УДК 519.65

# Об оценке степени однородности выборки в равномерно-регрессионной модели

*Т.П. Махаева[1], И.В. Пономарев[2]*

[1]Алтайский государственный педагогический университет (Барнаул, Россия)
[2]Алтайский государственный университет (Барнаул, Россия)

# On the Assessment of Homogeneity in a Uniform Regression Model

*T.P. Makhaeva[1], I.V. Ponomarev[2]*

[1]Altai State Pedagogical University (Barnaul, Russia)
[2]Altai State University (Barnaul, Russia)

При проведении разведочного анализа данных и последующего построения функциональных зависимостей между наблюдаемыми явлениями часто необходимо оценить степень зависимости между изучаемыми данными. В основу получения таких критериев при вероятностном подходе обычно закладывается корреляционная составляющая выборки. Выбор применяемого показателя напрямую зависит от методов изучения выборки, а также инструментов построения модели. В большинстве случаев на начальном этапе моделирования исследуются именно оценки однородности выборки, хороший подбор которых может сократить трудоемкость построения зависимости между данными.

В представленной работе изучается способ оценки однородности выборочных данных при построении равномерно-регрессионной модели. В первой части работы описывается коэффициент корреляции для $L_\infty$-регрессии, изучается интервал его изменения, описываются геометрическая интерпретация и алгоритм построения данного показателя. Во второй части работы исследуется метод построения показателя «сконцентрированности» выборки. Для этого выводятся формулы, связывающие коэффициент корреляции с размахом исходной выборки. В заключении приводится описание алгоритмов построения рассматриваемых показателей, делаются выводы о сложности данных алгоритмов.

***Ключевые слова***: равномерно-регрессионная модель, коэффициент корреляции, выпуклая оболочка, вычислительная сложность.

When conducting an exploratory analysis of the data and the subsequent construction of functional dependencies between the observed phenomena, it is often necessary to assess the degree of dependence between the studied data. The basis for obtaining such criteria with a probabilistic approach usually includes the correlation component of the sample. The choice of the used indicator directly depends on the methods of studying the sample, as well as the tools for constructing the model. In most cases, at the initial stage of modeling, it is precisely the homogeneity estimates of the sample that are studied, a good selection of which can reduce the complexity of constructing the relationship between the data.

In this paper, we study a method for assessing the uniformity of sample data when constructing a uniform regression model. The first part of the paper describes the correlation coefficient for the $L_\infty$ regression, studies the interval of its change, describes the geometric interpretation and the algorithm for constructing this indicator. In the second part of the paper, we study the method of constructing an indicator of "concentration" of the sample. For this, formulas are derived that relate the correlation coefficient to the magnitude of the original sample. In conclusion, a description is given of the algorithms for constructing the considered indicators, and conclusions are drawn about the complexity of these algorithms.

***Key words***: uniformly regression model, correlation coefficient, convex hull, computational complexity.

**DOI 10.14258/izvasu(2020)1-19**

**1. Введение.** Пусть $R^2$ – двумерное арифметическое евклидово пространство. Пусть $\Omega$ конечное подмножество точек:

$$\Omega = \{(x_i, y_i) : i = 1, \ldots, N\},$$

которое можно рассматривать как результат $N$ экспериментов. В работе [1] определен алгоритм построения линейной зависимости ($L_\infty$ регрессии) между координатами точек $\Omega$ на основе Чебышев-

ской нормы равномерного отклонения

$$\alpha_\infty(y) = 2 \cdot \min_{k;b} \left\{ \max_{i=1,\dots,N} |y_i - kx_i - b| \right\}.$$

С геометрической точки зрения величина $\alpha_\infty(y)$ равна минимуму ширины «полосы», ограниченной двумя параллельными прямыми, содержащими множества $\Omega$, ширина берется вдоль оси $y$. Эта величина также тесно связана с такими понятиями из выпуклой геометрии, как ширина выпуклого множества в данном направлении и широта выпуклого множества [2].

Уравнение гиперплоскости, на котором достигается $\alpha_\infty(y)$, называется уравнением $L_\infty$ регрессии. Очевидно, что для множества $\Omega$ возможно построение двух регрессий:

$$y = k_\infty x + b_\infty,$$
$$x = \overline{k}_\infty y + \overline{b}_\infty$$

с функционалами качества $\alpha_\infty = \alpha_\infty(\Omega, y)$ и $\overline{\alpha}_\infty = \alpha_\infty(\Omega, x)$ соответственно.

**2. Корреляция для $L_\infty$ регрессии.** Основным показателем линейной зависимости между одномерными выборками является коэффициент корреляции [3]. Рассмотрим задачу нахождения $\alpha_\infty$ (соответственно $\overline{\alpha}_\infty$) с геометрической точки зрения.

Решение сводится к нахождению полосы, заключенной между двумя параллельными прямыми и содержащей множество точек $\Omega$ такой, что существует треугольник $\triangle A_i A_j A_t$ с вершинами на прямых, у которого одна из вершин проектируется вдоль оси $OY$ (соответственно $OX$) на основание треугольника, как показано на рисунке 2.

**Определение 1.** Определим коэффициент корреляции $\mathrm{korr}_\infty(X,Y)$ для $L_\infty$ регрессии формулой:

$$\mathrm{korr}_\infty(X,Y) = k_\infty \cdot \overline{k}_\infty, \qquad (1)$$

где $k_\infty$, $\overline{k}_\infty$ угловые коэффициенты прямых $L_\infty$ регрессий $y$ на $x$ и $x$ на $y$ соответственно.

**Теорема 1.** Справедливо неравенство:

$$-1 \le \mathrm{korr}_\infty(X,Y) \le 1.$$

**Доказательство.** На рисунке 1 изображены полосы вертикальной и горизонтальной минимальной ширины для множества $\Omega$. Обозначим длины отрезков на осях $OX$ и $OY$, высекаемые этими полосами через $a$, $b$ и $a_1$, $b_1$ соответственно. Тогда

$$b = \alpha_\infty \le b_1, \quad a_1 = \overline{\alpha}_\infty \le a.$$

Перенося параллельно параллелограмм $\Omega$ в начало координат, как указано на рисунке, заметим, что

$$|k_\infty| = \frac{b}{a}, \quad |\overline{k}_\infty| = \frac{a_1}{b_1},$$

отсюда получим

$$|\mathrm{korr}_\infty(X,Y)| = \frac{b}{a} \cdot \frac{a_1}{b_1} \le 1.$$

Знак равенства достигается при $b = b_1$ и $a = a_1$

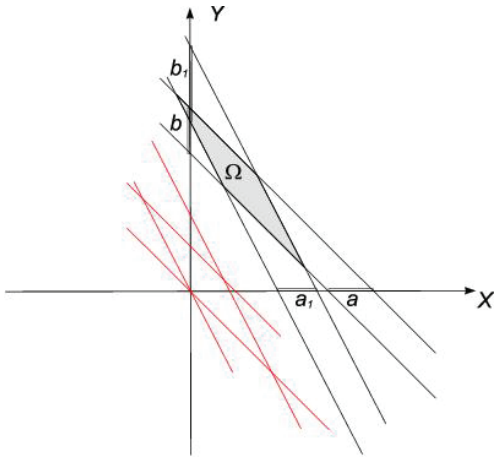


Рис. 1. Полосы вертикальной и горизонтальной минимальной ширины

когда либо полосы вертикальной и горизонтальной минимальной ширины совпадают, либо получены симметрией друг из друга относительно оси $OX$ или $OY$ и параллельным сдвигом. При условии регулярности экстремального симплекса второй случай невозможен [4].

**Замечание.** Совпадение полос минимальной вертикальной и горизонтальной ширины соответствует существованию двух экстремальных треугольников $\triangle A_i A_j A_t$ и $\triangle A_r A_s A_h$ с вершинами на прямых, у которых одна из вершин проектируется вдоль оси координат на основание треугольника, как показано на рисунке 2

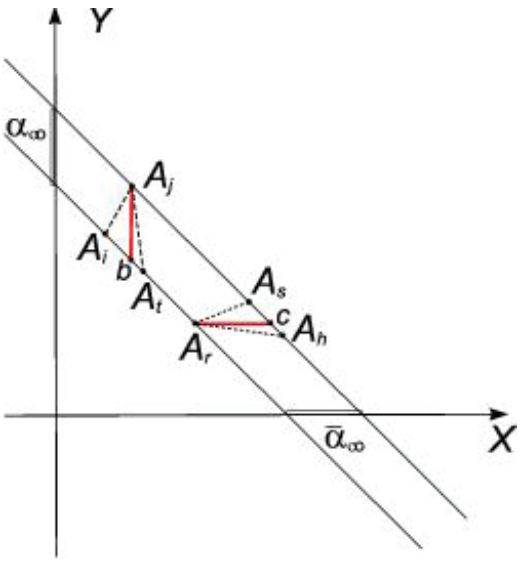


Рис. 2. Экстремальные $\triangle A_i A_j A_t$, $\triangle A_r A_s A_h$

Введенный таким образом коэффициент корреляции может быть использован в прикладных исследованиях как показатель причинности между признаками $x$ и $y$:
1. $\mathrm{korr}_\infty(X,Y) > 0$ – влияния $x$ на $y$ и $y$ на $x$ имеют одинаковое направление;
2. $\mathrm{korr}_\infty(X,Y) < 0$ – $x$ и $y$ оказывают друг на друга противоположное воздействие.

**3. Степень оценки «сконцентрированности» множества.** При использовании вероятностных методов построения регрессионной модели основной характеристикой разброса наблюдений в одномерном случае является среднеквадратичное отклонение, а в двумерном — матрица ковариации [5]. Эти характеристики показывают величину разброса наблюдений относительно среднего значения. Введем показатель «сконцентрированности» выборки в модели $L_\infty$ на основании следующей теоремы.

**Теорема 2.** Справедливо равенство:

$$S = \frac{\alpha_\infty \overline{\alpha}_\infty}{1 - \mathrm{korr}_\infty(X,Y)}, \qquad (2)$$

где $S$ — площадь параллелограмма, $\alpha_\infty$, $\overline{\alpha}_\infty$ — вертикальная и горизонтальная минимальная ширина.

Доказательство основывается на построении уравнений границ полос «в отрезках» и параллельном переносе полученного параллелограмма, содержащего $\Omega$, в начало координат.

**Следствие.** Геометрически (2) означает следующее равенство:

$$S = \frac{2S^*}{1 - \mathrm{korr}_\infty(X,Y)},$$

где $S$, $S^*$ — площади фигур, изображенных на рисунке 3

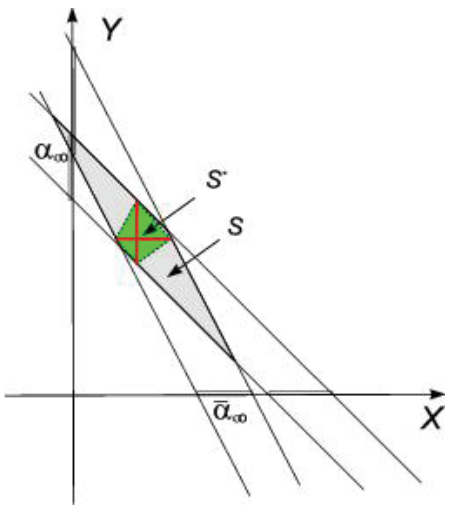


Рис. 3. Геометрическая интерпретация коэффициента корреляции $\mathrm{korr}_\infty(X,Y)$

Интерпретировать площадь множества $S^*$ можно как наименьшее допустимое подмножество $\Omega$, при котором не изменяются коэффициенты и другие характеристики $L_\infty$-регрессий.

**4. Вычислительная процедура.** Для вычисления коэффициента корреляции необходимо вычислить параметры равномерно-регрессионной модели. Это можно сделать с помощью алгоритмов построения выпуклой оболочки множества $\Omega$. Наиболее эффективными алгоритмами построения выпуклой оболочки являются:
• обход Грэхема, суть которого состоит в переводе точек множества $\Omega$ в полярную систему координат с последующим сравнением троек этих точек. Сложность алгоритма $O(N \log N)$ [6];
• алгоритм Quickhull. Главным преимуществом Quickhull является малая чувствительность к большому объему данных и погрешностям вычислений. Сложность данного алгоритма $O(N \log N)$ [7];
• алгоритм Чана, который является объединением алгоритмов Грэхема и Джарвиса и имеет более приемлемую сложность $O(N \log h)$ [8],
где $N$ — количество точек в $\Omega$; $h$ — количество точек в выпуклой оболочке.

После построения выпуклой оболочки требуется определить ширину наименьшей вертикальной и горизонтальной полос, содержащих множество $\Omega$. Эта процедура имеет сложность $O(h)$ и осуществляется по следующему алгоритму:

1. Выбирается одна из сторон полученной выпуклой оболочки $A_iA_{i+1}$ и определяется прямая, содержащая эту сторону $l_i: \ y = kx + b$.
2. Для оставшихся вершин оболочки $A_j$ находятся прямые $l_j: \ y = k(x - x_j) + y_j$ и проверяется условие, что полоса, ограниченная прямыми $l_i$ и $l_j$, содержит все точки выпуклой оболочки.
3. Операции 2)-3) повторяются до тех пор, пока не будут рассмотрены все стороны выпуклой оболочки.
4. Из найденных полос выбирается одна, имеющая наименьшую вертикальную и горизонтальную ширину.

Далее по формулам (1) и (2) находим коэффициент корреляции и показатель разброса выборки.

**5. Заключение и выводы.** Разработанные в статье методы оценки однородности выборки являются универсальными и могут применяться при проведении разведочного анализа статистических данных как показатели внутреннего строения исследуемого множества [9]. Стоит отметить, что оценка «сконцентрированности» выборки может применяться в процедурах нахождения выбросов аналогичных алгоритму, рассмотренному в [10].

## Библиографический список

1. Ponomarev I.V., Slavsky V.V. Uniformly fuzzy model of linear regression // Journal of Mathematical Sciences. 2012. Vol. 186, Issue 3.

2. Сантало Луи А. Интегральная геометрия и геометрические вероятности : пер. с англ. / под ред. Р.В. Амбарцумяна. М., 1983.

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

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

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

6. Берг М., Чеонг О., Кревельд М., Овермарс М. Вычислительная геометрия. Алгоритмы и приложения = Computational Geometry: Algorithms and Applications. М., 2016.

7. Barber C.B., Dobkin D.P., Huhdanpa H.T. The Quickhull Algorithm for Convex Hulls // ACM Transactions on Mathematical Software. 1996. Vol. 22, № 4.

8. David M. Mount. Computational Geometry. University of Maryland, 2002.

9. Брюс П., Брюс Э. Практическая статистика для специалистов Data Science : пер. с англ. СПб., 2018.

10. Пономарев И.В., Саженкова Т.В., Славский В.В. Метод поиска экстремальных наблюдений в задаче нечеткой регрессии // Известия Алт. гос. ун-та. 2018. №4 (102). DOI:10.14258/izvasu(2018)4-18.