшим образом аппроксимирующее множество Ω.

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

$$\alpha_2(x^1) = \min \sum_{i=1}^{N} \left(x_i^1 - \left(a_0 + a_2 \cdot x_i^2 + \ldots + a_k \cdot x_i^k \right) \right)^2,$$

аргументом x^1 будем подчеркивать тот факт, что результирующей переменной является x^1.

Поставим задачу найти форму зависимости между значениями функционалов качества регрессий до и после преобразования исходного множества Ω.

При решении данной задачи будем опираться на геометрический метод нахождения функционала качества (см. работу [5]). Таким образом, для регрессионной модели (1) значение функционала качества может быть найдено по формуле

$$\alpha_2(x^1) = (k!)^2 \frac{\frac{1}{(k)!} \sum\limits_{i_1, \ldots, i_k} V_{i_1, \ldots, i_k}^2}{\frac{1}{(k-1)!} \sum\limits_{i_1, \ldots, i_{k-1}} V_{i_1, \ldots, i_{k-1}}^2}, \qquad (2)$$

где $V_{i_1, \ldots, i_k}$ — ориентированный объем симплекса с вершинами $A_{i_1}(x_{i_1}^1, \ldots, x_{i_1}^k)$, …, $A_{i_k}(x_{i_k}^1, \ldots, x_{i_k}^k)$.

2. Преобразования плоскости и пространства. Заметим, что при любом значении k параллельный перенос не меняет объемов всевозможных симплексов, полученных из множества Ω. Следовательно, значения функционалов качества линейной регрессии меняться не будут, т.е.

$$\left(x_i^1, \ldots, x_i^k \right) \to \left(x_i^1 + \xi^1, \ldots, x_i^k + \xi^k \right) \Rightarrow$$
$$\Rightarrow \alpha_2(x^1) = \alpha_2(x^1 + \xi^1). \qquad (3)$$

Подвергнем множество Ω вращению и получим Ω':

$$\Omega' = \left\{ B_i \left(y_i^1, \ldots, y_i^k \right) : i = 1, \ldots, N \right\},$$

где

$$\left(y_i^1, \ldots, y_i^k \right) = \left(x_i^1, \ldots, x_i^k \right) \cdot O, \quad \det(O) = 1.$$

В случае $k = 2$ (плоскость) матрица $O = \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix}$. Формулы преобразования функционалов качества доказаны в [6] и имеют вид

$$\frac{1}{\alpha_2(y^1)} = \frac{\cos^2\beta}{\alpha_2(x^1)} + \frac{\sin^2\beta}{\alpha_2(x^2)} - \frac{\sin 2\beta \cdot r(X^1, X^2)}{\sqrt{\alpha_2(x^1) \cdot \alpha_2(x^2)}},$$
$$\frac{1}{\alpha_2(y^2)} = \frac{\sin^2\beta}{\alpha_2(x^1)} + \frac{\cos^2\beta}{\alpha_2^2(x^2)} + \frac{\sin 2\beta \cdot r(X^1, X^2)}{\sqrt{\alpha_2(x^1) \cdot \alpha_2(x^2)}},$$

где $r(X^1, X^2)$ — коэффициент корреляции между векторами $X^1 = \begin{pmatrix} x_1^1 \\ \vdots \\ x_N^1 \end{pmatrix}$ и $X^2 = \begin{pmatrix} x_1^2 \\ \vdots \\ x_N^2 \end{pmatrix}$.

В случае $k = 3$ (трехмерное пространство) рассмотрим вращение относительно какой-нибудь оси координат, например, относительно оси x^3. Тогда матрица вращения O будет иметь вид

$$O = \begin{pmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Следовательно, необходимо рассматривать три регрессионные модели на множестве Ω

$$x_i^1 = a_0 + a_2 x_i^2 + a_3 x_i^3,$$
$$x_i^2 = b_0 + b_1 x_i^1 + b_3 x_i^3,$$
$$x_i^3 = c_0 + c_1 x_i^1 + c_2 x_i^2,$$

с функционалами качества $\alpha_2(x^1)$, $\alpha_2(x^2)$ и $\alpha_2(x^3)$ соответственно. На множестве Ω' значения функционалов качества аналогичных регрессий от переменных y станут равны $\alpha_2(y^1)$, $\alpha_2(y^2)$ и $\alpha_2(y^3)$.

Рассмотрим один из статистических показателей связи между случайными переменными.

Определение 1. Коэффициент частной корреляции показывает степень взаимосвязи двух переменных относительно друг друга без учета влияния третьей переменной.

Коэффициент частной корреляции между случайными векторами X^1 и X^2 без учета влияния X^3 может быть вычислен по формуле [7]

$$r(X^1, X^2 | X^3) = \frac{r(X^1, X^2) - r(X^1, X^3) \cdot r(X^2, X^3)}{\sqrt{(1 - r^2(X^1, X^3)) \cdot (1 - r^2(X^2, X^3))}}. \qquad (4)$$

Заметим, что коэффициент частной корреляции можно геометрически интерпретировать как проектирование угла между переменными X^1 и X^2 в исходном пространстве на ортогональное подпространство с фиксированной переменной X^3 [8].

Теорема 1. Коэффициент частной корреляции на множестве Ω может быть найден по формуле

$$r(X^1, X^2 \big| X^3) = \frac{\sum\limits_{i,j,m} S_{1;i,j,m} \cdot S_{2;i,j,m}}{\sqrt{\sum\limits_{i,j,m} S^2_{1;i,j,m} \cdot \sum\limits_{i,j,m} S^2_{2;i,j,m}}}, \quad (5)$$

где $S_{1;i,j,m}$ и $S_{2;i,j,m}$ — площади треугольников, построенных на точках $C_i(x_i^1, x_i^3)$, $C_j(x_j^1, x_j^3)$, $C_m(x_m^1, x_m^3)$ и $D_i(x_i^2, x_i^3)$, $D_j(x_j^2, x_j^3)$, $D_m(x_m^2, x_m^3)$ соответственно.

Доказательство. Подставим вместо коэффициента корреляции в (4) формулу для вычисления корреляции (см. [6]):

$$\frac{\sum\limits_{i<j}(x_i^1 - x_j^1)(x_i^2 - x_j^2)}{\sqrt{\sum\limits_{i<j}(x_i^1 - x_j^1)^2 \cdot \sum\limits_{i<j}(x_i^2 - x_j^2)^2}} = r(X^1, X^2).$$

Перегруппируем полученное выражение с учетом формулы вычисления площади треугольника с вершинами $P_i(x_i^1, x_i^2)$, $P_j(x_j^1, x_j^2)$ и $P_m(x_m^1, x_m^2)$:

$$S_{i,j,m} = \frac{1}{2}\big|(x_i^1 - x_j^1)(x_i^2 + x_j^2) + \\ +(x_j^1 - x_m^1)(x_j^2 + x_m^2) + (x_m^1 - x_i^1)(x_m^2 + x_i^2)\big|.$$

Получим требуемый результат.

Теорема 2. Функционалы качества линейных регрессионных моделей на множествах Ω и Ω' связаны равенствами

$$\frac{1}{\alpha_2(y^1)} = \frac{\cos^2\beta}{\alpha_2(x^1)} + \frac{\sin^2\beta}{\alpha_2(x^2)} - \frac{\sin 2\beta \cdot r(X^1, X^2 \big| X^3)}{\sqrt{\alpha_2(x^1) \cdot \alpha_2(x^2)}},$$

$$\frac{1}{\alpha_2(y^2)} = \frac{\sin^2\beta}{\alpha_2(x^1)} + \frac{\cos^2\beta}{\alpha_2(x^2)} + \frac{\sin 2\beta \cdot r(X^1, X^2 \big| X^3)}{\sqrt{\alpha_2(x^1) \cdot \alpha_2(x^2)}},$$

$$\frac{1}{\alpha_2(y^3)} = \frac{1}{\alpha_2(x^3)},$$

где $r(X^1, X^2 \big| X^3)$ — частный коэффициент корреляции между векторами X^1 и X^2.

Доказательство. Воспользуемся второй геометрической интерпретацией. Заметим, что при повороте Ω относительно оси x^3 объемы всевозможных тетраэдров меняться не будут.

Докажем первую формулу. Заметим, что на множестве Ω' координаты проекций на плоскость регрессоров будут иметь вид $\left(-x_i^1 \sin\beta + x_i^2 \cos\beta; x_i^3\right)$. Тогда сумма квадратов площадей всевозможных треугольников равна

$$\cos^2\beta \sum_{i,j,m} S^2_{1;i,j,m} + \sin^2\beta \sum_{i,j,m} S^2_{2;i,j,m} - \\ - \sin 2\beta \sum_{i,j,m} S_{1;i,j,m} \cdot S_{2;i,j,m}.$$

Рассмотрим обратное значение функционала качества $\alpha_2(y^1)$ в представлении (2). т. е. разделим последнее выражение на сумму квадратов объемов всевозможных тетраэдров во множестве Ω. Первые два слагаемых преобразуются к виду

$$\frac{\cos^2\beta}{\alpha_2(x^1)} + \frac{\sin^2\beta}{\alpha_2(x^2)}.$$

Для приведения третьего слагаемого используем следствие из (2):

$$(3!) \sum_{i,j,m,l} V^2_{i,j,m,l} = \\ = \frac{1}{2!}\sqrt{\sum_{i,j,m} S^2_{1;i,j,m} \cdot \sum_{i,j,m} S^2_{2;i,j,m}}\sqrt{\alpha_2(x^1)\alpha_2(x^2)}.$$

Применяем к третьему слагаемому результат теоремы 1 и получаем требуемую формулу.

Справедливость вторая формулы доказывается аналогично с учетом того, что точки проекций будут иметь координаты $\left(x_i^1 \cos\beta + x_i^2 \sin\beta; x_i^3\right)$.

Для доказательства третьей формулы достаточно заметить, что геометрическая картинка для регрессии не изменится.

Доказательство закончено.

Следствие. Для регрессий на множествах Ω и Ω' справедливо равенство

$$\frac{1}{\alpha_2(y^1)} + \frac{1}{\alpha_2(y^2)} + \frac{1}{\alpha_2(y^3)} = \frac{1}{\alpha_2(x^1)} + \frac{1}{\alpha_2(x^2)} + \frac{1}{\alpha_2(x^3)}.$$

Справедливость этого утверждения доказывается суммированием равенств из теоремы 2.

3. Заключение и выводы. Известно, что любое вращение в n-мерном пространстве может быть представлено как композиция поворотов вокруг осей координат [9, 10]. Следовательно, для получения значения функционала качества при любом преобразовании множества Ω достаточно несколько раз последовательно воспользоваться формулами теоремы 2. Явное написание общих формул лишено смысла ввиду большого числа слагаемых.

Любое собственное движение пространства может быть представлено как композиция переноса и вращения. Значит, мы установили закон изменения значения функционала при воздействии группы преобразований n-мерного пространства на множество Ω.

Библиографический список

1. Greene W.H. Econometric Analysis. 5th edition. N.Y., 2008.

2. Дрейпер Н, Смит Г. Прикладной регрессионный анализ. Множественная регрессия // Applied Regression Analysis. 3-е изд. М., 2007.

3. Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М., 2010.

4. Amiri-Simkooei A.R., Jazaeri S. Weighted total least squares formulated by standard least squares theory // Journal of Geodetic Science. 2012. V. 2(2).

5. Пономарев И.В., Славский В.В. О геометрической интерпретации метода наименьших квадратов // Известия Алт. гос. ун-та. 2012. № 1-1(73).

6. Пономарев И.В. Геометрические преобразования модели линейной регрессии // Труды семинара по геометрии и математическому моделированию. 2018. №4.

7. Лагутин М.Б. Наглядная математическая статистика. В 2-х т. М., 2003.

8. Кендалл М., Стюарт А. Статистические выводы и связи. М., 1973. Т. 2.

9. Берже М. Геометрия. М., 1984. Т. 1.

10. Шафаревич И.Р., Ремизов А.О. Линейная алгебра и геометрия. М., 2009.