DOI: https://doi.org/10.32839/2304-5809/2019-6-70-30
УДК 681.5.015

Архіпова С.А.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

ПРО НЕМОЖЛИВІСТЬ ВІДНОВЛЕННЯ РОЗПОДІЛУ ПОХИБОК ВИХІДНИХ ДАНИХ В РЕГРЕСІЙНИХ МОДЕЛЯХ

Анотація. Розглянуто можливість отримання помилкових результатів при розв'язку задачі ідентифікації через втрати інформації, що мають місце при переході до параметризованого опису розподілу помилки вимірювань та оцінок параметрів моделі. Для вибору оптимального методу обробки використовується параметризоване представлення (модель) похибки, яка дозволяє використати її для виконання різного роду висновків і досліджень, вироблених аналітично, зокрема, оптимізації процедур статистичної обробки даних, включаючи структурно-параметричну ідентифікацію. При розв'язанні практичних задач значення шумів або параметри їх розподілу апріорно невідомі, що не дозволяє фактично, а не формально застосовувати засоби класичного регресійного аналізу.

Ключові слова: регресійний аналіз, похибка вимірювання, нормальність розподілу, методи оцінювання параметрів.

Arkhipova Sofiia
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

ABOUT THE IMPOSSIBILITY OF THE ERRORS DISTRIBUTION RESTORING OF INPUT DATA IN REGRESSION MODELS

Summary. The possibility of deriving of erroneous outcomes is considered at a solution of a task of identification because of loss of an information at passage to parameterized to exposition of an error distribution of measurements and estimations of parameters of a model. The task of optimization of the procedure of model parameters estimating by a sample data which is characterized by a some set of properties, in particular, a certain probability model of error, is sufficiently developed and usually solved within the framework of a general approach. Usually, a parameterized representation (model) of the error is used for the synthesis (inference) of the optimal method. This representation of error is often presented in the form of a probability distribution density. The parametrized form of the model representation makes it possible to use it for the implementation of various conclusions and analyzes performed analytically, in particular, for the optimization of procedures of the statistical data processing, including structural-parametric identification. However, the actual optimality of data processing is achieved only if the adopted parameterized model sufficiently fully reflects the basic properties and characteristics of the error. That is, in the case of absence of apriority information about the type of distribution, the problem of its a posteriori estimation is arises, in which the error distribution is restored through the distribution of residuals, and it is assumed that those distributions are close to each other. To obtain the residuals, preliminary estimates of the dependent variable are calculated, which, after optimizing the parameter estimation procedure, can be further refined. In this regard, it is advisable to investigate the possibility of selecting of parametrized models of error based on the results of a posteriori analysis of residuals of actual values, which are the estimates of the initial error of measurement.

Keywords: regression analysis, measurement error, normality of distribution, methods of parameter estimation

Постановка проблеми. Рішення задачі структурно-параметричної ідентифікації в класі лінійних за параметрами моделей, що реалізується тільки за допомогою залучення засобів класичного регресійного аналізу, в разі відмінності характеру розподілу від нормального призводить до вельми невизначених результатів.

Аналіз останніх досліджень і публікацій. У традиційному регресійному аналізі вектор оцінок параметрів моделі розраховується методом найменших квадратів (МНК) [2; 4]. Ефективність і незміщеність одержуваних МНК-оцінок залежить від нормальності розподілу похибки вихідних даних [2; 7; 8].

Виділення невирішених раніше частин загальної проблеми. На жаль, при розв'язанні практичних задач значення шумів або параметри їх розподілу апріорно невідомі, що не дозволяє фактично, а не формально застосовувати засоби класичного регресійного аналізу.

Доведення цього положення і є головною **метою цієї статті**.

Виклад основного матеріалу. Розглянемо задачу ідентифікації лінійної за параметрами регресійної моделі $\mu(A, x_j) = \sum_{j=1}^{m} a_j x_j$, $[a_1, ..., a_m]^T = A$ за емпірично отриманими даними, що включають матрицю плану $[x_{ij}]$, $i = \overline{1, n}$, $j = \overline{i, m}$ й вектор зашумлених значень залежної змінної $[z_1, ..., z_n]^T$; $z_i = y_i + e_i$, де e_i – значення центрованої випадкової величини E з обмеженою дисперсією σ_e^2.

Задача оптимізації процедури оцінювання параметрів моделі за вибіркою вихідних даних, що характеризуються певним набором властивостей, зокрема, деякою імовірнісною моделлю похибки E, достатньо розроблена й зазвичай вирішується в рамках загального підходу [1; 2; 3], в якому вибір (синтез) методу параметричної ідентифікації ω_{opt} – це результат вирішення оптимізаційної задачі, в якій ω_{opt} визначається у такий спосіб, що при відомих імовірнісних властивос-

тях похибки E одержувані цим методом оцінки (вектор $\tilde{A}_{opt}$) мають у деякому сенсі найкращі характеристики.

Якщо підхід до синтезу (виводу) оптимального методу оцінювання ґрунтується на принципі максимуму правдоподібності [4; 5], то для лінійної по параметрах моделі при нормальному розподілі похибки E максимум функції правдоподібності відповідає мінімуму показника

$$S^2 = \sum_{i=1}^{n} [z_i - \mu(\tilde{A}, x_{ij})]^2, \qquad (1)$$

де $\varepsilon = z_i - \mu(\tilde{A}, x_{ij})$ – нев'язка моделі, що в остаточному підсумку приводять до вибору в якості оптимального методу ω_{opt} метод найменших квадратів (МНК) [4; 5]. Якщо $f(e)$ описується законом Лапласа, то при всіх інших незмінних умовах вираз для S^2 набуде вигляду

$$S^2 = \sum_{i=1}^{n} | z_i - \mu(\tilde{A}, x_{ij}) | \qquad (2)$$

і оптимальним виявляється метод найменших модулів (МНМ), що мінімізує (2).

Якість параметричної ідентифікації звичайно оцінюється коваріаційною матрицею $D\{\tilde{A}\}$ оцінок коефіцієнтів (або її діагональними елементами – дисперсіями оцінок) і середнім значенням квадрата нев'язки моделі [6; 7].

Як правило, для синтезу (виводу) оптимального методу ω_{opt} використовується параметризоване представлення (модель) похибки E, часто у формі розподілу щільності ймовірності $f(e)$ [8]. Параметризована форма представлення моделі дозволяє використати її для виконання різного роду висновків і досліджень, вироблених аналітично, зокрема, оптимізації процедур статистичної обробки даних, включаючи структурно-параметричну ідентифікацію. Однак фактична оптимальність обробки досягається лише в тому випадку, якщо прийнята параметризована модель досить повно відображає основні властивості і характеристики похибки E. Тобто у випадку відсутності апріорної інформації про вид розподілу виникає задача його апостеріорного оцінювання, у якій розподіл $f(e)$ відновлюється через розподіл нев'язок $\varepsilon_i = z_i - \tilde{y}_i$, $i = \overline{1,n}$, у припущенні, що $f(e) \approx f(\varepsilon)$ [10], де $\tilde{y}_i = \mu(\tilde{A}, x_{ij})$ розраховані яким-небудь чином попередні оцінки залежної змінної, які після оптимізації процедури оцінювання параметрів можуть бути додатково уточнені. У зв'язку з цим доцільно дослідити можливості підбору параметризованих моделей похибки E за результатами апостеріорного аналізу залишків $\{\varepsilon_i\}$, $i = \overline{1,n}$, фактичних значень, які є оцінками вихідної похибки вимірів $\{e_i\}$.

Дослідження проведемо на тестовому прикладі, задавши модель $\propto$ співвідношенням:

$$y = a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + a_5x_1x_2^2, \qquad (3)$$

де в якості шуму використано вихід генератора псевдовипадкової величини E, що має в одному випадку рівномірний розподіл $f(e) = 1/2\Delta_e$, а в другому – розподіл $f(e)$, проміжне між розподілами Гауса $f^{(Г)}(e)$ та Лапласа $f^{(Л)}(e)$ (рис. 1), причому дисперсії σ_e^2 обох розподілів однакові. Згенеруємо L наборів вихідних даних виду $[z_{il}, x_{i1}, ..., x_{im}]$, де $z_{il} = y_i + e_{il}$, $l = \overline{1,L}$, що відрізняються тільки вибірками шумів $[e_{il}]$. Оцінимо за критерієм χ^2 близькість розподілів елементів вибірок розподілам $f^{(Г)}$ й $f^{(Л)}$, увівши наступні процентні співвідношення:

$r_Г$ – відносне число вибірок, для яких із двох конкуруючих гіпотез розподілу (Гауса й Лапласа) прийнята перша, тобто для яких справедливі співвідношення: $\chi^2_{(Г)} < \chi^2_{(Л)}$ і $\chi^2_{(Г)} \le \chi^2_{g\%}$, де рівень значимості $g\%$ не нижче 0.1%, тобто $r_Г = l_Г 100\% / L$, $l_Г$ – абсолютне число вибірок, для яких виконані зазначені співвідношення;

$r_Л$ – відносне число вибірок, для яких із двох конкуруючих гіпотез приймається гіпотеза розподілу Лапласа. Для таких вибірок справедливі співвідношення $\chi^2_{(Л)} < \chi^2_{(Г)}$, $\chi^2_{(Л)} \le \chi^2_{g\%}$, $r_Л = l_Л 100\% / L$, де $l_Л$ – абсолютне число вибірок, для яких виконуються наведені співвідношення;

r_- – відносне число вибірок, для яких відкидаються обидві гіпотези, що перевіряються: $\chi^2_{(Г)}$, $\chi^2_{(Л)} > \chi^2_{g\%}$.

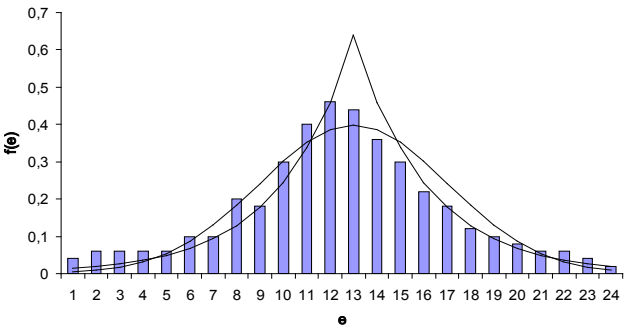

Рис. 1. Гістограмний розподіл вихідної похибки

Будемо розглядати в якості множини параметризованих моделей, що описують властивості похибки E, розподіли Гауса та Лапласа. Результати вибору параметризованої моделі на сукупності наборів похибок $\{e_i\}$, $\{\varepsilon_i\}_l$, $i = \overline{1,n}$, $l = \overline{1,L}$ опишемо трійкою значень $r_Г$, $r_Л$, r_-, уведених вище. Задача дослідження – визначити можливість відновлення значень $r_Г$, $r_Л$, r_- за результатами апостеріорного аналізу оцінок похибок $\{\varepsilon_i\}_l$. Очевидно, що ця можливість залежить від близькості оцінки $\{\varepsilon_i\}_l$ до відповідної фактичної вибірки похибок $\{e_i\}_l$, що у свою чергу визначається близькістю значень y_i та $\tilde{y}_i = \mu(\tilde{A}, x_{ij})$, тобто точністю знаходження модельних значень $\tilde{y}_i$.

Результати виконаних досліджень представлені в табл. 1.

У першому рядку табл. 1. з ідентифікатором $\{e_i\}$, містяться кількісні дані про істинний розподіл значень $r_Г$, $r_Л$, r_-, отриманих при перевірці гіпотез $f(e) = f^{(Г)}$, $f(e) = f^{(Л)}$ на множині вихідних даних $\{e_i\}_l$, $l = \overline{1,L}$, що можливо тільки в тестовому дослідженні. Наступні рядки отримані з результатів апостеріорного аналізу залишків $\{\varepsilon_i\}_l$, які є оцінками деякого шуму $\{e_i\}_l$, апріорно невідомого дослідникові. Одержувані значення $\varepsilon_i = z_i - \tilde{y}_i$ залежать винятково від точності знаходження модельних значень $\tilde{y}_i$, що визначається:

а) застосовуваним методом параметричної ідентифікації моделі μ: $\tilde{y}_i = \mu(x_{1i}, x_{1i}, ...) = a_0 + a_1x_{1i} + a_2x_{2i} + ...$, у цьому випадку застосовувалися методи найменших квадратів (МНК), найменших модулів (МНМ), метод середніх (МС) [9];

б) рівнем складності моделі μ, зокрема, ступенем ускладнення або спрощення моделі $\propto(x_{1i}, x_{1i}, ...)$ щодо істинної моделі $\propto_{ icm}$, обумовленої співвідношенням (3), $\mu_{уск} = a_1x_1 + a_2x_2 + a_3x_1x_2 + a_4x_1^2 + a_5x_2^2 + a_6x_1x_2^2$, $\mu_{спр} = a_1x_1 + a_2x_2$;

в) характером розподілу $f(e)$: перші три стовпці відповідають E із розподілом, що задані мал. 1, три наступних отримані для іншого типу апріорно невідомого розподілу, у якості якого використане рівномірний розподіл похибки E з $f(e) = 1/2\Delta_e$, дисперсія в обох випадках дорівнює $D(E) = \sigma_e^2$.

Висновок про можливість підбору параметризованої моделі похибки E за результатами апостеріорного аналізу залишків у першу чергу базується на збігу (близькості) значень трійки ($r_Г$, $r_Л$, r_-), розрахованих за фактичними значеннями вибірок $\{e_i\}$ похибки E та за оцінками $\{\varepsilon_i\}$. Останні відповідно залежать від заздалегідь невідомого виду розподілу $f(e)$, обраного методу параметричної ідентифікації та рівня складності моделі μ, тобто від факторів, що безпосередньо підлягають оптимізації на основі шуканої інформації про вид розподілу $f(e)$.

У зв'язку із цим у загальному випадку задача параметризації моделі за результатами апостеріорного аналізу оцінок $\{\varepsilon_i\}$ не може мати надійного рішення [1]. Більш того, можна вважати, що введення будь-якого модельного опису похибки супроводжується втратою частини інформації про її властивості та характеристики.

Тому орієнтований на прийняту модель даних оптимальний метод оцінювання параметрів на ділі призводить до субоптимальних результатів, причому характеристики оцінок, розраховані в рамках постульованої моделі даних, а саме: моменти та розподіли оцінок, їхня збіжність і т.п., можуть абсолютно не відображати реальну якість оцінок.

На відміну від традиційного (мал. 1) більш реалістичним представляється підхід [1], у якому метод ω_{opt} визначається як кращий за результатами аналізу якості рішень розглянутої реальної задачі, отриманих застосуванням ряду різних методів $\{\omega_1, ..., \omega_G\} = \{\omega_g\} = \Omega$, пріоритет яких підтверджений багаторазовою практичною апробацією. Ключовими при цьому стають питання вибору (формування) показника якості параметричної ідентифікації й обчислення цього показника за наявними даними, причому процедура його обчислення повинна бути вільна від виду розподілу похибки E, тобто знаходження чисельних значень показника не залежить від виду та параметрів розподілу $f(e)$, на практиці невідомого оброблювачу. Графічно цей підхід, який можна назвати прагматичним, зображений на мал. 1. Хоча вибір кращого ω^+ методу обробки із множини наявних $\Omega = \{\omega_g\}$ цілком залежить від об'єктивності застосовуваного критерію якості Q, ефективність знайдених оцінок у цілому визначається підбором методів, що становлять множину Ω. Проблема підбору методів звичайно пов'язана зі змістом та особливостями розв'язуваних задач і виходить за рамки даної роботи, накладаючи на показник якості Q одну додаткову умову: показник Q не повинен залежати від виду застосовуваних методів обробки, тобто повинен бути зовнішнім стосовно множини Ω.

Розглянута вище проблема, хоча і в дещо іншому формулюванні, проявляється і при вирішенні задач структурно-параметричної ідентифікації. Йдеться про об'єктивність результатів, одержуваних при застосуванні до оцінювання значущості коефіцієнтів регресії, інтервального підходу. Суть підходу у визначенні інтервалу Δa_j, в якому із заданою ймовірністю P може перебувати знайдена оцінка коефіцієнта $\tilde{a}_j$, і положення інтервалу на вісі дійсних чисел: якщо цей інтервал перекриває точку 0, правдоподібна гіпотеза про незначущість даного коефіцієнта і, отже, можливе виключення відповідної змінної x_j із структури моделі.

Вихідною інформацією для знаходження інтервалу Δa_j може служити гістограмна оцінка розподілу $\tilde{f}(\tilde{a}_j)$, одержувана, наприклад, методом варіювання вихідних даних [1], у результаті застосування якого утворяться вибірки $\{\tilde{a}_j\}$, $j = \overline{1,m}$, за якими і визначаються оцінки $\tilde{f}(\tilde{a}_j)$. Гістограмна оцінка апроксимується моделлю $f_M(\tilde{a}_j)$ – відомим з літератури типом безперервного розподілу, на підставі якого аналітично або за допомогою

Таблиця 1

Результати апостеріорного аналізу нев'язок при варіюванні методу ідентифікації та складності моделі

Умови параметризації			$f(e)$			$1/2\Delta_e$		
			$r_Г$	$r_Л$	r_-	$r_Г$	$r_Л$	r_-
$\{e_i\}$			**42**	**42**	**16**	**8**	**0**	**92**
$\{\varepsilon_i\}$	МНК	$\propto_{спр}$	20	71	9	54	0	46
		$\propto_{ icm}$	43	50	7	28	0	72
		$\propto_{уск}$	53	40	7	30	0	70
	МНМ	$\propto_{спр}$	16	70	14	38	1	61
		$\propto_{ icm}$	29	56	15	54	0	46
		$\propto_{уск}$	60	30	10	59	0	41
	МС	$\propto_{спр}$	100	0	0	100	0	0
		$\propto_{ icm}$	9	52	39	9	47	44
		$\propto_{уск}$	2	48	50	4	39	57

таблиць для заданої ймовірності P розраховується довжина інтервалу Δa_j. Для одномодальних розподілів $f(\tilde{a}_j)$, що звичайно зустрічаються в практиці обробки, інформація, одержувана з вибірки $\{\tilde{a}_j\}$, досить повно описує поведінку функції $f(\tilde{a}_j)$ лише в її центральній частині. Відомості про поведінку $f(\tilde{a}_j)$ на «хвостах» розподілу, особливо при малих m, вкрай бідні, тому при підборі моделі $f_M(\tilde{a}_j)$ дослідник фактично домислює периферійну частину («хвости») модельного розподілу. Тим часом саме «хвости» у першу чергу визначають довжину інтервалу Δa_j: задавши «тонкі хвости» (вибравши в якості $f_M(\tilde{a}_j)$ розподіл Гауса), одержимо свідомо більш короткий інтервал Δa_j, ніж для моделі з «товстими хвостами» (розподіл Лапласа), хоча рівень обох моделей від $f(\tilde{a}_j)$ може виявитися приблизно однаковий [11]. Таким чином і в цьому випадку заміна вихідної непараметризованої оцінки $\tilde{f}(\tilde{a}_j)$ деяким параметризованим розподілом $f_M(\tilde{a}_j)$ може призвести до фактичної невизначеності кінцевих результатів.

Висновки і пропозиції. Результати виконаних досліджень дозволяють зробити наступні висновки:

– процедура класичного регресійного аналізу не містить способів надійної селекції структури і оцінки параметрів моделі при відсутності апріорних відомостей про нормальність похибки E;

– апостеріорне твердження про можливість прийняття гіпотези нормальності залишків регресії не гарантує нормальності вихідної похибки E і, отже, не може служити обґрунтуванням прийняття положень регресійного аналізу, що спираються на нормальність розподілу E.

Список літератури:

1. Архипов А.Е., Архипова С.А. Анализ и оптимизация качества решения задачи идентификации. Праці П'ятої Української конференції з автоматичного управління «Автоматика-98»: Київ, 13-16 травня 1998 р. Ч. 3. Київ : видавництво НТУУ «Київський політехнічний інститут», 1998. С. 9–15.
2. Прикладная статистика: Классификация и снижение размерности : Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. Москва : Финансы и статистика, 1989. 607 с.
3. Сильвестров А.Н., Чинаев П.И. Идентификация и оптимизация автоматических систем. Москва : Атомэнергоиздат, 1987. 280 с.
4. Мудров В.И., Кушко В.Л. Методы обработки измерений. Москва : Сов. радио, 1976. 190 с.
5. Соболев В.И. Информационно-статистическая теория измерений. Учебник для вузов. Москва : Машиностроение, 1983. 224 с.
6. Льюнг Л. Идентификация систем. Теория для пользования. Москва : Наука, 1991. 432 с.
7. Себер Дж. Линейный регрессионный анализ. Москва : Мир, 1980. 456 с.
8. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. Москва : Финансы и статистика, 1983. 471 с.
9. Демиденко Е.З. Линейная и нелинейная регрессия. Москва : Финансы и статистика, 1981. 302 с.
10. Львовский Е.Н. Статистические методы построения эмпирических формул. Москва : Высш. школа, 1982. 224 с.
11. Архипова С.А. О применимости методов регрессионного анализа в задачах структурной идентификации при неполной информации о погрешностях измерений. *Вестник КМУЦА*. 1999. № 1. С. 315–322.

References:

1. Arhipov A.E., Arhipova S.A. (1998). Analiz i optimizacija kachestva reshenija zadachi identifikacii [Analysis and optimization of the quality of the identification problem solution]. Praci P'jatoi Ukrains'koї konferenciї z avtomatichnogo upravlinnja "Avtomatika-98": Kyiv, 13-16 travnja 1998. Vol. III. Kyiv: vidavnictvo NTUU "Kyivs'kij politehnichnij institut", pp. 9–15.
2. Ajvazjan S.A., Buhshtaber V.M., Enjukov I.S., Meshalkin L.D. (1989). Prikladnaja statistika: klassifikacija i snizhenie razmernosti [Applied statistics: classification and dimension reduction]: Sprav. izd. Moscow : Finansy i statistika. (in Russian)
3. Sil'vestrov A.N., Chinaev P.I. (1987). Identifikacija i optimizacija avtomaticheskih sistem [Identification and optimization of automatic systems]. Moscow : Atomjenergoizdat. (in Russian)
4. Mudrov V.I., Kushko V.L. (1976). Metody obrabotki izmerenij [Measurement Processing Methods]. Moscow : Sov. radio. (in Russian)
5. Sobolev V.I. (1983). Informacionno-statisticheskaja teorija izmerenij [Information and statistical measurement theory]. Uchebnik dlja vuzov. Moscow : Mashinostroenie. (in Russian)
6. L'jung L. (1991). Identifikacija sistem. Teorija dlja pol'zovanija [System identification. Theory to use]. Moscow : Nauka. (in Russian)
7. Seber Dzh. (1980). Linejnyj regressionnyj analiz [Linear regression analysis]. Moscow : Mir. (in Russian)
8. Ajvazjan S.A., Enjukov I.S., Meshalkin L.D. (1983). Prikladnaja statistika: Osnovy modelirovanija i pervichnaja obrabotka dannyh [Applied Statistics: Basics of Modeling and Primary Data Processing]. Spravochnoe izd. Moscow : Finansy i statistika. (in Russian)
9. Demidenko E.Z. (1981). Linejnaja i nelinejnaja regressija [Linear and nonlinear regression]. Moscow : Finansy i statistika. (in Russian)
10. L'vovskij E.N. (1982). Statisticheskie metody postroenija jempiricheskih formul [Statistical methods for constructing empirical formulas]. Moscow : Vyssha. shkola. (in Russian)
11. Arhipova S.A. (1999). O primenimosti metodov regressionnogo analiza v zadachah strukturnoj identifikacii pri nepolnoj informacii o pogreshnostjah izmerenij [On the applicability of regression analysis methods in problems of structural identification with incomplete information on measurement errors]. *Vestnik KMUCA,* no. 1, pp. 315–322.