УДК 510.5

Обобщенный алгоритм поиска выбросов в регрессионной модели

М.В. Куркина[1], И.В. Пономарев[2]

[1]Югорский государственный университет (Ханты-Мансийск, Россия)
[2]Алтайский государственный университет (Барнаул, Россия)

Generalized Algorithm for Finding Outliers in a Regression Model

M.V. Kurkina[1], I.V. Ponomarev[2]

[1]Ugra State University (Khanty-Mansiysk, Russia)
[2]Altai State University (Barnaul, Russia)

Одним из активно развивающихся направлений современных вычислительных задач является анализ данных. Изучаемые данные обладают различной структурой, что вызывает определенные трудности в процессе их сглаживания и анализа. Это влечет за собой потребность поиска новых универсальных алгоритмов обработки данных, создания компьютерных программ, обеспечивающих анализ данных различной природы. На сегодняшний день широко применяемым методом обработки данных является регрессионное моделирование. Оно применяется в задачах распознавания образов, классификации, снижения размерности и многих других. Очень важным требованием к качеству таких моделей является отсутствие в данных резко выделяющихся наблюдений (выбросов).

В представленной статье рассматривается метод исследования выборки на предмет выбросов. Полученный алгоритм может быть применен к регрессионным моделям, оцениваемым наиболее распространенными методами (метод наименьших квадратов, метод наименьших модулей). Математической основой данной процедуры является преобразование Лежандра, что обеспечивает при компьютерной реализации вычислительную точность. Адекватность полученного алгоритма была исследована на ряде тестовых выборок. Все испытания дали положительный результат с точки зрения определения выбросов. Был создан комплекс программ в системе MatLab, который позволяет строить различные регрессионные модели, а также оценивать исходную выборку на предмет резко выделяющихся наблюдений.

Ключевые слова: линейная регрессия, метод наименьших квадратов, метод наименьших модулей, статистические выбросы.

One of the actively developing areas of modern computational problems is data analysis. The studied data have a different structure, which causes certain difficulties in the process of smoothing and analysis. This fact entails the need to search for new universal algorithms for data processing and create computer programs that analyze data of various nature. Today, a widely used method of data processing is regression modeling. It is used in problems of pattern recognition, classification, dimensionality reduction, and many others. The literature describes various methods of constructing regression models, the basis of which is the optimization of a certain indicator — the quality functional. A very important requirement for the quality of such models is the absence of outliers (outliers) in the data.

This article discusses a method for examining a sample for outliers. The obtained algorithm can be applied to regression models estimated by the most common methods (least squares method, least modulus method). The mathematical basis of this procedure is the Legendre transformation, which provides computational accuracy in computer implementation. The adequacy of the obtained algorithm was investigated on a number of test samples. All tests were positive in terms of emissions. The MatLab system is used to develop a set of programs, which allows the building of various regression models and evaluation of the original sample for sharply distinguished observations.

Key words: linear regression, least squares, least modulus, statistical outliers.

DOI: 10.14258/izvasu(2021)4-16

Введение, постановка задачи. Регрессионное моделирование является одним из самых распространенных и часто используемых методов обработки и анализа информации. Постановка задачи линейной регрессии является стандартной и широко известной.

Пусть $\mathbb{R}^{k+1}$ – $k+1$-мерное евклидово пространство и $\Omega \subset \mathbb{R}^{k+1}$ конечное подмножество точек:

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

которое можно рассматривать как результат N экспериментов.

Задача линейной регрессии заключается в составлении уравнения

$$y_i = a_0 + a_1 \cdot x_i^1 + \ldots + a_k \cdot x_i^k + \varepsilon_i, \qquad (1)$$

наилучшим образом аппроксимирующего множество Ω, где ε_i — ошибка аппроксимации [1].

Различия в подходе к решению этой задачи начинают проявляться при выборе показателя качества. Так, например, если в функционале качества выбрана чебышевская метрика, то задача о нахождении параметров модели может быть записана в следующем виде:

$$\alpha\left(\Omega\right) = 2 \cdot \min_{a_p} \left\{ \max_{i=1,\ldots,N} |y_i - \sum_{j=1}^{k} a_j x_i^j - a_0| \right\}.$$

Также в исследованиях широкое применение получили:

- метод наименьших модулей [2,3];
- метод наименьших квадратов [4,5].

Задача о выбросах [6, 7] может быть сформулирована как задача об исключении из экспериментального множества данных Ω небольшого (обычно 5 % от исходного объема выборки) числа наблюдений таким образом, чтобы оставшиеся данные Ω_0 имели наименьшую величину разброса $\alpha(\Omega_0)$, т.е.

$$\alpha(\Omega_0) = \min \left\{ \alpha(\Omega') : \Omega' \subset \Omega, \, \#\left[\Omega'\right] = N_0 \right\}, \qquad (2)$$

где $\#\left[\Omega'\right]$ — число элементов в множестве Ω'; $N_0 < N$; $N - N_0 = m_0$ — число выбросов.

В работе [8] подробно описана процедура выявления выбросов в модели, основанной на чебышевской норме. При этом для пары натуральных чисел $1 \le r, s \le n$ определим две функции:

$$\underset{r}{MAX}\left[\{c_i\}_{i=1}^n\right] = c_{i_{r+1}}, \quad \underset{s}{MIN}\left[\{c_i\}_{i=1}^n\right] = c_{i_{n-s}},$$

где $\{c_{i_k}\}_{k=1}^n$ — перестановка последовательности $\{c_i\}_{i=1}^n$ в порядке убывания:

$$c_{i_1} \ge c_{i_2} \ge \ldots \ge c_{i_k} \ge \ldots \ge c_{i_n}.$$

Таким образом:

$$\underset{0}{MAX}\left[\{c_i\}_{i=1}^n\right] = \max\left[\{c_i\}_{i=1}^n\right]$$
$$\underset{0}{MIN}\left[\{c_i\}_{i=1}^n\right] = \min\left[\{c_i\}_{i=1}^n\right].$$

Имеющиеся в литературе методы оценки выбросов (см., например, [9,10]) привязаны к методу оценки модели. Поставим задачу обобщения этого алгоритма для использования в регрессионных моделях, оцениваемых другими методами.

Обобщение алгоритма для метода наименьших квадратов. Основой для построения регрессионной модели методом наименьших квадратов является минимизация величины

$$\alpha\left(\Omega\right) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{k} a_j x_i^j - a_0 \right)^2.$$

Поэтому точками, подозрительными на выбросы, можно считать те, которые максимально увеличивают этот функционал.

При заданных значениях параметров расположим величины квадратов отклонений по убыванию. Наибольший вклад в увеличение функционала будут вносить первые m_0 значений последовательности.

Для построения этой последовательности определим функции преобразования Лежандра. Данные функции будут иметь вид

$$f_r^+\left(a_p\right) = \underset{r}{MAX}\left\{ \left(a_0 + \sum_{j=1}^{k} a_j x_i^j - y_i \right)^2 \right\},$$

$$f_s^-\left(a_p\right) = \underset{s}{MIN}\left\{ \left(a_0 + \sum_{j=1}^{k} a_j x_i^j - y_i \right)^2 \right\},$$

где $i = 1, \ldots, N$.

Данные функции указывают $(r+1)$-й максимальный и $(s+1)$-й минимальный квадраты остатков в уравнении регрессии с параметрами a_p. Соответственно, лучшим уравнением регрессии можно признать то уравнение, в котором сумма квадратов остатков наименьшая. Варьируя значения параметров, можно определить минимальную величину этой суммы.

Теорема 1. Справедливо равенство:

$$\alpha(\Omega_0) = \min_{a_s} \sum_{0 \le r \le m_0 - 1} f_r^-\left(a_p\right) \qquad (3)$$

Доказательство. Рассмотрим правую часть равенства, и пусть минимум достигается при $a_p = a_p^*$ $(p = 0, \ldots, k)$. Перенумеруем последовательность квадратов остатков в порядке убывания

$$\left(a_0 + \sum_{j=1}^{k} a_j x_1^j - y_1 \right)^2 \ge \cdots \ge \left(a_0 + \sum_{j=1}^{k} a_j x_N^j - y_N \right)^2.$$

И так как последовательность упорядочена и состоит из неотрицательных чисел, то и сумма членов последовательности с номерами t,

где $N - m_0 < t \leq N$, будет минимальна. Это соответствует минимуму функционала $\alpha(\Omega')$.

Алгоритм описанной выше процедуры оценки выбросов имеет следующую структуру:

Шаг 1. Выбираются некоторые значения параметров модели a_p $(p = 0, \ldots, k)$.

Шаг 2. Для всех наблюдений выборки Ω вычисляются и упорядочиваются квадраты остатков.

Шаг 3. Вычисляют суммы первых m_0 квадратов остатков. Запоминается величина параметров, соответствующих этому значению.

Шаг 4. Шаги 1–3 повторяются для всевозможных параметров модели.

Шаг 5. Определяются номера наблюдений, для которых сумма квадратов остатков принимает наименьшее значение.

Данный алгоритм был запрограммирован в среде MatLab. Для проверки адекватности было проведено тестирование данной программы на ряде модельных примеров: выборки подвергались «искусственному» засорению. На рисунке 1 представлено изменение функционала качества при исключении из выборки одного наблюдения.

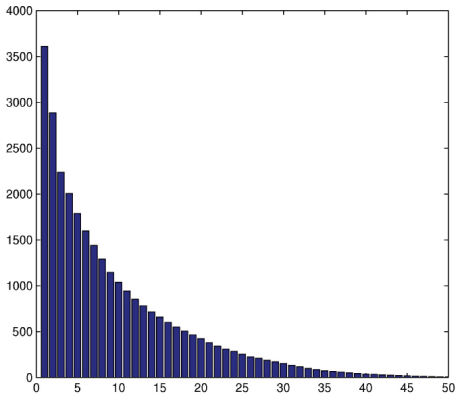

Рис. 1. Диаграмма изменения величины функционала $\alpha\,(\Omega)$

Из диаграммы видно, что на первых этапах происходит весомое изменение значения функционала качества (два первых наблюдения). Последующие изменения становятся более равномерными. Это свидетельствует о возможности исключения из выборки трех наблюдений. Заметим, что эти наблюдения и были включены в исходную выборку в качестве выбросов.

Обобщение алгоритма для метода наименьших модулей. При оценке параметров линейной регрессионной модели методом наименьших модулей видно:

$$\alpha\,(\Omega) = \sum_{i=1}^{N} \left| y_i - \sum_{j=1}^{k} a_j x_i^j - a_0 \right|.$$

При решении проблемы поиска выбросов можно воспользоваться аналогичным преобразованием. В этом случае функции преобразования Лежандра будут определятся как

$$f_r^+\,(a_p) = \underset{r}{MAX} \left\{ \left| a_0 + \sum_{j=1}^{k} a_j x_i^j - y_i \right| \right\},$$

$$f_s^-\,(a_p) = \underset{s}{MIN} \left\{ \left| a_0 + \sum_{j=1}^{k} a_j x_i^j - y_i \right| \right\},$$

где $i = 1, \ldots, N$.

Сформулируем теорему:

Теорема 2. Справедливо равенство:

$$\alpha(\Omega_0) = \min_{a_p} \sum_{0 \leq r \leq m_0 - 1} f_r^-\,(a_p)\,.$$

Доказательство аналогично доказательству теоремы 1.

Заметим, что алгоритм процедуры полностью соответствует алгоритму для метода наименьших квадратов. Изменения касаются только используемого функционала качества.

Проведенные тестирования на модельных выборках показали, что данный метод верно определяет 3–4 выброса при объеме выборки 60–100 наблюдений. Из рисунка 2 можно сделать вывод о наличии в выборке одного выброса.

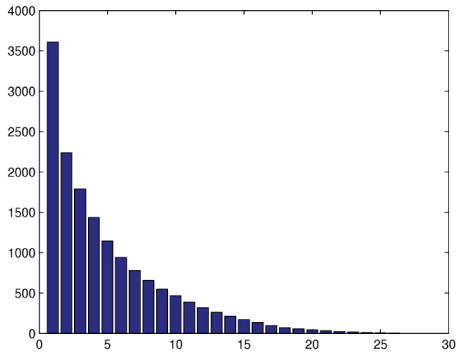

Рис. 2. Диаграмма изменения величины функционала в методе наименьших модулей

Заключение и выводы. Разработанный алгоритм поиска выбросов отличает универсальность подхода к регрессионным моделям, оцениваемым различными методами. Стоит заметить, что первичным приближением параметров при реализации алгоритма могут выступать оценки, полученные выбранным исследователем методом. В окрестности этой точки накладывается некоторая сетка и определяется следующий набор параметров, минимизирующих функцию. В окрестности полученной точки накладывается более мелкая сетка и т.д. Время процедуры заметно сокращается, если есть априорная информация о значении параметра.

Библиографический список

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

2. Мудров В.И., Кушко В.Л. Метод наименьших модулей. М., 1971.

3. Armstrong R.D., Kung D.S. Algorithm AS132: Least absolute value estimates for a simple linear regression problem // Appl. Stat. 1978. Vol. 7.

4. Weisberg S. Applied linear regression. 3rd ed. Jonh Wiley & Sans, Inc., 2005.

5. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия / пер. с англ. М., 1982. Вып. 1, 2.

6. Cook R.D. Detection of Influential Observation in Linear Regression // Technometrics. 1977. Vol. 19(1).

7. Andrews D.F., Pregibon D. Finding the outliers that matter // Journal of the Royal Statistical Society. 1978. Vol. 40.

8. 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(2021)1-17.

9. Arthur Zimek, Peter Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2018. Vol. 8. № 6. DOI: 10.1002/widm.1280.

10. Campello R.J.G.B., Moulavi D., Zimek A., Sander J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection // ACM Transactions on Knowledge Discovery from Data. 2015. Vol. 10. № 1. DOI: 10.1145/2733381.