УДК 519.85

О.В. СЕРАЯ, Т.И. КАТКОВА

ОЦЕНИВАНИЕ ПАРАМЕТРОВ УРАВНЕНИЯ РЕГРЕССИИ ПО НЕЧЕТКИМ ИСХОДНЫМ ДАННЫМ

Анотація. *Розглянуто технологію оцінювання параметрів рівняння регресії для випадку, коли вихідні дані – нечіткі числа з відомими функціями приналежності. Запропоновано метод розрахунку чітких значень шуканих оцінок, заснований на відшуканні чіткого рішення нечіткої системи лінійних алгебраїчних рівнянь.*
Ключові слова: *рівняння регресії, оцінювання параметрів, поліном, нечіткі числа, функція приналежності, система лінійних алгебраїчних рівнянь, критерій, оптимізація.*

Аннотация. *Рассмотрена технология оценивания параметров уравнения регрессии для случая, когда исходные данные – нечеткие числа с известными функциями принадлежности. Предложен метод расчета четких значений искомых оценок, основанный на отыскании четкого решения нечеткой системы линейных алгебраических уравнений.*
Ключевые слова: *уравнение регрессии, оценивание параметров, полином, нечеткие числа, функция принадлежности,система линейных алгебраических уравнений, критерий, оптимизация.*

Abstract. *A technology of the regress equation parameters estimation where initial data represents indistinct numbers with known accessory functions is considered. A method of calculation of accurate values of the required estimates, based on the search of accurate solution for indistinct system of linear algebraic equations, is offered.*
Key words: *the regress equation, estimation of parametres, polynom, indistinct numbers, accessory function, system of the linear algebraic equations, criterion, optimisation.*

1. Введение

Разнообразные технологии описания поведения технических, экономических, социальных и других систем, а также проблемы оценки их эффективности сводятся к однотипной математической задаче: найти аналитическое соотношение, связывающее численные значения наборов факторов, определяющих условия и режим функционирования системы, со значением некоторым образом выбранного результирующего показателя этой системы. По многим причинам такое соотношение, обычно называемое функцией отклика, удобно выбрать в форме так называемого полинома Колмогорова-Габора [1]:

$$y(X) = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n + a_{12} x_1 x_2 + \ldots + a_{n-1,n} x_{n-1} x_n , \qquad (1)$$

где x_j – значение j-го фактора, $j = 1,2,...,n$;

y – результирующий показатель.

Здесь максимальная учитываемая степень взаимодействия факторов равна двум.

Если для оценивания параметров полинома (1) используются результаты N экспериментов, то наилучший в смысле наименьших квадратов вектор $A^T = \begin{pmatrix} a_0 & a_1 & a_2 & \ldots & a_n & a_{12} & \ldots & a_{n-1,n} \end{pmatrix}$ определяется по формуле

$$A = \left(H^T H\right)^{-1} H^T Y , \quad H = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1n} & x_{11}x_{12} & \ldots & x_{1,n-1}x_{1n} \\ 1 & x_{21} & x_{22} & \ldots & x_{2n} & x_{21}x_{22} & \ldots & x_{2,n-1}x_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{Nn} & x_{N1}x_{N2} & \ldots & x_{N,n-1}x_{Nn} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{pmatrix}. \qquad (2)$$

ISSN 1028-9763. Математичні машини і системи, 2011, № 1

Здесь x_{ij} – значение j-го фактора в i-м эксперименте;

y_i – значение результирующего показателя в i-м эксперименте.

Этот стандартный подход усложняется, если значения результирующего показателя в каждом опыте оцениваются нечетко [2]. При этом, естественно, и оценки параметров уравнения регрессии будут нечеткими числами. Пусть заданы функции принадлежности $\mu_i(y_i)$, $i=1,2,...,N$ результатов измерений.

Введем матрицу $R=(r_{pi})=(H^TH)^{-1}H^T$, $\dim R = K\times N$, $K=1+n+n(n-1)/2$. Тогда, в соответствии с (2),

$$A=RY=\left(\sum_{i=1}^{N} r_{pi}y_i\right),\ p=1,2,...,K.$$

Теперь, используя правила выполнения операций над нечеткими числами [3,4], легко получить функции принадлежности компонентов вектора A. Пусть, например,

$$\mu_i(y_i)=\exp\left\{-\frac{(y_i-\bar{y}_i)}{2\sigma_i^2}\right\},\ i=1,2,...,N.$$

Тогда

$$\mu_p(a_p)=\exp\left\{-\frac{(a_p-\bar{a}_p)}{2D_p}\right\},\ p=1,2,...,K,\qquad \bar{a}_p=\sum_{i=1}^{N} r_{pi}\bar{y}_i,\ D_p=\sum_{i=1}^{N}\sigma_i^2 r_{pi}^2.$$

Гораздо более сложной становится задача, если не только результаты, но и условия проведения экспериментов, то есть значения факторов в каждом опыте, также нечеткие числа. Поставим задачу оценивания параметров уравнения регрессии (1) в этом случае более полной неопределенности.

2. Постановка задачи

Введем функции принадлежности $\mu_{ij}(x_{ij})$ значений факторов в каждом из опытов: $i=1,2,...,N$, $j=1,2,...,n$. При этом будем считать, что уровень неопределенности значений для каждого из факторов определяется характером этого фактора и механизмом оценивания его значений. Поэтому функции принадлежности нечетких значений одного и того же фактора в разных экспериментах отличаются только модами. Это же допущение примем и в отношении функций принадлежности значений результирующего показателя. С учетом (1) результатам N проведенных экспериментов соответствуют соотношения

$$\begin{aligned} &a_0+a_1x_{11}+a_2x_{12}+\ldots+a_nx_{1n}+a_{12}x_{11}x_{12}+\ldots+a_{n-1,n}x_{1,n-1}x_{1,n}=y_1,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &a_0+a_1x_{N1}+a_2x_{N2}+\ldots+a_nx_{Nn}+a_{12}x_{N1}x_{N2}+\ldots+a_{n-1,n}x_{N,n-1}x_{N,n}=y_N. \end{aligned} \qquad (3)$$

В этих соотношениях слева и справа находятся нечеткие числа и их равенство понимается в смысле равенства их функций принадлежности. Таким образом, задача оценивания параметров уравнения регрессии в случае, когда значения факторов и результаты экспериментов определены нечетко, сведена к отысканию наилучшего, в каком-либо естественном смысле, решения системы уравнений (3) с нечеткими параметрами. Рассмотрим возможный метод решения этой задачи.

3. Основные результаты

Пусть нечеткие значения x_{ij} и y_{ij} системы (3) имеют соответствующие функции принадлежности:

$$\mu_{ij}(x_{ij}) = \exp\left\{-\frac{(x_{ij}-\bar{x}_{ij})}{2\sigma_j^2}\right\},\ i=1,2,...,N,\ j=1,2,...,n,\ \mu_i(y_i) = \exp\left\{-\frac{(y_i-\bar{y}_i)}{2\sigma_y^2}\right\}. \quad (4)$$

Введем нечеткие числа

$$z_i = a_0 + a_1x_{i1} + a_2x_{i2} + \ldots + a_nx_{in} + a_{12}x_{i1}x_{i2} + \ldots + a_{n-1,n}x_{i,n-1}x_{i,n} - y_i \quad (5)$$

и запишем их функции принадлежности:

$$\mu(z_i) = \mu\left(a_0 + \sum_{j=1}^{n} a_jx_{ij} + \sum_{j_1=1}^{n-1}\sum_{j_2>j_1} a_{j_1j_2}x_{ij_1}x_{ij_2} - y_i\right) = \exp\left\{-\frac{(z_i-\bar{z}_i)^2}{2D(z_i)}\right\},$$

$$\bar{z}_i = a_0 + \sum_{j=1}^{n} a_j\bar{x}_{ij} + \sum_{j_1=1}^{n-1}\sum_{j_2>j_1} a_{j_1j_2}\bar{x}_{ij_1}\bar{x}_{ij_2} - \bar{y}_i,\ D(z_i) = \sum_{j=1}^{n} a_j^2\sigma_j^2 + \sum_{j_1=1}^{n-1}\sum_{j_2>j_1} a_{j_1j_2}^2\sigma_{ij_1}^2\sigma_{ij_2}^2 + \sigma_y^2,$$

$$i = 1,2,...,N.$$

Теперь решим четкую систему линейных алгебраических уравнений, порождаемую системой (3) в случае, если нечеткие числа x_{ij} заменить их модальными значениями. Так как в традиционной постановке задачи оценивания параметров уравнения регрессии число экспериментов превышает число оцениваемых параметров, то получаемая система переопределена. Решение таких систем отыскивается методом наименьших квадратов. При этом вектор A параметров уравнения регрессии определяется соотношением

$$\bar{A} = \left(\bar{H}^T\bar{H}\right)^{-1}\bar{H}^T\bar{Y}, \quad (6)$$

где матрица $\bar{H}$ по структуре совпадает с матрицей H, в которой нечеткие числа x_{ij} заменены их модальными значениями $\bar{x}_{ij}$, а $\bar{Y}^T = (\bar{y}_1 \quad \bar{y}_2 \quad \ldots \quad \bar{y}_N)$.

Рассмотрим общий подход к выбору четкого решения исходной нечеткой задачи. К этому решению естественно предъявить следующие требования. Во-первых, оно не должно слишком сильно отличаться от модального решения $\bar{A}$, получаемого при замене нечетких параметров задачи их модальными значениями. Во-вторых, функции принадлежности нечетких чисел z_i, вычисляемые при подстановке искомого решения в (5), должны быть как можно менее размытыми. При этом возможный вариант построения критериальной функции приводит к минимизации:

$$\Phi_1(A) = \sum_{j=1}^{N}\int_{-\infty}^{\infty}\mu(z_i)dz_i \,| + (A-\bar{A})^T(A-\bar{A}).$$

При этом, с учетом (6),

$$\int_{-\infty}^{\infty}\mu(z_i)dz_i = \int_{-\infty}^{\infty}\exp\left\{-\frac{(z_i-\bar{z}_i)^2}{2D(z_i)}\right\}dz_i = \sqrt{2\pi D(z_i)},$$

$$\Phi_1(A) = \sqrt{2\pi}\sum_{j=1}^{N}\left(\sum_{j=1}^{n} a_j^2\sigma_j^2 + \sum_{j_1=1}^{n-1}\sum_{j_2>j_1} a_{j_1j_2}^2\sigma_{j_1}^2\sigma_{j_2}^2 + \sigma_y^2\right)^{0.5} + \left[(A-\bar{A})^T(A-\bar{A})\right]^{0.5}. \quad (7)$$

Смысл этого критерия понятен. Первая группа слагаемых характеризует уровень компактности функций принадлежности нечетких чисел $z_1, z_2, ..., z_N$, соответствующих решению, а последнее – степень близости получаемого решения к модальному.

Второй вариант построения критерия реализует чебышевское, минимаксное приближение в искомом «идеальном»решении. При этом

$$\Phi_2(A) = \sqrt{2\pi}\max_i\left(\sum_{j=1}^{n} a_j^2\sigma_j^2 + \sum_{j_1=1}^{n-1}\sum_{j_2>j_1} a_{j_1j_2}^2\sigma_{j_1}^2\sigma_{j_2}^2 + \sigma_y^2\right)^{0.5} + \left[\left(A-\overline{A}\right)^T\left(A-\overline{A}\right)\right]^{0.5}. \quad (8)$$

Аналитическое выражение критериев (7) и (8) можно несколько упростить, введя одноиндексную нумерацию слагаемых в соотношении (1). С этой целью предварительно перепишем его следующим образом:

$$y(X) = a_{00}x_0x_0 + a_{01}x_0x_1 + a_{02}x_0x_2 + \ldots + a_{0n}x_0x_n + a_{12}x_1x_2 + \ldots + a_{n-1,n}x_{n-1}x_n =$$

$$= \sum_{j_1=0}^{n-1}\sum_{j_2=0}^{n} a_{j_1j_2}x_{j_1}x_{j_2}\ ,\ x_0 \equiv 1. \quad (9)$$

Введем теперь индекс $p = 0,1,2,...,K$, определяющий номер слагаемого в (9), через значения j_1 и j_2 следующим образом:

$$p = \begin{cases} 0, & j_1 = 0, \quad j_2 = 0, \\ j_2, & j_1 = 0, \quad j_2 = 1,2,...,n, \\ \sum_{s=0}^{j_1-1}(n-s) + (j_1 - j_2), & j_1 = 1,2,...,n-1, \quad j_2 = j_1+1, j_1+2,...,n. \end{cases} \quad (10)$$

Кроме того, зададим наборы

$$u_p = \begin{cases} x_0x_0, & j_1 = 0, \quad j_2 = 0, \\ x_0x_{j_2}, & j_1 = 0, \quad j_2 = 1,2,...,n, \\ x_{j_1}x_{j_2}, & j_1 = 1,2,...,n-1, \quad j_2 = j_1+1, j_1+2,...,n, \end{cases}$$

$$b_p = \begin{cases} 0, & j_1 = 0, \quad j_2 = 0, \\ \sigma_p^2, & j_1 = 0, \quad j_2 = 1,2,...,n, \\ \sigma_p^4, & j_1 = 1,2,...,n-1, \quad j_2 = j_1+1, j_1+2,...,n. \end{cases} \quad (11)$$

Теперь, с учетом (10), (11), запишем выражения для уравнения регрессии (9) и критериев (7) и (8):

$$y(x) = \sum_{p=0}^{K} a_p u_p$$

$$\Phi_1(A) = \sqrt{2\pi}\sum_{j=1}^{N}\left(\sum_{p=0}^{K} a_p^2 b_p + \sigma_y^2\right)^{0.5} + \left(\sum_{p=0}^{K}\left(a_p - \overline{a}_p\right)^2\right)^{0.5}, \quad (12)$$

$$\Phi_2(A) = \sqrt{2\pi}\max_i\left(\sum_{p=0}^{K} a_p^2 b_p + \sigma_y^2\right)^{0.5} + \left(\sum_{p=0}^{K}\left(a_p - \overline{a}_p\right)^2\right)^{0.5}. \quad (13)$$

Искомый вектор A в обоих случаях отыскивается с использованием любого прямого метода численной минимизации (12) или (13).

Таким образом, предложенный метод сводит исходную задачу оценивания параметров уравнения регрессии в условиях нечетких исходных данных к обычной четкой задаче математического программирования. При этом понятно, что результат решения данной задачи – четкий набор параметров уравнения регрессии зависит от того, как выбран критерий качества четкого решения. Неоднозначность выбора делает целесообразным рассмотрение иного подхода к этой задаче, позволяющего получить ортодоксальное нечеткое ее решение. С формальных позиций, технология решения состоит в следующем. Сначала искомые значения параметров уравнения регрессии (1) необходимо выразить через значения функции отклика y_i и факторов x_{ij} в соответствующих экспериментах, то есть получить соотношения

$$a_p = f_p\left((y_i),(x_{ij})\right),\ i=1,2,...,N\ ,\ \ j=1,2,...,n\ ,\ \ p=1,2,...,K\ . \qquad (14)$$

Далее с использованием правил выполнения операций над нечеткими числами для заданных функций принадлежности $\mu_i(y_i)$, $\mu_{ij}(x_{ij})$ нечетких чисел $(y_i),(x_{ij})$ непосредственно отыскиваются функции принадлежности параметров a_p. К сожалению, реализация этой технологии ввиду нелинейности (14) для задач практической размерности затруднена. Приближенное решение задачи может быть получено следующим образом.

Используем рассчитываемый в соответствии с (6) модальный набор $\overline{A}$ параметров уравнения регрессии. Построим теперь многошаговую процедуру, на каждом шаге которой будем считать, что только одна какая-либо из компонентов задачи является нечеткой. Значения остальных компонентов положим равными модальным. Для ясности изложения вернемся к двухиндексной нумерации переменных. Пусть нечетким является конкретный, например, j_0-й фактор. Запишем систему уравнений (3), выделив элементы, содержащие неопределенность:

$$\begin{aligned}
&a_0 + a_1\overline{x}_{11} + a_2\overline{x}_{12} + \ldots + a_{j_0}\overline{x}_{1j_0} + \ldots + a_n\overline{x}_{1n} + a_{12}\overline{x}_{11}\overline{x}_{12} + \ldots + \\
&+ a_{j_0-1,j_0}\overline{x}_{1,j_0-1}\overline{x}_{1,j_0} + a_{j_0,j_0+1}\overline{x}_{1,j_0}\overline{x}_{1,j_0+1} + \ldots + a_{n-1,n}\overline{x}_{1,n-1}\overline{x}_{1,n} = \overline{y}_1\,, \\
&a_0 + a_1\overline{x}_{21} + a_2\overline{x}_{22} + \ldots + a_{j_0}\overline{x}_{2j_0} + \ldots + a_n\overline{x}_{2n} + a_{12}\overline{x}_{21}\overline{x}_{22} + \ldots + \\
&+ a_{j_0-1,j_0}\overline{x}_{2,j_0-1}\overline{x}_{2,j_0} + a_{j_0,j_0+1}\overline{x}_{2,j_0}\overline{x}_{2,j_0+1} + \ldots + a_{n-1,n}\overline{x}_{2,n-1}\overline{x}_{2,n} = \overline{y}_2\,, \qquad (15) \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&a_0 + a_1\overline{x}_{N1} + a_2\overline{x}_{N2} + \ldots + a_{j_0}\overline{x}_{Nj_0} + \ldots + a_n\overline{x}_{Nn} + a_{12}\overline{x}_{N1}\overline{x}_{N2} + \ldots + \\
&+ a_{j_0-1,j_0}\overline{x}_{N,j_0-1}\overline{x}_{N,j_0} + a_{j_0,j_0+1}\overline{x}_{N,j_0}\overline{x}_{N,j_0+1} + \ldots + a_{n-1,n}\overline{x}_{N,n-1}\overline{x}_{N,n} = \overline{y}_N\ .
\end{aligned}$$

Используем эту систему для последовательного определения значений параметров уравнения регрессии. При этом для расчета параметра a_0 решим независимо N уравнений системы (15), считая остальные параметры a_p, $p=1,2,...,K$ равными модальным. Получаемое при решении каждого из этих уравнений значение a_0 является нечетким. Его функция принадлежности по результатам решения, например, i-го уравнения, имеет вид

$$\mu_i\left(a_0(j_0)\right) = \exp\left\{-\frac{\left(a_0 - \overline{a}_0^{(i)}(j_0)\right)}{2\sigma^2_{a_0(j_0)}}\right\},$$

$$\bar{a}_0^{(i)}(j_0) = \bar{y}_i - \sum_{j\neq 0} \bar{a}_j \bar{x}_{ij} - \sum_{j_1\neq 0} \sum_{j_2=j_1+1} \bar{a}_{j_1 j_2} \bar{x}_{ij_1} \bar{x}_{ij_2},$$

$$\sigma^2_{a_0(j_0)} = \left(\bar{a}^2_{j_0} + \bar{a}_{j_{0-1}} \bar{a}_{j_0} + \bar{a}_{j_{0-1}} \bar{a}_{j_{0+1}}\right)\sigma^2_{j_0}.$$

Полученные N функций принадлежности для параметра a_0 комплексируются, формируя при этом условную функцию принадлежности параметра a_0, соответствующую неопределенности фактора j_0:

$$\mu_i(a_0(j_0)) = \exp\left\{-\frac{\left(a_0 - \bar{a}_0(j_0)\right)}{2\sigma^2_{a_0(j_0)}}\right\},$$

$$\bar{a}_0(j_0) = \frac{1}{N}\sum_{i=1}^{N} a_0^{(i)}(j_0).$$

Аналогично рассчитываются условные функции принадлежности для остальных параметров уравнения регрессии.

Теперь с использованием полученных условных функций принадлежности для каждого из параметров уравнения регрессии сформируем их безусловные функции принадлежности. При этом для произвольного параметра a_p получим

$$\mu_i(a_p) = \exp\left\{-\frac{\left(a_p - \bar{a}_p\right)}{2\sigma^2_p}\right\},$$

$$\bar{a}_p = \frac{\sum_{j_0=1}^{n} \frac{\bar{a}_p^{(i)}(j_0)}{\sigma^2_p(j_0)}}{\sum_{j_0=1}^{n} \frac{1}{\sigma^2_p(j_0)}},\ \sigma^2_p = \frac{1}{n}\sum_{j_0=1}^{n} \sigma^2_p(j_0),\ p = 1,2,...,K.$$

4. Выводы

Таким образом, в статье предложены методы оценивания параметров уравнения регрессии для случая, когда условия проведения опытов, используемых для идентификации регрессии, а также их результаты – нечеткие числа. Описанные подходы позволяют получить четкий и нечеткий наборы искомых регрессионных коэффициентов путем оптимизации критериев, имеющих ясный, естественным образом трактуемый смысл.

СПИСОК ЛИТЕРАТУРЫ

1. Рао С.Р. Линейные статистические методы и их применение / Рао С.Р.; пер. с англ. – М.: Наука, 1968. – 547 с.
2. Серая О.В. Оценивание состояния с использованием нечеткой регрессии / О.В. Серая, Т.И. Каткова, Л.В. Бачкир // Вісник НТУ «КПІ». – Київ: ВЕК+, 2008. – № 49. – С. 140 – 145.
3. Дюбуа Д. Теория возможностей. Приложение к представлению знаний в информатике / Д. Дюбуа, А. Прад; пер. с франц. – М.: Радио и связь, 1990. – 286 с.
4. Раскин Л.Г. Нечеткая математика. Основы теории. Приложения / Л.Г. Раскин, О.В. Серая. – Х.: Парус, 2008. – 353 с.

Стаття надійшла до редакції 26.07.2010