УДК 519.233.2: 621.391.83 (045)

Кузьмин В.М., к.т.н.,
Соломенцев О.В., д.т.н.,
Заліський М.Ю., к.т.н.

МЕТОДИЧНИЙ ПІДХІД ДЛЯ ТЕСТУВАННЯ СТАТИСТИЧНИХ ДАНИХ НА ЛІНІЙНІСТЬ

Національний авіаційний університет

avsolomentsev@ukr.net
maximus2812@ukr.net

Розглянуто методичний підхід для тестування статистичних даних на лінійність під час вирішення задач обробки експериментальних даних в системах експлуатації технічних комплексів

Ключові слова: обробка статистичних даних, лінійність, система експлуатації, таблиця критичних значень

Вступ

Система експлуатації технічних комплексів формує та реалізує керуючі впливи для підтримки ефективного використання за призначенням цих комплексів. Для цього виконується ряд процедур статистичної обробки даних щодо визначальних параметрів. Ці параметри зазвичай є випадковими нестаціонарними процесами. Моделі процесів визначають структуру та зміст окремих процедур обробки даних.

До процедур обробки даних можна віднести процедури перевірки статистичних гіпотез, оцінювання параметрів, апроксимації, фільтрації, прогнозування, виявлення розладок тощо. Ефективність кожної процедури впливає на результуючу ефективність використання систем експлуатації технічних комплексів [1 – 5].

У цій статті розглядається одна із процедур обробки статистичних даних, що призначена для прийняття рішень щодо лінійного характеру в змінах значень випадкового процесу тренду визначального параметра.

Постановка завдання

Задача виявлення лінійного характеру в змінах визначального параметра може бути пов'язана з побудовою регресійних залежностей. Під час регресійного аналізу будують аналітичні залежності, що задовільно апроксимують досліджуваний набір експериментальних даних. Найчастіше у цьому випадку використовують лінійну залежність, що не кожного разу є обґрунтованим. Тому доречно було б кожного разу виконувати перевірку кореляційного поля експериментальних даних на лінійність.

Аналіз літератури [6] показав, що існує класичний метод Закса перевірки на лінійність. Проте цей метод має суттєвий недолік, пов'язаний з кратністю емпіричних точок, а саме: в кожному перетині необхідно мати два й більше експериментальних значень. Альтернативний варіант тестування на лінійність був запропонований у [7]. Проте таблиця критичних значень для тестування цим методом є обмежена та не має аналітичного опису. Тому у цій статті розглядаються питання узагальнення щодо апроксимаційної залежності критичних значень для тестування на лінійність.

Основна частина

У статті [7] був запропонований новий метод тестування експериментальних статистичних даних на лінійність (нелінійність). Аналіз літератури показав, що існують методи на основі використання багатократних вимірювань за одну процедуру отримання вибіркової сукупності. Головна особливість запропонованого методу – використання однократного спостереження, що найбільш часто зустрічається на практиці.

Таблиця критичних значень для тестування на лінійність в [7] – дискретна, при цьому доцільно отримати значення граничних параметрів у неперервному вигляді. Попередній аналіз значень цієї таблиці показав, що може бути знайдена апроксимаційна залежність типу

$$z(x,y) = a + bx + cy + dxy + ey^2, \qquad (1)$$

де x – логарифми обсягу вибірки, y – нормована довірча ймовірність.

Рівняння (1) є узагальненим рівнянням кривої другого порядку, в якому коефіцієнт при змінній x^2 є не значимим, тому цим коефіцієнтом можна знехтувати у подальших розрахунках.

Для визначення коефіцієнтів необхідно розв'язати систему рівнянь у матричному вигляді

$$\begin{pmatrix} nk & k\sum\limits_{i=1}^{n} x_i & n\sum\limits_{j=1}^{k} y_j & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i y_j & n\sum\limits_{j=1}^{k} (y_j)^2 \\ k\sum\limits_{i=1}^{n} x_i & k\sum\limits_{i=1}^{n} (x_i)^2 & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i y_j & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} (x_i)^2 y_j & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i (y_j)^2 \\ n\sum\limits_{j=1}^{k} y_j & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i y_j & n\sum\limits_{j=1}^{k} (y_j)^2 & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i (y_j)^2 & n\sum\limits_{j=1}^{k} (y_j)^3 \\ \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i y_j & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} (x_i)^2 y_j & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i (y_j)^2 & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} (x_i y_j)^2 & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i (y_j)^3 \\ n\sum\limits_{j=1}^{k} (y_j)^2 & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i (y_j)^2 & n\sum\limits_{j=1}^{k} (y_j)^3 & \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} x_i (y_j)^3 & n\sum\limits_{j=1}^{k} (y_j)^4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} z_{i,j} \\ \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} z_{i,j} x_i \\ \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} z_{i,j} y_j \\ \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} z_{i,j} x_i y_j \\ \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k} z_{i,j} (y_j)^2 \end{pmatrix}$$

По осі абсцис відкладаються значення обсягу вибірки (або їх логарифми), а по осі ординат значення

$$y = \ln\left(\frac{1}{1-p}\right),$$

де p – довірча ймовірність значущості лінійності або нелінійності, тобто

$$p = 1 - e^{-y}.$$

Введення логарифмічних перетворень дозволяє перейти до лінійного представлення критичних значень, взятих із [7], за віссю абсцисс. У результаті розв'язання системи рівнянь у матричному вигляді отримаємо таку апроксимаційну функцію

$$\begin{aligned} z = {} & 0{,}45063x + 0{,}06405y^2 - \\ & - 0{,}000108xy - 0{,}224423y + \\ & + 0{,}554146. \end{aligned} \qquad (2)$$

Рівняння (2) є основним апроксимаційним рівнянням для отримання критичних значень для тестування на лінійність.

Графік апроксимаційної функції наведено на рис. 1.

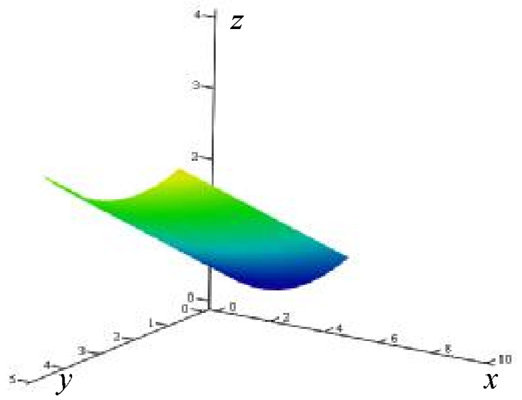

Рис. 1. Апроксимаційна функція для визначення критичних значень для тестування на лінійійність

За результатами розрахунку з використанням формули (2) можна побудувати таблиці критичних значень для тестування вибіркових сукупностей на лінійність (табл. 1).

Розглянемо методику для тестування на лінійність на конкретному прикладі, що представлений статистичними даними щодо залежності споживаної електроенергії від площі приміщення (табл. 2) [8].

Таблиця 1. Критичні значення для тестування вибіркових сукупностей на лінійність

z	P						
n	0,8	0,85	0,9	0,95	0,97	0,99	0,999
4	2.673	2.673	2.722	2.947	3.251	4.495	14.643
6	3.209	3.209	3.267	3.538	3.902	5.396	17.573
8	3.653	3.653	3.719	4.027	4.442	6.142	20.002
10	4.039	4.039	4.113	4.453	4.911	6.79	22.114
12	4.385	4.385	4.464	4.834	5.331	7.371	24.004
14	4.7	4.7	4.785	5.181	5.715	7.901	25.728
16	4.992	4.992	5.082	5.502	6.069	8.39	27.321
18	5.264	5.264	5.359	5.802	6.399	8.847	28.808
20	5.52	5.519	5.619	6.084	6.71	9.277	30.206
22	5.762	5.762	5.866	6.351	7.004	9.684	31.529
24	5.992	5.992	6.1	6.605	7.284	10.07	32.788
26	6.212	6.212	6.324	6.847	7.552	10.44	33.99
28	6.423	6.423	6.539	7.079	7.808	10.794	35.143
30	6.626	6.625	6.745	7.303	8.054	11.134	36.25
32	6.821	6.821	6.944	7.518	8.292	11.463	37.318
34	7.01	7.01	7.136	7.726	8.521	11.78	38.35
36	7.193	7.192	7.323	7.928	8.743	12.087	39.349
38	7.37	7.37	7.503	8.123	8.959	12.385	40.318
40	7.542	7.542	7.678	8.313	9.168	12.674	41.259
42	7.71	7.71	7.849	8.498	9.372	12.955	42.175
44	7.873	7.873	8.015	8.677	9.57	13.229	43.067
46	8.033	8.032	8.177	8.853	9.763	13.497	43.937
48	8.188	8.188	8.335	9.024	9.952	13.758	44.786
50	8.34	8.339	8.49	9.192	10.137	14.013	45.616
52	8.489	8.488	8.641	9.355	10.318	14.263	46.428
54	8.634	8.634	8.79	9.516	10.494	14.507	47.223
56	8.777	8.776	8.935	9.673	10.668	14.746	48.002
58	8.917	8.916	9.077	9.827	10.837	14.981	48.766
60	9.054	9.053	9.217	9.978	11.004	15.212	49.516
62	9.189	9.188	9.354	10.126	11.168	15.438	50.252
64	9.321	9.32	9.489	10.272	11.329	15.66	50.974
66	9.451	9.45	9.621	10.416	11.487	15.878	51.685
68	9.579	9.578	9.751	10.557	11.642	16.093	52.384
70	9.705	9.704	9.879	10.695	11.795	16.305	53.072
72	9.829	9.828	10.005	10.832	11.946	16.513	53.748
74	9.951	9.95	10.13	10.966	12.094	16.717	54.415
76	10.071	10.07	10.252	11.099	12.24	16.919	55.072
78	10.19	10.189	10.373	11.229	12.384	17.118	55.719
80	10.307	10.306	10.492	11.358	12.526	17.315	56.357
82	10.422	10.421	10.609	11.485	12.666	17.508	56.987
84	10.536	10.535	10.725	11.61	12.804	17.699	57.608
86	10.648	10.647	10.839	11.734	12.941	17.888	58.221
88	10.759	10.758	10.952	11.856	13.075	18.074	58.827
90	10.868	10.867	11.063	11.977	13.208	18.257	59.424
92	10.976	10.975	11.173	12.096	13.34	18.439	60.015
94	11.083	11.082	11.282	12.214	13.469	18.618	60.598
96	11.189	11.188	11.39	12.33	13.598	18.796	61.175
98	11.293	11.292	11.496	12.445	13.724	18.971	61.745
100	11.396	11.395	11.601	12.559	13.85	19.144	62.309

Таблиця 2. Залежність споживаної електроенергії від площі приміщення

Площа приміщення, S_i, кв. фути	1290	1350	1470	1600	1710	1840	1980	2230	2400	2930
Споживана електроенергія, P_i, кВтгод/місяць	1182	1172	1264	1493	1571	1711	1804	1840	1956	1954

Графічне представлення статистичних даних (залежність споживаної електроенергії від площі приміщення) зображено на рис. 2.

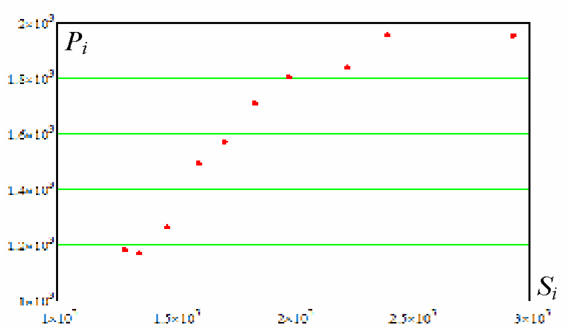

Рис. 2. Графічне представлення вихідних даних (залежність споживаної електроенергії від площі приміщення)

Виконаємо апроксимацію цих даних двома способами: прямою лінією та параболою другою порядку. Результати апроксимації наведені на рис. 3 та 4.

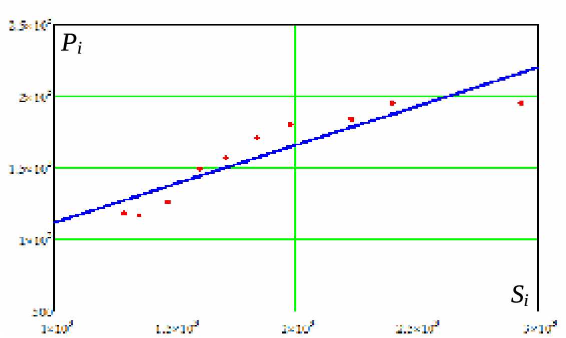

Рис. 3. Апроксимація вихідних даних прямою лінією

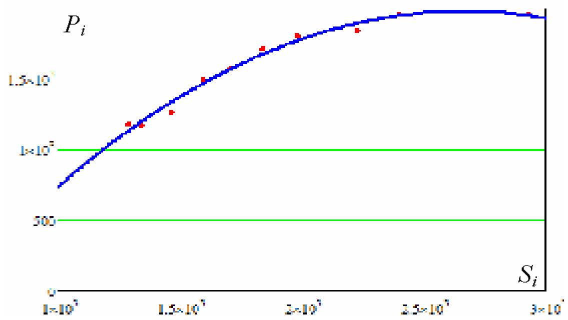

Рис. 3. Апроксимація вихідних даних параболою другого ступеня

Після використання методу найменших квадратів були отримані такі рівняння апроксимаційних залежностей:

$$y(x) = 0{,}5403x + 578{,}9,$$

$$y(x) = -0{,}00045x^2 + 2{,}3989x - 1216{,}1.$$

Далі виконаємо запропоновану в [7] процедуру тестування на лінійність. Для цього визначимо кумулятивну криву залишків відносно лінійної апроксимації та знайдемо її розмах, який у цьому прикладі становить $R = 547{,}4$. Також були знайдені стандартні відхилення для прямої лінії $\sigma = 133{,}44$ та для параболи $\sigma = 46{,}8$.

Стандартне відхилення було розраховане для параболи другого ступеню з метою уточнення його дійсного значення. Саме це значення і буде використовуватись у подальших розрахунках. Як видно з графіку (рис. 4), парабола має максимум, тобто збільшення площі приміщення буде відповідати зменшенню споживаної електроенергії, що не відповідає дійсності. Тому апроксимація параболою є недоречною.

Головним показником для тестування на лінійність є відношення розмаху до стандартного відхилення. У розглянутому прикладі цей показник дорівнює 11,698.

Для визначення дійсної величини ймовірності існування нелінійності в досліджуваних даних використаємо рівняння (2). Підставивши у це рівняння логарифм отриманого показника (замість z) та логарифм обсягу вибірки (замість x), у результаті отримаємо квадратне рівняння, одним із коренів якого є величина шуканої ймовірності.

$$y = 0{,}0084347\ln(n) + 1{,}751936 \pm$$
$$\pm\, 7{,}8064 \cdot \sqrt{\left(0{,}00108048\ln(n) + 0{,}224423\right)^2 - 0{,}11545\ln(n) + 0{,}2562\ln\left(\frac{R}{\sigma}\right) - 0{,}14197}.$$

У наведеному прикладі були отримані значення коренів рівняння: 0,996 та –6,477. Оскільки $0 \le p \le 1$, то можна стверджувати, що з ймовірністю 0,996 нелінійність у аналізованих даних має місце.

Як видно, наведений розрахунок доволі складний, проте дозволяє отримати значення ймовірності з високою точністю для довільних обсягів вибірки. Більш простий метод визначення ймовірності наявності нелінійності – це безпосереднє використання табл. 1.

Висновки

У статті вирішена задача побудови таблиці критичних значень та узагальнення апроксимаційної залежності, що описує ці значення, що дає можливість використовувати процедуру тестування на лінійність для довільних обсягів вибірки та довільних довірчих імовірностей. Отримані результати можуть бути використані під час розробки нових та модернізації діючих систем експлуатації технічних комплексів та їх підсистем обробки статистичних даних.

Список літератури

1. Байхельт Ф. Надежность и математическое обслуживание. Математический подход / Ф. Байхельт, П. Франкен. – М.: Радио и связь, 1988. – 392 с.
2. Диллон Б. Инженерные методы обеспечения надежности систем. Пер. с англ. / Б. Диллон, Ч. Сингх. – М.: Мир, 1984. – 318 с.
3. Левин Б.Р. Теория надёжности радиотехнических систем / Б.Р. Левин. – М.: Советское радио, 1978. – 274 с.
4. Hoyland, A. System reliability theory / A. Hoyland, M. Rausand. – N.Y.: John Wiley & Sons, Inc., 1994. –518 p.
5. Rausand, M. System reliability theory: models, statistical methods and applications / M. Rausand. – N.Y.: John Wiley & Sons, Inc., 2004. – 458 p.
6. Закс Л. Статистическое оценивание / Л. Закс. – М.: Статистика, 1976. – 598 с.
7. Kuzmin V. New statistical method for identification of nonlinearity of empirical data / V.N. Kuzmin. // Proceedings of the Fifth International Conference “Computer Data Analysis and Modeling” (June, 8–12, 1998, Minsk). – P. 159 – 164.
8. McClave, J.T. Statistics / J.T. McClave, F.H. Dietrich. – San Francisco: Dellen Publishing Company, 1991. – 624 p.

Статтю подано до редакції 21.12.2015