ВЕСТНИК ЮГОРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2021 г. Выпуск 1(60). С. 30-35

DOI: 10.17816/byusu20210130-35
УДК 519.237.5

И. В. Пономарев

АЛГОРИТМ ОБНАРУЖЕНИЯ ВЫБРОСОВ В МОДЕЛИ РАВНОМЕРНО-НЕЧЕТКОЙ РЕГРЕССИИ

При построении математических моделей по статистическим данным перед исследователем возникает необходимость оценки однородности выборки, в частности, изучение данных на выбросы. Наличие в выборке выбросов негативно сказывается на результатах моделирования и адекватности модели в целом. В данной работе разработан алгоритм, позволяющий количественно измерить эффект влияния каждого наблюдения на качество построенной модели. Приводится описание данного алгоритма. Ранее автором проводились аналогичные исследования для различных регрессионных моделей.

Ключевые слова: модель нечеткой линейной регрессии; расстояние Кука; статистические выбросы.

I. V. Ponomarev

ALGORITHM FOR DETECTING OUTLIERS IN THE MODEL UNIFORMLY FUZZY REGRESSION

When constructing mathematical models based on statistical data, the researcher faces the need to assess the homogeneity of the sample, in particular, the study of data on emissions. Availability in a sample of outliers negatively affects the modeling results and the adequacy of the model as a whole. In this work, an algorithm has been developed that allows one to quantitatively measure the effect of the influence of each observation on the quality of the constructed model. The description of this algorithm is given. Previously the author carried out similar studies for various regression models.

Key words: fuzzy linear regression model; Cook's distance; statistical outliers.

1. Введение

Актуальным направлением развития регрессионного моделирования является применение теории нечетких множеств. В этом направлении можно выделить работы [1-6]. В данной работе за основу выбрана равномерно-нечеткая регрессионная модель [7].

Целью исследования является получение алгоритма позволяющего проверить исходную статистическую выборку на наличие выбросов. Подобным исследованиям посвящены работы [8-11]. Отличительной особенность данной методики является использование "двойной" оценки – при построении регрессионной модели и при вычислении характеристики наблюдения. В результате работы алгоритма каждому наблюдению ставится в соответствие числовая характеристика – расстояние Кука. Данная характеристика будет полезна исследователю при проведении экспертного анализа выборки.

Рассмотрим равномерно-нечеткую регрессионную модель

$$f(X) = a_0 + \sum_{j=1}^{k} a_j x_{ij} \tag{1}$$

которая равна моде нечеткого числа $A = f(X)$. В данной модели $f \in \Phi$ – нечеткая числовая функция; $a_0, a_1, \ldots, a_k$ являются параметрами модели, а x_{ij} – регрессорами.

Предполагая, что функция принадлежности будет иметь конкретный вид

$$\mu_A(y) = \varphi\left(\frac{|f(X) - y|}{\sigma}\right), \tag{2}$$

где $\varphi : [0, \infty) \to [0, 1]$ – фиксированная убывающая функция, $\varphi(0) = 1$; $\sigma > 0$ – параметр, определяем из условия нормировки достоверности модели.

Пусть имеется некоторая выборка $\Omega = \{(x_{i1}, \ldots, x_{ik}, y_i) : i = 1, \ldots, N\}$. Достоверность модели будет определяться величиной

$$\delta(f) = \min_{i=1,\ldots,N} \{\mu_{A_i}(y_i)\}.$$

В виду того, что спецификация модели является линейной функцией, задача нахождения наиболее достоверной модели сводится к нахождению

$$\alpha_\infty(X, y) = \min_{a_j, j=\overline{0,k}} \max_{i=1,\ldots,n} \left| a_0 + \sum_{j=1}^{k} a_j x_{ij} - y_i \right|. \tag{3}$$

Задача нахождения (3) можно сформулировать в виде задачи линейного программирования

$$\begin{cases} \min\limits_{u;v;a_j, j=\overline{0,k}} (u - v), \\ u \geq \sum\limits_{j=1}^{k} a_j x_{sj} - y_s, \quad s = 1, \ldots, N, \\ v \leq \sum\limits_{j=1}^{k} a_j x_{tj} - y_t, \quad t = 1, \ldots, N, \end{cases}$$

где u – верхняя огибающая, v – нижняя огибающая.

Графическая реализация алгоритма решения представлен на рисунке 1.

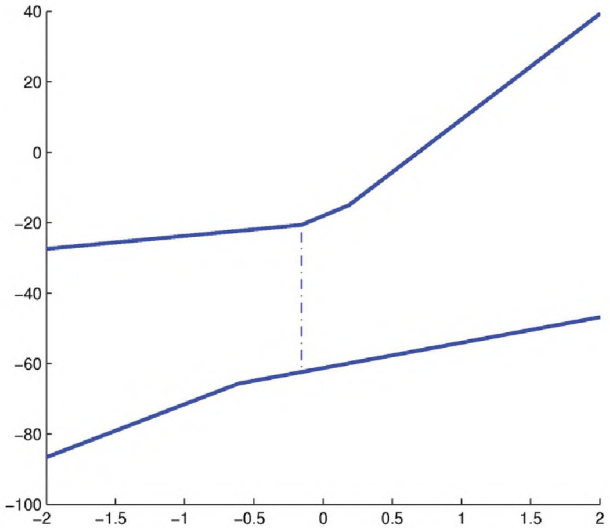

Рисунок 1. Минимальное значение разности двух огибающих в двумерном случае

2. Модификация расстояния Кука

В силу вероятностных предположений относительно регрессионной модели, все наблюдения имеют одинаковое значение, равнозначное влияние на результат моделирования. Поэтому в [9] предполагается, что удаление из выборки одного значения не должно в значительной мере изменять коэффициенты регрессии. Показателем изменения коэффициентов регрессии, влиянием наблюдения на результат будет служить расстояние Кука.

Определение 1. Пусть имеется регрессионная модель

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_k x_{ik} + \varepsilon_i,$$

где $\widehat{b}, \widehat{b}_{(i)}$ – оценки коэффициентов регрессии по исходным данным и после исключения i-го наблюдения.

Расстоянием Кука будем называть величину

$$CD(\widehat{Y}, \widehat{Y}_{(i)}) = \frac{(\widehat{a}_{(i)} - \widehat{a})^T (X^T X)(\widehat{a}_{(i)} - \widehat{a})}{(k+1)s^2},$$

где X – матрица регрессоров; k – количество регрессоров; s^2 – оценка дисперсии ошибок.

Предельным значением расстояния Кука считается значение статистики $F(\alpha, k+1, N-k-1)$.

Таким образом, имеем следующий алгоритм исследования исходных данных методом расстояния Кука:

1. Вычисляются оценки коэффициентов регрессии $\widehat{a}$ и дисперсия s^2.
2. Из набора наблюдений исключается i-ое наблюдение и находятся оценки $\widehat{a}_{(i)}$.
3. Определяется расстояние Кука $CD(\widehat{Y}, \widehat{Y}_{(i)})$ и сравнивается с $F(\alpha, k+1, N-k-1)$.
4. Если $CD(\widehat{Y}, \widehat{Y}_{(i)}) > F(\alpha, k+1, N-k-1)$, то делается заключение что i-ое наблюдение является выбросом.

Заметим, что расстояние Кука можно представить в виде

$$CD(\widehat{Y}, \widehat{Y}_{(i)}) = \frac{(X(\widehat{a}_{(i)} - \widehat{a}))^T (X(\widehat{a}_{(i)} - \widehat{a}))}{(k+1)s^2} = \frac{(\widehat{Y} - \widehat{Y}_{(i)})'(\widehat{Y} - \widehat{Y}_{(i)})}{(k+1)s^2}, \tag{4}$$

что позволяет рассматривать данную метрику как аналог обычного евклидова расстояния.

Определение 2. Евклидовым расстоянием между нечеткими числами $A = \{(z_i, \mu_A(z_i)), i = \overline{1,n}\}$ и $B = \{(z_i, \mu_B(z_i)), i = \overline{1,n}\}$ называется величина

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (\mu_A(z_i) - \mu_B(z_i))^2}. \tag{5}$$

Объединим метрики (4) и (5) и введем следующее определение.

Определение 3. Расстоянием Кука между нечеткими числами A и B называется величина

$$FDK(A,B) = \frac{\sqrt{\sum_{i=1}^{n} \left(\mu_A(z_i) - \mu_B(z_i)\right)^2}}{\frac{\sum_{i=1}^{n}(z_i-\overline{z})^2}{n-1}}, \tag{6}$$

где $\overline{z}$ – среднее значение.

3. Алгоритм исследования данных на выбросы

Рассмотрим модель (1) с треугольной функцией принадлежности, т.е. φ – линейная убывающая функция.

Проверка исходных данных будет заключаться в построении расстояния (6) между вектором теоретических $Y = \{y_1, \ldots, y_N\}$ и расчетных $\widehat{Y} = \{\widehat{y}_1, \ldots, \widehat{y}_N\}$ значений. Соответствующий алгоритм можно представить следующим образом.

1. Решая задачу (3), определяются оценки коэффициентов $\widehat{a}$ регрессии (1).
2. По формуле (2) вычисляются расчетные значения выходной переменной и соответствующие им функции принадлежности $\widehat{Y} = \{(\hat{y}_i, \mu_A(\hat{y}_i))\}$.
3. Из набора данных исключается j-ое наблюдение и повторяются шаги 1 и 2. Получаем значение $\widehat{Y}_{(j)} = \{(\hat{y}_i^*, \mu_A(\hat{y}_i^*))\}$, $i \neq j$.
4. Определяется расстояние Кука (6) $FDK(\widehat{Y}, \widehat{Y}_{(j)})$.
5. Шаги 3 и 4 повторяются для всех $j = \overline{1, N}$.

Для проверки работоспособности данного алгоритма был создан комплекс программ. Комплекс был запрограммирован в системе компьютерной математики MatLab и включает в себя три отдельные программы. Первая программа по заданному объему и размерности генерирует выборку одинаково распределенных случайных величин и искусственно "засоряет" ее небольшим (обычно 5% от объема исходной выборки) количеством дополнительных наблюдений. Эти наблюдения отличаются по распределению от основной выборки и играют роль выбросов.

Вторая программа строит модель равномерно-нечеткой регрессии. От пользователя требуется указать массивы входных и результирующей переменных. На выходе получаются два массива: коэффициентов и значений функции принадлежности для результирующей переменной $\mu_A(\hat{y}_i)$.

Третья программа производит пошаговое исключение из исходной выборки по одному наблюдению и вычисляет для оставшихся наблюдений новые значения функции принадлежности $\mu_A(\hat{y}_i^*)$ с использованием второй программы. На каждом таком шаге вычисляется расстояние Кука между полученными нечеткими числами. Результат записывается в массив и выводится графическая иллюстрация (диаграмма).

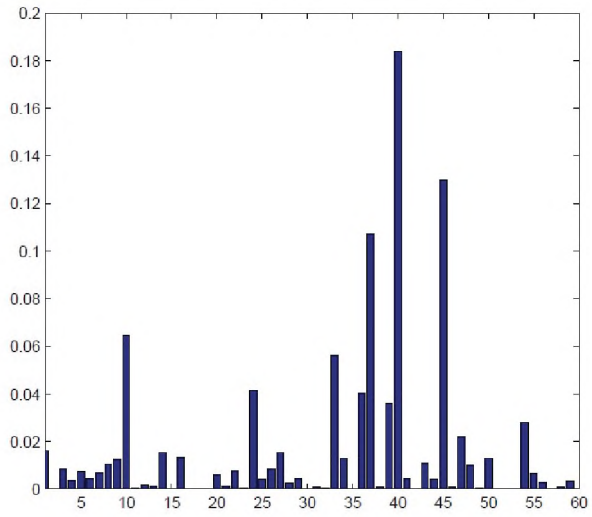

Рисунок 2. Диаграмма значений расстояний Кука для тестовой выборки

На рисунке 2 представлена диаграмма рассеяния расстояний Кука для одной из тестовых выборок. Легко заметить, что критическими могут быть признаны четыре наблюдения. Все эти наблюдения и были заранее введены в выборку.

С использованием разработанного программного комплекса был проведен ряд подобных испытаний с различными выборками. Результаты анализа показали, что данный метод верно определяет 3-4 выброса при объеме выборки 60-100 наблюдений. Уменьшение объема выборки влечет за собой увеличение разброса данных и тем самым осложняет процесс нахождения выбросов. При тестировании малых выборок представленный алгоритм верно определял 1-2 выброса. Таким образом, данный алгоритм пригоден для анализа результатов регрессионного моделирования. Полученный программный комплекс позволяет за разумное время обработать достаточное число наблюдений.

Заметим, что окончательный ответ на вопрос об отнесении наблюдения к выбросам дает непосредственно исследователь. Разработанный алгоритм и комплекс программ являются удобными инструментами для обнаружения "подозрительных" элементов и вычисляет соответствующую численную характеристику.

Литература

1. David, B. Alternativ Methods of Regression / B. David, D. Yadolah. - New York : Jonh Wiley & Sans, Inc., 1993. - 248 p.
2. Gomez, A. T. Applications Of Fuzzy Regression In Actuarial Analysis / A. T. Gomez, J. de A. Sanchez // Journal of Risk & Insurance. - 2003. - Vol. 30. - P. 665-699.
3. Tanaka, H. Linear regression analysis with fuzzy model / H. Tanaka, S. Uejima, K. Asai // IEEE Transactions on Systems, Man and Cybernetics. - 1982. - Vol. 12 (6). - P. 903-907.
4. Брюс, П. Практическая статистика для специалистов Data Science : перевод с английского / П. Брюс, Э. Брюс. - Санкт-Петербург : БХВ-Петербург, 2018. 304 с. - Текст : непосредственный.
5. Дрейпер, Н. Прикладной регрессионный анализ. Множественная регрессия = Applied Regression Analysis / Н. Дрейпер, Г. Смит. - 3-е издание. - Москва, 2007. - 369 с. - Текст : непосредственный.

6. Стрижов, В. В. Методы выбора регрессионных моделей / В. В. Стрижов, Е. А. Крымова. - Москва : ВЦ РАН, 2010. - 60 с. - Текст : непосредственный.
7. Пономарев, И. В. Нечеткая модель линейной регрессии / И. В. Пономарев, В. В. Славский. - Текст : непосредственный // Доклады Академии наук. - 2009. - Т. 428, № 5. - С. 598-600.
8. Andrews, D. F. Finding the outliers that matter / D. F. Andrews, D. Pregibon // Journal of the Royal Statistical Society. - 1978. - Vol. 40. - P. 85-93.
9. Cook, R. D. Detection of Influential Observation in Linear Regression / R. D. Cook // Technometrics. - 1977. - Vol. 42, № 1. - P. 15-18.
10. Weisberg, S. Applied linear regression / S. Weisberg. - 3rd editor. - New York : Jonh Wiley & Sans, Inc., 2005. - 260 p.
11. Пономарев, И. В. Метод поиска экстремальных наблюдений в задаче нечеткой регрессии / И. В. Пономарев, Т. В. Саженкова, В. В. Славский. - Текст : непосредственный // Известия Алтайского государственного университета. - 2018. - № 4 (102). - С. 98-101.