УДК 519.237.5

Метод поиска экстремальных наблюдений в задаче нечеткой регрессии

И.В. Пономарев[1], *Т.В. Саженкова*[1], *В.В. Славский*[2]

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

A Method of Searching for Extreme Observations in a Problem of Fuzzy Regression

I.V. Ponomarev[1], *T.V. Sazhenkova*[1], *V.V. Slavsky*[2]

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

Изучение статистических данных на предмет выбросов является актуальной задачей современной математики. От надежности данных методов напрямую зависит качество последующей обработки массива данных и адекватность получаемых выводов. В общем случае данная задача предусматривает проверку всех имеющихся наблюдений и сопоставление с ними некого числового индикатора. Дальнейший вывод делается на основе сопоставления этих индикаторов между собой.

В данной работе рассматривается методика поиск выбросов для одной из возможных регрессионных моделей, основанной на чебышевской норме. В основу предлагаемого подхода положено одно из известных преобразований, использующееся в выпуклом анализе, — преобразование Лежандра. Основанный на этом преобразовании алгоритм позволяет относить к группе выбросов не отдельные наблюдения, а множество наблюдений. Это отличает данный метод от большинства использующихся алгоритмов. Также это позволяет решить поставленную задачу за один проход и сокращает время выполнения алгоритма. Приводится пример исследования выборки на предмет выбросов. Возможность сравнения получаемых характеристик дает возможность решать задачу для различного количества предполагаемых экстремальных значений.

Ключевые слова: нечеткая регрессия, статистические выбросы, преобразование Лежандра, выпуклый анализ.

The study of statistical data for outliers is an urgent task of modern mathematics. The reliability of these methods directly affects the quality of the subsequent processing of statistical data sets and the adequacy of the resulting conclusions. In general, all available observations should be checked and compared with a certain numerical indicator. The further conclusion should be made by comparing these indicators among themselves.

In this paper, a technique to search for statistical outliers for one of the possible regression models based on the Chebyshev norm is considered. The proposed approach is based on the Legendre transformation, one of the known transformations used in the convex analysis. This algorithm allows us to refer to the group of statistical outliers for a set of observations and not for individual observations. This key point distinguishes this algorithm from the most of the commonly used algorithms. This way, the task can be solved in one pass with less required time. An example of the study of a sample for outliers is presented. The possibility to compare the obtained characteristics provides the opportunity to solve the problem for a different number of assumed extreme values.

Key words: fuzzy regression, statistical outliers, Legendre transformation, convex analysis.

DOI 10.14258/izvasu(2018)4-18

1. Введение, постановка задачи. В настоящее время одним из самых распространенных методов изучения закономерностей по статистическим данным является регрессионное моделирование. В математике разработано большое количество методов построения регрессионных зависимостей [1–4]. Стоит отметить, что при работе со статистическими данными исследователю приходится сталкиваться с проблемой выбросов — наблюдений, которые находятся аномально далеко от основной группы данных. Наличие выбросов может негативно сказываться на результатах

моделирования, делая полученную модель непригодной для практического использования. В работах [5–7] предложены методы проверки данных на предмет выбросов для классической линейной регрессионной модели.

В данной работе (следуя [8, 9]) будем рассматривать нечеткую линейную регрессионную модель L_∞.

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

$$\Omega = \{(x_{i,1}, \ldots, x_{i,m-1}, y_i) : i = 1, \ldots, n\},$$

которое можно рассматривать как результат n экспериментов.

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

$$\alpha_\infty(\Omega) = 2 \cdot \min_{k_s, b} \left\{ \max_{i=1,\ldots,n} |y_i - \sum_{j=1}^{m-1} k_s x_{i,s} - b| \right\}. \quad (1)$$

С геометрической точки зрения величина $\alpha_\infty(\Omega)$ равна минимуму ширины «полосы» ограниченной двумя параллельными гиперплоскостями и содержащей множество Ω, ширина берется вдоль оси y в R^m (т.е. длина пересечения полосы с осью y).

Определение 2. Уравнение гиперплоскости на котором достигается (1) назовем уравнением L_∞:

$$y = \sum_{j=1}^{m-1} k_s^0 x_s - b^0, \quad (2)$$

или уравнением регрессии относительно чебышевской нормы.

Задачу о выбросах сформулируем следующим образом: требуется из имеющегося экспериментального множества данных Ω отбросить фиксированный процент данных (например 5%) так, чтобы оставшиеся данные Ω_0 имели наименьшую величину разброса $\alpha_\infty(\Omega_0)$, т.е.

$$\alpha_\infty(\Omega_0) = \min\{\alpha_\infty(\Omega') : \Omega' \subset \Omega, \#[\Omega'] = n_0\}, \quad (3)$$

где $\#[\Omega']$ — число элементов в множестве Ω', $n_0 < n$, $n - n_0 = w_0$ — число выбросов.

2. Преобразование Лежандра. Преобразование Лежандра применяется в самых различных разделах чистой и прикладной математики: выпуклый анализ, механика, вариационное исчисление, геометрия, уравнения математической физики. В работе [10] определяется и исследуется обобщенное преобразование Лежандра для произвольного конечного подмножества евклидова пространства.

Пусть дана пара натуральных чисел $1 \le r, s \le n$. Определим две функции:

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

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:

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

Введенные функции обладают следующими свойства функции $\underset{r}{MAX}$.

1. Если для любого i $\{a_i \ge b_i\}$, то $\underset{r}{MAX}\{a_i\} \ge \underset{r}{MAX}\{b_i\}$.
2. Если $\mu \ge 0$, то $\underset{r}{MAX}\{\mu a_i\} = \mu \underset{r}{MAX}\{a_i\}$.
3. Для любой перестановки σ индексов $\{1, 2, \ldots, n\}$ такой, что $\sigma(k) = i_k$, выполняется неравенство
$$\underset{r}{MAX}\{a_i\} \ge a_{i_r}.$$
4. $\underset{r}{MAX}\{a_i + b_i\} \le \underset{r}{MAX}\{a_i\} + \underset{r}{MAX}\{b_i\}$.

Справедливость первых трех свойств непосредственно следует из определения функции $\underset{r}{MAX}$.

Для проверки четвертого свойства заметим, что не ограничивая общности можно считать, что

$$a_1 + b_1 \ge a_2 + b_2 \ge \ldots \ge a_p + b_p \ge \ldots \ge a_n + b_n.$$

Тогда из 3-го свойства следует искомое неравенство

$$\underset{r}{MAX}\{a_i + b_i\} = \{a_r + b_r\} \le$$
$$\le \underset{r}{MAX}\{a_i\} + \underset{r}{MAX}\{b_i\}.$$

Аналогичные свойства справедливы для $\underset{s}{MIN}$ с заменой в свойствах 3 и 4 знака неравенства на противоположное.

Определение 2. Обобщенным преобразованием Лежандра множества Ω назовем пару функций:

$$f_r^+(k) = \underset{r}{MAX}\left\{ \sum_{s=1}^{m-1} x_{i,s} k_s - y_i : i = 1, \ldots, n \right\},$$
$$f_s^-(k) = \underset{s}{MIN}\left\{ \sum_{s=1}^{m-1} x_{i,s} k_s - y_i : i = 1, \ldots, n \right\},$$

где $k = (k_1, \ldots, k_s)$.

Опираясь на свойства функций $\underset{r}{MAX}$ и $\underset{r}{MIN}$ заметим, что:

- функции f_r^+, f_s^- — выпуклые вниз и вверх соответственно;
- разность $f_r^+(k) - f_s^-(k)$ — неотрицательная функция, выпуклая вниз.

3. Основной результат. Докажем применимость введенного обобщенного преобразования Лежандра к поиску и устранению выбросов в регрессионной модели L_∞.

Теорема. Справедливо равенство

$$\alpha_\infty(\Omega_0) = \min_k \min_{0\le r\le w_0} \left[f_r^+(k) - f_{w_0-r}^-(k)\right]. \quad (4)$$

Доказательство. Пусть минимальное значение в правой части (4) достигается при $k = k_0$ и $r = r_0$, и пусть множество точек Ω занумеровано так, чтобы

$$x_1k_0 - y_1 \ge x_2k_0 - y_2 \ge \cdots \ge x_nk_0 - y_n,$$

где $x_ik_0 = \sum\limits_{s=1}^{m-1} x_{i,s}k_s$.

Возьмем в качестве выбросов множество $\Omega\backslash\Omega'$, состоящее из $w_0 = n - n_0$ точек. Так как функция $f_r^+(k) - f_{w_0-r}^-(k)$ выпукла вниз и неотрицательна, то

$$\alpha_\infty(\Omega') = f_r^+(k_0) - f_{w_0-r}^-(k_0).$$

Следовательно,

$$\alpha_\infty(\Omega_0) \le \min_k \min_{0\le r\le w_0} \left[f_{r_0}^+(k) - f_{w_0-r_0}^-(k)\right].$$

Докажем обратное неравенство. Пусть минимум левой части будет, если взять в качестве выбросов множество точек $\Omega\backslash\Omega'$, и пусть оставшееся множество точек $\Omega' = \{A_{i_1}, A_{i_2}, \ldots, A_{n_0}\}$. Тогда

$$\alpha_\infty(\Omega') = \min_k f'^+(k) - f'^-(k),$$

где $\{f'^+, f'^-\}$ — преобразование Лежандра множества Ω'. Пусть минимум достигается при $k = k_0$, и пусть множество точек Ω занумеровано так, чтобы

$$x_1k_0 - y_1 \ge x_2k_0 - y_2 \ge \cdots \ge x_nk_0 - y_n.$$

Тогда номера точек $[i_1 \le i_2 \le \cdots \le i_{n_0}]$ множества Ω' образуют связный (без пропусков) подынтервал в интервале номеров $[1, 2, \ldots, n]$, так в противном случае величину

$$f'^+(k_0) - f'^-(k_0)$$

можно было бы уменьшить за счет другого выбора выбросов. Доказательство закончено.

На основе полученного равенства (4) была создана компьютерная программа, помогающая выделению из данного множества наблюдений «подозрительных» на выбросы.

4. Примеры вычислений. Для иллюстрации предлагаемого метода рассмотрим множество точек плоскости. Построим уравнение регрессии L_∞, которое задается уравнениями

$$y = 1,794 - 0,848x,$$
$$y = 0,933 - 0,848x$$

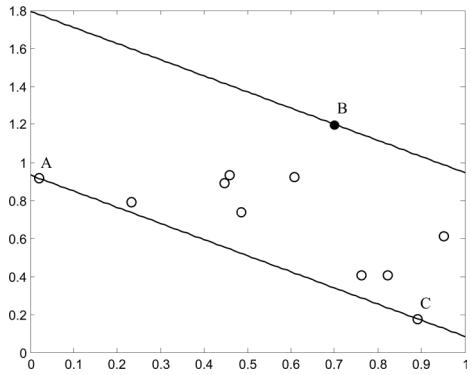

Рис. 1. График регрессии L_∞

и геометрически представлено на рисунке 1.

Точки A, B и C образуют экстремальный треугольник и, согласно теореме, являются «подозрительными» на выбросы. Визуально можно предположить, что точка B вносит больший эффект в вертикальную ширину данной полосы. Это подтверждается вычислением графика изменения функционала качества. На рисунке 2 первый столбец соответствует первоначальному набору данных, а следующие — данным после удаления одного наблюдения. Значит, самое существенное изменение происходит при удалении из данных точки B.

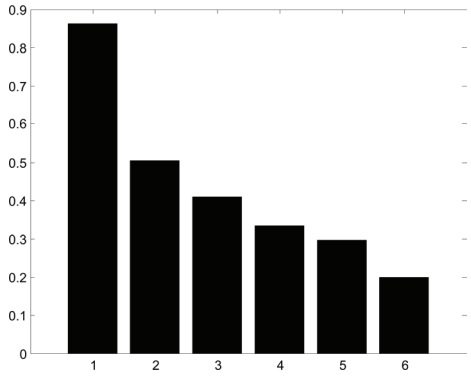

Рис. 2. Диаграмма изменения функционала α_∞

5. Заключение. Необходимо отметить, что указанный метод определяет только математическую характеристику наблюдений и никаким образом не иллюстрирует содержательную природу исследуемого наблюдения. Поэтому при проведении прикладных исследований возможно привлечение других методов, что будет способствовать разностороннему изучению проблемы и, как следствие, принятию более объективного решения.

Библиографический список

1. Tanaka H., Hayashi I., Watada J. Possibilistic Linear Regression Analysis with Fuzzy Model // European Journal of Operational Research. — 1989. — V. 40.

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

3. Gomez A.T., Sanchez, Jorge de Andres. Applications Of Fuzzy Regression In Actuarial Analysis // Journal of Risk & Insurance. — 2003. — V. 30.

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

5. Cook R.D. Detection of Influential Observation in Linear Regression // Technometrics. — 1977. — Vol. 19, № 1.

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

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

8. Пономарев И.В., Славский В.В. Нечеткая модель линейной регрессии // Доклады Академии наук. — 2009. — Т. 428, № 5.

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

10. Куркина М.В., Пономарев И.В. Система нечетких отношений равенств в банаховом пространстве // Дифференциальные уравнения. Функциональные пространства. Теория приближений. Международная конференция, посвященная 100-летию со дня рождения С. Л. Соболева (Новосибирск, 5–12 октября 2008 г.) : тезисы докладов. — Новосибирск, 2008.