УДК 519.281: 658

Джунь Й. В., д. ф.-м. н., професор (Міжнародний економіко-гуманітарний університет, м. Рівне)

НЕКЛАСИЧНИЙ РЕГРЕСІЙНИЙ АНАЛІЗ, ЙОГО ЗНАЧЕННЯ І ЗАСТОСУВАННЯ

Анотація. *В статті досліджено причини виникнення і правила застосування некласичного регресійного аналізу (НРА). Розкрито, що його слід застосовувати в тих випадках, коли не виконується головна умова класичного регресійного аналізу (КРА), що означає негаусів характер розподілу залишкових похибок, який, як правило, проявляється при їх числі n > 500. Наведено робочі формули НРА і розглянуто головні етапи його програмної реалізації. Зроблено висновок, що НРА є необхідною еволюцією НРА, яка обумовлена зміною уявлень про характер дійсних розподілів залишкових похибок при великих обсягах спостережень.*
Ключові слова: *некласичний регресійний аналіз, залишкові похибки, закон Пірсона-Джеффріса.*

Аннотация. *В статье исследованы причины возникновения и правила использования неклассического регрессионного анализа (НРА). Раскрыто, что его следует применять в том случае, когда не выполняется главное условие классического регрессионного анализа (КРА), которое состоит в негауссовом характере распределения остаточных ошибок, который обычно проявляется, когда их число n > 500. Приведены рабочие формулы НРА и рассмотрены главные этапы его программной реализации. Сделан вывод, что НРА есть результатом необходимой эволюции КРА, которая обусловлена изменением представлений о характере действительных распределений остаточных ошибок при больших объемах наблюдений.*
Ключевые слова: *неклассический регрессионный анализ, остаточные ошибки, закон Пирсона-Джеффриса.*

Annotation. *The article deals with the causes of origin and rules of the non-classical regression analysis (NRA) usage. It is shown that one should apply it in the case where the main condition of the classical regression analysis (CRA) is not realized, and it is in non-Gauss character of residual errors distribution, which usually snows itself when their number is n> 500. The NRA working formulas are presented and the main stages of its software support are considered. The author made the conclusion that the NRA is the result of necessary evolution of the CRA, which is due to a change of ideas about the nature of the residual errors actual distribution in large volumes of observations.*
Keywords: *non-classical regression analysis, residual errors, Pearson-Jeffrey's law.*

В регресійному аналізі, як правило, спостереження y_i розглядають як n випадкових величин, які є лінійними комбінаціями p невідомих сталих (факторів), плюс похибки $e_i, i = 1,2,..., n$:

$$y_i = x_{1i}\beta_j + x_{2i}\beta_2 + ... + x_{pi}\beta_p + e_i, \quad (1)$$

де x_{ij} – відомі значення факторних ознак, які діють на результативну ознаку y_i. Інтерес для дослідника представляють регресори β_j, j = 1, 2,…, p, які відображають силу дії кожного із досліджуваних факторів.

Найменші припущення щодо випадкових величин e_i полягають у такому:

$$M[e_i] = 0; \quad [e_i \; e_j] = \delta_{ij} \; \sigma^2, \quad (2)$$

де M – символ математичного сподівання, δ_{ij} – символ Кронекера; σ^2 – дисперсія випадкових незалежних похибок e_i. Оскільки σ^2 є вичерпною характеристикою тільки нормально розподілених похибок, то умови (2) фактично означають, що похибки e_i підкоряються закону Гауса з нульовим математичним сподіванням і дисперсією σ^2. Але в тому випадку, коли залишкові похибки e_i не підкоряються закону Гауса, застосовувати класичний регресійний аналіз не можна.

Актуальність нашого дослідження полягає в розробці нового, ще неіснуючого некласичного регресійного аналізу, який необхідно застосовувати у тому разі, коли розподіл залишкових похибок e_i суттєво відхиляється від закону Гауса.

Аналіз робіт за цією проблемою засвідчує, що класичний регресійний аналіз, основними вимогами якого є умови (2), блискуче зарекомендував себе на протязі більше ніж 200 років застосування при вирішенні найрізноманітніших проблем науки. Проте, переможна хода регресійного аналізу обумовлена зовсім не «вічною спроможністю принципів» (2), на яких цей аналіз побудовано, а обсягом вибірок. Вперше це помітив відомий англійський математик, професор Кембриджського університету сер Г. Джеффріс. У своїй знаменитій праці [1], яка витримала у Великобританії дев'ять перевидань, сер Г. Джеффріс, в розділі 5.7 «Дослідження нормального закону» пише: «Дійсні розподіли похибок спостережень, як правило, досить близько наближаються до нормального закону і відхилення від нього важко встановити, якщо число спостережень

n не більше 500». Тобто, нормальний закон є цілком адекватним практиці спостережень, але за умови якщо $n < 500$. При $n > 500$, зі збільшенням *n,* стають все більш і більш помітними суттєві відхилення дійсних розподілів похибок від закону Гауса. При цьому, ці відхилення є вражаюче типовими. Якщо для нормального розподілу асиметрія $A = 0$ і ексцес $\varepsilon = 0$, то при $n > 500$ дійсні розподіли похибок, при тій же нульовій асиметрії *A,* набувають стійкий додатній ексцес. При цьому практика показала, що кожен інструмент, метод чи навіть місце спостережень мають свій, лише їм присутній додатній ексцес. Сер Г. Джеффріс запропонував для математичного опису розподілу випадкових похибок спостережень, за умови, що $n > 500$, наступну щільність ймовірності:

$$f(y) = \frac{m}{\pi} \cdot \frac{\mathrm{B}(m,0.5)}{\sqrt{2(m-0.5)}} \cdot \frac{1}{\sigma}\left[1 + k\left(\frac{y-\lambda}{\sigma}\right)^2\right]^{-m}, \qquad (3)$$

де λ, σ – відповідно параметри положення і розсіювання розподілу; *m* – міра відхилення розподілу (3) від закону Гауса, яка є в той же час і мірою ексцесу форми (3), який може змінюватись в межах: $0 \le \varepsilon < \infty$; B (*m*, 0.5) – бета-функція; $k = 0{,}5/M$; $M = (m\text{-}0{,}5)^3 \cdot m^{-2}$.

На перший погляд здається незрозумілим, чому саме розподіл (3) сер. Г. Джеффріс рекомендує в якості похибок спостережень. Адже є інші симетричні розподіли, якими можна апроксимувати похибки з додатнім ексцесом і асиметрією $A = 0$, наприклад, *t*-розподіл, *Lp*-розподіл, розподіл Коші тощо. Проте, всі інші розподіли: Стьюдента, Коші, *Lp* – є математично недосконалими: їх інформаційна матриця не є діагональною [2]. Це на практиці означає, що їх параметри є залежними, що дуже ускладнює їх оцінку [2]. Крім того, у негаусових симетричних розподілів з $\varepsilon > 0$ є і інші суттєві недоліки, наприклад, сімейство *Lp* є нерегулярним, що само по собі виключає можливість побудови для цього розподілу границь нерівності Рао-Крамера для ефективних оцінок його параметрів. Єдиним розподілом, єдиною сучасною моделлю ідеального ймовірнісного хаосу, яка має діагональну інформаційну матрицю Фішера, є форма (3), яку сер Г. Джеффріс створив здійснивши дотепне перетворення класичної кривої Пірсона VII типу, яка мала недіагональну інформаційну матрицю. Але створивши новий розподіл похибок з цією унікальною особливістю, сер Г. Джеффріс, будучи незвичайно скромною людиною, продовжував називати форму (3) розподілом Пірсона VII типу, що приводить до змішання понять відносно останнього і форми (3). Для того, щоб уникнути такої плутанини, ми в межах даної статті, будемо називати розподіл (3) законом похибок Пірсона-Джеффріса, або просто розподілом Пірсона-Джеффріса. Діагональність інформаційної матриці форми (3) дозволяє визначати її параметри

найпростішим чином, що продовжує прекрасну традицію класичної теорії похибок, що започаткована Гаусом, а саме – забезпечення найпростішої методики оцінювання параметрів досліджуваних величин.

Таким чином, в теорії похибок є лише два розподіли, що забезпечують діагональність міри інформації Фішера – це розподіл Гауса і розподіл (3), запропонований сером Г. Джеффрісом.

Практика показує, що при $n > 500$ похибки e_i в (1), як правило, а точніше в 75% випадків, підкоряються розподілу (1). При цьому, ймовірність того, що величини e_i при $i = 1, 2,\ldots$ $n > 500$ є вибірками із генеральної сукупності з розподілом (3), знайдена по χ^2- критерію Пірсона, свідчить про достатньо хорошу його адекватність дійсній практиці багатократних спостережень [2].

Про що свідчить той факт, що переважна більшість дійсних розподілів значень e_i при $n > 500$ слідують закону Пірсона-Джеффріса (3)? Він свідчить про те, що кількість великих похибок e_i, які перевищують потрійну СКП, завжди набагато більше, ніж це випливає із закону Гауса. В тому разі, коли спостережень небагато, скажімо, їх кількість n знаходиться в межах $30 < n < 500$, то аномальні результати просто відкидають згідно рекомендації Лежандра. При названих обсягах вибірок число викидів, як правило, не перевищує 2–3. Якраз і усі критерії вибраковки розраховані саме на таке незначне число аномальних результатів. Але при великій кількості викидів, наприклад, при їх числі в кілька десятків і більше, звичні критерії їх вибраковки застосовувати вже не можна. Візьмемо для ілюстрації ряд регулярних спостережень зміни широти в Грінвічі [3]. Похибки цього ряду мають СКП $\sigma = 0{,}131''$, ексцес $\varepsilon = +6.00 \pm 0.07$, його обсяг $n = 4982$. Цей ряд має 9.1% аномальних значень, тобто похибок, які перевищують 3σ. Чи розумно в цьому випадку відбракувати 445 спостережень як аномальні, підганяючи їх під неіснуючий ідеал нормальності, який в даному випадку є цілком некоректним і чужим проблемі.

Мета і завдання дослідження полягає в розробці сучасних процедур неокласичного регресійного аналізу, який необхідно використовувати в сучасних експериментах високої наукової і технічної складності, основною особливістю яких є великі обсяги інформації.

Спочатку дамо відповідь на питання: чому так важливо в регресійному аналізі враховувати відхилення розподілу похибок e_i в (1) від закону Гауса? Справа в тому, *що всі результати спостережень* y_i *в (1) є однорідними, (мають однакові ваги чи однакову дисперсію), лише за однієї умови:* e_i *мають бути нормальними.* Будь-який інший розподіл значень

e_i, як це показано в роботі [2], означає неоднорідність значень y_i, тобто, невиконання головної умови регресійного аналізу. В той же час використання розподілу (3) дозволяє математично строго вирішити цю проблему за допомогою його вагової функції [2, с. 59].

$$p(e_i) = \left[\left(\frac{m - 0.5}{m} \right)^3 \sigma^2 + \frac{e_i^2}{2m} \right]^{-1}, \tag{4}$$

де $e_i = y_i - \lambda; \quad \lambda, \sigma, m$ – параметри закону (3). При $m = \infty$ (закон Гауса) вагова функція $p(e)$ розподілу Пірсона-Джеффріса набуває виду константи:

$$p(e_i) = \sigma^{-2}. \tag{5}$$

З (5) можна зробити висновок, що *унікальна особливість нормально розподілених результатів полягає в тому, що всі вони мають однакову вагу*. Для будь-якого іншого розподілу значень e_i ця властивість вже не має місця. В цьому разі формула (4) дає вихід: вона дозволяє по значенню e_i з використанням максимально правдоподібних оцінок параметрів m і σ визначити вагу кожного результату спостережень y_i.

Із формули (5) випливає, що вагова функція $p(e_i)$ має розмірність оберненої дисперсії. Таким чином, *при негаусовому розподілі залишкових похибок* e_i *вага* $p(e_i)$ *означатиме оцінку оберненої дисперсії похибки спостереження* y_i.

В застосуванні вагової функції (4) при числі спостережень y_i, i=1, 2,… $n > 500$ якраз і полягає суть некласичного регресійного аналізу, реалізація якого здійснюється в три етапи.

На першому етапі застосовують класичний регресійний аналіз і визначають мінімізовані залишкові похибки

$$e_i = y_i - Y_i, \tag{6}$$

де Y_i – рівняння регресії.

На другому етапі визначають асиметрію A і ексцес залишкових похибок e_i і будують для них 90% довірчі інтервали, методом, детально викладеним в [2, с. 81], перевіряючи гіпотези:

$$A = 0;\ \varepsilon = 0. \tag{7}$$

Якщо гіпотези (7) мають місце, то обмежуються застосуванням класичного регресійного аналізу – рішення вважається остаточним і подальші обчислення припиняють.

Некласичний регресійний аналіз виконують лише при підтвердженні гіпотез:

$$A = 0;\ \varepsilon > 0. \tag{8}$$

Всі випадки, коли $A < 0$ чи $\varepsilon < 0$ детально розглянуті в [2, с. 82] і свідчать про патологічні випадки оцінювання, тобто, про некоректність поставленого експерименту.

Виконання умов (8) означає, що економічний експеримент проведений коректно, а саме регресійне моделювання є несингулярним (невиродженим). За таких умов похибки e_i гарно апроксимуються розподілом Пірсона-Джеффріса, ефективні оцінки параметрів якого знаходимо методом максимальної правдоподібності.

Третій етап починається з нормування рівнянь y_i в (1) шляхом множення їх на $\sqrt{p(e_i)}$, де $p(e_i)$ ваги, обчислені по формулі (4). Оцінки регресорів β_j в другому наближенні, отримують за умови $\sum_{i=1}^{n} e_i^2 p(e_i) = \min$:

$$\beta_j = D_j / D, \tag{9}$$

де D – детермінант системи нормальних рівнянь; D_j визначник відповідного регресора β_j.

Дисперсії оцінок регресорів β_j знаходимо з формул:

$$\sigma_j^2 = \sigma_0^2 A_{jj} D^{-1};\quad \sigma_0^2 = \sum_{i=1}^{n} e_i^2 p(e_i) /(n - k), \tag{10}$$

де A_{jj} - мінори діагональних елементів системи нормальних рівнянь; $j = 1, 2,\ldots, к$.

Основні результати дослідження полягають у тому, що некласичний регресійний аналіз виконується без особливих порушень звичних процедур класичного регресійного аналізу, навіть в програмне забезпечення останнього вноситься лише незначне доповнення – отримання ММП-оцінок параметрів закону похибок Пірсона-Джеффріса (1).

Отримання таких оцінок займає буквально кілька хвилин, якщо скористатись програмним продуктом на мові С++, алгоритм якого наведений в [2, с. 159]. Продукт створено в середовищі візуальної розробки програм Borland++ Builder 6 – сучасного інструменту створення програм з графічним інтерфейсом. Продукт створений магістрантом МЕГУ С. А. Карпіком [4] і є зрозумілим і зручним в застосуванні, навіть для користувачів, які не пов'язані з програмуванням.

За результатами проведеного дослідження, варто зазначити, що некласичний регресійний аналіз зводиться по суті до використання вагової функції залишкових похибок e_i, отриманої на основі оцінок параметрів закону Пірсона-Джеффріса. Не потрібно бентежитись навіть тоді, коли застосування вагової функції буде приводити до оцінок регресорів β_j, що мало будуть відрізнятись від класичних. Важливим є те, що застосування вагової функції дозволяє знаходити набагато більш об'єктивні оцінки точності регресорів, оскільки дія аномальних e_i буде подавлена. Саме в цьому і полягає значення процедур НРА.

1. Jeffreys H. Theory of Probability / H. Jeffreys. Sec. Edition. – Oxford, 1940. – 468 p.
2. Джунь И. В. Неклассическая теория погрешностей измерений. / И. В. Джунь. – Ривне. Изд. дом ЕСТЕРО, 2015. – 168 с. 3. Hulme H. R. The Law of Errors and the Combinations of Observations / H. R. Hulme, L. S. T. Syms // Mon. Notic. Roy. Astron. Soc. – 1939. – № 8. – P. 642–658. 4. Джунь Й. В. Закон розподілу похибок багатократних спостережень великих обсягів і оцінка їхніх параметрів // Й. В. Джунь, С. О. Карпік / Психолого-педагогічні основи гуманізації навчально-виховного процесу в школі та ВНЗ : збірник наукових праць. – №1 (11). – Рівне : РВЦ МЕГУ ім. акад. С. Дем'янчука. – 2014. – С. 133–141.

Рецензент: д.т.н., професор Власюк А. П.