УДК 519.85

НЕЧЕТКАЯ ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ И КОРРЕЛЯЦИЯ

Сапкина Наталья Владимировна,
аспирант кафедры вычислительной математики и прикладных информационных технологий Воронежского государственного университета; natashasapkina@yandex.ru

В данной статье предложено применение метода наименьших квадратов для оценки неизвестных коэффициентов нечеткой-парной линейной регрессионной модели для нечетких чисел L-R-типа. Найдена формула для вычисления коэффициента корреляции между четкой объясняющей переменной и нечетким выходом. В заключении полученные теоретические результаты использованы для анализа информации.

Ключевые слова: коэффициент корреляции, метод наименьших квадратов, нечеткая парная линейная регрессионная модель, нечеткие числа L-R-типа, операции с нечеткими числами, расстояние между двумя нечеткими числами.

Введение

Мы часто используем регрессионный анализ для моделирования взаимодействий между зависимыми (переменные отклика) и независимыми (объясняющие переменные) величинами. В традиционном регрессионном анализе остатки считаются следствием случайных ошибок. Таким образом, для получения оценки и результатов в регрессионном анализе применяются статистические методы. Однако остатки возникают иногда ввиду неопределенности структуры модели или по причине неточности данных наблюдения. Неопределенность в данном случае становится нечеткостью, а не случайностью. С тех пор, как Л.А. Заде (1965) предложил теорию нечетких множеств, понятие нечеткости получило большее внимание, а анализ нечеткой информации стал особенно важен.

Для того чтобы применить нечеткость к регрессионному анализу, Танака и др. (1982) первыми предложили провести исследование нечеткой линейной регрессионной модели. Они рассматривали параметр оценки таких моделей по двум факторам: степень точности подбора модели и ее неопределенности. Проблемы оценки были тем самым сведены к линейному программированию, основанному на этих двух факторах. Диамонд (1988) разработал нечеткий анализ методом наименьших квадратов как нечеткое расширение обычного метода наименьших квадратов путем введения нового понятия расстояния на множестве нечетких чисел. В общем случае все нечеткие регрессионные методы могут быть грубо разделены на две группы: первая базируется на методе линейного программирования Танака, вторая – на методе наименьших квадратов Диамонда.

В данной статье представим нечеткую парную линейную регрессионную модель для нечетких чисел L-R-типа и найдем оценки параметров этой модели на основе метода наименьших квадратов, получим формулу для вычисления коэффициента корреляции в случае четкой объясняющей переменной и нечеткого выхода. Теоретические результаты используем для анализа данных.

Понятие нечеткого числа

Для описания объектов и явлений в условиях неопределенности широко используется понятие нечеткой переменной.

Нечеткая переменная задается тройкой $\langle \alpha, X, A \rangle$, где α – название переменной, X – область определения переменной α, A – нечеткое множество на X с функцией принадлежности $m_a(x)$, описывающее ограничения на значение нечеткой переменной α [3].

Нечеткая величина – это нечеткая переменная, определенная на множестве действительных чисел $\Re$. Нечеткую величину и соответствующее ей нечеткое множество обозначают одной и той же переменной. Множество нечетких величин на $\Re$ обозначают через $F(\Re)$ [2, 3].

Среди нечетких величин принято выделять два типа – нечеткое число и нечеткий интервал. Следует заметить, что нечеткое число в общем случае является частным случаем нечеткого интервала, что полностью согласуется с обычными числами и интервалами на множестве действительных чисел. Нечеткий интервал – форма представления неточных величин, более богатая информацией, чем обычный интервал [3].

Нечетким числом A называется нечеткая величина A, функция принадлежности $m_a(x)$ которой является выпуклой и унимодальной на $\Re$. Нечеткое число с модальным значением μ можно рассматривать как нечеткое значение высказывания «x приблизительно равно μ» [3, 4].

Нечеткие числа L-R-типа и операции над ними

В общем виде нечеткие числа и интервалы задаются с помощью, так называемых L-R-функций, обладающих одними и теми же свойствами.

L-функция есть отображение $L : \Re^+ \to [0,1]$, удовлетворяющее следующим условиям [3, 4]:

1. $L(x)$ – непрерывная, невозрастающая функция;
2. $L(x)$ – четная функция, т.е. $L(-x) = L(x)$;
3. $L(0)=1$ и $\forall x > 0$: $L(x) < 1$,

а также выполняется одно из двух условий:

$\forall x < 1 : L(x) > 0$ и $L(1)=0$;

$\forall x > 0 : L(x) > 0$ и $\lim\limits_{x\to\infty} L(x) = 0$.

Пусть $L(x)$ и $R(x)$– функции, удовлетворяющие перечисленным выше условиям. A есть унимодельное нечеткое число L-R-типа, если существуют константы $\alpha, \beta >0$ такие, что функция принадлежности нечеткого числа A имеет вид [4, 6]:

$$m_a(x) = \begin{cases} L\left(\dfrac{\mu - x}{\alpha}\right), \text{если } x \le \mu, \\ R\left(\dfrac{x-\mu}{\beta}\right), \text{если } x \ge \mu. \end{cases} \quad (1)$$

где α и β – соответственно левый и правый коэффициенты нечеткости, μ – модальное значение нечеткого числа. Условно нечеткое число A обозначается в виде тройки параметров (μ, α, β). Если $\alpha=\beta$, то нечеткое число A называется *симметричным нечетким числом L-R-типа* и обозначается как $A=(\mu, \alpha)$ [6].

Приведем некоторые определения, основываясь на работах [1, 4, 6].

Определение 1. Пусть число A задано с помощью L-R-представления $A=(\mu_a, \alpha_a, \beta_a)$. Тогда произведение нечеткого числа A на некоторую константу $\lambda \in \Re$ имеет вид:

$$\lambda A = (\lambda\mu_a, \lambda\alpha_a, \lambda\beta_a),\ \lambda > 0 \quad (2)$$

$$\lambda A = (\lambda\mu_a, -\lambda\beta_a, -\lambda\alpha_a),\ \lambda < 0. \quad (3)$$

Определение 2 (формула расстояния Евклида). Пусть $A=(\mu_a, \alpha_a, \beta_a)$ и $B=(\mu_b, \alpha_b, \beta_b)$ являются нечеткими числами L-R-типа. Тогда расстояние между ними определяется по формуле:

$$D = \sqrt{(\mu_a - \mu_b)^2 + (\alpha_a - \alpha_b)^2 + (\beta_a - \beta_b)^2}. \quad (4)$$

Определение 3. Суммой нечетких чисел A и B называется нечеткое число вида

$$A + B = (\mu_a + \mu_b, \alpha_a + \alpha_b, \beta_a + \beta_b). \quad (5)$$

Оценка параметров нечеткой парной линейной регрессионной модели

Рассмотрим метод оценки неизвестных параметров нечеткой линейной парной регрессионной модели. Положим, что исходные данные состоят из выборки $\{(x_i, Y_i)\}_{i=\overline{1,n}}$, где $x_i \in \Re$, а $Y_i = (\mu_{yi}, \alpha_{yi}, \beta_{yi})$ – нечеткие числа L-R-типа.

$$Y^{мод} = B_0 + B_1 x, \quad (6)$$

где x – независимая (объясняющая) переменная (фактор), $x \in \Re$; Y – зависимая (объясняемая) переменная, $Y = (\mu_y, \alpha_y, \beta_y)$ – нечеткое число L-R-типа; B_0, B_1 – неизвестные коэффициенты регрессии (параметры уравнения), $B_0 = (\mu_{b0}, \alpha_{b0}, \beta_{b0})$ и $B_1 = (\mu_{b1}, \alpha_{b1}, \beta_{b1})$ – нечеткие числа L-R-типа.

Пусть n раз измерены значения фактора x и соответствующие значения Y. Тогда получим следующую линейную парную регрессионную модель с нечеткими коэффициентами (теоретическое линейное уравнение регрессии):

$$Y_i = B_0 + B_1 x_i + E_i,\ i = \overline{1,n}, \qquad (7)$$

где $x_i \in \Re$, $Y_i = (\mu_{yi}, \alpha_{yi}, \beta_{yi})$ – нечеткие числа L-R-типа, $B_0 = (\mu_{b0}, \alpha_{b0}, \beta_{b0})$ и $B_1 = (\mu_{b1}, \alpha_{b1}, \beta_{b1})$ – теоретические коэффициенты (параметры) регрессии, $E_i = (\mu_{ei}, \alpha_{ei}, \beta_{ei})$ – случайные ошибки (отклонения), нечеткие числа L-R-типа, $i = \overline{1,n}$ – номер наблюдения.

Для получения наиболее подходящих оценок $\tilde{B}_0, \tilde{B}_1$ регрессионных параметров B_0, B_1 применим метод наименьших квадратов с использованием формулы расстояния (4) между двумя нечеткими числами [5].

Следует учитывать при этом, что $Y_i = (\mu_{yi}, \alpha_{yi}, \beta_{yi})$, $i = \overline{1,n}$, и $B_0 = (\mu_{b0}, \alpha_{b0}, \beta_{b0})$ и $B_1 = (\mu_{b1}, \alpha_{b1}, \beta_{b1})$ имеют одинаковую функцию принадлежности, и после надлежащего преобразования данных можно сделать все $x_i > 0$, $i = \overline{1,n}$.

После подстановки оценок параметров $\tilde{B}_0, \tilde{B}_1$ в уравнение регрессии (7) приходим к следующей эмпирической (оценочной) нечеткой линейной парной регрессионной модели:

$$\tilde{Y}_i = \tilde{B}_0 + \tilde{B}_1 x_i,\ i = \overline{1,n}, \qquad (8)$$

где $x_i \in \Re$ – наблюдаемые значения независимой (объясняющей) переменной, $\tilde{Y}_i = (\tilde{\mu}_{yi}, \tilde{\alpha}_{yi}, \tilde{\beta}_{yi})$ – оценки значений зависимой (объясняемой) переменной, $\tilde{B}_0 = (\tilde{\mu}_{b0}, \tilde{\alpha}_{b0}, \tilde{\beta}_{b0})$ и $\tilde{B}_1 = (\tilde{\mu}_{b1}, \tilde{\alpha}_{b1}, \tilde{\beta}_{b1})$ – оценки неизвестных параметров B_0, B_1, называемые эмпирическими (выборочными) коэффициентами регрессии.

В соответствии с методом наименьших квадратов оценки параметров регресии находятся из решения задачи минимизации функции:

$$F(\tilde{B}_0, \tilde{B}_1) = \sum_{i=1}^{n} D^2(Y_i, \tilde{Y}_i) \to \min, \qquad (9)$$

где $Y_i = (\mu_{yi}, \alpha_{yi}, \beta_{yi})$ и $\tilde{Y}_i = (\tilde{\mu}_{yi}, \tilde{\alpha}_{yi}, \tilde{\beta}_{yi})$ вычисляются по формулам (7), (8), $D(Y_i, \tilde{Y}_i)$ – функция расстояния между нечеткими переменными Y_i и $\tilde{Y}_i$ (формула (4)).

Согласно формулам (2)-(5) целевую функцию можно записать в виде:

$$F(\tilde{B}_0, \tilde{B}_1) = \sum_{i=1}^{n} \left(\tilde{B}_0 + \tilde{B}_1 x_i - Y_i\right)^2 =$$

$$= \sum_{i=1}^{n} \left(\left(\tilde{\mu}_{b0} + \tilde{\mu}_{b1} x_i - \mu_{yi}\right)^2 + \left(\tilde{\alpha}_{b0} + \tilde{\alpha}_{b1} x_i - \alpha_{yi}\right)^2 + \left(\tilde{\beta}_{b0} + \tilde{\beta}_{b1} x_i - \beta_{yi}\right)^2 \right) \to \min.$$

Нетрудно заметить, что функция F является квадратичной функцией двух параметров $\tilde{B}_0 = (\tilde{\mu}_{b0}, \tilde{\alpha}_{b0}, \tilde{\beta}_{b0})$ и $\tilde{B}_1 = (\tilde{\mu}_{b1}, \tilde{\alpha}_{b1}, \tilde{\beta}_{b1})$ ($F = F(\tilde{B}_0, \tilde{B}_1)$), поскольку $\{(x_i, Y_i)\}_{i=\overline{1,n}}$ – известные данные наблюдений. Так как функция F непрерывна, выпукла и ограничена снизу ($F>0$), то она имеет минимум.

Необходимым условием существования минимума функции (9) является равенство нулю ее частных производных по неизвестным величинам $\tilde{\mu}_{b0}, \tilde{\alpha}_{b0}, \tilde{\beta}_{b0}$ и $\tilde{\mu}_{b1}, \tilde{\alpha}_{b1}, \tilde{\beta}_{b1}$:

$$
\begin{cases}
\dfrac{\partial F}{\partial \tilde{\mu}_{b0}} = 2\sum\limits_{i=1}^{n}\left(\tilde{\mu}_{b0} + \tilde{\mu}_{b1}x_i - \mu_{yi}\right) = 0, \\
\dfrac{\partial F}{\partial \tilde{\alpha}_{b0}} = 2\sum\limits_{i=1}^{n}\left(\tilde{\alpha}_{b0} + \tilde{\alpha}_{b1}x_i - \alpha_{yi}\right) = 0, \\
\dfrac{\partial F}{\partial \tilde{\beta}_{b0}} = 2\sum\limits_{i=1}^{n}\left(\tilde{\beta}_{b0} + \tilde{\beta}_{b1}x_i - \beta_{yi}\right) = 0.
\end{cases} \quad (10)
$$

$$
\begin{cases}
\dfrac{\partial F}{\partial \tilde{\mu}_{b1}} = 2\sum\limits_{i=1}^{n}\left(\tilde{\mu}_{b0} + \tilde{\mu}_{b1}x_i - \mu_{yi}\right)x_i = 0, \\
\dfrac{\partial F}{\partial \tilde{\alpha}_{b1}} = 2\sum\limits_{i=1}^{n}\left(\tilde{\alpha}_{b0} + \tilde{\alpha}_{b1}x_i - \alpha_{yi}\right)x_i = 0, \\
\dfrac{\partial F}{\partial \tilde{\beta}_{b1}} = 2\sum\limits_{i=1}^{n}\left(\tilde{\beta}_{b0} + \tilde{\beta}_{b1}x_i - \beta_{yi}\right)x_i = 0.
\end{cases} \quad (11)
$$

После преобразований получим системы нормальных уравнений для определения параметров нечеткой линейной парной регрессии [5]:

$$
\begin{cases}
n\tilde{\mu}_{b0} + \tilde{\mu}_{b1}\sum\limits_{i=1}^{n}x_i = \sum\limits_{i=1}^{n}\mu_{yi}, \\
n\tilde{\alpha}_{b0} + \tilde{\alpha}_{b1}\sum\limits_{i=1}^{n}x_i = \sum\limits_{i=1}^{n}\alpha_{yi}, \\
n\tilde{\beta}_{b0} + \tilde{\beta}_{b1}\sum\limits_{i=1}^{n}x_i = \sum\limits_{i=1}^{n}\beta_{yi}.
\end{cases} \quad (12)
$$

$$
\begin{cases}
\tilde{\mu}_{b0}\sum\limits_{i=1}^{n}x_i + \tilde{\mu}_{b1}\sum\limits_{i=1}^{n}x_i^2 = \sum\limits_{i=1}^{n}x_i\mu_{yi}, \\
\tilde{\alpha}_{b0}\sum\limits_{i=1}^{n}x_i + \tilde{\alpha}_{b1}\sum\limits_{i=1}^{n}x_i^2 = \sum\limits_{i=1}^{n}x_i\alpha_{yi}, \\
\tilde{\beta}_{b0}\sum\limits_{i=1}^{n}x_i + \tilde{\beta}_{b1}\sum\limits_{i=1}^{n}x_i^2 = \sum\limits_{i=1}^{n}x_i\beta_{yi}.
\end{cases} \quad (13)
$$

Теперь, разделив обе части уравнений систем (12) и (13) на n, получим следующие системы нормальных уравнений:

$$
\begin{cases}
\tilde{\mu}_{b0} + \tilde{\mu}_{b1}\overline{x} = \overline{\mu}_y, \\
\tilde{\alpha}_{b0} + \tilde{\alpha}_{b1}\overline{x} = \overline{\alpha}_y, \\
\tilde{\beta}_{b0} + \tilde{\beta}_{b1}\overline{x} = \overline{\beta}_y,
\end{cases} \quad (14)
$$

$$
\begin{cases}
\tilde{\mu}_{b0}\overline{x} + \tilde{\mu}_{b1}\overline{x^2} = \overline{x\mu_y}, \\
\tilde{\alpha}_{b0}\overline{x} + \tilde{\alpha}_{b1}\overline{x^2} = \overline{x\alpha_y}, \\
\tilde{\beta}_{b0}\overline{x} + \tilde{\beta}_{b1}\overline{x^2} = \overline{x\beta_y},
\end{cases} \quad (15)
$$

где соответствующие средние определяются по формулам:

$$
\overline{x} = \frac{\sum\limits_{i=1}^{n}x_i}{n}, \quad (16)
$$

$$
\overline{x^2} = \frac{\sum\limits_{i=1}^{n}x_i^2}{n}, \quad (17)
$$

$$
\overline{\mu_y} = \frac{\sum\limits_{i=1}^{n}\mu_{yi}}{n}, \quad \overline{\alpha_y} = \frac{\sum\limits_{i=1}^{n}\alpha_{yi}}{n}, \quad \overline{\beta_y} = \frac{\sum\limits_{i=1}^{n}\beta_{yi}}{n}. \quad (18)
$$

$$\overline{x\mu_y} = \frac{\sum_{i=1}^{n} x_i \mu_{yi}}{n}, \ \overline{x\alpha_y} = \frac{\sum_{i=1}^{n} x_i \alpha_{yi}}{n}, \ \overline{x\beta_y} = \frac{\sum_{i=1}^{n} x_i \beta_{yi}}{n}. \tag{19}$$

Выразим $\tilde{\mu}_{b0}, \tilde{\alpha}_{b0}, \tilde{\beta}_{b0}$ в системе (14) через $\tilde{\mu}_{b1}, \tilde{\alpha}_{b1}, \tilde{\beta}_{b1}$:

$$\begin{cases} \tilde{\mu}_{b0} = \overline{\mu}_y - \tilde{\mu}_{b1}\overline{x}, \\ \tilde{\alpha}_{b0} = \overline{\alpha}_y - \tilde{\alpha}_{b1}\overline{x}, \\ \tilde{\beta}_{b0} = \overline{\beta}_y - \tilde{\beta}_{b1}\overline{x}. \end{cases} \tag{20}$$

Подставив выражения из системы (20) в уравнения (15), получим:

$$\begin{cases} \left(\overline{\mu}_y - \tilde{\mu}_{b1}\overline{x}\right)\overline{x} + \tilde{\mu}_{b1}\overline{x^2} = \overline{x\mu_y}, \\ \left(\overline{\alpha}_y - \tilde{\alpha}_{b1}\overline{x}\right)\overline{x} + \tilde{\alpha}_{b1}\overline{x^2} = \overline{x\alpha_y}, \\ \left(\overline{\beta}_y - \tilde{\beta}_{b1}\overline{x}\right)\overline{x} + \tilde{\beta}_{b1}\overline{x^2} = \overline{x\beta_y}. \end{cases} \tag{21}$$

Решая систему (21), находим:

$$\begin{cases} \tilde{\mu}_{b1} = \dfrac{\overline{x\mu_y} - \overline{\mu}_y\overline{x}}{\overline{x^2} - \overline{x}^2} = \dfrac{\operatorname{cov}\left(x, \mu_y\right)}{\sigma_x^2}, \\ \tilde{\alpha}_{b1} = \dfrac{\overline{x\alpha_y} - \overline{\alpha}_y\overline{x}}{\overline{x^2} - \overline{x}^2} = \dfrac{\operatorname{cov}\left(x, \alpha_y\right)}{\sigma_x^2}, \\ \tilde{\beta}_{b1} = \dfrac{\overline{x\beta_y} - \overline{\beta}_y\overline{x}}{\overline{x^2} - \overline{x}^2} = \dfrac{\operatorname{cov}\left(x, \beta_y\right)}{\sigma_x^2}, \end{cases} \tag{22}$$

где σ_x^2 – выборочная дисперсия переменной x

$$\sigma_x^2 = \overline{x^2} - \overline{x}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2, \tag{23}$$

$\operatorname{cov}\left(x, \mu_y\right)$, $\operatorname{cov}\left(x, \alpha_y\right)$, $\operatorname{cov}\left(x, \beta_y\right)$ – выборочные корреляционные моменты или выборочные ковариации:

$$\begin{cases} \operatorname{cov}\left(x, \mu_y\right) = \overline{x\mu_y} - \overline{\mu}_y\overline{x} = \dfrac{\sum_{i=1}^{n} x_i \mu_{yi}}{n} - \dfrac{\sum_{i=1}^{n} \mu_{yi}}{n} \cdot \dfrac{\sum_{i=1}^{n} x_i}{n}, \\ \operatorname{cov}\left(x, \alpha_y\right) = \overline{x\alpha_y} - \overline{\alpha}_y\overline{x} = \dfrac{\sum_{i=1}^{n} x_i \alpha_{yi}}{n} - \dfrac{\sum_{i=1}^{n} \alpha_{yi}}{n} \cdot \dfrac{\sum_{i=1}^{n} x_i}{n}, \\ \operatorname{cov}\left(x, \beta_y\right) = \overline{x\beta_y} - \overline{\beta}_y\overline{x} = \dfrac{\sum_{i=1}^{n} x_i \beta_{yi}}{n} - \dfrac{\sum_{i=1}^{n} \beta_{yi}}{n} \cdot \dfrac{\sum_{i=1}^{n} x_i}{n}. \end{cases} \tag{24}$$

Коэффициент корреляции

Перейдем к оценке тесноты корреляционной зависимости. Подставим полученные в (20) значения модальной величины и левой и правой границ параметра $\tilde{B}_0$ в уравнение регрессии (8). Тогда получим

$$\left(\tilde{\mu}_y, \tilde{\alpha}_y, \tilde{\beta}_y\right) = \left(\tilde{\mu}_{b0}, \tilde{\alpha}_{b0}, \tilde{\beta}_{b0}\right) + \left(\tilde{\mu}_{b1}, \tilde{\alpha}_{b1}, \tilde{\beta}_{b1}\right) x$$

или более подробно

$$\left(\tilde{\mu}_y, \tilde{\alpha}_y, \tilde{\beta}_y\right) = \left(\overline{\mu}_y - \tilde{\mu}_{b1}\overline{x}, \overline{\alpha}_y - \tilde{\alpha}_{b1}\overline{x}, \overline{\beta}_y - \tilde{\beta}_{b1}\overline{x}\right) + \left(\tilde{\mu}_{b1}, \tilde{\alpha}_{b1}, \tilde{\beta}_{b1}\right) x \, .$$

Воспользовавшись формулой сложения нечетких чисел (5), имеем:

$$\left(\tilde{\mu}_y, \tilde{\alpha}_y, \tilde{\beta}_y\right) = \left(\overline{\mu}_y - \tilde{\mu}_{b1}\overline{x} + \tilde{\mu}_{b1}, \overline{\alpha}_y - \tilde{\alpha}_{b1}\overline{x} + \tilde{\alpha}_{b1}, \overline{\beta}_y - \tilde{\beta}_{b1}\overline{x} + \tilde{\beta}_{b1}\right) =$$
$$= \left(\overline{\mu}_y + \tilde{\mu}_{b1}(x - \overline{x}), \overline{\alpha}_y + \tilde{\alpha}_{b1}(x - \overline{x}), \overline{\beta}_y + \tilde{\beta}_{b1}(x - \overline{x})\right).$$

Таким образом, приходим к следующей системе уравнений:

$$\begin{cases} \tilde{\mu}_y = \overline{\mu}_y + \tilde{\mu}_{b1}(x - \overline{x}), \\ \tilde{\alpha}_y = \overline{\alpha}_y + \tilde{\alpha}_{b1}(x - \overline{x}), \\ \tilde{\beta}_y = \overline{\beta}_y + \tilde{\beta}_{b1}(x - \overline{x}). \end{cases} \qquad (25)$$

Представим уравнения системы (25) в эквивалентном виде:

$$\begin{cases} \dfrac{\tilde{\mu}_y - \overline{\mu}_y}{\sigma_\mu} = \tilde{\mu}_{b1} \dfrac{\sigma_x}{\sigma_\mu} \dfrac{(x - \overline{x})}{\sigma_x}, \\ \dfrac{\tilde{\alpha}_y - \overline{\alpha}_y}{\sigma_\alpha} = \tilde{\alpha}_{b1} \dfrac{\sigma_x}{\sigma_\alpha} \dfrac{(x - \overline{x})}{\sigma_x}, \\ \dfrac{\tilde{\beta}_y - \overline{\beta}_y}{\sigma_\beta} = \tilde{\beta}_{b1} \dfrac{\sigma_x}{\sigma_\beta} \dfrac{(x - \overline{x})}{\sigma_x}, \end{cases} \qquad (26)$$

где σ_x^2 – выборочная дисперсия переменной x, определяемая по формуле (23), а $\sigma_\mu^2, \sigma_\alpha^2, \sigma_\beta^2$ – выборочные дисперсии величин μ_y, α_y, β_y:

$$\sigma_\mu^2 = \overline{\mu_y^{\,2}} - \overline{\mu_y}^{\,2} = \frac{\sum_{i=1}^{n} \mu_{yi}^2}{n} - \left(\frac{\sum_{i=1}^{n} \mu_{yi}}{n}\right)^2, \qquad (27)$$

$$\sigma_\alpha^2 = \overline{\alpha_y^{\,2}} - \overline{\alpha_y}^{\,2} = \frac{\sum_{i=1}^{n} \alpha_{yi}^2}{n} - \left(\frac{\sum_{i=1}^{n} \alpha_{yi}}{n}\right)^2, \qquad (28)$$

$$\sigma_\beta^2 = \overline{\beta_y^{\,2}} - \overline{\beta_y}^{\,2} = \frac{\sum_{i=1}^{n} \beta_{yi}^2}{n} - \left(\frac{\sum_{i=1}^{n} \beta_{yi}}{n}\right)^2. \qquad (29)$$

Из системы (26) находим величину

$$r = \left(\tilde{\mu}_{b1} \frac{\sigma_x}{\sigma_\mu}, \tilde{\alpha}_{b1} \frac{\sigma_x}{\sigma_\alpha}, \tilde{\beta}_{b1} \frac{\sigma_x}{\sigma_\beta}\right). \qquad (30)$$

Величина r показывает тесноту связи и является выборочным коэффициентом корреляции (или просто коэффициентом корреляции).

Если $r > 0$ ($\tilde{B}_1 > 0$), то корреляционную связь между переменными будем называть прямой, если $r < 0$ ($\tilde{B}_1 < 0$) – обратной.

Учитывая (22), формулу для r представим в виде

$$r = \left(\frac{\operatorname{cov}\left(x, \mu_y\right)}{\sigma_x \sigma_\mu}, \frac{\operatorname{cov}\left(x, \alpha_y\right)}{\sigma_x \sigma_\alpha}, \frac{\operatorname{cov}\left(x, \beta_y\right)}{\sigma_x \sigma_\beta} \right) \tag{31}$$

Отметим другие модификации формулы для r, полученные из формулы (31) с помощью формул (22)-(24):

$$r = \left(\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(\mu_{yi} - \overline{\mu_y}\right)}{n\sigma_x \sigma_\mu}, \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(\alpha_{yi} - \overline{\alpha_y}\right)}{n\sigma_x \sigma_\alpha}, \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(\beta_{yi} - \overline{\beta_y}\right)}{n\sigma_x \sigma_\beta} \right), \tag{32}$$

$$r = \left(\mu_r, \alpha_r, \beta_r\right), \tag{33}$$

где μ_r, α_r, β_r вычисляются по формулам:

$$\mu_r = \frac{n\sum_{i=1}^{n} x_i \mu_{yi} - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \mu_{yi}\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} \mu_{yi}^2 - \left(\sum_{i=1}^{n} \mu_{yi}\right)^2}},$$

$$\alpha_r = \frac{n\sum_{i=1}^{n} x_i \alpha_{yi} - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \alpha_{yi}\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} \alpha_{yi}^2 - \left(\sum_{i=1}^{n} \alpha_{yi}\right)^2}},$$

$$\beta_r = \frac{n\sum_{i=1}^{n} x_i \beta_{yi} - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \beta_{yi}\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} \beta_{yi}^2 - \left(\sum_{i=1}^{n} \beta_{yi}\right)^2}}.$$

Для практических расчетов наиболее удобна формула (33), так как по ней r находится непосредственно из данных наблюдений и на значении r не скажутся округления данных, связанные с расчетом средних и отклонений от них.

Анализ данных

Используем полученные теоретические результаты для анализа нечеткой информации. Рассмотрим следующие примеры.

Пример 1. По данным табл. 1 найти уравнение регрессии Y по x.

Таблица 1

Исходные данные

	11	12	13	14	15	16	17	18	19	20	21	22	23	24
	11,2	12,5	12,9	14,1	14,8	16,1	17,5	18,9	18,9	20	21,1	22,2	23,3	24,4
	0,9	0,1	1	0,5	1,1	1,2	0,1	0,5	0,3	1,3	1,1	0,2	1,44	0,7
	0,2	0,4	0,5	1,1	0,7	0,1	0,08	1,2	0,7	0,48	1,9	0,4	0,8	0,71

Решение. Вычислим все необходимые суммы:

$$\sum_{i=1}^{14} x_i = 11 + 12 + 13 + 14 + 15 + 16 + 17 + 18 + 19 + 20 + 21 + 22 + 23 + 24 = 245;$$

$$\sum_{i=1}^{14} x_i^2 = 11^2 + 12^2 + 13^2 + 14^2 + 15^2 + 16^2 + 17^2 + 18^2 + 19^2 + 20^2 + 21^2 + 22^2 + 23^2 +$$
$$+24^2 = 4515;$$
$$\sum_{i=1}^{14} \mu_{yi} = 11,2 + 12,5 + 12,9 + 14,1 + 14,8 + 16,1 + 17,5 + 18,9 + 18,9 + 20 + 21,1 + 22,2 +$$
$$23,3 + 24,4 = 247,9;$$
$$\sum_{i=1}^{14} \alpha_{yi} = 0,9 + 0,1 + 1 + 0,5 + 1,1 + 1,2 + 0,1 + 0,5 + 0,3 + 1,3 + 1,1 + 0,2 + 1,44 + 0,7 =$$
$$= 10,44;$$
$$\sum_{i=1}^{14} \beta_{yi} = 0,2 + 0,4 + 0,5 + 1,1 + 0,7 + 0,1 + 0,08 + 1,2 + 0,7 + 0,48 + 1,9 + 0,4 + 0,8 +$$
$$+0,71 = 9,27;$$
$$\sum_{i=1}^{14} x_i\mu_{yi} = 11\cdot 11,2 + 12\cdot 12,5 + 13\cdot 12,9 + 14\cdot 14,1 + 15\cdot 14,8 + 16\cdot 16,1 + 17\cdot 17,5 +$$
$$+18\cdot 18,9 + 19\cdot 18,9 + 20\cdot 20 + 21\cdot 21,1 + 22\cdot 22,2 + 23\cdot 23,3 + 24\cdot 24,4 = 4567,7;$$
$$\sum_{i=1}^{14} x_i\alpha_{yi} = 11\cdot 0,9 + 12\cdot 0,1 + 13\cdot 1 + 14\cdot 0,5 + 15\cdot 1,1 + 16\cdot 1,2 + 17\cdot 0,1 + 18\cdot 0,5 +$$
$$+19\cdot 0,3 + 20\cdot 1,3 + 21\cdot 1,1 + 22\cdot 0,2 + 23\cdot 1,44 + 24\cdot 0,7 = 186,62;$$
$$\sum_{i=1}^{14} x_i\beta_{yi} = 11\cdot 0,2 + 12\cdot 0,4 + 13\cdot 0,5 + 14\cdot 1,1 + 15\cdot 0,7 + 16\cdot 0,1 + 17\cdot 0,08 + 18\cdot 1,2 +$$
$$+19\cdot 0,7 + 20\cdot 0,48 + 21\cdot 1,9 + 22\cdot 0,4 + 23\cdot 0,8 + 24\cdot 0,71 = 171.$$

Затем по формулам (16)-(24) находим выборочные характеристики уравнения регрессии:

$\overline{x} = \dfrac{245}{14} = 17,5$; $\overline{x^2} = \dfrac{1415}{14} = 322,5$;

$\overline{\mu_y} = \dfrac{247,9}{14} = 17,7$; $\overline{\alpha_y} = \dfrac{10,44}{14} = 0,75$, $\overline{\beta_y} = \dfrac{9,27}{14} = 0,662$;

$\overline{x\mu_y} = \dfrac{4567,7}{14} = 326,3$; $\overline{x\alpha_y} = \dfrac{186,62}{14} = 13,3$; $\overline{x\beta_y} = \dfrac{171}{14} = 12,2$;

$\sigma_x^2 = \overline{x^2} - \overline{x}^2 = 322,5 - 306,25 = 16,25$;

$$\begin{cases} \operatorname{cov}\left(x, \mu_y\right) = \overline{x\mu_y} - \overline{\mu}_y\overline{x} = 326,3 - 17,7\cdot 17,5 = 16,55, \\ \operatorname{cov}\left(x, \alpha_y\right) = \overline{x\alpha_y} - \overline{\alpha}_y\overline{x} = 13,3 - 0,75\cdot 17,5 = 0,175, \\ \operatorname{cov}\left(x, \beta_y\right) = \overline{x\beta_y} - \overline{\beta}_y\overline{x} = 12,2 - 0,662\cdot 17,5 = 0,615. \end{cases}$$

Теперь вычисляем параметры уравнения регрессии:

$$\begin{cases}\tilde{\mu}_{b1}=\dfrac{\operatorname{cov}\left(x,\mu_y\right)}{\sigma_x^2}=\dfrac{16,55}{16,25}=1,02,\\ \tilde{\alpha}_{b1}=\dfrac{\overline{x\alpha_y}-\overline{\alpha}_y\overline{x}}{\overline{x^2}-\overline{x}^2}=\dfrac{0,175}{16,25}=0,01,\\ \tilde{\beta}_{b1}=\dfrac{\overline{x\beta_y}-\overline{\beta}_y\overline{x}}{\overline{x^2}-\overline{x}^2}=\dfrac{0,615}{16,25}=0,037.\end{cases}$$

$$\begin{cases}\tilde{\mu}_{b0}=\overline{\mu}_y-\tilde{\mu}_{b1}\overline{x}=17,7-1,02\cdot 17,5=-0,15,\\ \tilde{\alpha}_{b0}=\overline{\alpha}_y-\tilde{\alpha}_{b1}\overline{x}=0,75-0,01\cdot 17,5=0,575,\\ \tilde{\beta}_{b0}=\overline{\beta}_y-\tilde{\beta}_{b1}\overline{x}=0,662-0,037\cdot 17,5=0,0145.\end{cases}$$

Итак, уравнение регрессии Y по x:

$$\tilde{Y}=\left(-0.15,0.575,0.0145\right)+\left(1.02,0.01,0.037\right)x.$$

Пример 2. По данным табл. 1 вычислить коэффициент корреляции между переменными Y и x.

Решение. В примере 1 были вычислены
$\sum_{i=1}^{14}x_i=245$; $\sum_{i=1}^{14}x_i^2=4515$; $\sum_{i=1}^{14}\mu_{yi}=247,9$; $\sum_{i=1}^{14}\alpha_{yi}=10,44$; $\sum_{i=1}^{14}\beta_{yi}=9,27$; $\sum_{i=1}^{14}x_i\mu_{yi}=4567,7$;
$\sum_{i=1}^{14}x_i\alpha_{yi}=186,62$; $\sum_{i=1}^{14}x_i\beta_{yi}=171.$

Вычислим теперь суммы:

$$\sum_{i=1}^{14}\mu_{yi}^2=11,2^2+12,5^2+12,9^2+14,1^2+14,8^2+16,1^2+17,5^2+18,9^2+18,9^2+20^2+$$
$$+21,1^2+22,2^2+23,3^2+24,4^2=4622,13;$$

$$\sum_{i=1}^{14}\alpha_{yi}^2=0,9^2+0,1^2+1^2+0,5^2+1,1^2+1,2^2+0,1^2+0,5^2+0,3^2+1,3^2+1,1^2+0,2^2+$$
$$+1,44^2+0,7^2=10,5736;$$

$$\sum_{i=1}^{14}\beta_{yi}^2=0,2^2+0,4^2+0,5^2+1,1^2+0,7^2+0,1^2+0,08^2+1,2^2+0,7^2+0,48^2+1,9^2+$$
$$+0,4^2+0,8^2+0,71^2=9,2409;$$

По формуле (33) найдем коэффициент корреляции между переменными Y и x.

$$\mu_r=\frac{14\cdot 4567,7-245\cdot 247,9}{\sqrt{14\cdot 4515-245^2}\cdot\sqrt{14\cdot 4622,13-247,9^2}}=\frac{3212,3}{3220,01}=0,998,$$

$$\alpha_r=\frac{14\cdot 186,62-245\cdot 10,44}{\sqrt{14\cdot 4515-245^2}\cdot\sqrt{14\cdot 10,5736-10,44^2}}=\frac{54,88}{352,605}=0,156,$$

$$\beta_r=\frac{14\cdot 171-245\cdot 9,27}{\sqrt{14\cdot 4515-245^2}\cdot\sqrt{14\cdot 9,2409-9,27^2}}=\frac{122,85}{371,91}=0,33,$$

т.е. $r=\left(0.998,0.156,0.33\right)$.

Заключение

В статье был рассмотрен нечеткий регрессионный анализ для парной линейной регрессионной модели с нечеткими коэффициентами. Для оценки параметров модели был использован метод наименьших квадратов. Как показывают результаты практических экспериментов, нечеткая линейная регрессия позволяет найти неизвестные коэффициенты модели с достаточно высокой точностью, близкой к единице.

Список источников

1. Конышева Л.К. Основы теории нечетких множеств [текст] /Л.К. Конышева, Д.М. Назаров. – Спб.: Питер, 2011. – 192 с.
2. Кофман, А. Введение в теорию нечетких множеств [текст] / А. Кофман. – М.: Радио и связь, 1982. – 432 с.
3. Леденева, Т.М. Обработка нечеткой информации [текст] / Т.М. Леденева. – Воронеж: ВГУ, 2006. – 233 с.
4. Пегат, А. Нечеткое моделирование и управление [текст] / А. Пегат. – М.: Бином, 2009. – 798 с.
5. Сапкина, Н.В. Нечеткий парный линейный регрессионный анализ [текст] / Н.В. Сапкина, А.А. Татаринцев // Материалы всероссийской молодежной научной школы «Инженерия знаний. Представление знаний: состояние и перспективы». – 2012. – С. 260 – 261.
6. Сапкина, Н.В. Свойства операций над нечеткими числами [текст] / Н.В. Сапкина // Вестник Воронежского государственного университета. Сер. Системный анализ и информационные технологии. – Воронеж: ВГУ, 2013. –№1. – С. 23 – 28.

INDISTINCT PAIR RECTILINEAR REGRESSION AND CORRELATION

Sapkina Natalia Vladimirovna,
Post-graduate student of the Chair of computing mathematic and application information technologies of Voronezh State University; natashasapkina@yandex.ru

In this article application of a least square method for an assessment of unknown coefficients of an indistinct steam room of rectilinear regression model for indistinct numbers L-R-type is offered. The formula for calculation of coefficient of correlation between an accurate explaining variable and an indistinct exit is found. In the conclusion the received theoretical results are used for information analysis.
Keywords: correlation, least square method, method of indistinct pair rectilinear regression, indistinct numbers L-R-type, operation with indistinct numbers, distance between two indistinct numbers.