УДК 528.11:519.218

Джунь Й. В., д.ф.-м.н., професор (Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука, м. Рівне)

УМОВИ ЗАСТОСУВАННЯ РЕГРЕСІЙНОГО АНАЛІЗУ

Анотація. *В статті розглянуто умови застосування класичного регресійного аналізу, який до цього часу є одним із головних засобів математичного моделювання. Показано, що відсутність аналізу таких умов є головним недоліком, який спостерігається при застосуванні програмних продуктів для оцінок регресорів.*
Ключові слова: *математичне моделювання, регресійний аналіз, програмні продукти.*

Аннотация. *В статье рассмотрены условия применения классического регрессионного анализа, который до настоящего времени является одним из главных методов математического моделирования. Показано, что отсутствие анализа таких условий есть главным недостатком, который наблюдается при применении программных продуктов для оценок регрессоров.*
Ключевые слова: *математическое моделирование, регрессионный анализ, программные продукты.*

Annotation. *In the article the terms of the classic regressive analysis application which is the one of the main methods of mathematical modeling are considered. It is shown that absence of analysis of such terms is the main defect observed during application of software products for the regressors estimations.*
Keywords: *mathematical modeling, regressive analysis, programmatic products.*

Внаслідок широкого застосування математичних методів в різних галузях науки і виробництва виникає проблема, яка полягає в необхідності їх професійного використання навіть в суміжних з математикою областях. Зокрема це стосується такого потужного і широкого застосовного математичного апарату як регресійний аналіз.

Основним недоліком, який спостерігається при застосуванні регресійного аналізу в наукових дослідженнях, для прогнозних розрахунків чи в дипломному проектуванні є, практично, повна відсутність перевірки фундаментальних вимог, які закладені в його основу.

Виконання цих вимог забезпечує надійність оцінок регресорів, адекватність моделі і її професійне застосування. В той же час переважна більшість дослідників, навіть фахівців з інформатики, кому доводиться використовувати програмні продукти, що забезпечують регресійний

аналіз, не лише не перевіряють необхідні умови його застосування, а навіть і не здогадуються про їх існування. Програмуючи регресійні моделі фахівці з інформатики вважають, що математики, пропонуючи метод, добре розібрались з усіма тонкощами його застосування. Математики же навпаки, вважають що програмісти всі ці тонкощі знають самі. А результат отримуємо плачевний: ніхто з програмістів про ці тонкощі не здогадується і використовує той чи інший готовий програмний продукт без усякої перевірки адекватності принципів, покладених в математичний фундамент методу. Назвати таке відношення до цього методу непрофесійним – не зовсім правильно. Таке відношення є звичайним невіглаством, яке вже набуло достатнього поширення.

Отже, метою і завданням нашого дослідження є узагальнення наявного досвіду регресійного моделювання і чітке викладення тих фундаментальних математичних принципів, дотримання і перевірка яких є необхідним для того, щоб його застосування було в достатній мірі професійним.

Регресійне моделювання зазвичай зводиться до оцінок параметрів $a_j, a_0, a_1, a_2, ..., \ a_k, \ j = 0.1, ..., \quad k$ такої моделі:

$$Y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + ... + a_k x_{ki}, \qquad (1)$$

де Y_i – модельне значення досліджуваної ознаки y_i, i=1,2,…,n; x_{ji} – значення факторних ознак. Отже, значення результативної, досліджуваної ознаки y_i можна представити наступним чином:

$$y_i = Y_i + e_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + ... + a_k x_{ki} + e_i, \qquad (2)$$

де $e_i = y_i - Y_i$ – випадкові незалежні похибки моделі (в ЕС і США ці похибки позначають символом О-С: «Observation-Calculation»).

Так в чому ж суть головних умов застосування регресійного аналізу? Перерахуємо в порядку їх значення.

1. Попередньо побудовані кореляційні поля залежностей для пар $y_i \to x_{ji}$ мають підтвердити їх лінійний характер. Проте, цієї важливої вимоги майже ніколи не перевіряють і навіть в програмних продуктах відсутні необхідні засоби для цього.

2. Випадкові похибки e_i в моделі (2) мають підкорятись закону Гауса. Один із авторів методу найменших квадратів (МНК) – Лагранж, а саме цей метод лежить в основі класичного регресійного моделювання, наполегливо рекомендував вилучати ті спостереження, які є аномальними і не вписуються в закон Гауса. Та майже ніхто з дослідників, не перевіряє розподіл значень e_i на нормальність, незважаючи на те, що саме на цьому фундаментальному принципі побудовані процедури регресійного аналізу. Насправді ж, як показали чисельні дослідження, нормальність похибок не підтверджуються більше ніж в 90 % випадків [1-3].

3. Величезний досвід застосування класичного регресійного моделювання, показує, що його доцільно застосовувати при кількості спостережень *n* в межах:

$$25 \le n < 500 \qquad (3)$$

При *n*<25 оцінки точності параметрів регресійної моделі є досить завищеними, оскільки число спостережень мале і їх оцінки ще не набувають стійкого статистичного характеру. Насправді ж є дуже багато випадків навіть у підручниках, коли регресійне моделювання залюбки демонструється при 15, 10, чи навіть при 5 спостереженнях. При *n*>500 похибки e_i стають стійконегаусовими. В цьому разі класичний регресійний аналіз не можна розглядати в якості кінцевого рішення. Для отримання ефективних оцінок регресорів необхідно застосовувати вже неокласичні процедури [4].

4. Кореляційна матриця, яка побудована для факторних ознак x_{ji}, обов'язково має бути діагональною. Якщо якісь дві факторні ознаки є залежними, то одна з них повинна бути вилучена. А після цього знову потрібно перевірити діагональність кореляційної матриці для ознак x_{ji}, які залишились. Оскільки число регресорів*k*зазвичай менше 25, то оцінку значущості коефіцієнтів кореляційної матриці факторних ознак необхідно здійснюватистрогим методом на основі z-перетворення Фішера. Якщо в регресійній моделі (1), принаймні 2 параметра залежні, то така модель буде неадекватною. Проте, важко зустріти роботи, де б проводилась ця необхідна перевірка. Крім того, при використанні регресійного моделювання потрібно враховувати наступне. Це моделювання було розроблено для активного експерименту у якому зміни факторів x_{ji}

можна було б регулювати в потрібних і досить великих межах, наприклад: тиск, температуру, напругу, тощо.В цьому разі число факторних ознак, вплив яких на досліджуване явище вивчається, може складати 10, 20 чи може бути навіть більшим.

Проте при застосуванні регресійного аналізу в економіці, педагогіці біології, медицині, спорті, тощо,ми не можемо змінювати значення факторів x_{ji} в будь-яких межах. А це означає, що моделювання проводиться в умовах пасивного експерименту, який не може забезпечити високу надійність оцінок. Це відбувається через те, що фактори x_{ji} змінюються в незначних межах на які ми не можемо впливати. За умов пасивного експерименту регресійна модель повинна включати не більше чотирьох, але найголовніших факторів. Часто дослідники цього не враховують. Маючи в розпорядженні результати пасивного експерименту, вони змагаються в тому, хто введе в свою модель більше параметрів, не розуміючи, що в такому разі більшість цих параметрів можна віднести до шумового поля. Крім того, таке змагання є прямим наслідком непрофесійного підходу до регресійного моделювання. Першим свідченням цього є те, що багато дослідників, отримавши регресори a_j не завжди обтяжують себе оцінками їх середньої квадратичної похибки. Значення цих похибок, написані під кожним регресором, відразу вкажуть на ті фактори, які можна віднести до шумів. Відсутність таких оцінок приводить до безглуздого змагання моделей по кількості факторних ознак, коли насправді більшість із них просто шумові і не випливають істотно на досліджуваний процес.

Окрім того, обов'язковим елементом регресійного моделювання є перевірка істотності виявлених зв'язків, наприклад, за допомогою коефіцієнта детермінації. Та, часто цей необхідний елемент регресійного моделювання не використовується і є відсутнім навіть в деяких підручниках по математичній статистиці.

Необхідно також зазначити наступне. Всі, хто знайомий з теорією регресійного аналізу знають правило складання дисперсій, тобто, загальна дисперсія досліджуваної ознаки мусить дорівнювати сумі факторної і залишкової дисперсій. Але чомусь ніхто з програмістів не використовує це правило за призначенням, а саме – для контролю правильності обчислень в регресійному аналізі. Така операція є вкрай необхідною, оскільки основною особливістю сучасних експериментів є використання значних обсягів інформації. За таки умов жоден фахівець і інформатики, який

проводить обчислення не застрахований від впливу вірусів, чи просто від технічного збою або неуважності, що може спотворити результат. Дуже важливо знати, що для забезпечення згаданого контролю необхідно використовувати оцінки параметрів регресійної моделі з числом знаків більшим, ніж в моделі призначеній для використання. Так, якщо в моделі доцільно використовувати регресори, що мають три знаки після коми, то контроль моделювання здійснюють за допомогою незаокруглених регресійних коефіцієнтів, або принаймні, таких, що мають п'ять знаків після коми. На жаль, такий метод контролю навіть не зазначений у відповідних програмних продуктах, тому більшість фахівців з інформатики не здогадуються про його існування.

Міжнародний досвід не лише регресійного, а будь-якого математичного моделювання свідчить про те, що його програма і проведення має забезпечуватись за участі не менше трьох високопрофесійних фахівців. Якщо це моделювання економічних процесів, то його алгоритм має створюватись як результат співпраці трьох професіоналів: економіста, математика і програміста. Якщо досліджуються педагогічні явища, то в такій тріаді економіст змінюється професіоналом педагогом. В іншому разі моделювання буде непрофесійним і приводитиме до неправильних висновків.

Підводячи підсумки проведеного дослідження сучасних проблем регресійного моделювання слід зазначити, що з метою його професійного проведення необхідно:

1. Доповнити відповідні програмні продукти з регресійного аналізу додатковими модулями, у яких чітко були б вказані порядок перевірки вказаних вище фундаментальних принципів, які забезпечують його професійну постановку і використання.

2. Фахівцями різних галузей можна рекомендувати проводити моделювання не самотужки, а у тісній співпраці з професіоналами з математики і програмування.

1. Орлов А. И. Часто ли распределение результатов наблюдений является нормальным? / А. И. Орлов // Заводская лаборатория. – 1991. – № 7. – С. 64-66. **2.** Джунь Й. В. Проблеми застосування імовірністних методів в економіці / Й. В. Джунь // Евріка. – РЕГІ. – Рівне, 1998. – № 1. – с. 43-45. **3.** Gazda V. Normal probability Distribution in financial Theory – false Assumption and Consequences / V. Gazda // Department of Economics, University of Economics, Faculty of Businness Economics, Kosice, 1999. – P. 5-8. **4.** Джунь Й. В. Математическая обработка астрономической и космической информации при негауссовых ошибках наблюдении / Й. В. Джунь. – Киев: ГАО АННУ. – 1992. – С. 46.

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