О методе наименьших квадратов при регрессии с нечеткими данными

Вельдяксов В.Н., Шведов А.С.

Данные, используемые при регрессионном анализе, могут быть неточными или неоднозначными. Неопределенность данных может вытекать из случайности или из нечеткости. Регрессия, основанная на вероятностных моделях, широко распространена. Но трудности могут возникать, например, когда множество наблюдений слишком мало или предположения о виде вероятностных распределений недостоверны. При обычном эконометрическом оценивании предполагается, что и зависимые, и независимые переменные даны в форме действительных чисел. Но во многих прикладных задачах доступны лишь нечеткие данные. Существующие статистические методы могут быть обобщены и на случай такой неопределенности.

Методы нечеткой регрессии основаны на теории нечетких множеств. Такая регрессия достаточно широко применяется в финансах, деловом администрировании и других областях. Регрессионная модель с нечеткими данными может рассматриваться с различных точек зрения, переменные могут считаться нечеткими, или отношение между переменными может считаться нечетким.

По моделям нечеткой регрессии опубликовано много работ. При этом рассматриваются различные варианты моделей: с нечеткими регрессорами и с четкими коэффициентами регрессии, с четкими регрессорами и с нечеткими коэффициентами регрессии, с нечеткими регрессорами и с нечеткими коэффициентами регрессии.

В настоящей работе рассматривается задача повышения точности регрессионной модели, когда некоторые или все наблюдения – нечеткие, при этом коэффициенты модели остаются действительными числами. Предлагается новый способ оценивания свободного члена в регрессионной модели, при этом свободный член представляет собой нечеткое число. Этот способ основан на решении задачи вариационного исчисления. На примерах показано, что включение в модель нечеткого свободного члена позволяет повысить предсказательную силу регрессионной модели.

Ключевые слова: нечеткая линейная регрессия; оценивание параметров.

Вельдяксов Василий Николаевич – аспирант кафедры математической экономики и эконометрики НИУ ВШЭ. E-mail: veldyaksov@gmail.com

Шведов Алексей Сергееевич – профессор кафедры математической экономики и эконометрики НИУ ВШЭ. E-mail: ashvedov@hse.ru

Статья поступила в Редакцию в апреле 2014 г.

1. Введение

Нечеткие числа, являющиеся обобщением действительных чисел, предложены в работе [Zadeh, 1965] и с тех пор нашли применение во многих областях. Включение в математическую модель нечетких чисел дает возможность другой передачи неопределенности, чем при вероятностном подходе, в котором используются случайные величины. Если цель использования случайных величин – включить в модель некоторое множество различных значений неизвестных показателей и вероятности этих значений, то цель использования нечетких чисел – передать расплывчатость, неопределенность самих значений. Эти два подхода к моделированию, вероятностный и нечеткий, успешно применяются, как независимо друг от друга, так и объединенно, в том числе и при построении регрессий.

Нечеткая регрессия – это направление, относительно недавно возникшее и интенсивно развивающееся. Хотя число публикаций по нечеткой регрессии и уступает числу публикаций по вероятностной регрессии, но все же очень велико, и исчерпывающий обзор работ по нечеткой регрессии не является задачей настоящей статьи. Так, в работе [Abdalla, Buckley, 2007] указывается, что по запросу «fuzzy linear regression» авторами получено 579000 ссылок. Мы укажем лишь на несколько публикаций по нечетким и нечетко-случайным регрессионным моделям.

Одной из первых работ, где изучается задача нечеткой регрессии, является работа [Tanaka, Uegima, Asai, 1982]. В этой работе рассматриваются нечеткие объясняемые переменные, четкие регрессоры и нечеткие коэффициенты регрессии. Для нахождения коэффициентов регрессии решается задача математического программирования. Дальнейшее развитие этого подхода представлено, например, в работе [Tanaka, Hayashi, Watada, 1989].

Метод наименьших квадратов при построении нечеткой регрессии используется в работе [Celmiņš, 1987]. Также этот метод, когда и объясняющие переменные, и объясняемые переменные нечеткие, а коэффициенты – четкие числа, изучается в работе [Diamond, 1988], причем в этой работе рассматриваются модели и с четким, и с нечетким свободными членами, но при некоторых упрощающих предположениях относительно вида нечетких чисел. Данный подход развивается в работе [Diamond, Körner, 1997]. Также метод наименьших квадратов при регрессии с нечеткими данными изучается в работе [Bargiela, Pedrycz, Nakashima, 2007]; отказ от использования функций принадлежности как основного способа определения нечетких чисел, принятый в этой работе, дает ощутимые преимущества (такой же подход применяется в работе [Шведов, 2013], где приводится новое определение нечетко-случайных величин). Однако в работе [Bargiela, Pedrycz, Nakashima, 2007] рассматриваются лишь модели с четким свободным членом. В работе [Kao, Chyu, 2002] авторы предлагают двухшаговую процедуру построения нечеткой регрессионной модели; на первом шаге все нечеткие наблюдения подвергаются процедуре дефазификации, и регрессия оценивается обычным методом наименьших квадратов; на втором шаге происходит отдельная оценка параметра нечеткости исходя из требования минимизации расстояния между значениями наблюдаемой переменной и предсказанными значениями. В работе [Yang, Lin, 2002] исследуется модель с нечеткими наблюдаемыми переменными, нечеткими регрессорами и нечеткими параметрами модели и используется метод наименьших квадратов.

Нечетко-случайная регрессия рассматривается, например, в работах [González-Rodríguez, Blanco, Colubi, Lubiano, 2009; Nather, 2006].

Для изучения экономических задач нечеткая регрессия применяется, например, в работах [de Sánchez, Gómez, 2003; Lin, Zhuang, Huang, 2012].

Целью настоящей работы является совершенствование способов построения регрессионной модели, включающей нечеткие данные, для увеличения предсказательной силы модели. В разделе 2 содержатся некоторые предварительные сведения, относящиеся к нечетким числам и операциям над ними. В разделе 3 метод наименьших квадратов для регрессии с нечеткими данными из работы [Bargiela, Pedrycz, Nakashima, 2007] обобщается таким образом, чтобы допускать и нечеткие свободные члены. Это обобщение не является прямолинейным, оказывается необходимым использовать методы вариационного исчисления. Приводится и некоторый анализ, относящийся к случаю четких свободных членов. В разделе 4 для тестовых данных сравниваются подходы с нечетким свободным членом и с четким свободным членом. Точность регрессионной модели при использовании нечеткого свободного члена оказывается выше. В разделе 5 приводятся выводы.

2. Нечеткие числа и операции над ними

Существуют различные подходы к определению нечетких чисел и к операциям над такими числами. В настоящей статье используется определение нечеткого числа то же, что и в работе [Шведов, 2013]. Компактное подмножество $K \subseteq \mathbb{R}^2$ (координаты в пространстве $\mathbb{R}^2$ будем обозначать (ξ, η)) называется нечетким числом, если выполнены следующие условия: при $t \notin [0,1]$ пересечение множества K с прямой $\eta = t$ пусто; при $t \in [0,1]$ пересечение множества K с прямой $\eta = t$ имеет вид

$$\{(\xi, \eta) : k_1(t) \le \xi \le k_2(t), \eta = t\},$$

где k_1 – монотонно неубывающая непрерывная слева функция аргумента t; k_2 – монотонно невозрастающая непрерывная слева функция аргумента t. Функции k_1 и k_2 будем называть левым и правым индексом нечеткого числа соответственно (см. рис. 1). Если $k_1(t) = k_2(t)$ при любом $t \in [0,1]$, то нечеткое число вырождается в обычное действительное число.

Нечеткое число называется трапецеидальным, если функции k_1 и k_2 линейные, и $k_1(1) < k_2(1)$. Нечеткое число называется треугольным, если функции k_1 и k_2 линейные, и $k_1(1) = k_2(1)$. Обычно треугольное нечеткое число задают в виде упорядоченной тройки $(k_1(0), k_1(1), k_2(0))$.

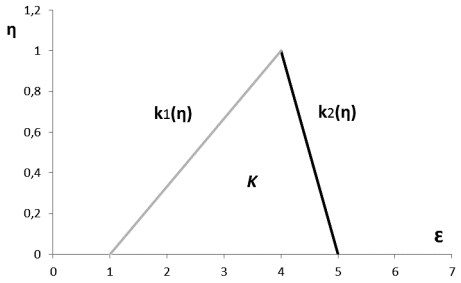

Рис. 1. Нечеткое число K, левый индекс k_1 и правый индекс k_2

Пусть A и B – нечеткие числа, λ – действительное число; a_1 и a_2 соответственно левый и правый индексы нечеткого числа A; b_1 и b_2 соответственно левый и правый индексы нечеткого числа B. Суммой нечетких чисел A и B называется нечеткое число $A+B$, обладающее тем свойством, что для любого $t\in[0,1]$ пересечение множества $A+B$ с прямой $\eta=t$ имеет вид

$$\{(\xi,\eta): a_1(t)+b_1(t)\le\xi\le a_2(t)+b_2(t),\eta=t\}.$$

Произведением действительного числа $\lambda\ge 0$ и нечеткого числа A называется нечеткое число λA, обладающее тем свойством, что для любого $t\in[0,1]$ пересечение множества λA с прямой $\eta=t$ имеет вид

$$\{(\xi,\eta): \lambda a_1(t)\le\xi\le\lambda a_2(t),\eta=t\}.$$

Произведением действительного числа $\lambda<0$ и нечеткого числа A называется нечеткое число λA, обладающее тем свойством, что для любого $t\in[0,1]$ пересечение множества λA с прямой $\eta=t$ имеет вид

$$\{(\xi,\eta): \lambda a_2(t)\le\xi\le\lambda a_1(t),\eta=t\}.$$

В качестве расстояния $d(A,B)$ между нечеткими числами A и B примем

$$d(A,B)=\sqrt{\int_0^1\left(a_1(t)-b_1(t)\right)^2dt+\int_0^1\left(a_2(t)-b_2(t)\right)^2dt}.$$

Такое же определение расстояния используется в работе [Bargiela, Pedrycz, Nakashima, 2007]. Другие подходы к определению расстояния между нечеткими числами приведены, например, в исследованиях [Шведов, 2013; Tran, Duckstein, 2002].

3. Построение регрессий для нечетких данных

Допустим, что наблюдения Y_i, X_i, $i = 1...n$ – это наборы нечетких чисел, заданных индексами $y_{i1}(t), y_{i2}(t)$ и $x_{i1}(t), x_{i2}(t)$ соответственно, $0 \le t \le 1$. Будем исследовать регрессионную модель следующего вида:

$$Y_i = aX_i + B.$$

Требуется найти действительное число a и нечеткое число B, которые доставляют минимум функционалу H:

$$H(a,B) = \sum_{i=1}^{n} d^2(Y_i, aX_i + B).$$

В силу определения расстояния d между нечеткими числами, при $a \ge 0$ функционал H можно записать в следующем виде:

$$H_p(a,B) = \sum_{i=1}^{n} \int_0^1 \left(y_{i1}(t) - ax_{i1}(t) - b_1(t)\right)^2 dt + \sum_{i=1}^{n} \int_0^1 \left(y_{i2}(t) - ax_{i2}(t) - b_2(t)\right)^2 dt. \quad (1)$$

При $a < 0$ функционал H выглядит следующим образом:

$$H_n(a,B) = \sum_{i=1}^{n} \int_0^1 \left(y_{i1}(t) - ax_{i2}(t) - b_1(t)\right)^2 dt + \sum_{i=1}^{n} \int_0^1 \left(y_{i2}(t) - ax_{i1}(t) - b_2(t)\right)^2 dt. \quad (2)$$

Далее необходимо решить отдельно две задачи для $a \ge 0$ и для $a < 0$, получить два набора параметров (a_p, B_p) и (a_n, B_n), сравнить $H_p(a_p, B_p)$ и $H_n(a_n, B_n)$ и затем выбрать такой итоговый набор параметров, при которых значение функционала наименьшее.

Будем считать, что $a \ge 0$. Первый этап: выбор функции $b_1(t)$ при фиксированном a для минимизации первой суммы и выбор функции $b_2(t)$ при фиксированном a для мини-

мизации второй суммы. Это задача вариационного исчисления (см., например: [Эльсгольц, 2006]). При заданных z_0 и z_1 в классе гладких на [0,1] функций z, удовлетворяющих условиям $z(0)=z_0, z(1)=z_1$, требуется найти функцию $z(t)$, доставляющую минимум функционалу:

$$\int_0^1 L\left(\mathrm{t}, z(t), z'(t)\right)dt.$$

Если функция $z(t)$ является локальным минимумом, то при условии гладкости функции L должно выполняться условие Эйлера

$$L_z - \frac{d}{dt}L_{z'} = 0.$$

Если функция L не зависит от z', то уравнение Эйлера принимает вид

$$L_z = 0.$$

Однако, как отмечается в работе [Эльсгольц, 2006], в этом случае z_0 и z_1 уже не могут быть произвольными. Введем следующие обозначения:

$$z(t)=b_1(t), \quad f_i(t)=y_{i1}(t)-ax_{i1}(t), \quad i=1,\ldots,n.$$

Функции $f_i(t)$ на первом этапе считаются известными, $y_{i1}(t)$ и $x_{i1}(t)$ – известные данные, параметр a фиксирован. Тогда требуется найти функцию $z(t)$, доставляющую минимум функционалу:

$$\sum_{i=1}^{n}\int_0^1 \left(f_i(t)-z(t)\right)^2 dt.$$

Таким образом, $L\left(t, z(t), z'(t)\right)=\sum_{i=1}^{n}\left(f_i(t)-z(t)\right)^2$. Очевидно, что $L_{z'}=0$. Уравнение $L_z=0$ принимает вид

$$z(t)=\frac{1}{n}\sum_{i=1}^{n} f_i(t). \qquad (3)$$

При выборе

$$z_0=\frac{1}{n}\sum_{i=1}^{n} f_i(0), \; z_1=\frac{1}{n}\sum_{i=1}^{n} f_i(1)$$

функция $z(t)$ удовлетворяет необходимому условию локального минимума. Получаем, что

$$(4) \qquad b_1(t)=\frac{1}{n}\sum_{j=1}^{n}\left(y_{j1}(t)-ax_{j1}(t)\right),\ b_2(t)=\frac{1}{n}\sum_{j=1}^{n}\left(y_{j2}(t)-ax_{j2}(t)\right).$$

Второй этап: выбор числа a для минимизации (1) при условиях (4). Введем обозначения:

$$u_{i1}(t)=x_{i1}(t)-\frac{1}{n}\sum_{j=1}^{n}\mathrm{x}_{j1}(t),\ v_{i1}(t)=y_{i1}(t)-\frac{1}{n}\sum_{j=1}^{n}y_{j1}(t),$$

$$u_{i2}(t)=x_{i2}(t)-\frac{1}{n}\sum_{j=1}^{n}x_{j2}(t),\ v_{i2}(t)=y_{i2}(t)-\frac{1}{n}\sum_{j=1}^{n}y_{j2}(t).$$

С учетом введенных обозначений и формул (4) функционал (1) принимает вид

$$F(a)=\sum_{i=1}^{n}\int_0^1\left(v_{i1}(t)-au_{i1}(t)\right)^2dt+\sum_{i=1}^{n}\int_0^1\left(v_{i2}(t)-au_{i2}(t)\right)^2dt.$$

После введения обозначений

$$I_1=\sum_{i=1}^{n}\int_0^1v_{i1}(t)u_{i1}(t)dt,\ I_2=\sum_{i=1}^{n}\int_0^1v_{i2}(t)u_{i2}(t)dt,$$

$$K_1=\sum_{i=1}^{n}\int_0^1u_{i1}^2(t)dt,\ K_2=\sum_{i=1}^{n}\int_0^1u_{i2}^2(t)dt,$$

$$L=\sum_{i=1}^{n}\int_0^1\left(v_{i1}^2(t)+v_{i2}^2(t)\right)dt$$

получаем, что функционал $F(a)$ может быть записан в следующем виде:

$$F(a)=a^2(K_1+K_2)-2a(I_1+I_2)+L.$$

И из необходимого условия минимума $dF(a)/da=0$ следует, что

$$a=\frac{I_1+I_2}{K_1+K_2}.$$

Окончательно получаем

$$(5) \qquad a=max\left(0,\frac{I_1+I_2}{K_1+K_2}\right).$$

Анализ для случая $a \le 0$ можно провести аналогично, но в этом нет необходимости. Достаточно заметить, что во всех формулах должны только поменяться местами $x_{i1}(t)$ и $x_{i2}(t)$ и соответственно $u_{i1}(t)$ и $u_{i2}(t)$. Таким образом, вместо формул (4) получаем формулы

$$(6) \qquad b_1(t) = \frac{1}{n}\sum_{j=1}^{n}\left(y_{j1}(t) - ax_{j2}(t)\right), \quad b_2(t) = \frac{1}{n}\sum_{j=1}^{n}\left(y_{j2}(t) - ax_{j1}(t)\right)$$

вместо формулы (5) получаем формулу

$$(7) \qquad a = min\left(0, \frac{J_1 + J_2}{K_1 + K_2}\right),$$

где

$$J_1 = \sum_{i=1}^{n}\int_0^1 v_{i1}(t)u_{i2}(t)dt, \;\; J_2 = \sum_{i=1}^{n}\int_0^1 v_{i2}(t)u_{i1}(t)dt.$$

Таким образом, по формулам (4), (5) определяются a_p и B_p, и по формуле (1) расчитывается $H_p(a_p, B_p)$. По формулам (6) и (7) определяются a_n и B_n, и по формуле (2) расчитывается $H_n(a_n, B_n)$. Окончательно получаем, что если $H_p(a_p, B_p) \le H_n(a_n, B_n)$, то $(a, B) = (a_p, B_p)$, а если $H_p(a_p, B_p) > H_n(a_n, B_n)$, то $(a, B) = (a_n, B_n)$.

Заметим, что после проведения расчетов $b_1(t)$ и $b_2(t)$ не всегда задают нечеткое число (т.е. не всегда удовлетворяют определению индексов нечеткого числа). Рассмотрим пример, где в выборке имеется два наблюдаемых значения и два регрессора, все они являются нечеткими числами треугольного типа. Вид индексов этих нечетких чисел представлен на рис. 2. После проведения расчетов в модели $Y_i = aX_i + B$ получим $a = 2$. График полученных по формулам (4) функций $b_1(t) = 2 - t$ и $b_2(t) = t$ представлен на рис 3. Монотонность функций b_1 и b_2 противоположна требуемой.

Следовательно, необходима корректировка функций b_1 и b_2. Возможный алгоритм корректировки включает два шага.

Шаг 1. Если $b_1(1) > b_2(1)$, положить $\frac{b_1(1) + b_2(1)}{2}$ и как новое значение для $b_1(1)$, и как новое значение для $b_2(1)$.

Шаг 2. При отсутствии требуемой монотонности функции $b_1(t)$ или функции $b_2(t)$ заменить эту функцию на константу.

Данный подход в дальнейшем будем называть МНК с нечетким свободным членом (МНК – метод наименьших квадратов).

Для удобства численного сравнения приведем также описание известного подхода, который будем называть МНК с четким свободным членом. Однако наше обоснование данного метода является, по-видимому, более полным, чем имеющиеся обоснования. Несколько «кустарный» подход к определению точки минимума функционала, применяемый в настоящей работе, является полностью строгим. Когда же точка минимума определяется из необходимого условия, состоящего в равенстве нулю двух первых частных производных, это оставляет вопросы, поскольку случаи $a \geq 0$ и $a < 0$ рассматриваются отдельно, и примерно в половине расчетов точка минимума попадает на границу $a = 0$.

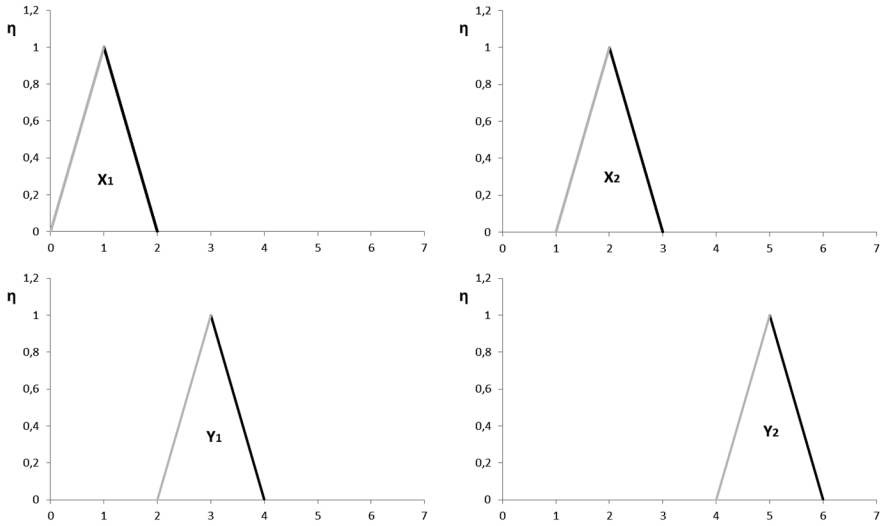

Рис. 2. Пример нечетких регрессоров X_1, X_2 и нечетких наблюдаемых переменных Y_1, Y_2 при $n = 2$

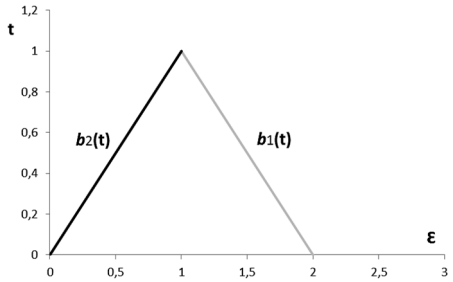

Рис. 3. Пример, когда построенные функции $b_1(t)$ и $b_2(t)$ не являются индексами нечеткого числа

Когда и вместо функции $b_1(t)$, и вместо функции $b_2(t)$ в функционалах (1) и (2) стоит одно и то же число b,

$$H_p(a,b) = a^2(K_1+K_2) + 2ab(L_1+L_2) + 2nb^2 - 2a(I_1+I_2) - 2b(M_1+M_2) + N$$

при $a \geq 0$,

$$H_n(a,b) = a^2(K_1+K_2) + 2ab(L_1+L_2) + 2nb^2 - 2a(J_1+J_2) - 2b(M_1+M_2) + N$$

при $a < 0$. Здесь

$$I_1 = \sum_{i=1}^{n} \int_0^1 y_{i1}(t)\, x_{i1}(t)\, dt, \quad I_2 = \sum_{i=1}^{n} \int_0^1 y_{i2}(t)\, x_{i2}(t)\, dt,$$

$$J_1 = \sum_{i=1}^{n} \int_0^1 y_{i1}(t)\, x_{i2}(t)\, dt, \quad J_2 = \sum_{i=1}^{n} \int_0^1 y_{i2}(t)\, x_{i1}(t)\, dt,$$

$$K_1 = \sum_{i=1}^{n} \int_0^1 x_{i1}^2(t)\, dt, \quad K_2 = \sum_{i=1}^{n} \int_0^1 x_{i2}^2(t)\, dt,$$

$$L_1 = \sum_{i=1}^{n} \int_0^1 x_{i1}(t)\, dt, \quad L_2 = \sum_{i=1}^{n} \int_0^1 x_{i2}(t)\, dt,$$

$$M_1 = \sum_{i=1}^{n} \int_0^1 y_{i1}(t)\, dt, \quad M_2 = \sum_{i=1}^{n} \int_0^1 y_{i2}(t)\, dt,$$

$$N = \sum_{i=1}^{n} \int_0^1 \left(y_{i1}^2(t) + y_{i2}^2(t) \right) dt.$$

Заметим, что $H_p(a,b)$ при фиксированном a как функция аргумента b является многочленом второй степени с положительным коэффициентом при b^2. Минимум $H_p(a,b)$ достигается при

$$b = \frac{1}{2n}(M_1+M_2) - \frac{1}{2n} a (L_1+L_2). \tag{8}$$

H_p как функция аргумента a (после подстановки полученного по формуле (8) выражения для b) является многочленом второй степени с коэффициентом при a^2

$$(K_1+K_2) - \frac{1}{2n}(L_1+L_2)^2.$$

Применяя дважды неравенство Коши – Буняковского, сначала для интегралов, а затем для суммы, получаем

$$|L_1+L_2| \le \sum_{i=1}^{n}\int_0^1 |x_{i1}(t)|\,dt + \sum_{i=1}^{n}\int_0^1 |x_{i2}(t)|\,dt \le$$
$$\le \sum_{i=1}^{n}\left(\int_0^1 x_{i1}^2(t)dt\right)^{1/2} + \sum_{i=1}^{n}\left(\int_0^1 x_{i2}^2(t)dt\right)^{1/2} \le \sqrt{K_1+K_2}\cdot\sqrt{2n},$$

причем последнее неравенство является строгим, если не все интегралы, входящие в K_1 и K_2, равны между собой, что мы и будем предполагать. Тогда коэффициент при a^2 положителен, и минимум многочлена второй степени достигается при

$$a = \frac{2n(I_1+I_2)-(L_1+L_2)(M_1+M_2)}{2n(K_1+K_2)-(L_1+L_2)^2}.$$

Окончательно для случая $a \ge 0$ получаем

$$a = max\left(0, \frac{2n(I_1+I_2)-(L_1+L_2)(M_1+M_2)}{2n(K_1+K_2)-(L_1+L_2)^2}\right). \quad (9)$$

Аналогично для случая $a < 0$ имеем

$$a = min\left(0, \frac{2n(J_1+J_2)-(L_1+L_2)(M_1+M_2)}{2n(K_1+K_2)-(L_1+L_2)^2}\right). \quad (10)$$

Таким образом, по формулам (8) и (9) определяются a_p и b_p, затем расчитывается $H_p(a_p,b_p)$. По формулам (8) и (10) определяются a_n и b_n, затем расчитывается $H_n(a_n,b_n)$. Окончательно получаем, что если $H_p(a_p,b_p) \le H_n(a_n,b_n)$, то $(a,b)=(a_p,b_p)$, а если $H_p(a_p,b_p) > H_n(a_n,b_n)$, то $(a,b)=(a_n,b_n)$. Если все нечеткие числа вырождаются в четкие, построенная модель совпадает с обычной моделью, построенной методом наименьших квадратов (см., например: [Магнус, Катышев, Пересецкий, 2004]).

Вместо регрессионной модели $Y_i = aX_i + B$ по описанной схеме можно исследовать и регрессионную модель

$$Y_i = a_1X_{i1} + \ldots + a_kX_{ik} + B,$$

где $a_1,\ldots,a_k$ – действительные числа. Однако при больших k, как отмечается в работе [Bargiela, Pedrycz, Nakashima, 2007], возникают проблемы вычислительного характера из-за

необходимости рассматривать 2^k возможных комбинаций положительных и отрицательных значений для коэффициентов $a_1,...,a_k$.

4. Численные результаты

В примерах 1 и 2 будем оценивать регрессионные модели $Y_i = aX_i + b$ и $Y_i = aX_i + B$, где $a \in \mathbb{R}$, $b \in \mathbb{R}$, B является нечетким числом. В случае, когда регрессор – нечеткое число, прогнозное значение тоже будет нечетким числом. Пусть $\hat{Y}_i$ – прогнозное значение наблюдаемой переменной Y_i, расчитаное после оценивания регрессионной модели. Обозначим индексы нечеткого числа $\hat{Y}_i$ через $\hat{y}_{i1}$ и $\hat{y}_{i2}$. При $a \geq 0$

$$\hat{y}_{i1}(t) = ax_{i1}(t) + b_1(t), \;\; \hat{y}_{i2}(t) = ax_{i2}(t) + b_2(t).$$

При $a < 0$

$$\hat{y}_{i1}(t) = ax_{i2}(t) + b_1(t), \;\; \hat{y}_{i2}(t) = ax_{i1}(t) + b_2(t).$$

Если рассматривается модель с четким свободным членом, то в последних формулах вместо функций $b_1(t)$ и $b_2(t)$ стоит действительное число b.

Определим ошибку прогноза для конкретного наблюдения $Err\left(Y_i, \hat{Y}_i\right)$ как расстояние между нечеткими числами Y_i и $\hat{Y}_i$:

$$Err\left(Y_i, \hat{Y}\right) = \sqrt{\int_0^1 \left(y_{i1}(t) - \hat{y}_{i1}(t)\right)^2 dt + \int_0^1 \left(y_{i2}(t) - \hat{y}_{i2}(t)\right)^2 dt}, i = 1,...,n.$$

Пусть n – общее число наблюдений. В качестве оценки качества подгонки регрессионной модели примем следующий показатель:

$$Error = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Err^2\left(Y_i, \hat{Y}_i\right)}.$$

Пример 1. Имеется набор данных [Tanaka, Uegima, Asai, 1982] с четкими регрессорами и нечеткой наблюдаемой переменной. Наблюдаемые переменные являются нечеткими числами треугольного типа.

Y_i	X_i
(6,2; 8,0; 9,8)	(1,0; 1,0; 1,0)
(4,2; 6,4; 8,6)	(2,0; 2,0; 2,0)
(6,9; 9,5; 12,1)	(3,0; 3,0; 3,0)
(10,9; 13,5; 16,1)	(4,0; 4,0; 4,0)
(10,6; 13,0; 15,4)	(5,0; 5,0; 5,0)

Ниже в таблице даны рассчитанные коэффициенты регрессии для моделей с использованием МНК с четким свободным членом и МНК с нечетким свободным членом. Также в таблице приведены значения функции *Error* для каждой из моделей.

Модель	Параметр a	Свободный член (b – для первой строки, B – для второй строки)	*Error*
МНК с четким свободным членом	1,710	4,950	2,706
МНК с нечетким свободным членом	1,710	(2,630; 4,950; 7,270)	1,932

Пример 2. Имеется набор данных [Kao, Chyu, 2003], где и регрессор, и наблюдаемая переменная являются нечеткими числами треугольного типа.

Y_i	X_i
(3,5; 4,0; 4,5)	(1,5; 2,0; 2,5)
(5,0; 5,5; 6,0)	(3,0; 3,5; 4,0)
(6,5; 7,5; 8,5)	(4,5; 5,5; 6,5)
(6,0; 6,5; 7,0)	(6,5; 7,0; 7,5)
(8,0; 8,5; 9,0)	(8,0; 8,5; 9,0)
(7,0; 8,0; 9,0)	(9,5; 10,5; 11,5)
(10,0; 10,5; 11,0)	(10,5; 11,0; 11,5)
(9,0; 9,5; 10,0)	(12,0; 12,5; 13,0)

Ниже в таблице даны рассчитанные коэффициенты регрессии для моделей с использованием МНК с четким свободным членом и МНК с нечетким свободным членом. Также в таблице приведены значения функции *Error* для каждой из моделей.

Модель	Параметр a	Свободный член (b – для первой строки, B – для второй строки)	*Error*
МНК с четким свободным членом	0,525	3,530	1,163
МНК с нечетким свободным членом	0,520	(3,268; 3,568; 3,868)	1,137

В обоих примерах точность модели с нечетким свободным членом оказывается выше.

5. Выводы

В работе рассмотрен подход, основанный на методах вариационного исчисления, когда и наблюдаемая переменная, и регрессор, и свободный член могут представлять собой нечеткие числа. Коэффициенты регрессии при этом остаются четкими числами. Показано, что на рассмотренных примерах предложенный метод наименьших квадратов с нечетким свободным членом дает выигрыш с точки зрения качества подгонки модели по сравнению с методом наименьших квадратов с четким свободным членом.

* *
*

СПИСОК ЛИТЕРАТУРЫ

Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. М.: ДЕЛО, 2004.

Шведов А.С. О нечетко-случайных величинах: препринт WP2/2013/02. М.: НИУ ВШЭ, 2013.

Эльсгольц Л.Э. Вариационное исчисление. М.: URSS, 2006.

Abdalla A., Buckley J.J. Monte Carlo Methods in Fuzzy Linear Regression // Soft Computing. 2007. 11. P. 991–996.

Bargiela A., Pedrycz W., Nakashima T. Multiple Regression with Fuzzy Data // Fuzzy Sets and Systems. 2007. 158. P. 2169–2188.

Celmiņš A. Least Squares Model Fitting to Fuzzy Vector Data // Fuzzy Sets and Systems. 1987. 22. P. 245–269.

De Sánchez A.J., Gómez A.T. Applications of Fuzzy Regression in Actuarial Analysis // Journal of Risk and Insurance. 2003. 70. P. 665–699.

Diamond P. Fuzzy Least Squares // Information Sciences. 1988. 46. P. 141–157.

Diamond P., Körner R. Extended Fuzzy Linear Models and Least Squares Estimates // Computers Math. Applic. 1997. 33(9). P. 15–32.

González-Rodr´ıguez G., Blanco A., Colubi A., Lubiano M.A. Estimation of a Simple Linear Regression Model for Fuzzy Random Variables // Fuzzy Sets and Systems. 2009. 160. P. 357–370.

Kao C., Chyu C. A Fuzzy Linear Regression Model with Better Explanatory Power // Fuzzy Sets and Systems. 2002. 126. P. 401–409.

Kao C., Chyu C. Least-squares Estimates in Fuzzy Regression Analysis // European Journal of Operational Research. 2003. 148. P. 426–435.

Lin J.-G., Zhuang Q.-Y., Huang C. Fuzzy Statistical Analysis of Multiple Regression with Crisp and Fuzzy Covariates and Applications in Analyzing Economic Data of China // Computational Economics. 2012. 39. P. 29–49.

Nather W. Regression with Fuzzy Random Data // Computational Statistics and Data Analysis. 2006. 51. P. 235–252.

Tanaka H., Uegima S., Asai K. Linear Regression Analysis with Fuzzy Model // IEEE Trans. on Systems, Man and Cybernetics. 1982. 12. P. 903–907.

Tanaka H., Hayashi I., Watada J. Possibilistic Linear Regression Analysis with Fuzzy Model // European Journal of Operational Research. 1989. 40. P. 389–396.

Tran L., Duckstein L. Comparison of Fuzzy Numbers Using a Fuzzy Distance Measure // Fuzzy Sets and Systems. 2002. 130. P. 331–341.

Yang M., Lin T. Fuzzy Least-squares Linear Regression Analysis for Fuzzy Input-output Data // Fuzzy Sets and Systems. 2002. 126. P. 389–399.

Zadeh L.A. Fuzzy Sets // Information and Control. 1965. 8. P. 338–353.

On Fuzzy Least-squares Regression Analysis

Veldyaksov Vasily[1], Shvedov Alexey[2]

[1] National Research University Higher School of Economics,
20, Myasnitskaya ul., Moscow, 101990, Russian Federation.
E-mail: veldyaksov@gmail.com

[2] National Research University Higher School of Economics,
20, Myasnitskaya ul., Moscow, 101990, Russian Federation.
E-mail: ashvedov@hse.ru

The data used in regression analysis may be inexact or uncertain. Uncertainty of data comes from randomness and from fuzziness. Statistical regression has many applications. But problems can occur, for instance, if the data set is too small, or there is difficulty verifying distribution assumptions. The standard econometric estimation is used when both the independent and dependent variables are given as real numbers. However, in many real-life situations only fuzzy data is available. The statistical techniques can be extended to include ambiguity of events.

Fuzzy linear regression is a modelling techniques based on fuzzy set theory. It is applied to different areas such as finance, business administration and so on. The regression model with fuzzy data has been treated from diffferent points of view. Models where the variables are fuzzy or models where the relation of the variables is fuzzy may be considered.

Significant amount of research has been conducted on fuzzy regression models. One can consider models with fuzzy observations and crisp parameters, crisp observations and fuzzy parameters, fuzzy observations and fuzzy parameters,

In this paper, we apply calculus of variations methods in fuzzy regression analysis. The fuzzy regression model is considered to be fuzzy outputs, fuzzy inputs and crisp parameters. In order to include fuzzy constant term into regression model, we solve the calculus of variations problem. The results show that the regression model with fuzzy constant term has better performance than the regression model with crisp constant term.

***Key words*:** fuzzy linear regression; least-squares estimates.

JEL Classification: C14, C32.

* *
*

References

Magnus Ja.R., Katyshev P.K., Pereseckij A.A. (2004) *Jekonometrika. Nachal'nyj kurs*. [Econometrics. Basic Course]. Moscow: DELO.

Shvedov A.S. (2013) *O nechetko-sluchajnyh velichinah* [On a Vaguely-random Variables]. Working Paper WP2/2013/02, Moscow: HSE.

Jel'sgol'c L.Je. (2006) *Variacionnoe ischislenie* [Calculus of Variations].Moscow: URSS.

Abdalla A., Buckley J.J. (2007) Monte Carlo Methods in Fuzzy Linear Regression. *Soft Computing*, 11, pp. 991–996.

Bargiela A., Pedrycz W., Nakashima T. (2007) Multiple Regression with Fuzzy Data. *Fuzzy Sets and Systems*, 158, pp. 2169–2188.

Celmiņš A. (1987) Least Squares Model Fitting to Fuzzy Vector Data. *Fuzzy Sets and Systems*, 22, pp. 245–269.

De Sánchez A.J., Gómez A.T. (2003) Applications of Fuzzy Regression in Actuarial Analysis. *Journal of Risk and Insurance*, 70, pp. 665–699.

Diamond P. (1988) Fuzzy Least Squares. *Information Sciences*, 46, pp. 141–157.

Diamond P., Körner R. (1997) Extended Fuzzy Linear Models and Least Squares Estimates. *Computers Math. Applic*, 33(9), pp. 15–32.

González-Rodr´ıguez G., Blanco A., Colubi A., Lubiano M.A. (2009) Estimation of a Simple Linear Regression Model for Fuzzy Random Variables. *Fuzzy Sets and Systems*, 160, pp. 357–370.

Kao C., Chyu C. (2002) A Fuzzy Linear Regression Model with Better Explanatory Power. *Fuzzy Sets and Systems*, 126, pp. 401–409.

Kao C., Chyu C. (2003) Least-squares Estimates in Fuzzy Regression Analysis. *European Journal of Operational Research*, 148, pp. 426–435.

Lin J.-G., Zhuang Q.-Y., Huang C. (2012) Fuzzy Statistical Analysis of Multiple Regression with Crisp and Fuzzy Covariates and Applications in Analyzing Economic Data of China. *Computational Economics*, 39, pp. 29–49.

Nather W. (2006) Regression with Fuzzy Random Data. *Computational Statistics and Data Analysis*, 51, pp. 235–252.

Tanaka H., Uegima S., Asai K. (1982) Linear Regression Analysis with Fuzzy Model. *IEEE Trans. on Systems, Man and Cybernetics*, 12, pp. 903–907.

Tanaka H., Hayashi I., Watada J. (1989) Possibilistic Linear Regression Analysis with Fuzzy Model. *European Journal of Operational Research,* 40, pp. 389–396.

Tran L., Duckstein L. (2002) Comparison of Fuzzy Numbers Using a Fuzzy Distance Measure. *Fuzzy Sets and Systems*, 130, pp. 331–341.

Yang M., Lin T. (2002) Fuzzy Least-squares Linear Regression Analysis for Fuzzy Input-output Data. *Fuzzy Sets and Systems*, 126, pp. 389–399.

Zadeh L.A. (1965) Fuzzy Sets. *Information and Control*, 8, pp. 338–353.