УДК 519.6

Г. Н. КОДОЛА (УГХТУ)

СТРУКТУРНОЕ МОДЕЛИРОВАНИЕ, КОГДА ЧИСЛО ОПЫТОВ МЕНЬШЕ, ЧЕМ ЧИСЛО ПОКАЗАТЕЛЕЙ

На підставі методу ε-ортогоналізації наведена методика побудови математичних моделей з мінімальною похибкою за дослідними даними, коли кількість показників перевищує кількість досвідів.

На основании метода ε-ортогонализации предложена методика построения математических моделей с минимальной погрешностью по опытным данным, когда число показателей превышает число опытов.

On the basis of the method of ε-orthogonalization, a technique of constructing mathematical models with a minimum error has been proposed under the test data, when the number of indices exceeds the number of tests.

Деятельность железных дорог Украины и Укрзалізниці в целом описывается очень большим числом показателей: грузооборот, пассажирооборот, количество груженых вагонов, количество разгруженных вагонов, продуктивность локомотива, оборот груженого вагона, простой груженого вагона на одной технической станции, простой вагона под одной грузовой операцией, участковая скорость, объем отправленных грузов, количество отправленных пассажиров, средняя численность работников на перевозках, потребная доля электротяги в грузообороте и многие другие. При описании деятельности дорог с помощью математических моделей возникает множество вопросов: какие же из этих показателей можно использовать для описания грузовой или же пассажирской деятельности дорог? А какие для описания деятельности дороги в целом? Какие из них наиболее существенные для построения модели? Какие параметры более точно позволят описать работу и в дальнейшем помогут производить прогноз на последующие года?

Если рассматривать данные переменные как простые количественные показатели, не учитывая ту смысловую информацию, какую они несут, возникает задача исследования влияния одних переменных на другие. Даже тогда, когда не очевидна связь между переменными, мы можем стремиться к тому, чтобы выявить ее с помощью математического моделирования.

Один из методов выбора математической модели с заданным набором переменных – Метод регрессионного анализа [1]. Существенным недостатком этого метода является то, что необходимо построить все модели, а за тем среди них выбрать математическую модель, осуществляющую прогноз с минимальной погрешностью. При большом количестве переменных – это очень громоздкая работа.

В случае, когда переменных больше, чем опытов, возникает вопрос выбора набора переменных, которые можно взять в качестве независимых. В методе наименьших квадратов [1; 2] количество опытов больше числа переменных, однако, очень часто возникает ситуация, когда переменных больше, чем опытов. Так как не удалось найти в литературе рассмотренной ситуации, то настоящая работа посвящается ее рассмотрению.

ε-ортогональность и ее применение в задачах математического моделирования

Традиционно при построении математических моделей линейных по параметрам, то есть при раскрытии зависимости вида

$$y = \sum_{i=1}^{k} \beta_i \phi_i(x),$$

где β_i – параметры модели; $\phi_i(x)$ – заданные функции; x – вектор размерности n.

Существенно используется понятие ортогональности векторов, так, например, в методе наименьших квадратов [2] рассматривается задача минимизации функции

$$S^2(\beta) = \sum_{j=1}^{N} \left[y_j - \sum_{i=1}^{k} \beta_i \phi_i(x_j) \right]^2,$$

где y_j – значения отклика в j-ом эксперименте; x_j – вектор предикторных переменных в j-ом эксперименте.

Необходимые и достаточные условия минимума функции $S2$ представляют собой

$$\frac{\partial S^2}{\partial \beta_\nu} = -2\sum_{j=1}^{N}\left[y_j - \sum_{i=1}^{k}\beta_i \phi_i\left(x_j\right)\right]\phi_\nu\left(x_j\right) = 0,$$

$$\nu = \overline{1,k}$$

или в нормальной форме получаем систему уравнений:

$$\sum_{i=1}^{k}\beta_i \overline{\varphi_i \varphi_\nu} = \overline{y\varphi_\nu},$$

где

$$\overline{\varphi_i \varphi_\nu} = \frac{1}{N}\sum_{i=1}^{N}\varphi_i\left(x_i\right)\varphi_\nu\left(x_j\right),$$

$$\overline{y\varphi_\nu} = \frac{1}{N}\sum_{j=1}^{N} y_j \varphi_\nu\left(x_j\right).$$

Таким образом, если рассмотреть вектора:

$$Y = \frac{1}{\sqrt{N}}\begin{pmatrix} y_1 \\ y_2 \\ \hdashline y_N \end{pmatrix} \quad \Phi_i = \frac{1}{\sqrt{N}}\begin{pmatrix} \varphi_i\left(x_1\right) \\ \varphi_i\left(x_2\right) \\ \hdashline \varphi_i\left(x_N\right) \end{pmatrix},$$

то $\overline{y\varphi_\nu}$ и $\overline{\varphi_i\varphi_\nu}$ можно рассмотреть как скалярное произведение введенных векторов. Если вектора Φ_i ортогональны, тогда система нормальных уравнений принимает простой вид

$$\beta_i < \Phi_i, \Phi_i >=< Y, \Phi_i >,$$

откуда

$$\beta_i = \frac{< Y, \Phi_i >}{< \Phi_i, \Phi_i >}.$$

Однако ортогональность в расчетах на ЭВМ не может быть строго реализована из-за конечности разрядной сетки ЭВМ и естественно возникает задача ε-ортогональности и ее последствий при построении математических моделей.

ε-ортогональность

Для определенности изложения введем формальное понятие ε-ортогональности.

Определение 1. Пусть имеется два вектора *a*, *b*, тогда будем говорить, что эти вектора ε-ортогональны если имеет место

$$\left| < a, b > \right| \le \varepsilon,$$

где ε – наперед заданное положительное число.

Понятие ортогональности позволяет ввести базис [3]. Обобщая это понятие, введем в рассмотрение ε – базис.

Определение 2. ε-базисом некоторого пространства *X* будем называть максимальный набор попарно ε-ортогональных векторов из пространства *X*. То есть, если $e_1, e_2, \ldots e_m$ – вектора из пространства *X* и такие, что $\left|\left\langle e_i, e_j \right\rangle\right| \le \varepsilon$, при чем этот набор не пополняем, то его, в силу определения, и будем называть ε-базисом.

Разложение (представление) вектора в ε-ортогональном базисе

Пусть X – векторное пространство, а $\{e_i\}, i = \overline{1,m}$ его ортогональный базис. Для любого вектора $x \in X$ представление

$$x = x_1 e_1 + x_2 e_2 + \ldots + x_m e_m$$

называют разложением вектора x по базисным векторам $\{e_i\}$, а числа $x_1, x_2, \ldots, x_m$ называют компонентами вектора x в рассматриваемом базисе [3]. Для определения чисел x_i используется система:

$$x_1 < e_1, e_1 > + x_2 < e_2, e_1 > + \ldots + x_m < e_m, e_1 >=< x, e_1 >;$$

$$x_1 < e_1, e_2 > + x_2 < e_2, e_2 > + \ldots + x_m < e_m, e_2 >=< x, e_2 >;$$

. . . . . .

$$x_1 < e_1, e_m > + x_2 < e_2, e_m > + \ldots + x_m < e_m, e_m >=< x, e_m >.$$

Не ограничивая общность рассмотрения, считаем, что ε – базис нормирован, то есть

$$< e_i, e_i >= 1.$$

Обозначим скалярное произведение векторов e_i и e_j через $\varepsilon_{ij} =< e_i, e_j >$, тогда вышеприведенная система примет вид

$$x_1 + \varepsilon_{12} x_2 + \ldots + \varepsilon_{1m} x_m = b_1;$$

$$\varepsilon_{21} x_1 + x_2 + \ldots + \varepsilon_{2m} x_m = b_2;$$

. . . . . .

$$\varepsilon_{m1} x_1 + \varepsilon_{m2} x_2 + \ldots + x_m = b_m.$$

где $b_i = \left\langle x, e_i \right\rangle, \quad i = \overline{1,m}$.

В матричной форме имеем

$$\left(\breve{I}+\breve{E}\right)x=b, \qquad (1)$$

где

$$\breve{I}=\left(\begin{array}{ccc|c}1&0&0&0\\0&1&0&0\\\hline 0&0&0&1\end{array}\right);$$

$$\breve{E}=\left(\begin{array}{ccc|c}0&\varepsilon_{12}&\varepsilon_{13}&\varepsilon_{1m}\\\varepsilon_{21}&0&\varepsilon_{23}&\varepsilon_{2m}\\\hline \varepsilon_{m1}&\varepsilon_{m2}&\varepsilon_{m3}&0\end{array}\right).$$

В силу определения скалярного произведения матрица $\breve{E}$ является симметричной, и ее элементы удовлетворяют условию $\left|\varepsilon_{ij}\right|\le\varepsilon$, при $i\ne j$.

Решение системы (1) формально имеет вид

$$x=\sum_{k=0}\left(-1\right)^k\breve{E}^k b. \qquad (2)$$

Естественно возникает вопрос о сходимости ряда (2). Прежде, чем рассмотрим этот вопрос в общем виде, исследуем частный случай, когда число базисных векторов равно 2, то есть в качестве X выступает плоскость. Матрица $\breve{E}$ будет следующей:

$$\breve{E}=\begin{pmatrix}0&\varepsilon_{12}\\\varepsilon_{21}&0\end{pmatrix}$$

и так как $\varepsilon_{12}=\varepsilon_{21}$, то имеет место

$$\breve{E}=\varepsilon_{12}\begin{pmatrix}0&1\\1&0\end{pmatrix}.$$

Квадрат матрицы $\breve{E}$ будет равен

$$\breve{E}^2=\varepsilon^2{}_{12}\begin{pmatrix}1&0\\0&1\end{pmatrix},$$

а для куба получаем

$$\breve{E}^3=\varepsilon_{12}^3\begin{pmatrix}0&1\\1&0\end{pmatrix}.$$

Тогда

$$x=\frac{1}{1+\varepsilon_{12}}\begin{pmatrix}0&1\\1&0\end{pmatrix}b,$$

откуда следует, что если $\varepsilon_{12}\ne-1$, то это условие является необходимым и достаточным, чтобы разложение вектора x по базису e_1, e_2 было единственным.

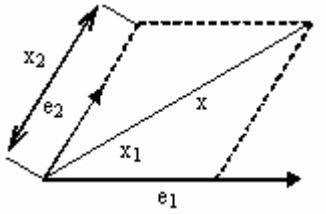

Рис. 1

Геометрическая интерпретация данной ситуации представляет собой разложении вектора x в косоугольной системе координат (рис. 1). И если угол между векторами e_1 и e_2 незначительно отличается от прямого, т. е. $\left|\varepsilon_{12}\right|<<1$, то в качестве координат можно взять

$$x_1\approx b_1-\varepsilon_{12}b_2;$$

$$x_2\approx b_2-\varepsilon_{12}b_1.$$

Точность этого представления будет иметь порядок $\mathrm{O}\left(\varepsilon_{12}^2\right)$.

В общем случае можно утверждать, что если

$$\det\left(\breve{I}+\breve{E}\right)\ne 0,$$

то разложение вектора x по ε-базису $\left\{e_i\right\}$ единственно, а при

$$\varepsilon=\max_{\substack{i,j\\ i\ne j}}\left|\varepsilon_{ij}\right|<<1$$

в качестве приближенного значения получаем

$$\begin{pmatrix}x_1\\x_2\\\hline x_m\end{pmatrix}\cong\begin{pmatrix}b_1\\b_2\\\hline b_m\end{pmatrix}-\left(\begin{array}{ccc|c}0&\varepsilon_{12}&\varepsilon_{13}&\varepsilon_{1m}\\\varepsilon_{21}&0&\varepsilon_{23}&\varepsilon_{2m}\\\hline \varepsilon_{m1}&\varepsilon_{m2}&\varepsilon_{m3}&0\end{array}\right)\begin{pmatrix}b_1\\b_2\\\hline b_m\end{pmatrix}. \qquad (3)$$

Для оценки погрешности данного приближения воспользуемся следующей леммой.

Лемма. Если $\varepsilon m<1$, то ряд (2) сходится и погрешность разложения (3) равна $\mathrm{O}\left[\left(\varepsilon m\right)^2\right]$.

Доказательство. Для доказательства этой леммы рассмотрим матрицу

$$\breve{E}=\left(\begin{array}{ccc|c}0&\varepsilon_{12}&\varepsilon_{13}&\varepsilon_{1m}\\\varepsilon_{21}&0&\varepsilon_{23}&\varepsilon_{2m}\\\hline \varepsilon_{m1}&\varepsilon_{m2}&\varepsilon_{m3}&0\end{array}\right).$$

Порядок элементов матрицы $\breve{E}^k$ найдем, используя порядок элементов матрицы $\breve{E}^k$, где

$$\breve{E} = \left(\begin{array}{ccc|c} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ \hdashline 1 & 1 & 1 & 0 \end{array}\right).$$

Используя метод математической индукции, получаем

$$\breve{E}^2 = \left(\begin{array}{cc|c} m-1 & m-2 & m-2 \\ m-2 & m-1 & m-2 \\ \hdashline m-2 & m-2 & m-1 \end{array}\right)$$

и

$$\breve{E}^3 = \left(\begin{array}{cc|c} (m-1)(m-2) & (m-1)+(m-2)^2 & (m-1)+(m-2)^2 \\ (m-1)+(m-2)^2 & (m-1)(m-2) & (m-1)+(m-2)^2 \\ \hdashline (m-1)+(m-2)^2 & (m-1)+(m-2)^2 & (m-1)(m-2) \end{array}\right)$$

и т. д. то есть элементы матрицы $\breve{E}^k$ будут иметь порядок $\mathrm{O}\left(m^{k-1}\right)$.

Следовательно, элементы матрицы $\breve{E}^k$ будут иметь порядок $\mathrm{O}\left(\varepsilon^k m^{k-1}\right)$, потому что $\breve{E}_{ij}^k \le \varepsilon^k \breve{E}_{ij}^k \le \varepsilon^k m^{k-1}$, а с учетом условия леммы $\left(\varepsilon m < 1\right)$ получаем, что ряд (2)сходится, но т. к. он является знакочередующимся, то условие $\varepsilon m < 1$ является не только необходимым, но и достаточным [3].

Пусть $X = \{x_1, x_2, ..., x_m\}$ – набор векторов; $x_k \subset E_N$, $k = \overline{1,m}$; $Y = \begin{pmatrix} y_1 \\ y_2 \\ \hdashline y_N \end{pmatrix}$ – значения откликов; $Y \subset E_N$; N – размерность каждого вектора (число опытов); ε – наперед заданное положительное число.

Определение 3. Множество $M \subset X$ назовем ε-базисом если выполняются следующие два условия:

1. $\forall x_i; x_j \subset M, i \ne j$

$$\left|\left\langle x_i, x_j \right\rangle\right| \le \varepsilon;$$

2. $\left(\forall x_k \in X/M\right)\exists x_v \in M$

$$\left|\langle x_k, x_v \rangle\right| > \varepsilon. \qquad (4)$$

Модель с минимальной погрешностью

Рассмотрим модель вида:

$$y = a_0 x_0 + a_1 x_1 + \ldots + a_m x_m, \qquad (5)$$

где a_i, $i = \overline{0,m}$ определяются по методу наименьших квадратов.

Положим

$$\varepsilon_0 = \max_{1 \le k \le N} \left| y_k - \sum_{i=1}^{m} a_i x_{ik} \right|. \qquad (6)$$

Пусть $\Omega = \{x_1, x_2, \ldots x_m\}$, а $V \subseteq \Omega$, тогда модель типа (5) по показателям $x \in \Omega/V$ будет равна

$$y(\Omega/V) = \sum_{x_i \in \Omega/V} a_i x_i$$

погрешность этой модели будет определяться

$$\varepsilon(V) = \max_{1 \le k \le N} \left| y_k - \sum_{i \in \Omega/V} a_i x_{ik} \right|.$$

Очевидно, что

$$\varepsilon(V) = \max_{1 \le k \le N} \left| y_k - \sum_{i=0}^{m} a_i x_{ik} + \sum_{i \in V} a_i x_{ik} \right| \le \varepsilon_0 + \max_{1 \le k \le N} \left| \sum_{i \in V} a_i x_{ik} \right|.$$

Положим, $\tilde{\varepsilon} = \max_{1 \le k \le N} \left| \sum_{i \in V} a_i x_{ik} \right|$ представляет собой оценку сверху приращения погрешности при удалении из модели (5) показателя из перечня V.

Возникает задача

$$\tilde{\varepsilon}(V) \to \min, \quad |V| \to \max,$$
$$V \subseteq \Omega, \qquad (7)$$

где $|V|$ – число элементов во множестве V.

Решение этой задачи и позволит определить наборы предикторных переменных для построения математических моделей, с погрешностью, не превышающей заданную.

Пример. Используя среднесуточные статистические данные [4; 5] по годам за период 1991–2001 гг., построить математическую модель с погрешностью, не превышающей заданной, для описания грузовой деятельности Приднепровской железной дороги.

Решение данной задачи разложим на три этапа:

1. Применяя метод ε-ортогонализации определим наборы предикторных переменных типа $M \subset X$, удовлетворяющие условию (4), которые будут являться базисом для построения математической модели с погрешностью, не превышающей заданную.

2. Строятся математические модели для описания грузовой деятельности Приднепровской железной дороги и определяются максимальные относительные погрешности для каждой модели.

3. Используя усовершенствованный метод регрессионного анализа отбора предикторных переменных по заданной точности математической модели, определить наборы предикторных переменных для построения математических моделей, описывающих грузовую деятельность дороги с погрешностью, не превышающей заданную.

В исходной информации представлено 17 параметров (грузооборот, пассажирооборот, количество груженых вагонов, количество разгруженных вагонов, производительность локомотива, оборот груженого вагона, простой груженого вагона на одной технической станции, простой вагона под одной грузовой операцией, участковая скорость, объем отправленных грузов, количество отправленных пассажиров, средняя численность работников на перевозках, потребная доля электротяги в грузообороте и т. д.) за 11 лет, то есть количество опытов меньше, чем количество показателей.

Определяя наборы переменных, которые удовлетворяют условию (4), рассмотрим один из них:

$$M = \{x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{12}, x_{13}, x_{17}\}.$$

Построим математическую модель, описывающую грузооборот Приднепровской железной дороги (показатель x_1) по методу наименьших квадратов, исключив сначала показатель x_2, и определим максимальную относительную погрешность для данной модели.

$$x(t) = 7,927x_3(t) + 13,596x_4(t) + \\ +0,109x_5(t) + 9,461x_7(t) - 1,734x_8(t) + \\ +6,818x_9(t) + 0,027x_{12}(t) + 0,002x_{13}(t) - \\ -1,720x_{17}(t) - 547,393.$$

Максимальная относительная погрешность составляет 1,81 %.

Теперь построим математические модели, выражающие показатель x_1, через набор переменных M, исключив из него сначала показатель x_3 затем x_4 и т. д. Для каждой модели определим максимальные относительные погрешности и результат сведем в табл. 1.

Таблица 1

Исключенная переменная	Коэффициент при										Свободный член	Относительная погрешность, %
	x_2	x_3	x_4	x_5	x_7	x_8	x_9	x_{12}	x_{13}	x_{17}		
x_2	–	7,927	13,60	0,109	9,461	–1,734	6,818	0,027	–0,002	1,720	–547,390	1,81
x_3	–1,296	–	18,51	0,222	5,731	–2,172	5,100	–0,042	–0,011	0,679	–133,790	1,47
x_4	–1,256	11,250	–	0,350	12,760	–3,508	8,134	–0,107	–0,015	0,280	–60,675	1,37
x_5	–0,827	7,389	21,15	–	6,484	–1,044	6,492	0,081	0,001	2,950	–741,430	1,78
x_7	–1,335	–1,056	25,62	0,081	–	–0,709	3,475	0,024	–0,005	1,278	–285,720	2,77
x_8	–0,772	5,021	24,55	–0,030	3,411	–	8,075	0,086	0,007	3,123	–755,970	2,27
x_9	–1,287	–7,464	29,16	0,169	–2,500	–0,618	–	–0,029	–0,007	0,198	54,166	5,32
x_{12}	–1,09	6,935	14,85	0,520	8,290	–2,051	6,686	–	–0,006	1,825	–427,780	0,53
x_{13}	–0,909	7,637	16,37	0,102	8,005	–1,403	6,606	0,025	–	2,502	–588,330	1,12
x_{17}	–1,263	4,835	8,95	0,293	8,662	–2,919	6,343	–0,080	–0,145	–	–13,209	0,75

На третьем этапе полученный результат сведем в табл. 2.

Погрешность, указанная в табл. 2, представляет собой оценку сверху погрешности математической модели, построенной на основе соответствующего набора предикторных переменных.

Таблица 2

Вариант	Набор предикторных переменных	Исключенные переменные	Погрешность, %
1	$\{x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{13}, x_{17}\}$	$\{x_{12}\}$	0,53
2	$\{x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{13}\}$	$\{x_{12}, x_{17}\}$	1,28
3	$\{x_2, x_3, x_4, x_5, x_7, x_8, x_9\}$	$\{x_{12}, x_{13}, x_{17}\}$	2,40
4	$\{x_2, x_3, x_5, x_7, x_8, x_9\}$	$\{x_4, x_{12}, x_{13}, x_{17}\}$	3,77
5	$\{x_2, x_5, x_7, x_8, x_9\}$	$\{x_3, x_4, x_{12}, x_{13}, x_{17}\}$	5,24
6	$\{x_2, x_7, x_8, x_9\}$	$\{x_3, x_4, x_5, x_{12}, x_{13}, x_{17}\}$	7,02
7	$\{x_7, x_8, x_9\}$	$\{x_2, x_3, x_4, x_5, x_{12}, x_{13}, x_{17}\}$	8,83
8	$\{x_7, x_9\}$	$\{x_2, x_3, x_4, x_5, x_8, x_{12}, x_{13}, x_{17}\}$	11,10
9	$\{x_9\}$	$\{x_2, x_3, x_4, x_5, x_7, x_8, x_{12}, x_{13}, x_{17}\}$	13,87
10	$\{\ \}$	$\{x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{12}, x_{13}, x_{17}\}$	19,19

Итак, для того чтобы построить математическую модель с заданной погрешностью, необходимо выбрать из табл. 2 соответствующий набор предикторных переменных, у которого погрешность не превосходит заданную.

Для того, чтобы убедиться в правильности данного утверждения построим для каждого полученного набора, представленного в таблице 2, математические модели, подсчитаем погрешности. Результат сведем в табл. 3. Как видно из табл. 3 математические модели с первой по четвертую удовлетворяют погрешности, представленной в табл. 2, а уже начиная с набора 5…9, погрешности намного превышают оценку сверху из табл. 2.

Это происходит за счет того, что математические модели строятся на основе исходной информации, у которой ε-базис не является нормированным. Для того чтобы избежать превышение оценок из табл. 2 необходимо ортонормировать исходную информацию, то есть привести к виду, чтобы удовлетворялись следующие два условия:

$$\left|\left(z_i, z_j\right)\right| = \varepsilon_*, \quad i \neq j,$$

$$\left(z_i, z_j\right) = 1, \quad i = j, \tag{8}$$

где ε_* – машинная точность.

Рассмотрим тот же пример, но только для ортонормированных данных. В качестве базиса рассмотрим прежний:

$$M = \{z_2, z_3, z_4, z_5, z_7, z_8, z_9, z_{12}, z_{13}, z_{17}\}.$$

Результат сведем в табл. 4.

Таблица 3

Исключенная переменная	Коэффициент при										Свободный член	Относительная погрешность, %
	x_2	x_3	x_4	x_5	x_7	x_8	x_9	x_{12}	x_{13}	x_{17}		
1	–1,090	6,935	14,85	0,152	8,290	–2,05	6,686	–	–0,006	1,83	–427,78	0,53
2	–0,808	3,167	14,85	0,145	5,808	–2,09	5,739	–	–0,011	–	–173,04	1,21
3	–0,352	2,595	13,94	0,122	4,776	–0,91	4,968	–	–	–	–226,29	1,80
4	0,022	13,260	–	0,134	10,410	–1,64	7,938	–	–	–	–347,60	1,98
5	1,457	–	–	0,304	27,240	–5,19	17,910	–	–	–	–805,15	20,84
6	–2,637	–	–	–	19,570	–3,56	25,170	–	–	–	–589,55	29,27
7	–	–	–	–	23,690	–3,93	25,180	–	–	–	–687,97	29,06
8	–	–	–	–	–4,870	–	22,660	–	–	–	–566,20	32,02
9	–	–	–	–	–	–	27,350	–	–	–	–745,67	30,38

Таблица 4

Исключенная переменная	Коэффициент при										Свободный член	Относительная погрешность, %
	z_2	z_3	z_4	z_5	z_7	z_8	z_9	z_{12}	z_{13}	z_{17}		
z_2	–	0,029	0,015	–0,032	0,0030	0,012	–0,017	–0,003	–0,0003	–0,004	0,2883	46,16
z_3	–0,311	–	0,017	–0,035	0,0033	0,013	–0,018	–0,003	–0,0004	–0,005	0,3108	8,14
z_4	–0,311	0,032	–	–0,035	0,0033	0,013	–0,018	–0,003	–0,0004	–0,005	0,3161	3,81
z_5	–0,311	0,032	0,017	–	0,0033	0,013	–0,018	–0,003	–0,0004	–0,005	0,3158	5,57
z_7	–0,311	0,032	0,017	–0,035	–	0,013	–0,018	–0,003	–0,0004	–0,005	0,3162	1,08
z_8	–0,311	0,032	0,017	–0,035	0,0033	–	–0,018	–0,003	–0,0004	–0,005	0,3161	4,67
z_9	–0,311	0,032	0,017	–0,035	0,0033	0,013	–	–0,003	–0,0004	–0,005	0,3161	4,85
z_{12}	–0,311	0,032	0,017	–0,035	0,0033	0,013	–0,018	–	–0,0004	–0,005	0,3162	0,54
z_{13}	-0,311	0,032	0,017	-0,035	0,0033	0,013	-0,018	-0,003		-0,005	0,3162	0,08
z_{17}	–0,311	0,032	0,017	–0,035	0,0033	0,013	–0,018	–0,003	–0,0004	–	0,3162	1,11

Коэффициенты при показателях в построенных моделях (табл. 4) по преобразованным данным мало изменяются.

Решая задачу (7), усовершенствуем метод регрессионного анализа. Результаты сведем в табл. 5. По данным наборам предикторных переменных построим математические модели и подсчитаем для каждой модели погрешности (табл. 6). Все построенные модели (табл. 6) имеют погрешность, не превышающую погрешность для соответствующего набора предикторных переменных, представленного в табл. 5.

На первом этапе моделирования использовалась ε-ортогональность для определения базиса, набора переменных, на основе которого должен проводиться анализ. Такой набор является не единственным (в качестве примера для исследования был представлен только один набор). Возникает вопрос: какому набору отдать предпочтение?

Таблица 5

Вариант	Набор предикторных переменных	Исключенные переменные	Погрешность, %
1	$\{z_2, z_3, z_4, z_5, z_7, z_8, z_9, z_{12}, z_{17}\}$	$\{z_{13}\}$	0,08
2	$\{z_2, z_3, z_4, z_5, z_7, z_8, z_9, z_{17}\}$	$\{z_{12}, z_{13}\}$	0,62
3	$\{z_2, z_3, z_4, z_5, z_8, z_9, z_{17}\}$	$\{z_7, z_{12}, z_{13}\}$	1,70
4	$\{z_2, z_3, z_4, z_5, z_8, z_9\}$	$\{z_7, z_{12}, z_{13}, z_{17}\}$	2,81
5	$\{z_2, z_3, z_5, z_8, z_9\}$	$\{z_4, z_7, z_{12}, z_{13}, z_{17}\}$	6,62
6	$\{z_2, z_3, z_5, z_9\}$	$\{z_4, z_7, z_8, z_{12}, z_{13}, z_{17}\}$	11,29
7	$\{z_2, z_3, z_5\}$	$\{z_4, z_7, z_8, z_9, z_{12}, z_{13}, z_{17}\}$	16,14
8	$\{z_2, z_3\}$	$\{z_4, z_5, z_7, z_8, z_9, z_{12}, z_{13}, z_{17}\}$	21,71
9	$\{z_2\}$	$\{z_3, z_4, z_5, z_7, z_8, z_9, z_{12}, z_{13}, z_{17}\}$	29,85
10	$\{\ \}$	$\{z_2, z_3, z_4, z_5, z_7, z_8, z_9, z_{12}, z_{13}, z_{17}\}$	76,01

Таблица 6

Исключенная переменная	Коэффициент при										Свободный член	Относительная погрешность, %
	z_2	z_3	z_4	z_5	z_7	z_8	z_9	z_{12}	z_{13}	z_{17}		
1	–0,311	0,032	0,017	–0,035	0,0033	0,013	–0,018	–0,003	–	–0,005	0,3162	0,080
2	–0,311	0,032	0,017	–0,035	0,0033	0,013	–0,018	–	–	–0,005	0,3162	0,549
3	–0,311	0,032	0,017	–0,035	–	0,013	–0,018	–	–	–0,005	0,3162	1,440
4	–0,311	0,032	0,017	–0,035	–	0,013	–0,018	–	–	–	0,3162	1,560
5	–0,311	0,032	–	–0,035	–	0,013	–0,018	–	–	–	0,3161	3,690
6	–0,311	0,032	–	–0,035	–	–	–0,018	–	–	–	0,316	6,850
7	–0,311	0,032	–	–0,035	–	–	–	–	–	–	0,3159	3,070
8	–0,31	0,032	–	–	–	–	–	–	–	–	0,3156	9,480
9	–0,310	–	–	–	–	–	–	–	–	–	0,3152	10,480

Проделав исследования на нескольких базисах, можно сказать, что предпочтение можно отдать тому набору, у которого сумма погрешностей для построенных моделей на втором этапе (представленных в табл. 1 и 4) минимальна.

Нашу задачу можно записать теперь так.

Исходные данные:

$$\left[\begin{array}{c|ccc} y_1 & x_{11} & x_{12} & x_{1m} \\ \hdashline y_2 & x_{21} & x_{22} & x_{2m} \\ \hdashline y_N & x_{N1} & x_{N2} & x_{Nm} \end{array}\right],$$

где N – количество опытов, m – количество показателей.

Используя ε-ортогонализацию, определяем наборы типа $M \subset X$, удовлетворяющие условию (4). $|M| = p$ – мощность множества M.

Ортонормируем исходные данные для каждого набора и строим модели типа:

$$y^{\nu} = \sum_{\substack{i=0 \\ i \neq \nu}}^{k} c_i^{\nu} z_i ,$$

где c_i^{ν}, $i = \overline{1,k}$ определяются по методу наименьших квадратов; $\nu = \overline{1,p}$ – номер модели; k – количество независимых переменных.

Далее определим погрешности для каждой модели

$$\varepsilon_0^{\nu}\left(y^{\nu}\right) = \max_{1 \le j \le N} \left| y_j - \sum_{\substack{i=1 \\ i \neq \nu}}^{k} c_i^{\nu} z_{ij} \right|, \quad \nu = \overline{1,p}$$

и выбираем такой базис, для которого

$$\sum_{\nu=1}^{p} \varepsilon_0^{\nu} \to \min .$$

Решаем задачу (7). Вернемся от ортонормированных данных к исходным. В этом случае модель примет вид

$$y^{\nu} = \sum_{\substack{i=0 \\ i \neq \nu}}^{k} c_i^{\nu} z_i = \sum_{i=1}^{k} c_i^{\nu} \sum_{j=1}^{k} \alpha_{ij} xj = \sum_{j=1}^{k} x_j \sum_{i=1}^{k} c_i \alpha_{ij}$$

или в общем виде:

$$\Upsilon^{\nu} = C^T A X ,$$

где X – набор предикторных переменных; C – вектор коэффициентов в ортонормированных моделях; A – матрица коэффициентов перехода к ортонормированным данным.

Данная методика позволяет значительно сократить количество переборов построения моделей в регрессионном анализе и тем его усовершенствовать.

Благодарность

Автор благодарит проф. А. А. Босова за помощь в постановке и решении задачи.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Дрейпер Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. – М.: Финансы и статистика. Т. 1, 1986. – 366 с.
2. Наконечний С. І. Економетрія / С. І. Наконечний, Т. О Терещенко, Т. П. Романюк. – К.: КНЕУ, 2000, – 296 с.
3. Гельфанд И. М. Лекции по линейной алгебре. – М.: Наука, 1971. – 272 с.
4. Пасечник В. І. Аналіз динаміки показників залізниць України // Залізничний транспорт України, 2002. – № 5. – С. 2–6.
5. Железнодорожный транспорт – ведущая отрасль экономики Украины: (Информ.-стат. и аналит. материалы) / Под ред. Т. Мукминовой. – К.: Транспорт України, 2003. – 32 с.

Поступила в редколлегию 11.06.2004.