ДК 681.5.015

С.А. Архипова

ПРИМЕНИМОСТИ МЕТОДОВ РЕГРЕССИОННОГО АНАЛИЗА ЗАДАЧАХ СТРУКТУРНОЙ ИДЕНТИФИКАЦИИ ПРИ НЕПОЛНОЙ НФОРМАЦИИ О ПОГРЕШНОСТЯХ ИЗМЕРЕНИЙ

Рассматривается возможность использования методов регрессивного анализа при решении задач идентификации в условиях отсутствия информации про особенности погрешности измерений. Показана недостаточность этих методов для получения качественных результатов.

Рассматривается задача построения наилучшей модели $y=\mu(x_1,x_2,...)$ в классе линейных регрессионных моделей по экспериментально полученным данным: матрице плана $[x_{ij}]$ и вектору $[z_i]=[y_i]+[e_i]$ результатов измерения значений зависимой переменной y, наблюдаемо в присутствии аддитивного случайного шума E. Исследования выполняются на двухфакторном тестовом примере.

Последовательность значений $x_{i1}, i=\overline{1,n}$ задается программным генератором псевдослучайных чисел. Вектор точных значений переменной рассчитывается соответствии с истинным уравнением регрессии

$$y = x_1 + x_2 + 0{,}04x_1^2 + 0{,}05x_2^2 + 0{,}003x_1x_2^2.$$

Для селекции наилучшей модели применим один из вариантов процедуры регрессионного анализа, описанный в работе [1]. Оценивание параметров моделей производится методом наименьших квадратов (МНК). Процедура селекции сводится следующему (табл.1):

1. Модели регрессии, в зависимости от количества m_j входящих в эти модели элементо разбиваются на ряд классов. Класс А представлен моделью $y = a_0$.
2. Внутри каждого класса модели упорядочиваются в соответствии с убыванием значен $S^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \tilde{y}_i)^2$ среднего квадрата ошибки аппроксимации, где $\tilde{y}_i$ - модельное значени зависимой переменной $\tilde{y}_i = \tilde{\mu}(x_{1i}, x_{2i},...)$.
3. В каждом классе отбираются одна или несколько моделей, имеющих минимальные д этого класса значения S^2.
4. Производится сопоставление и исследование этих моделей, включающее: провер адекватности моделей исходным данным, проверку необходимости усложнения эт моделей. Конечная цель этого анализа - выбор структуры аппроксимативной модели.

На рис.1 показана зависимость $S^2(m_j)$, при построении которой в качестве ордина использовались минимальные в классах Б – И значения S_j^2 (в табл. 1 выделены рамками).

С введением новых переменных в модель скорость уменьшения показателя S замедляется. С этих позиций (согласно [1] – [4]) можно полагать адекватными моде μ_9, μ_{10} либо μ_{11} (рис.1).

Таблица 1

j	Класс	Модель μ_j	m_j	S_j^2	$\tilde{F}_j$
1	А	$y(1)$	1	568,05	
2	Б	$y(x_1)$	1	257,41	2,21
3		$y(x_2)$	1	259,79	2,19
4	В	$y(x_1, x_2)$	2	197,82	2,86
5		$y(x_1, x_1^2)$	2	191,96	2,94
6		$y(x_2, x_2^2)$	2	201,37	2,81
7		$y(x_1, x_2^2)$	2	199,09	2,84
8	Г	$y(x_1, x_2, x_1x_2)$	3	146,08	3,85
9		$y(x_1, x_2, x_1^2)$	3	142,92	3,93
10		$y(x_1, x_2, x_2^2)$	3	142,92	3,93
11	Д	$y(x_1, x_2, x_1^2, x_2^2)$	4	132,79	4,21
12		$y(x_1, x_2, x_1x_2, x_2^2)$	4	138,38	4,04
13		$y(1, x_1, x_2, x_1x_2)$	4	142,20	3,93
14		$y(1, x_1, x_2, x_2^2)$	4	141,51	3,95
15	Е	$y(x_1, x_2, x_1^2, x_2^2, x_1^2x_2)$	5	123,24	4,52
16		$y(x_1, x_2, x_1^2, x_2^2, x_1x_2^2)$	5	121,49	4,57
17		$y(x_1, x_2, x_1^2, x_2^2, x_1^2x_2^2)$	5	127,14	4,38
18		$y(x_1, x_2, x_1x_2, x_2^2, x_1x_2^2)$	5	126,75	4,40
19		$y(x_1, x_2, x_1^2, x_2^2, x_2^3)$	5	124,02	4,49
20		$y(x_1, x_2, x_1^2, x_2^2, x_1^3)$	5	125,29	4,53
21	Ж	$y(x_1, x_2, x_1^2, x_2^2, x_1x_2^2, x_2^3)$	6	121,44	4,56
22		$y(x_1, x_2, x_1^2, x_2^2, x_1x_2^2, x_1^3)$	6	121,25	4,66
23		$y(1, x_1, x_2, x_1^2, x_2^2, x_1x_2^2)$	6	121,15	4,57
24		$y(x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1x_2^2)$	6	121,06	4,57
25	З	$y(x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1^2x_2, x_1x_2^2)$	7	121,01	4,55
26		$y(x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1x_2^2, x_2^3)$	7	121,01	4,55
27		$y(1, x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1x_2^2)$	7	120,24	4,58
28	И	$y(1, x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1^2x_2, x_1x_2^2)$	8	120,19	4,56
29		$y(1, x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1x_2^2, x_2^3)$	8	120,29	4,55

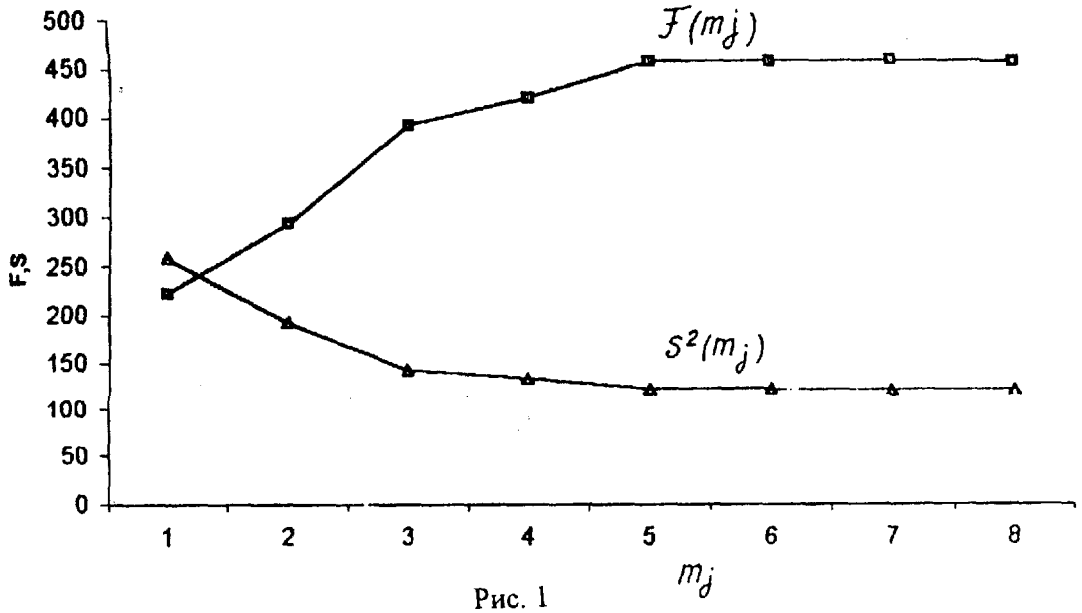

Рис. 1

Эмпиризм подобного подхода обычно побуждает исследователя к поиску других особов селекции модели. При этом прибегают к априорному постулированию рмальности распределения погрешностей измерения значения зависимой переменной, что зволяет сразу решить две проблемы: обоснованность оптимального применения МНК для енивания параметров модели и использовать имеющиеся в классическом регрессионном ализе наработки по оценке качества линейной регрессии для селекции структуры модели.

Для проверки значимости уравнений регрессии применяют статистику, которая ределяет, во сколько раз исследуемая j-я модель лучше предсказывает результаты сперимента, чем среднее $\bar{y}$, т.е. модель $y(1) = a_0$ [2], [4]:

$$\tilde{F}_j = \frac{S_1^2/(n-1)}{S_j^2/(n-m_j)},$$

Если табличное значение F-критерия для g%-го уровня значимости с ν_1, ν_2 степенями боды оказывается меньше $\tilde{F}_j$: $\tilde{F}_{g,\nu_1,\nu_2} < \tilde{F}_j$, проверяемая j-я модель предполагается ачимой.

Значения $\tilde{F}_j$ при g=5%, n=200 приведены в табл. 1 (а также на рис.1) и максимальное ачение $\tilde{F}_{27}$=4,58 соответствует семиэлементной модели μ_{27}, что, однако, не означает тимальности структуры этой модели: высокие значения $\tilde{F}_j$ следует рассматривать только как статистически подтвержденную значимость соответствующих моделей μ_j по отношению к модели среднего $y = a_0$ (модель μ_1).

Дальнейшая селекция модели с наилучшей структурой основана на применении более сложной системы попарного сопоставительного анализа моделей, в основе которого - проверка значимости изменения показателя $\tilde{F}$ при поэлементном введении изменений в структуру модели [3] по статистике

$$\widetilde{F}_{j\to r} = \frac{S_j^{\,2} - S_r^{\,2}}{S_r^{\,2}} \frac{n - m_r}{m_r - m_j},$$

имеющей F - распределение со степенями свободы $\nu_1 = m_r - m_j$, $\nu_2 = n - m_r$. Если теоретическое значение $F_{g,\nu_1,\nu_2} < \widetilde{F}_{j\to r}$, следует признать, что усложнение модели приводит к значимому росту ее качества.

По результатам исследований можно предположить, что наиболее адекватны исходным данным модели класса Ж: μ_{21}, μ_{22}, μ_{24}.

Если сопоставить итог селекции моделей по показателю $\widetilde{F}_{j\to r}$ (модели μ_{21}, μ_{22}, μ_{24}) с моделями μ_9, μ_{10}, μ_{11}, отобранными по показателю S^2, очевидно наличие существенных различий в сложности приведенных групп моделей. Даже если предположить, что по каким-либо причинам эти две группы моделей соответствуют противоположным границам области возможных решений задачи структурной идентификации, признать удовлетворительным подобный результат нельзя из-за содержащейся в нем весьма значительной неопределенности: область возможных решений перекрывает четыре класса моделей Г, Д, Е, Ж. Необходимо проведение дополнительных исследований, позволяющих радикально сузить эту область.

Один из возможных способов уточнения вида модели состоит в проверке значимости найденных оценок коэффициентов моделей [2], [4], целиком базирующейся на гипотезе нормальности погрешности.

Поэтому актуальным представляется вопрос о возможности проверки гипотезы нормальности распределения погрешности значений зависимой переменной. Рекомендуется [2] проведение апостериорной проверки гипотезы нормальности распределения погрешности путем проверки нормальности остатков (невязок), определяемых выражением:

$$\varepsilon_i = z_i - \widetilde{y}_i. \qquad (1)$$

Получение последовательности независимых значений e_1, e_2,..., e_n реализуется программным генератором псевдослучайных чисел (ГПСЧ), закон распределения которых задается в непараметрическом виде. Его графическое представление дано на рис. 2 в виде ступенчатой гистограммы $f(e)$. Моментные характеристики этого распределения: математическое ожидание $M\{e\} \approx 0$, дисперсия $D\{e\} = \sigma_e^2 = 121{,}32$. На этом же рисунке представлены графики распределения плотности вероятностей для нормального закона $f^{(\Gamma)}(e)$ и закона Лапласа $f^{(\Pi)}(e)$, имеющих одинаковые дисперсии σ_e^2 и нулевые математические ожидания.

Используя в качестве меры попарной близости распределений $f(e)$ и $f^{(\Gamma)}(e)$, $f(e)$ и $f^{(\Pi)}$ статистику

$$\chi_{\{\}}^2 = \sum_{j=1}^{24} \frac{\left[f_j(e) - f_j^{\{\}}\right]^2}{f_j^{\{\}}},$$

на достаточно высоком 50%-м уровне значимости можно полагать в равной степени правдоподобной гипотезу принадлежности эмпирического распределения $f(e)$ как к распределению Гаусса (нормальному), так и к распределению Лапласа.

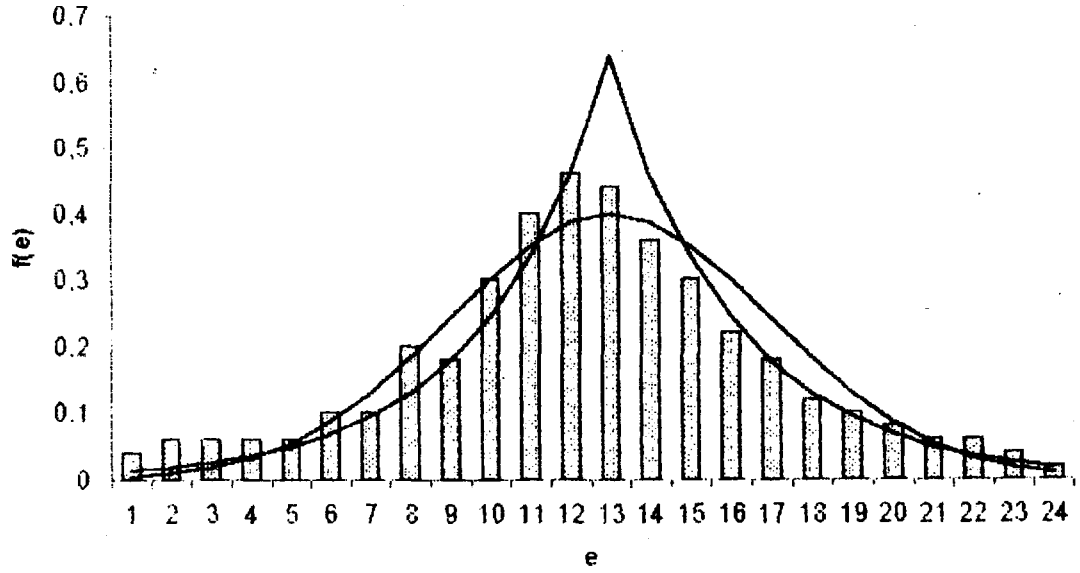

Рис. 2

Оценим возможность и объективность апостериорной проверки гипотезы нормальности через исследование остатков. Сгенерируем L матриц исходных данных $\{z_i,x_i\}_1,\ldots,\{z_i,x_i\}_L$. Рассчитаем по формуле (1) L выборок остатков $\{\varepsilon_1\}_1,\ldots,\{\varepsilon_i\}_L$ и проверим близость распределения ее элементов модели Гаусса или Лапласа, осуществим апостериорную проверку гипотезы о законе распределения погрешности E.

Чтобы исключить возможные искажения оценок погрешностей, будем полагать известной истинную структуру модели

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

Результаты представлены в табл. 2:

$r_Г$ - относительное число выборок, для которых из двух конкурирующих гипотез распределения (Гаусса и Лапласа) принята первая;

$r_Л$ - относительное число выборок, для которых из двух конкурирующих гипотез принимается гипотеза распределения Лапласа;

r_- - относительное число выборок, для которых отвергаются обе проверяемые гипотезы: $\chi^2_{(Г)}, \chi^2_{(Л)} > \chi^2_{g\%,\nu}$;

$r_{Г+}$ - относительное число выборок, для которых гипотеза нормальности является правдоподобной, т.е. $\chi^2_{(Г)} < \chi^2_{g\%,\nu}$

$\overline{\chi}^2_{(Г)}$ - среднее значение статистики χ^2.

Таблица 2

Выборки	$r_Г$	$r_Л$	r_-	$r_{Г+}$	$\overline{\chi}^2_{(Г)}$
$\{e_i\}$	42%	42%	16%	72%	27.6
$\{\varepsilon_i\}$	55.8%	42%	2.2%	90%	22.3

Первая значащая строка табл. 2 содержит численные значения введенных выш показателей, рассчитанные на множестве сгенерированных выборок $\{e_i\}_1,...,\{e_i\}_L$, втор строка – аналогичные показатели, рассчитанные по остаткам, т.е. по апостериорн найденным оценкам погрешностей $\{\varepsilon_i\}_1,...,\{\varepsilon_i\}_L$. Табличная информация свидетельствует возрастании доли принимаемых гипотез нормальности при апостериорном анализе данны Интересно, что помимо роста показателей r_Γ (с 42% до 55,8%) и $r_{\Gamma+}$ (с 72% до 90% происходит убывание показателя $\chi^2_{(\Gamma)}$ с 27,6 до 22,3, характеризующего уровень отличи исследуемых выборочных распределений от нормального, что говорит о нормализаци данных, о большей близости функции распределения остатков нормальному распределени по сравнению с исходным распределением.

Высокий уровень значений показателя $r_{\Gamma+}$, означает, что при отсутстви рассмотрения конкурирующих либо альтернативных нормальному распределению гипоте нормальность будет признаваться правдоподобной в подавляющем большинстве случаев тогда как наличие только одной альтернативной гипотезы резко снижает (с 90% до 55,8%) число случаев принятия гипотезы нормальности.

Таким образом, если следовать рекомендациям [2] и принять гипотезу нормальног распределения погрешностей в значениях зависимой переменной, следует ожидать нормальность распределения оценок параметров исследуемой модели (2).

Проверим это утверждение. Для полученных совокупностей оценок $\{\tilde{a}_{jl}\}$, $j=\overline{1,5}$, $l=\overline{1,L}$ построим эмпирические функции распределения $\tilde{F}\{\tilde{a}_j\}$ и исследуем их близость к нормальному распределению и к распределению Лапласа (табл. 3). Гипотеза нормальности распределения оценок подтверждается лишь для коэффициента a_2. Гипотеза о законе Лапласа оказывается более приемлемой, причем для коэффициентов a_4, a_5 она принимается при высоких уровнях значимости.

Таблица 3

Оцени-ваемый пара-метр	Проверка гипотезы распределения оценки по критерию χ^2							
	Нормальное распределение				Распределение Лапласа			
	χ^2	$\chi^2_{g\%,\nu}$	g%	ν	χ^2	$\chi^2_{g\%,\nu}$	g%	ν
a_1	34.56	-	-	-	27.97	29.141	1%	16
a_2	18.65	21.064	10%	14	38.79	-	-	-
a_3	468.8	-	-	-	19.43	21.026	5%	12
a_4	290.1	-	-	-	14.78	15.987	15%	10
a_5	56.72	-	-	-	13.60	14.631	20%	11

Результаты выполненных исследований позволяют сделать следующие выводы:

– процедура классического регрессионного анализа не содержит способов надежной селекции структуры модели при отсутствии априорных сведений о нормальности погрешности E;

– апостериорное утверждение о возможности принятия гипотезы нормальности остатков регрессии не гарантирует нормальности исходной погрешности E и, следовательно, не может служить обоснованием принятия положений регрессионного анализа, опирающихся на нормальность распределения E, в частности, гипотезы о нормальности распределения оценок параметров и базирующихся на ней методов проверки значимости оценок и модели регрессии.

Список литературы

1. *Себер Дж.* Линейный регрессионный анализ.– М.: Мир, 1980.– 456 с.

2. *Львовский Е.Н.* Статистические методы построения эмпирических формул. – М.: Высш. шк., 1982. – 224 с.

3. *Пугачев В.С.* Теория вероятностей и математическая статистика. – М.: Наука, 1979. – 496 с.

4. *Вучков И., Л.Бояджиева, Е.Солаков* и др. Прикладной линейный регрессионный анализ. – М.: Финансы и статистика, 1987. – 239 с.

Стаття надійшла до редакції 26 березня 1998 року

Софія Анатоліївна Архіпова (1961) закінчила Київський політехнічний інститут в 1984 році. Здобувач факультету авіаційного обладнання Київського міжнародного університету цивільної авіації. Автор 15 публикацій в галузі ідентифікації, моделювання та обробки даних.

Sofiya A. Arkhipova (b.1961) graduated from Kyiv Polytechnical Institute (1984) post competition graduate of Kyiv International University of Civil Aviation. Author of 15 publication in the field of identification, simulation and data processing.