

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 – це головний недолік, застосування 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 – главный недостаток применения 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.

Унаслідок широкого застосування математичних методів у різних галузях науки і виробництва виникає проблема, яка полягає в необхідності їхнього професійного використання навіть у суміжних із математикою сферах. Зокрема, це стосується такого потужного і широкого застосовного математичного апарату, як регресійний аналіз.

Основним недоліком застосування регресійного аналізу в наукових дослідженнях для прогнозних розрахунків чи в дипломному проектуванні є, практично, повна відсутність перевірки фундаментальних вимог, які слугують його основою.

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 їхнє існування. Програмуючи регресійні моделі, фахівці з інформатики вважають, що математики, пропонуючи метод, ґрунтовно опрацювали усі тонкощі його застосування. Математики, навпаки, вважають, що програмісти всі ці тонкощі знають самі. Результат отримуємо плачевний: ніхто з програмістів про ці тонкощі не здогадується і використовує той чи інший готовий програмний продукт без усякої перевірки адекватності принципів, що слугують математичним фундаментом методу. Назвати таке ставлення до цього методу непрофесійним – не зовсім правильно, оскільки йдеться про звичайне невігластво, яке набуло достатнього поширення.

Отже, метою і завданням нашого дослідження визначено узагальнення наявного досвіду регресійного моделювання і чітке викладення тих фундаментальних математичних принципів, дотримання і перевірку яких визнано необхідним для того, щоб його застосування було достатньою мірою професійним.

Регресійне моделювання, зазвичай, зводиться до оцінок параметрів $a_j, a_0, a_1, a_2, \dots, a_k, j = 0, 1, \dots, k$ такої моделі:

$$Y_i = a_0 + a_1x_{1i} + a_2x_{2i} + \dots + a_kx_{ki}, \quad (1)$$

де Y_i – модельне значення досліджуваної ознаки $Y_i, i=1, 2, \dots, n$; x_{ji} – значення факторних ознак. Тобто, значення результативної, досліджуваної ознаки Y_i можна представити так:

$$y_i = Y_i + e_i = a_0 + a_1x_{1i} + a_2x_{2i} + \dots + a_kx_{ki} + e_i, \quad (2)$$

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

Так, у чому ж сутність головних умов застосування регресійного аналізу? Перерахуємо у порядку їхні значення.

1. Попередньо побудовані кореляційні поля залежностей для пар $Y_i \rightarrow x_{ji}$ мають підтвердити їх лінійний характер. Проте цієї важливої вимоги майже ніколи не перевіряють, і навіть у програмних продуктах не передбачено необхідних засобів для цього.

2. Випадкові похибки e_i в моделі (2) мають підкорятись закону Гауса. Один з авторів методу найменших квадратів (МНК) – Лагранж, а саме цей метод слугує основою 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 побудовано процедури регресійного аналізу. Насправді ж, як визначено в ході численних досліджень, нормальність похибок не підтверджено більше ніж у 90 % випадків [1-3].

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

$$25 \leq n < 500 \quad (3)$$

При $n < 25$ оцінки точності параметрів регресійної моделі є досить завищеними, оскільки число спостережень мале і їхні оцінки ще не набувають стійкого статистичного характеру. Насправді ж описано дуже багато випадків, навіть у підручниках, коли регресійне моделювання можна легко демонструвати 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, яку побудовано 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 залишилися. Оскільки число регресорів k зазвичай менше 25, то оцінку значущості коефіцієнтів кореляційної матриці факторних ознак необхідно здійснювати строгим методом на основі z-перетворення Фішера. Якщо в регресійній моделі (1) принаймні 2 параметра залежні, то така модель буде неадекватною. Проте важко знайти роботи, де б проводилась ця необхідна перевірка. Крім того, при використанні регресійного моделювання потрібно враховувати таке. Це моделювання було розроблено для активного експерименту, в якому зміни факторів x_{ji} можна було б регулювати у потрібних і досить великих межах, наприклад, тиск, температуру, напругу тощо. В цьому разі число факторних ознак, вплив

яких на досліджуване явище вивчаємо, 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 складатися із не більше чотирьох, але найголовніших факторів. Часто дослідники цього не враховують. Маючи в розпорядженні результати пасивного експерименту, вони змагаються в тому, хто введе у свою модель більше параметрів, не розуміючи, що в такому разі більшість цих параметрів можна зарахувати до шумового поля. Крім того, таке змагання є безпосереднім наслідком непрофесійного підходу до регресійного моделювання. Першим доказом цього є те, що багато дослідників, отримавши регресори a_j не завжди обтяжують себе оцінками їхньої середньої квадратичної похибки. Значення цих похибок, написані під кожним регресором, відразу вкажуть на ті фактори, які можна зарахувати до шумів. Відсутність таких оцінок призводить до безглузкого змагання моделей за кількістю факторних ознак, коли насправді більшість із них просто шумові і не впливають 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 використовують, і він відсутній навіть в деяких підручниках з математичної статистики.

Необхідно також зазначити таке. Всі, хто ознайомлений 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 основна особливість сучасних експериментів – це використання значних обсягів інформації. За таких умов жоден фахівець з 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 у моделі, призначеній для використання. Так, якщо у 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. Втім, про такий метод контролю навіть не зазначено у відповідних програмних продуктах, тому більшість фахівців з інформатики не здогадуються про його наявність.

З огляду на міжнародний досвід не лише регресійного, а будь-якого математичного моделювання можна стверджувати про те, що його програму і проведення має бути забезпечено за участі не менше трьох високопрофесійних фахівців. Якщо йдеться про моделювання економічних процесів, то його алгоритм потрібно створюватись як результат співпраці трьох професіоналів: економіста, математика і програміста. Якщо досліджують педагогічні явища, то в такій тріаді економіст змінюється професіоналом педагогом. В іншому разі моделювання буде непрофесійним і призводитиме до неправильних висновків.

Як підсумок проведеного дослідження сучасних проблем регресійного моделювання слід зазначити, що для його професійного проведення необхідно:

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

2. Фахівцям різних галузей можна рекомендувати проводити моделювання не самостійно, 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. Джуль Й. В. Проблеми застосування імовірнісних методів в економіці / Й. В. Джуль // Еврика. – 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 Business Economics, Kosice, 1999. – P. 5—.
4. Джуль Й. В. Математическая обработка астрономической и космической информации при негауссовых ошибках наблюдения / Й. В. Джуль. – К. : ГАО АННУ, 1992. – С. 46.

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