С.М. Лапач, С.Г. Радченко Проблеми визначення структури рівняння регресії в множинному регресійному аналізі / Наукові вісті НТУУ «КПІ», №1(51), 2007. С.150–155.

УДК 519.237.5

Лапач С.Н., Радченко С.Г. Проблемы определения структуры уравнения регрессии в множественном регрессионном анализе

Сформулированы нерешенные проблемы регрессионного анализа. Приведена формализованная структура многофакторной полиномиальной математической модели. На примере решения известной задачи – регрессии Дж. Лонгли – с использованием программного средства ПРИАМ показана технология формирования устойчивой структуры регрессионной модели, получение которой специалисты считали невозможным.

Ил. 2. Табл. 2. Библиогр.: 14 наз.

UDC 519.237.5

Lapach S.N., Radchenko S.G. Problems of Determining the Structure of the Regression Equation in the Multiple Regression Analysis.

Unresolved problems of regression analysis have been formulated. A formalized structure of the multifactor polynomial mathematical model is presented. The process of formation of a stable structure of the regression model which obtaining was considered impossible by specialists was shown on the example of solution of the well-known problem – that was J. Longley's regression – with the use of the software PRIAM.

Figs 2. Tabl.: 2. Refs.: 14 titles.

Кл. сл..

Линейная регрессия, регрессия Лонгли, структура уравнения регрессии, ПРИАМ, устойчивая структура регрессии

Linear regression, J. Longley's regression, stable structure of the regression model, software PRIAM.

УДК 519.237.5

Лапач С.Н., Радченко С.Г.

Проблеми визначення структури рівняння регресії в множинному регресійному аналізі

Введення

Значне збільшення кількості користувачів статистичних методів і зростання складності вирішуваних задач привели до певних проблем як користувачів, так і розробників. Це одночасно й проблеми застосування методів, і проблеми їх внутрішнього розвитку. У сучасній статистиці існує ряд невирішених теоретичних і практичних проблем. Проблеми пов'язані з невиконанням передумов застосування статистичних методів при вирішенні реальних задач [1]. Причина виниклих утруднень полягає в тому, що коректне застосування будь-якого статистичного методу або критерію зв'язано з виконанням деякого набору допущень і передумов, які прийняті в математичному формулюванні задачі. Якщо вони не виконуються, то ніяких обґрунтованих висновків із застосування методу зробити, строго кажучи, не можна. Допущення й передумови вводяться статистиками для того, щоб можна було теоретично обґрунтувати рішення, що приймається. На жаль, у реальних умовах найчастіше досить важко перевірити виконання всіх передумов і допущень або забезпечити їх виконання.

Крім того, зросла складність формалізації розв'язуваної задачі, внаслідок чого найчастіше відбувається не розв'язання прикладної задачі, а застосування статистичних методів. Це приводить до отримання багатоваріантних і суперечливих результатів або до результатів, щодо яких невідомий ступінь їхньої вірогідності.

Причиною багатьох невдач є неусвідомленість того факту, що процес отримання інформації, в цьому випадку побудова математичних моделей, є по суті таким же технологічним процесом, як і будь-який інший у різних областях техніки. Відступ від технології веде до браку, а при відсутності технології промисловий випуск продукції неможливий. Побудова математичних моделей за експериментальними даними є складною й багатоетапною процедурою, в якій кожний наступний етап залежить від результатів попереднього.

Постановка проблеми

Складність процедур у множинному регресійному аналізі така, що кожний окремий етап є предметом різних наукових напрямків. Це привело до того, що проблеми кожного

етапу вирішувалися у відриві від кінцевої мети й часто досить відома інформація, визнана в одному з наукових напрямків, не приймалася в другому. Так, у чисельних методах розв'язання систем лінійних рівнянь добре відомий вплив властивостей вихідної матриці на стійкість одержаних результатів [2], але в класичному регресійному аналізі це вважається несуттєвим. Концепція *D*-, *A*-, *E*-, *G*- оптимальності багатофакторних планів експериментів не доводиться до прикладного використання в задачах високої складності; використання системи ортогональних поліномів Чебишева в регресійному аналізі практично зустрічається досить рідко; перевірки одержаних статистичних моделей за критеріями їхньої якості не завжди проводяться тощо.

Основними невирішеними проблемами в області планування експерименту й рег- ресійного аналізу є [3, с. 14-16]:

- формування найкращих вихідних умов одержання багатофакторних статистичних моделей у вигляді послідовного планування багатофакторних експериментів з використанням робастних планів експериментів – залежно від особливостей задачі необхідно вибирати процедури обробки даних і їхні параметри на кожному етапі рішення, що часто може виявитися під силу тільки фахівцеві-статистику;
- відсутність ефективних алгоритмів виділення семантичної структури рівняння регресії – користувач зацікавлений в одержанні структури рівняння регресії, яка повинна відповідати істинній структурі зв'язків, статистики ж в більшості випадків говорять про "найбільш інформативну підмножину регресорів";
- стійка оцінка коефіцієнтів моделі в умовах вихідної мультиколлинеарности факторів.

Концепція розв'язання

Розглянемо один із широко застосовуваних методів прикладної статистики – регрессійний аналіз. Одна з невирішених проблем статистики – це визначення частинної структури рівняння регресії або, як соромливо кажуть статистики, – визначення найкращої підмножини регресорів. Проблема тут принципова – ті, хто користуються статистикою, уважають, що рівняння регресії повинне відбивати "істинну" структуру зв'язків у досліджуваному процесі або системі й бути придатним для якісного аналізу явищ, що відбуваються. Статистики (переважна більшість) говорять лише про деякий набір статистичних властивостей і не хочуть розглядати одержання "істинної" структури. У підсумку створюється суперечлива ситуація між одержаним результатом і прикладними вимогами до нього.

Формалізований пошук структурного зв'язку факторів з відгуком базується на теоремі К. Вейєрштраса, а також теоремах М. Стоуна й Д. Джексона [3, с. 171].

При одержанні багатофакторної математичної моделі необхідно вибрати вид зв'язку (структуру) між умовами експерименту – факторами $X_1, X_2, \ldots, X_k$ – і його результатами – відгуками $y_1, y_2, \ldots, y_w, \ldots, y_m$ модельованої системи.

$$\hat{y}_w = f_w(X_1, X_2, \ldots, X_k),$$

де $\hat{y}_w$ – модель w-го відгуку системи (процесу, об'єкта);

f_w – структура математичної моделі;

k – загальна кількість факторів.

У більшості розв'язуваних прикладних завдань вид структури статистичної моделі дослідникові заздалегідь не відомий.

В [3, с. 106–108] запропонована й обґрунтована формалізована структура багатофакторної поліноміальної математичної моделі, лінійної щодо параметрів, у вигляді множини ефектів схеми повного факторного експерименту:

$$(1+X_1+X_1^2+\ldots+X_1^{s_1-1})\times\ldots\times(1+X_k+X_k^2+\ldots+X_k^{s_k-1}) \to N_{\Pi}, \qquad (1)$$

де 1 – значення фіктивної незалежної змінної $x_0 \equiv 1$;

$X_1, \ldots, X_k$ – фактори шуканої математичної моделі в натуральних значеннях;

$s_1, \ldots, s_k$ – кількість рівнів факторів $X_1, \ldots, X_k$ відповідно;

k – загальна кількість факторів;

N_{Π} – кількість дослідів повного факторного експерименту, рівна кількості структурних елементів його схеми $N_{\Pi\,СТР}$.

Вираз (1) використовується для створення системи базисних функцій. Він являє собою сукупність всіх багаточленів факторів $X_1, X_2, \ldots, X_i, \ldots, X_k$, кожний ступеня від 1 до найвищого s_i-1, і різних взаємодій цих факторів (точніше ступенів цих факторів) по два, три і далі до максимально можливих взаємодій з k різних елементів. Сукупність всіх зазначених елементів утворює дійсний евклідів простір.

У виразу (1) наведені всі можливі ефекти в поліноміальному виді, необхідні й достатні для адекватної апроксимації вихідних даних схем повних і дробових факторних експериментів.

При переході від натуральних значень факторів $X_1, \ldots, X_k$ до системи ортогональних поліномів Чебишева (системи ортогональних контрастів) структура математичній моделі має вигляд

$$(1+x_1^{(1)}+x_1^{(2)}+\ldots+x_1^{(s_1-1)})\times\ldots\times(1+x_k^{(1)}+x_k^{(2)}+\ldots+x_k^{(s_k-1)}) \to N_{\Pi},$$

де $x_1^{(1)}, \ldots, x_1^{(s_1-1)}; \ldots; x_k^{(1)}, \ldots, x_k^{(s_k-1)}$ – ортогональні контрасти факторів X_1, ..., X_k ; опис інших позначень наведено вище.

Кількість структурних елементів $N_{\text{П СТР}}$ схеми повного факторного експерименту дорівнює кількості дослідів повного факторного експерименту $N_{\text{П}}$: $N_{\text{П СТР}} = N_{\text{П}}$. Всі ефекти для схеми повного факторного експерименту ортогональні один до одного. Отже, будь-який повний факторний експеримент описується стійкою математичною моделлю.

У дробовому факторному експерименті кількість дослідів $N_{\text{Д}}$ менше кількості структурних елементів $N_{\text{П СТР}}$ і, отже, з $N_{\text{П СТР}}$ можливих базисних функцій необхідно вибрати k' ($k' < N_{\text{Д}}$) для одержання шуканої математичної моделі. З приведеного виходить, що у дробовому факторному експерименті деякі ефекти (взаємодії) – будуть корельовані з іншими ефектами, і задача стає, у загальному випадку, некоректно поставленою.

Обчислювальний експеримент

Для ілюстрації складності визначення структури рівняння регресії приведемо класичний приклад, відомий практично всім, хто використовував або вивчав регресійний аналіз – регресію Дж. Лонглі (J.W. Longley) [5, с. 109–114].

Це регресія від шести факторів: X_1 – дефлятор (індекс) цін (%); X_2 – валовий національний продукт (млрд. дол.); X_3 – загальне число безробітних (тис. чол.); X_4 – кількість військовослужбовців (тис. чол.); X_5 – непрацююче населення від 14 років (тис. чол.); X_6 – рік; відгук y – загальна кількість зайнятих в економіці США (тис. чол.). У постановці завдання сьомим фактором Дж. Лонглі назвав фіктивний фактор x_0, тобто $x_0 = X_7 \equiv 1$.

Вихідні дані за значеннями факторів $X_1 ... X_6$ і відгуку y наведені в табл. 1

Таблиця 1. Вихідні дані для визначення регресії Лонглі

X_1	X_2	X_3	X_4	X_5	X_6	y
83	234,289	2356	1590	107608	1947	60323
88,5	259,426	2325	1456	108632	1948	61122
88,2	258,054	3682	1616	109773	1949	60171
89,5	284,599	3351	1650	110929	1950	61187
96,2	328,975	2099	3099	112075	1951	63221
98,1	346,999	1932	3594	113270	1952	63639
99	365,385	1870	3547	115094	1953	64989
100	363,112	3578	3350	116219	1954	63761
101,2	397,469	2904	3048	117388	1955	66019
104,6	419,18	2822	2857	118734	1956	67857
108,4	442,769	2936	2798	120445	1957	68169
110,8	444,546	4681	2637	121950	1958	66513
112,6	482,704	3813	2552	123366	1959	68655

114,2	502,601	3931	2514	125368	1960	69564
115,7	518,173	4806	2572	127852	1961	69331
116,9	554,894	4007	2827	130081	1962	70551

Дані були взяті за період з 1947 по 1962 рік. Виявилося, що матриця даних погано обумовлена: число обумовленості cond $\approx 4{,}8\times10^{9}$ [6, с. 313].

Розглянемо табл. 2. З неї добре видно, що більшість факторів корельовані один з одним з величиною коефіцієнта парної кореляції r_{ij}, близьким до 1. Крім того, ці коефіцієнти більші, ніж коефіцієнти кореляції факторів з відгуком. Це говорить про те, що структура моделі нестійка [7–10].

Таблиця 2. Кореляційна матриця для регресії Лонглі

	Значення r_{ij}						
	y	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0,970899	1					
X_2	0,983552	0,991589	1				
X_3	0,502498	0,620633	0,604261	1			
X_4	0,457307	0,464744	0,446437	–0,17742	1		
X_5	0,960391	0,979163	0,99109	0,686552	0,364416	1	
X_6	0,971329	0,991149	0,995273	0,668257	0,417245	0,993953	1

При розрахунку за різними програмами і на різних ЕОМ оцінки коефіцієнтів регресії виходили не тільки з різними значеннями, але й з різними знаками. За допомогою спеціального обчислювального пристрою були виконані розрахунки з 40 значущими цифрами й одержана наступна модель [5, с. 110]:

$$\hat{y} = 15{,}0619X_1 - 0{,}0358X_2 - 2{,}0202X_3 - 1{,}0332X_4 - 0{,}0511X_5 + 1829{,}15X_6 - 3482258{,}635. \quad (2)$$

Ця регресія дотепер застосовується для перевірки точності роботи програм по регресійному аналізу. І це незважаючи на те, що деякі теоретики показали її безглуздість (A.E. Beaton, D.B. Rubin, J.L. Barone в 1976 р. [11] и Є.З. Демиденко в 1981 р. [5, с. 111, 114]), не пропонуючи, щоправда, іншого рішення.

Авторами ця задача вирішувалася за допомогою програмного засобу "Планування, регресія і аналіз моделей" (ПЗ ПРІАМ) [12], розробленого авторами, в автоматичному режимі, тобто структура рівняння регресії формувалася програмою. Була отримана модель

$$\hat{y} = 65317{,}00 + 5810{,}43x_2, \quad (3)$$

де x_2 = 0,00598102(X_2 – 387,698).

Всі інші фактори X_1, X_3, …, X_6 не були включені в математичну модель, тому що участь у моделі кожного ефекту було обмежено по мінімальному коефіцієнту кореляції з відгуком значенням 0,01 і по мінімальній частці розсіювання значенням 0,005. Після включення в рівняння регресії лінійного ефекту фактора X_2 зазначені умови для факторів X_1, X_3, …, X_6 не виконувалися.

Аналіз кореляційної матриці для регресії Лонглі показує, що кожний з факторів X_1, X_3, X_5, X_6 статистично взаємозв'язаний з y менше, ніж з фактором X_2. Фактор X_4 статистично пов'язаний з фактором X_2 приблизно так само, як з y. Із цієї причини введення в модель усіх факторів, крім X_2, викликає сумнів з позиції їхніх причинно-структурних зв'язків з y: фактори X_1, X_3, …, X_6 виражають лінійно перетворений фактор X_2, тобто виражають під різними назвами величину фактора X_2.

Якщо ефекти x_i, x_j ($1 \le i < j \le k$; k – загальна кількість факторів) статистично взаємозв'язані, то дисперсії коефіцієнтів b_i, b_j для відповідних ефектів будуть рівні

$$D(b_i; b_j) = \frac{\sigma^2_{\text{від}}}{N(1 - R^2_{ij})},$$

де $\sigma^2_{\text{від}}$ – дисперсія відтворюваності результатів експериментів;

N – загальна кількість проведених експериментів;

R^2_{ij} – квадрат коефіцієнту множинної кореляції між ефектами x_i, x_j і всіма іншими ефектами в одержаній моделі.

При $R^2_{ij} \to 1$ множник $[1/(1 - R^2_{ij})] \to \infty$ і дисперсії коефіцієнтів необмежено збільшуються. При $R^2_{ij} \to 0$ множник $[1/(1 - R^2_{ij})] \to 1$ і при $R^2_{ij} = 0$, тобто у випадку ортогональності всіх ефектів один до одного, $[1/(1 - R^2_{ij})] = 1$. Тоді дисперсія коефіцієнтів моделі b_i, b_j стає мінімально можливої, або оцінка ефективна в статистичному сенсі.

Для моделі (3) число обумовленості інформаційної матриці Фішера X^TX cond = 1. Модель пояснює 96,74 % розсіювання. Коефіцієнт множинної кореляції R = 0,984, критерій Бокса й Веца дорівнює 5, тобто інформативність моделі висока. За цією моделлю похибка апроксимації перебуває в межах 0,36...2,13 %, тоді як за моделлю Дж. Лонглі 13,45...28,43%, тобто на порядок більше.

На рис. 1. наведено діаграму розсіювання для y і X_2, з якої видно, що одержана модель (3) стійка й придатна для прогнозування, навіть якщо вона буде побудована по частині даних.

Рис. 1. Діаграма розсіювання для кількості зайнятих в економіці США і валового національного продукту

Побудуємо, наприклад, лінійну модель по першим вісьмох роках 1947...1954 ($\hat{y}$ = 62301,6 + 2384,4 x_2, де x_2 = 0,0141211(X_2 – 305,105)), і використаємо її для завбачення наступних восьми років 1954...1962 (рис. 2). Залежність по осі ?ординат не від року, а від істинного фактора – валового національного продукту (X_2).

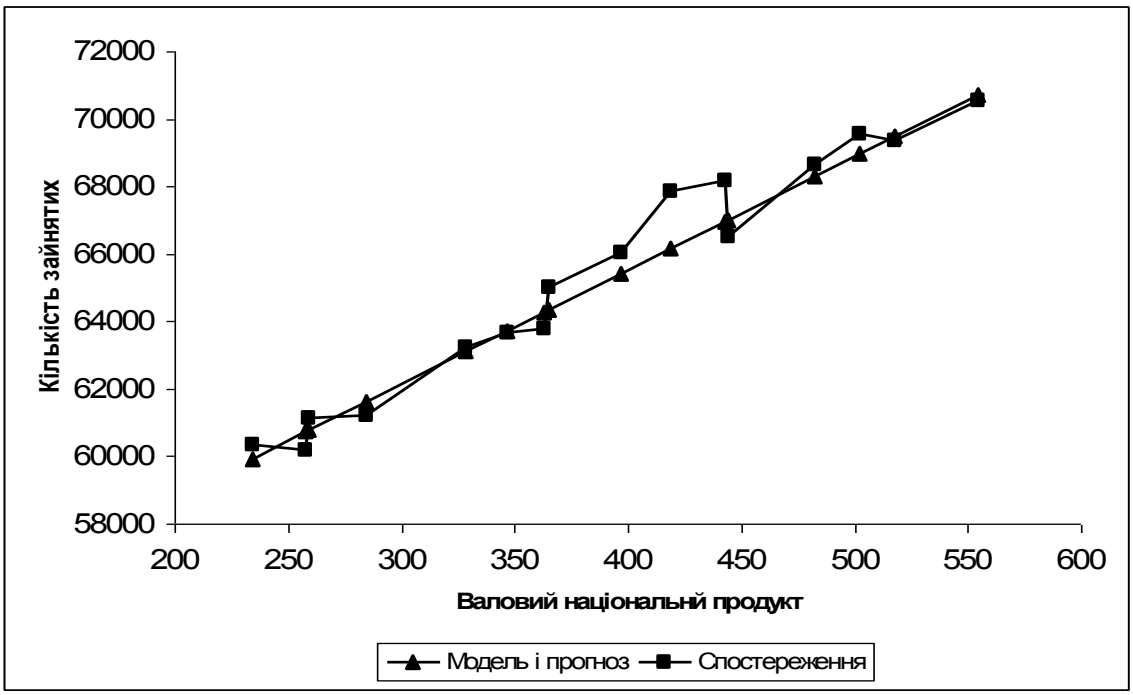

Рис. 2. Лінійна модель і експеримент

Як бачимо, цілком задовільні передбачаючі можливості (помилка завбачення від 0,21 % до 2,52 %), та ще на досить тривалий проміжок часу.

Цей приклад ясно показує, що потрібно вирішувати завдання, а не застосовувати методи, інакше ми одержуємо відповідь на питання, що не задавали, а наше питання залишається без відповіді.

Проблема в даному випадку в тому, що пропущено етап специфікації (визначення структури) моделі й відразу виконується ідентифікація, що неприпустимо [8, 9, 13, 14].

Зрозуміло, для рішення таких задач необхідні спеціальні алгоритми й програмні засоби. Розроблені алгоритми визначення структури рівняння регресії [3, 4, 7–10] і інформаційна технологія побудови математичних моделей по експериментальним даним [12] були створені й успішно апробовані на розв'язанні декількох сотень технічних, технологічних, матеріалознавських, вимірювальних, економічних і інших задач.

Висновки

1. Регресійні моделі одержують за вихідними даними, що містять випадкові похибки. Метод найменших квадратів чутливий до порушень передумов множинного регресійного аналізу, зокрема, до мультиколінеарності факторів. Необхідно забезпечити стійкість початкових умов і обчислювального методу одержання структури й коефіцієнтів моделі.

2. Основний підхід у забезпеченні одержання стійких структур статистичних моделей і їхніх коефіцієнтів полягає в стійкому (робастному) плануванні, тобто у виборі початкових умов, плану багатофакторного експерименту. Однак у реальних задачах таке планування експерименту не завжди можливе.

3. При відсутності можливості планувати експеримент необхідно в структуру статистичної моделі включати тільки ортогональні або слабко корельовані ($r_{ij}(x_i, x_j) \le 0{,}3\ldots0{,}4$) ефекти, використовувати систему ортогональних поліномів Чебишева, перетворювати вихідні дані й наводити ефекти в ортогональному виді.

4. Формальне використання методу найменших квадратів без логічно-професійного аналізу умов одержання статистичних моделей, у загальному випадку, не може дати гарних результатів. Численні спроби формального рішення регресії Лонглі різними дослідниками не дали позитивних результатів.

Література

1. Загоруйко Н.Г., Орлов А.И. Некоторые нерешенные математические задачи прикладной статистики // Современные проблемы кибернетики. (Прикладная статистика). – М.: Знание, 1981. – 64 с.
2. Райс Дж. Матричные вычисления и математическое обеспечение / Пер. с англ. О.Б. Арушаняна; Под ред. В.В. Воеводина. – М.: Мир, 1984. – 264с.
3. Радченко С.Г. Устойчивые методы оценивания статистических моделей: Монография. – К.: ПП «Санспарель», 2005. – 504 с.
4. Радченко С.Г. Формализованный поиск структуры многофакторного уравнения регрессии в технологических исследованиях // Надежность режущего инструмента и оптимизация технологических систем: Сб. ст. В 2 т. / Донецк. Гос. машиностроит. акад.; Предс. редсовета Г.Л. Хает. – Краматорск, 1997. Т. 2. – С. 40–46.
5. Демиденко Е.З. Линейная и нелинейная регрессии. – М.: Финансы и статистика, 1981. – 302 с.
6. Себер Дж. Линейный регрессионный анализ / Пер. с англ. В.П. Носко; Под. ред. М.Б. Малютова. – М.: Мир, 1980. – 456 с.
7. Лапач С.Н. Определение структуры уравнения регрессии при неортогональной матрице эксперимента // Применение вычислительной техники и математических методов в научных исследованиях. – К.: – 1986. – С. 95–97.
8. Лапач С.Н., Пасечник М.Ф., Чубенко А.В. Статистические методы в фармакологии и маркетинге фармацевтического рынка. – К.: ЗАО «Укрспецмонтажпроект», 1999. – 312 с.
9. Лапач С.Н. Проблемы построения математических моделей экспериментально-статистическими методами // Прогресивна техніка і технологія машинобудування, приладобудування і зварювального виробництва. Праці НТУУ «КПІ», – Т. 2, –К.: НТУУ «КПІ», – 1998. – С. 25–29.
10. Лапач С.Н. Оптимизация режимов обработки жаропрочных никелевых сплавов инструментом из сверхтвердых материалов // Надежность режущего инструмента и оптимизация технологических систем: Сб. ст. В 2 т. / Донецк. Гос. машиностроит. акад.; Предс. редсовета Г.Л. Хает. – Краматорск, 1997. Т. 2. – С. 122–128.
11. Beaton A.E., Rubin D.B., Barone J.L. The acceptability of regression solutions: another look at computational accuracy. – JASA, 1976, v. 71, № 353.
12. Лапач С.Н., Радченко С.Г., Бабич П.Н. Планирование, регрессия и анализ моделей PRIAM (ПРИАМ). SCMC–90; 325, 660, 668 // Программные продукты Украины: Каталог. = Software of Ukraine: Catalog. – К., 1993. – С. 24–27.
13. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей: Справ. изд. / Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1985. – 487 с.

14. Трофимов В.П. Логическая структура статистических моделей. – М.: Финансы и статистика, 1985. – 191 с.