УДК 519.688

Метод нахождения выбросов в регрессионной модели L_1

И.В. Пономарев
АлтГУ, г. Барнаул

Результаты статистического моделирования напрямую зависят от качества исходных данных. Основными критериями статистической выборки являются такие характеристики как репрезентативность, точность, объем и т.п. Однако, даже при наличии всех вышесказанных условий возможно наличие в выборке наблюдений резко выделяющихся на фоне остальных – выбросов. Поиску выбросов и устранению их влиянию посвящено большое количество исследований [1-3].

Данная работа продолжает исследования [4-6] и ставит задачу отыскания выбросов в процессе построения регрессии L_1.

L_1 регрессией будем называть линейную регрессионную модель

$$y = a_0 + a_1 x_1 + \ldots + a_k x_k + \varepsilon, \tag{1}$$

где y – зависимая переменная; x_i $(i = \overline{1,k})$ – независимые переменные; ε – ошибка; a_i $(i = \overline{0,k})$ – параметры модели.

Функционал качества данной модели имеет вид

$$\alpha_1 = \min_{a_0,\ldots,a_k} \left| y - a_0 + a_1 x_1 + \ldots + a_k x_k \right| \tag{2}$$

Задачу о нахождении выбросов сформулируем следующим образом: пусть из данного множества наблюдений $\Omega = \left\{ (x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = \overline{1,N} \right\}$ требуется исключить фиксированный процент наблюдений так, чтобы оставшиеся Ω_0 данные имели наименьшую величину разброса $\alpha_p(\Omega_0)$, т.е.

$$\alpha_p(\Omega_0) = \min \left\{ \alpha_p(\Omega') : \Omega' \subset \Omega, \#[\Omega'] = N_0 \right\}, \tag{3}$$

где $\#[\Omega']$ – число элементов во множестве Ω'; $N - N_0 = M_0$ – число выбросов.

Для решения этой задачи для величины $\alpha_1(\Omega)$ можно воспользоваться преобразованием Лежандра. Аналогичное преобразование для нечеткой регрессионной модели было рассмотрено в работе [6]. Точ-

ками подозрительными на выбросы можно считать те, которые максимально увеличивают функционал $\alpha_1(\Omega)$. При заданных значениях параметров расположим величины квадратов отклонений по убыванию. Наибольший вклад в увеличение функционала будут вносить первые M_0 значений последовательности.

Функции преобразования Лежандра определим как

$$f_r^+(a_0,\ldots,a_k) = MAX_r\left\{\left|y_i - a_0 + a_1x_{i1} + \ldots + a_{ik}x_{ik}\right| : i = \overline{1,N}\right\}$$

$$f_s^-(a_0,\ldots,a_k) = MIN_s\left\{\left|y_i - a_0 + a_1x_{i1} + \ldots + a_{ik}x_{ik}\right| : i = \overline{1,N}\right\}$$

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

Теорема. Справедливо равенство:

$$\min\left\{\alpha_2(\Omega') : \Omega' \subset \Omega, \#[\Omega'] = N_0\right\} = \min_{a_0,\ldots,a_k} \sum_{0 \le r \le M_0 - 1} f_r^-(a_0,\ldots,a_k).$$

Доказательство.

Рассмотрим правую часть равенства. Пусть минимум достигается при $a_0 = a_0^*,\ldots,a_k = a_k^*$. Перенумеруем последовательность модулей остатков в порядке возрастания

$$\left|y_1 - a_0^* + a_1^*x_{11} + \ldots + a_k^*x_{1k}\right| \ge \left|y_2 - a_0^* + a_1^*x_{21} + \ldots + a_k^*x_{2k}\right| \ge \ldots$$

$$\ldots \ge \left|y_N - a_0^* + a_1^*x_{N1} + \ldots + a_k^*x_{Nk}\right|$$

Так как последовательность упорядочена и состоит из неотрицательных чисел, то и сумма членов последовательности, с номерами k, где $N - M_0 < k \le N$, будет минимальна. Это соответствует минимуму функционала $\alpha_1(\Omega')$.

Библиографический список

1. Weisberg S. Applied linear regression. – 3th ed. – Jonh Wiley & Sans, Inc., 2005.

2. Cook R.D. Detection of Influential Observation in Linear Regression // Technometrics. _ 1977. _ Vol. 19, No. 1. – P. 15–18.

3. Andrews D.F., Pregibîn D. Finding the outliers that matter // Journal of the Royal Statistical Society. – 1978. – Vol. 40. – P. 84–93.

4. Пономарев И.В. Исследование статистических данных на выбросы // Сборник трудов всероссийской конференции по математике.

МАК: Математики - Алтайскому краю. – Барнаул: Изд-во Алт. ун-та, 2017. – С. 133–135.

5. Пономарев И.В. Об одном методе проверки статистических данных на наличие выбросов // Сборник трудов всероссийской конференции по математике. МАК: Математики - Алтайскому краю. – Барнаул: Изд-во Алт. ун-та, 2018. – С. 203–205.

6. Пономарев И.В., Саженкова Т.В., Славский В.В. Метод поиска экстремальных наблюдений в задаче нечеткой регрессии // Известия Алтайского государственного университета. – 2018. - №4(102). – С. 98–101. – https://doi.org/10.14258/izvasu(2018)4-18.

УДК 519.87

Агентно-ориентированные имитационные модели для реальных городских процессов

С.П.Пронь[1], С.П.Семенов[2], А.О.Ташкин[2], Е.В.Токарева[1]
[1]АлтГУ, г. Барнаул ,[2]ЮГУ, г. Ханты-Мансийск

Рассмотрены агентно-ориентированные имитационные модели для реальных процессов, характерных для каждого города. В частности, представлены агентно-ориентированные имитационные модели ГИС для МГН geowheel.ru и платежеспособности и финансовой устойчивости фонда регионального оператора МКД в среде AnyLogic.

Имитационная модель – это формальное описание логической структуры и динамики взаимодействия отдельных элементов реального объекта с учётом стохастических факторов, реализованное как программа для компьютера [1-3].

Концептуальная модель ГИС для МГН geowheel.ru, представлена в статье [2]. В настоящем исследовании представлено продолжение работы по развитию гибридного подхода к построению имитационных моделей, включающего в себя характерные черты дискретно-событийного и агентно-ориентированного подходов [1]. Смысл объединения двух подходов заключается в том, что с помощью дискретно-событийного подхода возможно построить логическую структурно-функциональную схему модели, а с помощью агентно-ориентированного подхода отобразить динамику взаимодействия заявки и элементов системы. Для построения и исследования компьютерных моделей различных процессов существует большое количество программных пакетов. Для реализации гибридной модели было выбрано ПО AnyLogic, которое позволяет создавать агентно-ориентированные модели, отражающие общее представление о поведении моделируемой системы посредством