

## Проверка гипотез при регрессии с нечеткими данными

Вельдяков В.Н., Шведов А.С.

Регрессионный анализ широко применяется в научных разработках, и нечеткая линейная регрессия является активно развивающейся областью исследований. Это связано с тем, что во многих реальных задачах зависимые или независимые переменные не представляют собой действительные числа. Регрессионные модели с нечеткими данными рассматриваются при различных типах зависимых и независимых переменных.

В настоящей работе изучается модель регрессии  $y_i = A + bx_i + \varepsilon_i$ ,  $i = 1, \dots, n$ , где  $A, x_1, \dots, x_n$  – нечеткие числа;  $b$  – действительное число;  $\varepsilon_1, \dots, \varepsilon_n, y_1, \dots, y_n$  – нечетко-случайные величины.

В предыдущей работе авторов [Вельдяков, Шведов, 2014] с использованием метода наименьших квадратов построены оценки для коэффициентов  $A, b$ . При построении этих оценок используются методы вариационного исчисления. Указанные оценки являются развитием ранее известных оценок, относящихся к случаю, когда  $A$  – действительное число.

Основной акцент в работе [Вельдяков, Шведов, 2014] делается на построении оценки для коэффициента  $A$ . Однако получена и некоторая оценка для коэффициента  $b$ . В первой части настоящей работы доказывается, что оценка для коэффициента  $b$ , полученная в статье [Вельдяков, Шведов, 2014], обладает свойством несмещенности. При доказательстве существенную роль играет новое определение нечетко-случайных величин из работы [Шведов, 2013].

Во второй части настоящей работы на ряде расчетов проводится сравнение доверительных интервалов для коэффициента  $b$  и бутстреп процентных интервалов для этого коэффициента. Установлено, что совпадение длин этих интервалов улучшается при увеличении размера выборки  $n$ .

Данный вывод, а также несмещенность оценки для коэффициента  $b$  позволяют предложить процедуру проверки гипотезы о конкретном значении для коэффициента  $b$  в приведенной регрессионной модели.

**Ключевые слова:** простая регрессия; нечетко-случайные величины; проверка гипотез; доверительные интервалы; бутстреп-процентные интервалы.

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

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

Статья поступила в Редакцию в сентябре 2014 г.

## 1. Введение

Математическое описание неопределенности имеет большое практическое значение, в том числе и в экономических задачах. Традиционно для этого используется теория вероятностей, основное понятие в которой – это случайная величина. Другой подход к описанию неопределенности дан в 1965 г. в работе [Zadeh, 1965], где предлагается понятие нечеткого множества. Если теория случайных величин направлена на передачу в математической модели вероятностей различных значений, то теория нечетких множеств применяется для моделирования размытости самих значений. После опубликования работы [Zadeh, 1965] теория нечетких множеств стала интенсивно развиваться и в настоящее время используется в различных прикладных областях. Естественным является объединение обоих подходов, случайного и нечеткого. Понятие нечетко-случайной величины дает один из путей такого объединения. Изучение нечетко-случайных величин начато в работах [Kwakernaak, 1978; Puri, Ralescu, 1986]. Нами используется определение нечетко-случайной величины из работы [Шведов, 2013]. Общая идея этого определения совпадает с идеей, используемой в предшествующих работах. Нечетко-случайная величина – это та же случайная величина, т.е. измеримая функция, только значениями этой функции являются не обычные действительные числа, а нечеткие числа. Однако детали определения в данном случае существенны (подробнее см.: [Шведов, 2013]).

Обычно теория статистического вывода основывается на применении случайных величин. Однако в последние десятилетия получили распространение и нечеткие подходы в статистике (см., например: [Colubi, 2009; Filzmoser, Viertl, 2004; Gil, Montenegro, González-Rodríguez, Colubi, Casals, 2006; Taheri, 2003; Viertl, 2006]). Если говорить о нечетком регрессионном анализе, то, по-видимому, число публикаций, где обсуждаются способы статистического вывода относительно коэффициентов регрессии, значительно уступает числу публикаций, в которых строятся оценки для этих коэффициентов. Некоторый обзор публикаций, основным направлением которых является построение оценок для коэффициентов нечеткой регрессии, дается в статье [Вельдяксов, Шведов, 2014]. Из работ, где обсуждаются способы статистического вывода относительно этих коэффициентов, назовем исследования [Akbari, Mohammadalizadeh, Rezaei, 2012; Arnold, Gerke, 2003; González-Rodríguez, Blanco, Corral, Colubi, 2007; Näther, 2006]. В работе [Lin, Zhuang, Huang, 2012] тест о нулевом значении коэффициента при независимой переменной в уравнении нечеткой регрессии из работы [González-Rodríguez, Blanco, Corral, Colubi, 2007] применяется для анализа экономических данных. Одним из основных подходов к задачам статистического вывода при нечеткой постановке в указанных, а также в других работах является бутстреп.

В разделе 2 настоящей работы доказывается несмещенность оценки из статьи [Вельдяксов, Шведов, 2014] для коэффициента при независимой переменной. В разделе 3 проводится численное сравнение доверительных интервалов с последующим применением результатов этого сравнения для проверки гипотез о конкретном значении коэффициента при независимой переменной в модели нечеткой регрессии.

## 2. Несмещенность оценки для коэффициента при независимой переменной

Изучается модель регрессии  $y_i = A + bx_i + \varepsilon_i, i = 1, \dots, n$ , где  $A, x_1, \dots, x_n$  – нечеткие числа;  $b$  – действительное число;  $\varepsilon_1, \dots, \varepsilon_n, y_1, \dots, y_n$  – нечетко-случайные величины.

Определение нечеткого числа  $K$ , его левого индекса  $k_1(\eta)$  и правого индекса  $k_2(\eta)$  даются в работе [Вельдяков, Шведов, 2014] и здесь повторяться не будут. Также не будут повторяться приводимые в указанной работе определения сложения нечетких чисел и умножения нечеткого числа на действительное число. Результатом каждой из этих операций является нечеткое число.

Ожиданием нечеткого числа  $K$  называется действительное число

$$E(K) = \int_0^1 \frac{k_1(\eta) + k_2(\eta)}{2} d\eta.$$

Такой способ дефазификации является широко употребительным.

В сжатом виде приведем определения нечетко-случайной величины, нечеткого ожидания и ожидания из работы [Шведов, 2013]. Пусть  $\Omega$  – вероятностное пространство. Нечетко-случайная величина – это функция  $A$ , определенная на множестве  $\Omega$  такая, что  $A(\omega)$  является нечетким числом при любом  $\omega \in \Omega$ . В работе [Шведов, 2013] доказывается, что левый индекс  $a_1(\omega, \eta)$  и правый индекс  $a_2(\omega, \eta)$  нечетко-случайной величины  $A$  при фиксированном  $\eta$  являются измеримыми функциями от аргумента  $\omega$ , и накладывается условие ограниченности этих функций. Нечеткое ожидание нечетко-случайной величины  $A$  – это нечеткое число с левым и правым индексами

$$a_{e1}(\eta) = \int_{\Omega} a_1(\omega, \eta) dP, \quad a_{e2}(\eta) = \int_{\Omega} a_2(\omega, \eta) dP$$

соответственно. Ожиданием нечетко-случайной величины  $A$  называется ожидание ее нечеткого ожидания и обозначается  $E(A)$ .

Если  $A$  и  $B$  – нечетко-случайные величины,  $\lambda$  – действительное число, то

$$E(A + B) = E(A) + E(B), \quad E(\lambda A) = \lambda E(A).$$

Поскольку при умножении нечеткого числа на отрицательное действительное число левый и правый индексы нечеткого числа меняются местами, а при умножении на положительное действительное число не меняются, рассуждения для регрессионной модели с  $b \geq 0$  и для регрессионной модели с  $b < 0$  оказываются не совсем одинаковыми.

Индексы нечеткого числа  $x_i$  обозначим  $x_{i1}(\eta)$  и  $x_{i2}(\eta)$ . Определим нечеткое число  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , индексы этого нечеткого числа обозначим  $\bar{x}_1(\eta)$  и  $\bar{x}_2(\eta)$ . Индексы не-

четко-случайных величин  $y_i$  и  $\varepsilon_i$  обозначим соответственно  $y_{i1}(\eta)$ ,  $y_{i2}(\eta)$  и  $\varepsilon_{i1}(\eta)$ ,  $\varepsilon_{i2}(\eta)$ . Определим нечетко-случайные величины

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i.$$

Индексы этих нечетко-случайных величин обозначим  $\bar{y}_1(\eta)$ ,  $\bar{y}_2(\eta)$ ,  $\bar{\varepsilon}_1(\eta)$ ,  $\bar{\varepsilon}_2(\eta)$ .

Тогда  $y_{i1}(\eta)$ ,  $y_{i2}(\eta)$ ,  $\varepsilon_{i1}(\eta)$ ,  $\varepsilon_{i2}(\eta)$ ,  $\bar{y}_1(\eta)$ ,  $\bar{y}_2(\eta)$ ,  $\bar{\varepsilon}_1(\eta)$ ,  $\bar{\varepsilon}_2(\eta)$  при фиксированном  $\eta \in [0,1]$  являются случайными величинами.

В разделе 3 будем считать, что нечетко-случайные величины  $\varepsilon_1, \dots, \varepsilon_n$  независимы (определение независимости нечетко-случайных величин см. в работе [Шведов, 2013]), но в данном разделе независимость этих нечетко-случайных величин не используется, здесь используется лишь условие, что при любом  $\eta \in [0,1]$

$$E(\varepsilon_{11}(\eta)) = \dots = E(\varepsilon_{n1}(\eta)), \quad E(\varepsilon_{12}(\eta)) = \dots = E(\varepsilon_{n2}(\eta)).$$

Пусть

$$\begin{aligned} I_1 &= \sum_{j=1}^n \int_0^1 v_{j1}(\eta) u_{j1}(\eta) d\eta, \quad I_2 = \sum_{j=1}^n \int_0^1 v_{j2}(\eta) u_{j2}(\eta) d\eta, \\ J_1 &= \sum_{j=1}^n \int_0^1 v_{j1}(\eta) u_{j2}(\eta) d\eta, \quad J_2 = \sum_{j=1}^n \int_0^1 v_{j2}(\eta) u_{j1}(\eta) d\eta, \\ K_1 &= \sum_{j=1}^n \int_0^1 u_{j1}^2(\eta) d\eta, \quad K_2 = \sum_{j=1}^n \int_0^1 u_{j2}^2(\eta) d\eta, \end{aligned}$$

где

$$\begin{aligned} u_{jk}(\eta) &= x_{jk}(\eta) - \bar{x}_k(\eta), \quad v_{jk}(\eta) = y_{jk}(\eta) - \bar{y}_k(\eta), \\ j &= 1, \dots, n; \quad k = 1, 2. \end{aligned}$$

Оценка  $\bar{a}, \bar{b}$  для коэффициентов регрессии  $(A, b)$  в работе [Вельдяков, Шведов, 2014] строится следующим образом. При

$$b_p = \max\left(0, \frac{I_1 + I_2}{K_1 + K_2}\right), \quad b_m = \min\left(0, \frac{J_1 + J_2}{K_1 + K_2}\right),$$

$$a_{p1}(\eta) = \frac{1}{n} \sum_{j=1}^n (y_{j1}(\eta) - b_p x_{j1}(\eta)),$$

$$a_{p2}(\eta) = \frac{1}{n} \sum_{j=1}^n (y_{j2}(\eta) - b_p x_{j2}(\eta)),$$

$$a_{m1}(\eta) = \frac{1}{n} \sum_{j=1}^n (y_{j1}(\eta) - b_m x_{j2}(\eta)),$$

$$a_{m2}(\eta) = \frac{1}{n} \sum_{j=1}^n (y_{j2}(\eta) - b_m x_{j1}(\eta)),$$

$$\begin{aligned} H_p &= \sum_{j=1}^n \int_0^1 (y_{j1}(\eta) - b_p x_{j1}(\eta) - a_{p1}(\eta))^2 d\eta + \\ &+ \sum_{j=1}^n \int_0^1 (y_{j2}(\eta) - b_p x_{j2}(\eta) - a_{p2}(\eta))^2 d\eta, \\ H_m &= \sum_{j=1}^n \int_0^1 (y_{j1}(\eta) - b_m x_{j2}(\eta) - a_{m1}(\eta))^2 d\eta + \\ &+ \sum_{j=1}^n \int_0^1 (y_{j2}(\eta) - b_m x_{j1}(\eta) - a_{m2}(\eta))^2 d\eta \end{aligned}$$

принимается  $\overline{\overline{A, b}} = (A_p, b_p)$ , если  $H_p \leq H_m$ , и принимается  $\overline{\overline{A, b}} = (A_m, b_m)$ , если  $H_p > H_m$ . (Может оказаться необходимой корректировка какой-то из функций  $a_{p1}(\eta)$ ,  $a_{p2}(\eta)$ ,  $a_{m1}(\eta)$ ,  $a_{m2}(\eta)$ , если эта функция не удовлетворяет условиям, которым должен удовлетворять индекс нечеткого числа.) В последних формулах  $y_1, \dots, y_n$  – это нечеткие числа, являющиеся реализациями нечетко-случайных величин  $y_1, \dots, y_n$ . Здесь  $A_p$  – нечеткое число с индексами  $a_{p1}(\eta)$ ,  $a_{p2}(\eta)$ ;  $A_m$  – нечеткое число с индексами  $a_{m1}(\eta)$ ,  $a_{m2}(\eta)$ .

Несмещенность оценки  $\overline{\overline{b}}$  будем понимать в том смысле, что при  $b \geq 0$

$$E \left( \frac{I_1 + I_2}{K_1 + K_2} \right) = b$$

и при  $b < 0$   $E \left( \frac{J_1 + J_2}{K_1 + K_2} \right) = b$ .

Свойство несмещенности оценки  $\overline{\overline{b}}$  является очень важным. Например, если рассмотреть оценку коэффициента  $b$  для случая, когда  $A$  – действительное число (эта

оценка приводится и в работе [Вельдяков, Шведов, 2014], формулы (9), (10)), то расчеты показывают, что данная оценка является сильно смещенной. И при этом оказывается невозможной замкнутая процедура построения доверительных интервалов и проверки гипотез.

Доказательство несмещенности оценки  $\bar{b}$  проведем для случая, когда все нечеткие числа, в том числе и являющиеся значениями нечетко-случайных величин, трапецеидальные (хотя с применением теоремы Фубини результат верен и не только для трапецеидальных нечетких чисел). В случае трапецеидальных нечетких чисел все индексы являются линейными функциями аргумента  $\eta$ , поэтому

$$\begin{aligned}
 I_1 &= \frac{1}{3} \sum_{j=1}^n \left( v_{j1}(0)u_{j1}(0) + \frac{1}{2}v_{j1}(0)u_{j1}(1) + \frac{1}{2}v_{j1}(1)u_{j1}(0) + v_{j1}(1)u_{j1}(1) \right), \\
 I_2 &= \frac{1}{3} \sum_{j=1}^n \left( v_{j2}(0)u_{j2}(0) + \frac{1}{2}v_{j2}(0)u_{j2}(1) + \frac{1}{2}v_{j2}(1)u_{j2}(0) + v_{j2}(1)u_{j2}(1) \right), \\
 J_1 &= \frac{1}{3} \sum_{j=1}^n \left( v_{j1}(0)u_{j2}(0) + \frac{1}{2}v_{j1}(0)u_{j2}(1) + \frac{1}{2}v_{j1}(1)u_{j2}(0) + v_{j1}(1)u_{j2}(1) \right), \\
 J_2 &= \frac{1}{3} \sum_{j=1}^n \left( v_{j2}(0)u_{j1}(0) + \frac{1}{2}v_{j2}(0)u_{j1}(1) + \frac{1}{2}v_{j2}(1)u_{j1}(0) + v_{j2}(1)u_{j1}(1) \right), \\
 K_1 &= \frac{1}{3} \sum_{j=1}^n \left( u_{j1}(0)u_{j1}(0) + \frac{1}{2}u_{j1}(0)u_{j1}(1) + \frac{1}{2}u_{j1}(1)u_{j1}(0) + u_{j1}(1)u_{j1}(1) \right), \\
 K_2 &= \frac{1}{3} \sum_{j=1}^n \left( u_{j2}(0)u_{j2}(0) + \frac{1}{2}u_{j2}(0)u_{j2}(1) + \frac{1}{2}u_{j2}(1)u_{j2}(0) + u_{j2}(1)u_{j2}(1) \right).
 \end{aligned}$$

Пусть  $b \geq 0$ . Тогда из регрессионной модели следует, что при любом  $\eta \in [0, 1]$

$$\begin{aligned}
 E(y_{i1}(\eta)) &= a_1(\eta) + bx_{i1}(\eta) + E(\varepsilon_{i1}(\eta)), \\
 E(y_{i2}(\eta)) &= a_2(\eta) + bx_{i2}(\eta) + E(\varepsilon_{i2}(\eta)),
 \end{aligned}$$

$i = 1, \dots, n$ . А также

$$\begin{aligned}
 E(\bar{y}_1(\eta)) &= a_1(\eta) + b\bar{x}_1(\eta) + E(\bar{\varepsilon}_1(\eta)), \\
 E(\bar{y}_2(\eta)) &= a_2(\eta) + b\bar{x}_2(\eta) + E(\bar{\varepsilon}_2(\eta)).
 \end{aligned}$$

С учетом определения  $v_{j1}(\eta), v_{j2}(\eta)$  и условий, наложенных на нечетко-случайные величины  $\varepsilon_i$ , получаем

$$E(v_{i1}(\eta)) = b(x_{i1}(\eta) - \bar{x}_1(\eta)) = bu_{i1}(\eta),$$

$$E(v_{i2}(\eta)) = b(x_{i2}(\eta) - \bar{x}_2(\eta)) = bu_{i2}(\eta).$$

Поэтому для трапецеидальных нечетких чисел

$$E(I_1) = bK_1, \quad E(I_2) = bK_2.$$

И несмещенность оценки  $\bar{b}$  при  $b \geq 0$  доказана.

Пусть  $b < 0$ . Тогда из регрессионной модели следует, что при любом  $\eta \in [0, 1]$

$$E(y_{i1}(\eta)) = a_1(\eta) + bx_{i2}(\eta) + E(\varepsilon_{i1}(\eta)),$$

$$E(y_{i2}(\eta)) = a_2(\eta) + bx_{i1}(\eta) + E(\varepsilon_{i2}(\eta)),$$

$i = 1, \dots, n$ .

$$E(\bar{y}_1(\eta)) = a_1(\eta) + b\bar{x}_2(\eta) + E(\bar{\varepsilon}_1(\eta)),$$

$$E(\bar{y}_2(\eta)) = a_2(\eta) + b\bar{x}_1(\eta) + E(\bar{\varepsilon}_2(\eta)).$$

Следовательно,

$$E(v_{i1}(\eta)) = b(x_{i2}(\eta) - \bar{x}_2(\eta)) = bu_{i2}(\eta),$$

$$E(v_{i2}(\eta)) = b(x_{i1}(\eta) - \bar{x}_1(\eta)) = bu_{i1}(\eta).$$

Поэтому для трапецеидальных нечетких чисел  $E(J_1) = bK_2$ ,  $E(J_2) = bK_1$ .

Несмещенность оценки  $\bar{b}$  доказана и при  $b < 0$ .

### 3. Доверительные интервалы, проверка гипотез

В первое десятилетие, после того как бутстреп был предложен, были предприняты серьезные попытки дать математическое обоснование этого метода, получены глубокие результаты, в том числе и относящиеся к задачам регрессии (см., например: [Singh, 1981; Bickel, Freedman, 1981; Freedman, 1981]). Но, как это часто бывает и с другими серьезными и важными вычислительными методами, область применения метода оказывается шире, строгое математическое обоснование имеется лишь для некоторых частных случаев. Правомочность применения метода в других случаях проверяется расчетами. Разумеется, при этом должна продолжаться и работа по расширению той области, для которой метод строго математически обоснован.

В классической эконометрике (при отсутствии нечеткости) для проверки гипотез, связанных с коэффициентами регрессии, может быть сделано предположение о нормальном распределении ошибок, может использоваться центральная предельная теорема

(при достаточно больших размерах выборки), может и бутстреп. В задачах нечеткой регрессии бутстреп выходит на первый план.

План этого раздела такой. Сначала при известных коэффициентах уравнения регрессии  $A, b$  и при известном распределении ошибок  $\varepsilon_i$  с использованием всей этой ин-

формации методом симулирования определяются квантили порядка  $\frac{\alpha}{2}$  и порядка  $\left(1 - \frac{\alpha}{2}\right)$

распределения вероятностей случайной величины  $\bar{b}$ . Здесь  $\alpha$  – некоторое малое положительное число. Проверяется, что истинное значение коэффициента  $b$  близко к середине интервала, концами которого являются эти квантили (здесь ключевую роль играет установленный в разделе 2 результат о несмещенности оценки  $\bar{b}$ ). Затем для этих же значений коэффициентов регрессии и этого же распределения ошибок симулируется не-

четкая выборка  $\left(\begin{matrix} x_1 \\ y_1 \end{matrix}\right), \dots, \left(\begin{matrix} x_n \\ y_n \end{matrix}\right)$ , которая объявляется основной выборкой для процедуры

бутстрепа (сделаем уточнение, что симулируются лишь нечеткие числа  $y_1, \dots, y_n$ ; нечеткие числа  $x_1, \dots, x_n$  остаются фиксированными для всего расчета). С использованием

только этой основной выборки строятся бутстреп-процентные интервалы для  $\bar{b}$  при тех же значениях  $\frac{\alpha}{2}$  и  $\left(1 - \frac{\alpha}{2}\right)$ . Проверяется совпадение длин доверительных интервалов

для коэффициента  $b$ , построенных на первом этапе расчета, и длин бутстреп-процентных интервалов, построенных на втором этапе расчета; это совпадение оказывается в ряде случаев удовлетворительным, в ряде случаев хорошим (в зависимости от длины выборки  $n$ ). На основании этого вывода длины бутстреп-процентных интервалов оказывается возможным использовать в качестве длин доверительных интервалов при проверке гипотез о конкретном значении коэффициента при независимой переменной в уравнении регрессии.

В этом разделе приводится описание результатов для четырех серий расчетов. В каждой серии расчетов рассматриваются выборки размера  $n = 125, n = 500, n = 2000$ . Степень нечеткости и для независимых переменных  $x_1, \dots, x_n$ , и для ошибок  $\varepsilon_1, \dots, \varepsilon_n$  характеризуется положительным числом  $C$ ; рассматриваются два значения  $C = 0,4$  и  $C = 0,8$ . Рассматриваются три формы распределения вероятностей ошибок  $\varepsilon_1, \dots, \varepsilon_n$ : с легкими хвостами, с тяжелыми хвостами, и с очень тяжелыми хвостами. Кроме того, распределение вероятностей ошибок характеризуется некоторым параметром  $\sigma$ , который можно было бы назвать стандартным отклонением, если бы речь шла о случайных величинах, а не о нечетко-случайных величинах. Однако выяснилось, что отношения длин интервалов зависят от этого параметра слабо, поэтому во всех расчетах используется одно значение:  $\sigma = 0,25$ . Таким образом, каждая серия содержит 18 расчетов; в каждом расчете определяются одно значение для длины доверительного интервала и одно значение для длины бутстреп-процентного интервала.



Все нечеткие числа, используемые и в качестве независимых переменных  $x_1, \dots, x_n$ , и в качестве значений нечетко-случайных величин  $\varepsilon_1, \dots, \varepsilon_n$ , являются треугольными.

В каждой из четырех серий расчетов при данном  $n$  независимые переменные  $x_1, \dots, x_n$  остаются фиксированными. Именно набором независимых переменных одна серия расчетов отличается от другой. Для построения нечетких чисел  $x_1, \dots, x_n$  используются  $2n$  равномерно распределенных на отрезке  $[0,1]$  случайных чисел  $\xi_1, \dots, \xi_{2n}$ . Нечеткое число  $x_i$  – это равнобедренный треугольник, вершина которого имеет абсциссу  $10\xi_i - 5$  и ординату 1, основанием треугольника является отрезок  $[10\xi_i - 5 - C\xi_{n+i}, 10\xi_i - 5 + C\xi_{n+i}]$ , расположенный на оси абсцисс;  $i = 1, \dots, n$ .

Во всех расчетах в качестве нечетко-случайной величины  $\varepsilon_i$  берется дискретная нечетко-случайная величина, принимающая  $M$  значений; однако распределение вероятностей этой нечетко-случайной величины имитирует некоторое известное вероятностное распределение;  $M = 161$ . Все используемые нечетко-случайные величины независимы. При каждом  $i = 1, \dots, n$  нечетко-случайная величина  $\varepsilon_i$  с вероятностью  $p_m$  принимает значение, равное равнобедренному треугольнику, вершина которого имеет абсциссу  $\mu_m = -10\sigma + 20\sigma(m-1)/M + 10\sigma/M$  и ординату 1, основанием треугольника является отрезок  $[\mu_m - C/2, \mu_m + C/2]$ , расположенный на оси абсцисс;  $m = 1, \dots, M$ .

При использовании распределения с легкими хвостами принимается (имитация нормального распределения)

$$p_m^0 = \exp\left(-\frac{1}{2} \frac{\mu_m^2}{\sigma^2}\right), \quad p_m = p_m^0 / \sum_{m=1}^M p_m^0.$$

При использовании распределения с тяжелыми хвостами принимается (имитация t-распределения с тремя степенями свободы)

$$p_m^0 = \left(1 + \frac{1}{3} \frac{\mu_m^2}{\sigma^2}\right)^{-2}, \quad p_m = p_m^0 / \sum_{m=1}^M p_m^0.$$

При использовании распределения с очень тяжелыми хвостами принимается (имитация равномерного распределения)

$$p_m = \frac{1}{M}, \quad m = 1, \dots, M.$$

На первом этапе расчета для изучения распределения статистики  $\bar{b}$  необходимо использовать конкретные значения коэффициентов  $A$  и  $b$ . Во всех расчетах принято  $A = 1,2$  (четкое число),  $b = 0,75$ . Также во всех расчетах берется  $\alpha = 0,05$ . При  $N = 1000$  следующие действия повторяются  $N$  раз:

- симулирование значений нечетко-случайных величин  $\varepsilon_1, \dots, \varepsilon_n$ ;
- вычисление нечетких чисел  $y_1, \dots, y_n$  по формуле  $y_i = A + bx_i + \varepsilon_i$ ,  $i = 1, \dots, n$ ;
- расчет  $\overline{\overline{b}}$  по методу наименьших квадратов, как это описано в разделе 2.

Полученные  $N$  значений статистики  $\overline{\overline{b}}$  упорядочиваются:

$$\overline{\overline{b}}_{(1)} \leq \overline{\overline{b}}_{(2)} \leq \dots \leq \overline{\overline{b}}_{(N)}.$$

В качестве  $100(1-\alpha)$ -процентного доверительного интервала для  $b$  принимается  $(\overline{\overline{b}}_{(N_1)}, \overline{\overline{b}}_{(N_2)})$ , где  $N_1 = \left[ (N+1) \frac{\alpha}{2} \right]$ ,  $N_2 = \left[ (N+1) \left( 1 - \frac{\alpha}{2} \right) \right]$ . Здесь  $[z]$  – ближайшее целое число к действительному числу  $z$ .

На втором этапе расчета все вычисления производятся лишь исходя из основной выборки  $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$ . Данная выборка симулируется при помощи известной регрессионной модели тем же способом, что и на первом этапе расчета. При  $J = 1000$  следующие действия повторяются  $J$  раз:

- взятие бутстреп-выборки  $\begin{pmatrix} x_1^* \\ y_1^* \end{pmatrix}, \dots, \begin{pmatrix} x_n^* \\ y_n^* \end{pmatrix}$  из основной выборки. Для этого используются  $n$  равномерно распределенных на отрезке  $[0,1]$  случайных чисел  $\xi_1, \dots, \xi_n$ . При  $j = 1, \dots, n$  в случае попадания  $\xi_j$  в полуинтервал  $\left( \frac{k-1}{n}, \frac{k}{n} \right]$  принимается  $\begin{pmatrix} x_j^* \\ y_j^* \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix}$ ;

- расчет  $\overline{\overline{b}}$  по методу наименьших квадратов, как это описано в разделе 2.

Полученные  $J$  значений статистики  $\overline{\overline{b}}$  упорядочиваются:

$$\overline{\overline{b}}_{(1)} \leq \overline{\overline{b}}_{(2)} \leq \dots \leq \overline{\overline{b}}_{(J)}.$$

В качестве  $100(1-\alpha)$ -процентного доверительного интервала для  $b$  принимается  $(\overline{\overline{b}}_{(J_1)}, \overline{\overline{b}}_{(J_2)})$ , где  $J_1 = \left[ (J+1) \frac{\alpha}{2} \right]$ ,  $J_2 = \left[ (J+1) \left( 1 - \frac{\alpha}{2} \right) \right]$ .

Результаты расчетов показаны в табл. 1–4. Используется переменная  $t$ , показывающая форму распределения нечетко-случайной величины  $\varepsilon_i$ :  $t = 1$  – распределение с легкими хвостами,  $t = 2$  – распределение с тяжелыми хвостами,  $t = 3$  – распределение с очень тяжелыми хвостами. Через  $\Delta$  обозначается  $100(1-\alpha)$ -процентный доверительный интервал для  $b$ ; через  $\Delta_b$  обозначается бутстреп  $100(1-\alpha)$ -процентный интервал.

$|\Delta|, |\Delta_b|$  – длины соответствующих интервалов;  $\bar{b}$  – значение статистики  $\bar{b}$  для основной выборки.

Результаты оказываются мало зависящими от параметра  $C$ . Поэтому приводятся только результаты, относящиеся к случаю  $C = 0,4$ .

Таблица 1.

## Результаты для первой серии расчетов

$t$	$n$	$\Delta$	$ \Delta $	$\bar{b}$	$\Delta_b$	$ \Delta_b $	$\frac{ \Delta_b }{ \Delta }$
1	125	(0,7330;0,7666)	0,0336	0,7522	(0,7360;0,7677)	0,0317	0,9429
1	500	(0,7430;0,7583)	0,0153	0,7529	(0,7444;0,7615)	0,0171	1,1150
1	2000	(0,7462;0,7539)	0,0077	0,7488	(0,7448;0,7527)	0,0080	1,0379
2	125	(0,7242;0,7749)	0,0506	0,7520	(0,7305;0,7739)	0,0434	0,8579
2	500	(0,7394;0,7626)	0,0232	0,7534	(0,7388;0,7677)	0,0289	1,2424
2	2000	(0,7444;0,7561)	0,0117	0,7487	(0,7423;0,7547)	0,0124	1,0630
3	125	(0,6472;0,8458)	0,1986	0,7638	(0,6655;0,8568)	0,1913	0,9633
3	500	(0,7086;0,7982)	0,0897	0,7665	(0,7200;0,8142)	0,0942	1,0500
3	2000	(0,7271;0,7720)	0,0449	0,7408	(0,7176;0,7625)	0,0449	0,9996

Таблица 2.

## Результаты для второй серии расчетов

$t$	$n$	$\Delta$	$ \Delta $	$\bar{b}$	$\Delta_b$	$ \Delta_b $	$\frac{ \Delta_b }{ \Delta }$
1	125	(0,7351;0,7648)	0,0297	0,7660	(0,7492;0,7814)	0,0322	1,0851
1	500	(0,7417;0,7572)	0,0155	0,7528	(0,7454;0,7604)	0,0150	0,9675
1	2000	(0,7463;0,7537)	0,0074	0,7510	(0,7473;0,7550)	0,0078	1,0493
2	125	(0,7272;0,7734)	0,0462	0,7756	(0,7519;0,7985)	0,0466	1,0102
2	500	(0,7372;0,7618)	0,0246	0,7537	(0,7426;0,7647)	0,0221	0,8999
2	2000	(0,7441;0,7559)	0,0118	0,7517	(0,7460;0,7577)	0,0117	0,9951
3	125	(0,6624;0,8367)	0,1742	0,8336	(0,7328;0,9286)	0,1957	1,1236
3	500	(0,7061;0,7927)	0,0866	0,7680	(0,7234;0,8120)	0,0885	1,0222
3	2000	(0,7278;0,7715)	0,0437	0,7547	(0,7330;0,7776)	0,0446	1,0219

Таблица 3.

## Результаты для третьей серии расчетов

$t$	$n$	$\Delta$	$ \Delta $	$\bar{b}$	$\Delta_b$	$ \Delta_b $	$\frac{ \Delta_b }{ \Delta }$
1	125	(0,7353;0,7661)	0,0309	0,7386	(0,7254;0,7524)	0,0271	0,8766
1	500	(0,7429;0,7575)	0,0146	0,7543	(0,7468;0,7616)	0,0147	1,0096
1	2000	(0,7464;0,7537)	0,0074	0,7481	(0,7442;0,7520)	0,0078	1,0615
2	125	(0,7261;0,7749)	0,0487	0,7342	(0,7164;0,7539)	0,0375	0,7708
2	500	(0,7395;0,7618)	0,0223	0,7549	(0,7436;0,7657)	0,0220	0,9874
2	2000	(0,7444;0,7559)	0,0114	0,7477	(0,7417;0,7536)	0,0119	1,0364
3	125	(0,6658;0,8446)	0,1787	0,6792	(0,5979;0,7636)	0,1657	0,9273
3	500	(0,7084;0,7918)	0,0834	0,7764	(0,7339;0,8181)	0,0842	1,0098
3	2000	(0,7295;0,7726)	0,0431	0,7378	(0,7162;0,7597)	0,0435	1,0091

Таблица 4.

## Результаты для четвертой серии расчетов

$t$	$n$	$\Delta$	$ \Delta $	$\bar{b}$	$\Delta_b$	$ \Delta_b $	$\frac{ \Delta_b }{ \Delta }$
1	125	(0,7349;0,7654)	0,0305	0,7461	(0,7339;0,7589)	0,0250	0,8203
1	500	(0,7427;0,7575)	0,0148	0,7445	(0,7371;0,7521)	0,0150	1,0140
1	2000	(0,7464;0,7540)	0,0076	0,7502	(0,7466;0,7539)	0,0073	0,9606
2	125	(0,7258;0,7735)	0,0477	0,7444	(0,7268;0,7626)	0,0358	0,7490
2	500	(0,7388;0,7613)	0,0225	0,7411	(0,7304;0,7517)	0,0213	0,9484
2	2000	(0,7443;0,7561)	0,0118	0,7502	(0,7446;0,7559)	0,0113	0,9544
3	125	(0,6607;0,8345)	0,1739	0,7261	(0,6498;0,8110)	0,1611	0,9268
3	500	(0,7092;0,7914)	0,0823	0,7245	(0,6813;0,7697)	0,0884	1,0748
3	2000	(0,7288;0,7730)	0,0442	0,7529	(0,7312;0,7752)	0,0440	0,9958

В небольшом числе расчетов различие в длинах доверительных интервалов и в длинах бутстреп-процентных интервалов составляет около 20%. (Расчет с  $t = 2$ ,  $n = 500$  из первой серии; расчет с  $t = 2$ ,  $n = 125$  из третьей серии; расчет с  $t = 1$ ,  $n = 125$  из чет-

вертой серии; расчет  $ct = 2$ ,  $n = 125$  из четвертой серии.) В других расчетах различие меньше. И такого большого различия нет ни в одном расчете с  $n = 2000$ . Это подтверждает теоретические результаты о сходимости для бутстреп-метода.

Как и для доверительных интервалов, длины бутстреп-процентных интервалов уменьшаются примерно вдвое при увеличении размера выборки в четыре раза.

Отметим также достаточно хорошее соответствие середин построенных доверительных интервалов и теоретического значения  $b = 0,75$  для всех расчетов.

Очень интересным является то, насколько хорошо в бутстреп-методе воспроизводится зависимость длины интервала от распределения вероятностей ошибок. Так, во второй серии расчетов при  $n = 2000$  длины доверительных интервалов и бутстреп-процентных интервалов составляют соответственно 0,0074 и 0,0078 для распределения с легкими хвостами, 0,0118 и 0,0117 для распределения с тяжелыми хвостами, 0,0437 и 0,0446 для распределения с очень тяжелыми хвостами. При том, что, разумеется, никакой информации о распределении вероятностей при построении бутстреп-процентных интервалов не используется, эти интервалы строятся только по основной выборке. Закономерно, что длины доверительных интервалов возрастают при увеличении тяжести хвостов.

Для проверки гипотезы о конкретном значении для коэффициента при независимой переменной в уравнении регрессии может быть использована обычная двойственность между проверкой гипотез и построением доверительных интервалов, если за длину доверительного интервала принять длину бутстреп-процентного интервала.

В литературе значительное внимание уделяется ускорению сходимости для бутстреп-метода, чтобы результаты хорошей точности получались для выборок меньшего размера  $n$ . Если говорить о доверительных интервалах для коэффициента при независимой переменной в обычной (не нечеткой) регрессии, хорошие результаты по ускорению сходимости дает стьюдентизация. Исследование этого вопроса в случае нечеткой регрессии остается предметом для дальнейшей работы.

\* \*  
\*

### СПИСОК ЛИТЕРАТУРЫ

- Вельдяков В.Н., Шведов А.С. О методе наименьших квадратов при регрессии с нечеткими данными // Экономический журнал ВШЭ. 2014. Т. 18. № 2.
- Шведов А.С. О нечетко-случайных величинах: препринт WP2/2013/02. М.: НИУ ВШЭ, 2013.
- Akbari M.G., Mohammadalizadeh R., Rezaei M. Bootstrap Statistical Inference about the Regression Coefficients Based on Fuzzy Data // International Journal of Fuzzy Systems. 2012. 14. P. 549–556.
- Arnold B.F., Gerke O. Testing Fuzzy Linear Hypotheses in Linear Regression Models // Metrika. 2003. 57. P. 81–95.
- Bickel P.J., Freedman D.A. Some Asymptotic Theory for the Bootstrap // Annals of Statistics. 1981. 9. P. 1196–1217.
- Colubi A. Statistical Inference about the Means of Fuzzy Random Variables: Applications to the Analysis of Fuzzy- and Real-valued Data // Fuzzy Sets and Systems. 2009. 160. P. 344–356.

- Filzmoser P., Viertl R.* Testing of Hypotheses with Fuzzy Data: The Fuzzy P-value // *Metrika*. 2004. 59. P. 21–29.
- Freedman D.A.* Bootstrapping Regression Models // *Annals of Statistics*. 1981. 9. P. 1218–1228.
- Gil M.A., Montenegro M., González-Rodríguez G., Colubi A., Casals M.R.* Bootstrap Approach to the Multi-sample Test of Means with Imprecise Data // *Computational Statistics & Data Analysis*. 2006. 51. P. 148–162.
- González-Rodríguez G., Blanco A., Corral N., Colubi A.* Least Squares Estimation of Linear Regression Models for Convex Compact Random Sets // *Advanced Data Reporting and Analysis Private Class*. 2007. 1. P. 67–81.
- Kwakernaak H.* Fuzzy Random Variables – I. Definitions and Theorems // *Information Sciences*. 1978. 15. P. 1–29.
- 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.
- Näther W.* Regression with Fuzzy Random Data // *Computational Statistics and Data Analysis*. 2006. 51. P. 235–252.
- Puri M.L., Ralescu D.A.* Fuzzy Random Variables // *Journal of Mathematical Analysis and Applications*. 1986. 114. P. 409–422.
- Singh K.* On the Asymptotic Accuracy of Efron's Bootstrap // *Annals of Statistics*. 1981. 9. P. 1187–1196.
- Taheri S.M.* Trends in Fuzzy Statistics // *Austrian Journal of Statistics*. 2003. 32. P. 239–257.
- Viertl R.* Univariate Statistical Analysis with Fuzzy Data // *Computational Statistics and Data Analysis*. 2006. 51. P. 133–147.
- Zadeh L.A.* Fuzzy Sets // *Information and Control*. 1965. 8. P. 338–353.

## Hypothesis Testing in Regression Models with Fuzzy Data

Veldyaksov Vasily<sup>1</sup>, Shvedov Alexey<sup>2</sup>

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

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

Regression analysis is in wide use in scientific investigation. Fuzzy linear regression is an actively developing area of research since in many real-life situations dependent or independent variables are not given as real numbers. The regression problem with fuzzy data is treated in the literature with different kinds of input-output data.

We consider the model  $y_i = A + bx_i + \varepsilon_i$ ,  $i = 1, \dots, n$ , where  $A, x_1, \dots, x_n$  – fuzzy numbers;  $b$  – real number;  $\varepsilon_1, \dots, \varepsilon_n$ ,  $y_1, \dots, y_n$  – fuzzy random variables.

In [Veldyaksov, Shvedov, 2014]  $A, b$  estimates were proposed, using ordinary least squares approach. The estimates rely on calculus of variations, and on previous research conducted for the case when  $A$  is a crisp (real) number.

Estimate for  $b$  is also proposed in [Veldyaksov, Shvedov, 2014]. In first part of this paper, we prove that this estimate is unbiased. We use new fuzzy random variables definition from [Shvedov, 2013].

In second part of this paper we refer to a number of numerical tests to compare confidence intervals for  $b$  coefficient, calculated both using traditional approach, and bootstrap approach. We show that the intervals become closer, as number of observations grows. We also propose a procedure for hypothesis testing for  $b$  coefficient in regression models with fuzzy data.

**Key words:** simple regression; fuzzy random variables; hypothesis testing; confidence intervals; bootstrap percentile intervals.

**JEL Classification:** C14, C32.

\* \*  
\*

### References

- Vel'djakov V.N., Shvedov A.S. (2014) O metode naimen'shikh kvadratov pri regressii s nechetkimi dannymi [On Fuzzy Least-squares Regression Analysis]. *Ekonomicheskii zhurnal VSE*, vol. 18, no 2.
- Shvedov A.S. (2013) *O nechetko-sluchajnykh velichinah* [On Fuzzy Random Variables]. Working Paper WP2/2013/02, Moscow: HSE.
- Akbari M.G., Mohammadalizadeh R., Rezaei M. (2012) Bootstrap Statistical Inference about the Regression Coefficients Based on Fuzzy Data. *International Journal of Fuzzy Systems*, 14, pp. 549–556.
- Arnold B.F., Gerke O. (2003) Testing Fuzzy Linear Hypotheses in Linear Regression Models. *Metrika*, 57, pp. 81–95.
- Bickel P.J., Freedman D.A. (1981) Some Asymptotic Theory for the Bootstrap. *Annals of Statistics*, 9, pp. 1196–1217.
- Colubi A. (2009) Statistical Inference about the Means of Fuzzy Random Variables: Applications to the Analysis of Fuzzy- and Real-valued Data. *Fuzzy Sets and Systems*, 160, pp. 344–356.
- Filzmoser P., Viertl R. (2004) Testing of Hypotheses with Fuzzy Data: The Fuzzy P-value. *Metrika*, 59, pp. 21–29.
- Freedman D.A. (1981) Bootstrapping Regression Models. *Annals of Statistics*, 9, pp. 1218–1228.
- Gil M.A., Montenegro M., González-Rodríguez G., Colubi A., Casals M.R. (2006) Bootstrap Approach to the Multi-sample Test of Means with Imprecise Data. *Computational Statistics & Data Analysis*, 51, pp. 148–162.
- González-Rodríguez G., Blanco A., Corral N., Colubi A. (2007) Least Squares Estimation of Linear Regression Models for Convex Compact Random Sets. *Advanced Data Reporting and Analysis Private Class*, 1, pp. 67–81.
- Kwakernaak H. (1978) Fuzzy Random Variables – I. Definitions and Theorems. *Information Sciences*, 15, pp. 1–29.
- 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.
- Näther W. (2006) Regression with Fuzzy Random Data. *Computational Statistics and Data Analysis*, 51, pp. 235–252.
- Puri M.L., Ralescu D.A. (1986) Fuzzy Random Variables. *Journal of Mathematical Analysis and Applications*, 114, pp. 409–422.
- Singh K. (1981) On the Asymptotic Accuracy of Efron's Bootstrap. *Annals of Statistics*, 9, pp. 1187–1196.
- Taheri S.M. (2003) Trends in Fuzzy Statistics. *Austrian Journal of Statistics*, 32, pp. 239–257.
- Viertl R. (2006) Univariate Statistical Analysis with Fuzzy Data. *Computational Statistics and Data Analysis*, 51, pp. 133–147.
- Zadeh L.A. (1965) Fuzzy Sets. *Information and Control*, 8, pp. 338–353.



