УДК. 519.24

ПОСТРОЕНИЕ МНОГОМЕРНОЙ ПОЛИНОМИАЛЬНОЙ РЕГРЕССИИ. АКТИВНЫЙ ЭКСПЕРИМЕНТ

А.А. ПАВЛОВ, А.В. ЧЕХОВСКИЙ

Рассматривается конструктивный метод восстановления многомерной полиномиальной регрессии, представленной избыточным описанием. Распределение помехи является произвольным с неизвестной, но конечной дисперсией. Решение задачи основано на возможности проведения активного эксперимента. Приводятся практические рекомендации по использованию метода.

ВВЕДЕНИЕ

Задача конструктивного восстановления по статистическим данным регрессионной модели (детерминированной закономерности) — предмет исследования прикладного регрессионного анализа [1 – 8]. Наиболее употребляемым является метод наименьших квадратов. Практические проблемы реализации метода наименьших квадратов при построении многомерной полиномиальной регрессии заключаются в необходимости обращения плохо обусловленных матриц и отсутствии эффективных процедур восстановления истинной многомерной полиномиальной регрессии по ее избыточному описанию. Предлагаемый авторами метод в целом эффективно справляется с этими проблемами. Основы его построения сформулированы в работе [11].

ОДНОМЕРНЫЙ СЛУЧАЙ. ИССЛЕДОВАНИЕ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ С ИСПОЛЬЗОВАНИЕМ НОРМИРОВАННЫХ ОРТОГОНАЛЬНЫХ ПОЛИНОМОВ [9]

Постановку задачи и анализ известных результатов приведем в соответствие с работой [9].

Модель регрессии имеет вид

$$Y(x) = \theta_0 + \theta_1 x + \ldots + \theta_r x^r + E, \qquad (1)$$

где x — детерминированная переменная, значение которой в экспериментах может задаваться произвольно; θ_i, $i = \overline{0,r}$ — неизвестные коэффициенты; E — случайная величина с произвольным распределением; $ME = 0$

(M — знак математического ожидания); δ_E^2 (дисперсия) ограничена, ее значение неизвестно либо известна верхняя оценка.

Проведено n экспериментов, результатом которых являются две выборки объема n $(x_i,\ i=\overline{1,n};\ Y(x_i)=y_i,\ i=\overline{1,n})$.

В соответствии с (1)

$$y_i=\sum_{j=0}^{n}\theta_j x_i^j+\delta_i,\ \ i=\overline{1,n}, \tag{2}$$

где δ_i — неизвестная реализация случайной величины E в i-м эксперименте. Числа y_i, δ_i можно считать реализациями случайных величин Y_i, $i=\overline{1,n}$; Δ_i, $i=\overline{1,n}$, где Δ_i имеет распределение случайной величины E, а Y_i и Δ_i связаны соотношением

$$Y_i=\sum_{j}\theta_j x_i^j+\Delta_i. \tag{3}$$

Оценки неизвестных коэффициентов θ_j, $j=\overline{0,r}$ находятся из минимизации выражения

$$\min_{\theta_j,\,j=0,r}\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{r}\theta_j x_i^j\right)^2. \tag{4}$$

Введем матричные обозначения

$$A=\begin{pmatrix}1x_1,\ldots,x_1^r\\ 1x_n,\ldots,x_n^r\end{pmatrix},\ \ y=\left(y_1,\ldots y_n\right)^{\text{т}},$$

$$y=\left(Y_1,\ldots,Y_n\right)^{\text{т}},\ \ \theta=\left(\theta_0,\ldots,\theta_r\right)^{\text{т}},$$

$$\hat{\theta}=\left(\hat{\theta}_1,\ldots,\hat{\theta}_r\right)^{\text{т}},$$

где $\hat{\theta}_j$ оценки θ_j, $j=\overline{0,r}$ в соответствии с (4).

Тогда [9]

$$\hat{\theta}=\left(A^{\text{т}}A\right)^{-1}A^{\text{т}}y \tag{5}$$

либо

$$\hat{\theta}=\left(A^{\text{т}}A\right)^{-1}A^{\text{т}}Y,$$

если θ_j, $j=\overline{0,r}$ считать случайными величинами. Сложности, связанные с обращением матрицы $\left(A^{\text{т}}A\right)^{-1}$ исчезают, если от модели (1) перейти к модели регрессии, заданной с помощью нормированных ортогональных полиномов [9].

$$Y(x)=w_0\theta_0(x)+w_1\theta_1(x)+\ldots+w_2\theta_2(x)+E, \tag{6}$$

где $\theta_j(x),\ \ j=\overline{0,r}$ — нормированные ортогональные полиномы.

$$\theta_j(x)=q_{j_0}+q_{j_1}x+\ldots+q_{jj}x^j, \tag{7}$$

$$\sum_{i=1}^{n}\theta_j^2(x_i)=1,\ \sum_{i=1}^{n}\theta_j(x_i)\theta_l(x_i)=0,\ \forall j\neq l,\ q,l=\overline{0,r}.$$

Дж. Форсайт [10] предложил рекуррентную формулу для нахождения нормированных ортогональных полиномов

$$\lambda\theta_j(x)=x\theta_{j-1}(x)-\alpha\theta_{j-1}(x)-\beta\theta_{j-2}(x), \tag{8}$$

$$\alpha=\sum_{i=1}^{n}x_i\theta_{j-1}^2(x_i),\ \beta=\sum_{i=1}^{n}x_i\theta_{j-1}(x_i)\theta_{j-2}(x_i).$$

λ определяется из условия $\sum_{i=1}^{n}\theta_j^2(x_i)=1$.

$$\lambda=\sqrt{\sum_{i-1}^{n}\left(x_i\theta_{j-1}(x_i)-\alpha\theta_{j-1}(x_i)-\beta\theta_{j-2}(x_i)\right)^2}\ .$$

Для использования рекуррентной формулы (8) необходимо построить нормированные ортогональные полиномы $\theta_0(x)$ и $\theta_1(x)$. Очевидно, ими являются

$$\theta_0(x)=\frac{1}{\sqrt{n}},\ \theta_1(x)=-\frac{\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}+\frac{x}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}},$$

где $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$.

Применение метода наименьших квадратов к модели (6) приводит к следующим результатам [9].

Пусть $w=(w_0,...,w_r)^{\mathrm{T}}$, $\hat{w}=(\hat{w}_0,...,\hat{w}_r)^{\mathrm{T}}$, $\hat{w}_j$, $j=\overline{0,r}$ — оценки w_j, полученые методом наименьших квадратов $\hat{W}=(\hat{W}_0,...,\hat{W}_r)^{\mathrm{T}}$, $\hat{W}_j$, $j=\overline{0,r}$ — случайные величины, для которых $\hat{w}_j$ являются соответствующими реализациями.

Тогда

$$\hat{w}_j=\sum_{i=1}^{n}y_i\theta_j(x_i),\ j=\overline{0,r},\ \hat{W}_j=\sum_{i=1}^{n}Y_i\theta_j(x_i). \tag{9}$$

$$M\hat{W}_j=w_j,\ j=\overline{0,r},\ \operatorname{cov}(\hat{W}_j,\hat{W}_l)=0,\ \forall\ j\neq l.$$

$$D\hat{W}_j=\delta_E^2. \tag{10}$$

$$M\frac{\sum_{i=1}^{n}Y_i^2-\sum_{j=0}^{r}\hat{W}^2}{n-(r+1)}=\delta_E^2. \tag{11}$$

Связь моделей (1) и (6) следующая:

$$\theta_j=w_rq_{rj}+w_{r-1}q_{r-1}+\ldots+w_jq_{jj} \tag{12}$$

и соответственно

$$\hat{\theta}_j = \hat{w}_r q_{rj} + ... + \hat{w}_j q_{jj}, \quad j = \overline{0,r}, \tag{13}$$

либо

$$\hat{\theta}_j = \hat{W}_r q_{rj} + ... + \hat{W}_j q_{jj}, \tag{14}$$

если $\hat{\theta}_j$ считать случайной величиной.

При исследовании модели (1) либо эквивалентной ей (6) в работе [9] предполагалось, что r — степень полинома регрессии — известна заранее. Если это не так, то принято считать [1–9], что для произвольного распределения E нахождение истинного r является проблемой. Если E имеет нормальное распределение, то нахождение r сводится к проверке статических гипотез по критериям с известным распределением Фишера [9].

Покажем, что на самом деле проблема нахождения r имеет конструктивное решение для произвольного распределения E, а также покажем, как можно эффективно связать имеющиеся экспериментальные данные с точностью оценок неизвестных коэффициентов θ_j, $j = \overline{0,r}$.

Условие $\sum_{i=1}^{n} \theta_j^2(x_i) = 1$ с учетом (7) перепишем следующим образом:

$$\sum_{i=1}^{n} \left(\sum_{l=0}^{j} q_{jl} x_i^l \right)^2 = 1. \tag{15}$$

Найдем дисперсию $\hat{\theta}_j$. Учитывая, что $\operatorname{cov}(\hat{W}_l, \hat{W}_p) = 0$, из (10) и (14) $l \neq p$

получаем

$$D\hat{\theta}_j = \delta^2 \sum_{l=r}^{j} q_{lj}^2 . \tag{16}$$

Так как при неограниченном возрастании числа испытаний n минимум (4) асимптотически должен достигаться на истинных значениях коэффициентов θ_j, из анализа (15) и (16) следует, что при увеличении n модули значений коэффициентов $|q_{lj}|$, $l = \overline{r,j}$, $j = \overline{0,r}$ должны уменьшаться.

Аналитически в общем виде затруднительно связать числа x_i, $i = \overline{1,n}$; n; j $(j = \overline{0,r})$ с величиной q_{lj}, $l = \overline{r,j}$, $j = \overline{0,r}$. Тем не менее в случае активного эксперимента для эффективного решения прикладных задач (заданная точность, необходимое число вычислений, определение чисел $x_1, \ldots, x_n$) можно создать соответствующие статистические таблицы (табл. 1).

Таблица построена для линии регрессии, заданной полиномом пятого порядка. В первой колонке фиксируются различные значения n (количество значений детерминированного аргумента x). В колонках с номером $j\left(j = \overline{0,5}\right)$ заданы дисперсии коэффициентов $\hat{\theta}_j$, $j = \overline{0,5}$ как функция δ^2

(δ^2 — это дисперсия E либо ее верхняя оценка). Для построения таблицы были найдены все нормированные ортогональные полиномы $\theta_j(x)$, $j=\overline{0,5}$ (используются формулы (7), (8)). По формуле (16) определены соответствующие дисперсии. Значения x_i, $i=\overline{1,n}$ распределены с равным шагом по отрезку (–50,0; 50,0).

Таблица 1. Фрагмент возможной ситуации

n	0	1	2	3	4	5
10	$\sigma^2\cdot 0{,}400466$	$\sigma^2\cdot 0{,}0024855$	$\sigma^2\cdot 4{,}26\cdot 10^{-06}$	$\sigma^2\cdot 7{,}55\cdot 10^{-09}$	$\sigma^2\cdot 1{,}41\cdot 10^{-12}$	$\sigma^2\cdot 1{,}28\cdot 10^{-15}$
50	$\sigma^2\cdot 0{,}0706426$	$\sigma^2\cdot 0{,}0004607$	$\sigma^2\cdot 4{,}53\cdot 10^{-07}$	$\sigma^2\cdot 1{,}15\cdot 10^{-09}$	$\sigma^2\cdot 9{,}28\cdot 10^{-14}$	$\sigma^2\cdot 1{,}43\cdot 10^{-16}$
100	$\sigma^2\cdot 0{,}0351973$	$\sigma^2\cdot 0{,}0002298$	$\sigma^2\cdot 2{,}22\cdot 10^{-07}$	$\sigma^2\cdot 5{,}68\cdot 10^{-10}$	$\sigma^2\cdot 4{,}47\cdot 10^{-14}$	$\sigma^2\cdot 7{,}02\cdot 10^{-17}$
200	$\sigma^2\cdot 0{,}0175833$	$\sigma^2\cdot 0{,}0001149$	$\sigma^2\cdot 1{,}10\cdot 10^{-07}$	$\sigma^2\cdot 2{,}84\cdot 10^{-10}$	$\sigma^2\cdot 2{,}21\cdot 10^{-14}$	$\sigma^2\cdot 3{,}50\cdot 10^{-17}$
300	$\sigma^2\cdot 0{,}0117203$	$\sigma^2\cdot 7{,}66\cdot 10^{-05}$	$\sigma^2\cdot 7{,}36\cdot 10^{-08}$	$\sigma^2\cdot 1{,}89\cdot 10^{-10}$	$\sigma^2\cdot 1{,}47\cdot 10^{-14}$	$\sigma^2\cdot 2{,}33\cdot 10^{-17}$
500	$\sigma^2\cdot 0{,}0070316$	$\sigma^2\cdot 4{,}59\cdot 10^{-05}$	$\sigma^2\cdot 4{,}41\cdot 10^{-08}$	$\sigma^2\cdot 1{,}13\cdot 10^{-10}$	$\sigma^2\cdot 8{,}82\cdot 10^{-15}$	$\sigma^2\cdot 1{,}40\cdot 10^{-17}$
1000	$\sigma^2\cdot 0{,}0035157$	$\sigma^2\cdot 2{,}30\cdot 10^{-05}$	$\sigma^2\cdot 2{,}21\cdot 10^{-08}$	$\sigma^2\cdot 5{,}67\cdot 10^{-11}$	$\sigma^2\cdot 4{,}41\cdot 10^{-15}$	$\sigma^2\cdot 6{,}99\cdot 10^{-18}$
5000	$\sigma^2\cdot 0{,}0007031$	$\sigma^2\cdot 4{,}59\cdot 10^{-06}$	$\sigma^2\cdot 4{,}41\cdot 10^{-09}$	$\sigma^2\cdot 1{,}13\cdot 10^{-11}$	$\sigma^2\cdot 8{,}82\cdot 10^{-16}$	$\sigma^2\cdot 1{,}40\cdot 10^{-18}$
10000	$\sigma^2\cdot 0{,}0003516$	$\sigma^2\cdot 2{,}30\cdot 10^{-06}$	$\sigma^2\cdot 2{,}21\cdot 10^{-09}$	$\sigma^2\cdot 5{,}67\cdot 10^{-12}$	$\sigma^2\cdot 4{,}41\cdot 10^{-16}$	$\sigma^2\cdot 6{,}99\cdot 10^{-19}$

На качественном уровне анализ табл. 1 не зависит от величины $a>1$ отрезка разбиения $(-a,a)$ и величин r – степени полинома. Изложенные ниже выводы подтверждены экспериментально.

1. Приведенные значения дисперсий $\hat{\theta}_j$, $j=\overline{0,5}$ становятся конструктивными, если известна верхняя оценка δ^2 дисперсии E. Порядок δ_E^2 можно определить по реализации случайной величины [9]

$$\frac{R^{\text{т}}R}{n-(r+1)}=\frac{\sum_{i=1}^{n}Y_i^1-\sum_{j=0}^{r}W_j^2}{n-(r+1)},\ \text{так как}\ M\frac{R^{\text{т}}R}{n-(r+1)}=\delta_E^2.$$

Далее будет показано, что истинное значение r находят очевидным образом.

2. Чем больше j, тем меньше $\hat{\theta}_j$ при фиксированном n. Действительно, при $n=10$ $D\hat{\theta}_0=\delta^2\cdot 0{,}400466$; $D\hat{\theta}_1=\delta^2\cdot 0{,}0024855$; $D\hat{\theta}_2=\delta^2\times$ $\times 4{,}2610^{-6}\ldots D\hat{\theta}_5=\delta^2\cdot 128\cdot 10^{-15}$, т.е. с увеличением j значение $D\hat{\theta}_j$ уменьшается на порядок.

3. По минимальному количеству испытаний можно определить истинную степень полинома линии регрессии. В нашем примере при $n=10$ дисперсия оценки коэффициента при x^2 уже равна $\delta^2\cdot 4{,}26\cdot 10^{-6}$. Т.е. если истинная линия регрессии прямая, то реально оценками $\hat{\theta}_2,\hat{\theta}_3,\hat{\theta}_4,\hat{\theta}_5$ будут нули с точностью до соответствующих знаков после запятой (закон трех сигм).

4. Необходимое количество испытаний n определяется заданной точностью для нахождения $\hat{\theta}_j$ с наименьшим $j\,(j=0)$. Если эксперименты до-

рогие, то реально эффективно оценивать $\hat{\theta}_j$, начиная с $j=1$ (из анализа табл. 1 видно, что значения дисперсий $D\hat{\theta}_0$ и $D\hat{\theta}_1$ одного порядка достигаются на числе экспериментов, отличающихся на два порядка).

Таким образом, точность оценки θ_0 необходимо связывать с самой числовой оценкой θ_0 (чем больше по модулю это значение, тем достовернее результат). Если оценка θ_0 оказывается недостаточно точной, то полученное выражение для линии регрессии необходимо использовать в тех задачах, для решения которых величина θ_0 не имеет значения (например, сравнение значений линии регрессии для различных значений ее аргумента).

В некоторых задачах массив x_i, $i=\overline{1,n}$ может быть задан заранее и экспериментатор не может его изменить. Тогда до проведения эксперимента по формулам (7), (8), (16) можно найти дисперсии $\hat{\theta}_j$, $j=0,r$ (r можно задать избыточным) и провести предварительный анализ будущих результатов.

Пример 1.

Истинная модель имеет вид

$$y(x)=5+4x+3x^2+2x^3+x^4+0\cdot x^5+E\,. \qquad (17)$$

Регрессионная модель всегда должна задаваться избыточной. В нашем примере исследователь знает, что регрессионная модель является полиномом не выше пятой степени. Случайная величина E имеет нулевое математическое ожидание, равномерное распределение, $\delta_E=50$. Входные значения x_i равномерно распределены по отрезку (−50,0; 50,0) с шагом $\frac{100}{n}$.

Имитируются эксперименты с помощью реализаций случайной величины E. Исходные данные x_i, $y_i=5+4x_i+3{x_i}^2+2{x_i}^3+{x_i}^4+E_i$, $i=\overline{1,n}$. E_i — реализация случайной величины E.

Для генерации случайных чисел использована часть библиотеки расширений для C++ boost. http://www.boost.org/doc/libs/1_36_0/libs/random/index.html.

Для восстановления зависимости (17) по формуле (8) строятся шесть нормированных ортогональных полиномов, а по (13) — оценки коэффициентов в линии регрессии (16). Результаты эксперимента показаны ниже.

Одномерная регрессия, равномерное распределение $(\sigma_E=50)$.

Исходные коэффициенты: 5 , 4 , 3 , 2 , 1 , 0 ; цифры в квадратных скобках — количество чисел в круглых скобках. Количество испытаний $n=10$.

Ортогональные полиномы

[1] (0,316228).

[2] (0,0550482; 0,0110096).

[3] (−0,348155; 0,00435194; 0,000435194).

[4] (−0,12955; −0,0250104; 0,000269896; $1{,}79931\cdot10^{-5}$).

[5] (0,336581; −0,0155824; −0,00148033; $1{,}55824\cdot10^{-5}$; $7{,}79122\cdot10^{-7}$).

[6] (0,214834; 0,0384315; −0,00134272; $-8{,}35467\cdot10^{-5;}$ $8{,}95144\cdot10^{-7;}$ $3{,}58057\cdot10^{-8}$).

Оценки коэффициентов

[6] (8,21467; 4,40095; 2,99323; 1,99954; 1; $1{,}00514\cdot10^{-7}$).

Дисперсии коэффициентов

[6] (1001,17; 6,21366; 0,0106413; $1{,}88666\cdot10^{-5}$; $3{,}52078\cdot10^{-9}$; $3{,}20513\cdot10^{-12}$).

Количество испытаний $n = 50$.

Ортогональные полиномы

[1] (0,141421).

[2] (0,00489996; 0,00489996).

[3] (−0,158019; 0,000379853; 0,000189927).

[4] (−0,0112385; −0,0112235; $2.2513\cdot10^{-5}$;$7{,}50433\cdot10^{-6}$).

[5] (0,15878; −0,00127706; −0,000637333; $1{,}19511\cdot10^{-6}$; $2{,}98776\cdot10^{-7}$).

[6] (0,0176422; 0,0175761; $-9{,}91938\cdot10^{-5}$; $-3{,}29849\cdot10^{-5}$; $5{,}97553\cdot10^{-8}$; $1{,}19511\cdot10^{-8}$).

Оценки коэффициентов

[6] (5,15354; 3,91832; 3,00231; 2,00013; 0,999999; $-3{,}05835\cdot10^{-8}$).

Дисперсии коэффициентов

[6] (176,607; 1,15167; 0,00113153; $2{,}86437\cdot10^{-6}$; $2{,}32095\cdot10^{-10}$; $3{,}57069\cdot10^{-13}$).

В табл. 2 приведены оценки коэффициентов, точное значение которых равно 5,4,3,2,1,0, соответственно, для количества испытаний *n*.

Таблица 2. Оценки коэффициентов

n	$\theta_0 = 5$	$\theta_1 = 4$	$\theta_2 = 3$	$\theta_3 = 2$	$\theta_4 = 1$	$\theta_5 = 0$
10	8,21467	4,40095	2,99323	1,99954	1	$1{,}100510^{-7}$
50	5,15354	3,91832	3,00231	2,00013	0,999999	$-3{,}05835\cdot10^{-8}$
100	4,13942	3,92085	3,00063	2,00003	0,999999	$1{,}21093\cdot10^{-8}$
200	4,5745	3,92143	3,00406	2,00005	0,999999	$-2{,}61741\cdot10^{-9}$
300	5,39774	3,89625	2,99871	2,00013	1	$-3{,}14599\cdot10^{-8}$
500	5,23945	4,03909	3,00081	1,99996	1	$11189\cdot10^{-8}$
1000	4,8637	4,00444	2,99924	2	1	$-3{,}30327\cdot10^{-9}$
5000	5,20518	4,01987	2,99942	1,99997	1	$5823\cdot10^{-9}$
10000	4,95045	3,98802	3,0003	2,00002	1	$-8{,}36543\cdot10^{-9}$

Для количества испытаний $n \geq 50$ погрешность оценки коэффициентов не превышает (табл. 2) для $\theta_0 - 0{,}15$; $\theta_1 - 0{,}08$; $\theta_2 - 0{,}002$; $\theta_3 - 1{,}3\cdot10^{-4}$; $\theta_4 - 10^{-6}$; $\theta_5 - 10^{-7}$.

Таким образом, в результате моделирования достаточного количества испытаний можно создать статистические таблицы, каждая из которых составлена для фиксированных концов отрезка принадлежности аргумента x и содержит для ряда статистически обоснованных вероятностей значения погрешностей нахождения коэффициентов θ_j, $j = \overline{0,r}$ (зависящих от δ_E^2 или ее верхней оценки δ^2) для различного количества испытаний. Такие таблицы позволят заранее определить минимально необходимое число ис-

пытаний, для которых оценки коэффициентов θ_j (кроме, возможно, θ_0) находятся с приемлемой для практики точностью.

Примечание. Очевидно, чем больше длина интервала изменений аргумента x, тем меньше минимально необходимое число испытаний.

МНОГОМЕРНЫЙ СЛУЧАЙ

Возможность для одномерного случая практически гарантировано находить степень полинома линии регрессии и вычислять с допустимой вероятностью с заданной погрешностью коэффициенты этого полинома позволяют предложить достаточно эффективную процедуру восстановления многомерной полиномиальной линии регрессии (при условии реализации активного эксперимента).

Пусть многомерная модель задается в виде

$$y(\bar{x}) = \sum_{\forall(i_1,\ldots,i_t)\in \mathrm{K}} \sum_{\forall(j_1,\ldots,j_t)\in \mathrm{K}(i_1,\ldots,i_t)} b_{i_1,\ldots,i_t}^{j_1,\ldots,j_t} \left(x_{i_1}\right)^{j_1} \left(x_{i_2}\right)^{j_2} \ldots \left(x_{i_t}\right)^{j_t} + E, \qquad (18)$$

где $\bar{x} = (x_1,\ldots,x_n)^{\mathrm{T}}$ — детерминированный вектор входных переменных; x_i — i-я компонента вектора $\bar{x}$; $b_{i_1,\ldots,i_t}^{j_1,\ldots,j_t}$ — неизвестные коэффициенты; j_l — натуральные числа; j_l, i_l — натуральные индексы из множества $\{1,\ldots,n\}$; E — случайная величина с нулевым математическим ожиданием и ограниченной неизвестной дисперсией δ_E^2 (как и в одномерном случае может быть известна верхняя оценка δ_E^2).

Модель (18) является избыточной (возможно, некоторые из коэффициентов $b_{i_1,\ldots,i_t}^{j_1,\ldots,j_t}$ равны нулю). Для удобства линию регрессии модели (18) представим иначе.

$$\sum_{l=1}^{n} \sum_{\forall(i_1,\ldots,i_t)\in \mathrm{K}_l} \sum_{\forall(j_1,\ldots,j_t)\in \mathrm{K}_l(i_1,\ldots,i_t)} b_{i_1,\ldots,i_t}^{j_1,\ldots,j_t} \left(x_{i_1}\right)^{j_1} \left(x_{i_2}\right)^{j_2} \ldots \left(x_{i_t}\right)^{j_t}. \qquad (19)$$

Составляющие

$$\sum_{\forall(i_1,\ldots,i_t)\in \mathrm{K}_1} \sum_{\forall(j_1,\ldots,j_t)\in \mathrm{K}_1(i_1,\ldots,i_t)} b_{i_1,\ldots,i_t}^{j_1,\ldots,j_t} \left(x_{i_1}\right)^{j_1} \left(x_{i_2}\right)^{j_2} \ldots \left(x_{i_t}\right)^{j_t} \qquad (20)$$

содержат все слагаемые из (18), в каждое из которых входит компонента x_1. Составляющие

$$\sum_{\forall(i_1,\ldots,i_t)\in \mathrm{K}_l} \sum_{\forall(j_1,\ldots,j_t)\in \mathrm{K}_l(i_1,\ldots,i_t)} b_{i_1,\ldots,i_t}^{j_1,\ldots,j_t} \left(x_{i_1}\right)^{j_1} \left(x_{i_2}\right)^{j_2} \ldots \left(x_{i_t}\right)^{j_t}, \quad l = \overline{2,n} \qquad (21)$$

содержат все слагаемые из (18), в каждое из которых входит компонента x_l, за исключением тех составляющих, которые вошли в (20) и (21) для

$$\forall (i_1,\ldots,i_t) \in \mathrm{K}_m, \quad \forall (j_1,\ldots,j_t) \in \mathrm{K}_m(i_1,\ldots,t_t), \quad m = \overline{1,l-1}.$$

Рассмотрим составляющую (20).

Обозначим M_j^1, $j=\overline{1,n_1}$ количество слагаемых, каждое из которых содержит x_1 в j-й степени.

$\mathrm{M}^1=\max\limits_j M_j^1$, $j=\overline{1,n_1}$, n_1 — максимальная степень полинома от переменной x_1.

Фиксируем M^1 наборов значений компонент $x_2^s,\ldots,x_n^s$, $s=\overline{1,M^1}$. На числа x_i^s, $i=\overline{2,n}$, $s=\overline{1,M^1}$ накладывается единственное условие — определенные ниже квадратные матрицы должны быть невырожденными.

Реализуем M^1 экспериментов, в каждом из которых (s-м $s=\overline{1,M^1}$) переменные $x_2,\ldots,x_n$ принимают фиксированные значения x_i^5 $(i=\overline{2,n})$, а x_1 изменяется, как при построении одномерной полиномиальной регрессии. При фиксированных значениях переменных $x_2,\ldots,x_n$ в s-м эксперименте $(s=\overline{1,M^1})$ многомерная линия регрессии превращается в полином от переменной x_1 степени n_1.

Для каждого s-го эксперимента $(s=\overline{1,M^1})$ находим (16) значения дисперсий $D\hat{\theta}_j^s$, $j=\overline{1,n_1}$ и эти числа ранжируем по возрастанию их значений при фиксированном j. Получим n_1 проранжированных последовательностей оценок коэффициентов $\theta_j^{s_1},\ldots,\theta_j^{s_{M^1}}$ $(j=\overline{1,n_1})$.

Эти результаты позволяют сформировать n_1 систем линейных уравнений, решениями которых являются значения всех коэффициентов $b_{i_1\ldots i_t}^{j_1\ldots j_t}$ в выражении (20).

Действительно, в каждом из s экспериментов неизвестные коэффициенты θ_j^s $(j=\overline{1,n_1})$ одномерной полиномиальной регрессии степени n_1 от переменной x_1 определяются следующим образом: необходимо из всех членов выражения (содержащих переменную x_1 в степени j) вынести x_1^j. Полученное выражение для θ_j^s содержит только M_j^1 неизвестных коэффициентов вида $b_{i_1\ldots i_t}^{j_1\ldots j_t}$, так как в каждом s-м эксперименте при изменении значений переменной x_1 переменные x_i, $i=\overline{2,n}$ принимают одно и то же фиксированное значение x_i^s, $i=\overline{2,n}$. Таким образом, при построении системы линейных уравнений для нахождения M_1^1 коэффициентов вида $b_{i_1\ldots i_t}^{j_1\ldots j_t}$ надо использовать M_1^1 чисел $\hat{\theta}_1^{s_1},\ldots,\hat{\theta}_1^{s_{M_1^1}}$ (они имеют наименьшую дисперсию).

Для определения верхних статистических оценок точности нахождения M_1^1 коэффициента вида $b_{i_1\ldots i_t}^{j_1\ldots j_t}$ полученную систему линейных уравнений условно запишем так:

$$A\begin{pmatrix} x_1 \\ \vdots \\ x_{M_1^1} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_1^{s_1} \\ \vdots \\ \hat{\theta}_1^{s_{M_1^1}} \end{pmatrix}, \qquad (22)$$

где x_i, $i = \overline{1, M_1^1}$ — переменные (соответствующие M_1^1 переменным вида $b_{i_1 \ldots i_t}^{j_1 \ldots j_t}$).

Оценки $\hat{\theta}_1^{s_l}$, $l = \overline{1, M_1^1}$ с заданной статистически значимой вероятностью p оценивают $\theta_1^{S_l}$ с погрешностью, по модулю не превышающей чисел $\Delta_1^{S_l}$, $l = \overline{1, M_1^1}$. Тогда с вероятностью p максимальная величина погрешности нахождения точных значений M_1^1 соответствующих коэффициентов $b_{i_1 \ldots i_t}^{j_1 \ldots j_t}$ имеет вид

$$\max_{j=\overline{1,M_1^1}} \left\{ \max \left(\sum\nolimits^{(+)} a_{jl}^{-1} \Delta_1^{s_l}, \sum_l {}^{(-)} \left| a_{jl}^{-1} \right| \Delta_1^{s_l} \right) \right\}, \qquad (23)$$

где $\sum_l {}^{(+)} a_{jl}^{-1} \Delta_1^{s_l}$ берется по всем $l = \overline{1, M_1^1}$, для которых $a_{jl}^{-1} \geq 0$; $\sum_l {}^{(-)} a_{jl}^{-1} \Delta_1^{s_l}$ берется по всем $l = \overline{1, M_1^1}$, для которых $a_{jl}^{-1} < 0$; a_{jl}^{-1} — jl-й элемент матрицы A^{-1}.

Как указывалось выше, предполагается x_i^s, $i = \overline{2, n}$, $s = \overline{1, M_1^1}$ выбраны так, что матрица A^{-1} существует.

Аналогично строятся системы линейных уравнений (правыми частями которых являются столбцы $(\hat{\theta}_l^{s_1}, \ldots, \hat{\theta}_l^{s_{M_l^1}})^{\mathrm{T}}$, $l = \overline{2, n_1}$) для нахождения остальных коэффициентов $b_{i_1 \ldots i_t}^{j_1 \ldots j_t}$ из выражения (20). Аналогично строятся все оценки вида (23).

Процедуры нахождения всех неизвестных коэффициентов $b_{i_1 \ldots i_t}^{j_1 \ldots j_t}$ из выражений (21) для $l = \overline{2, n_1}$ полностью повторяют процедуру, изложенную для выражения (20).

Оценка константы в выражении (18) может быть получена как среднее арифметическое по всем проведенным испытаниям разностей $y_i - (\hat{y}(\bar{x}_i) - \theta_0)$, где y_i — значение выходной переменной модели, когда на вход подается векторное значение $\bar{x}_i$, а $\hat{y}(\bar{x}_i) - \theta_0$ — значение выражения (19) для x_i, из которого исключен коэффициент θ_0, и вместо $b_{i_1 \ldots i_t}^{j_1 \ldots j_t}$ подставлены их оценки.

Если верхняя оценка δ_E^2 неизвестна, то ее можно эффективно оценить как среднее арифметическое оценок δ_E^2 (11) по всем одномерным регрессиям.

ОБОБЩЕНИЯ

1. Очевидно, что полученные результаты применимы для случая, когда выражение (18) вместо переменных $x_1,\ldots,x_n$ содержит переменные $z_1,\ldots,z_m$, $m<n$, где $z_j=f_j(\bar{x}_j)$, $j=\overline{1,m}$, а компонентами векторов $\bar{x}_j$ являются компоненты вектора $\bar{x}$, и множества компонент векторов $\bar{x}_j$, $j=\overline{1,m}$ не пересекаются; f_j — непрерывные функции, ограниченные при ограниченных значениях своих аргументов.

2. Задача построения многомерной регрессии очевидным образом обобщается на случай, когда при построении одномерных регрессий на модель действуют разные случайные величины E_l (l номер одномерной регрессии), $ME_l=0$, $DE_l=\delta^2_{E_l}<\infty$. В общем виде распределения случайных величин E_l (при фиксированном l) могут не совпадать между собой. Анализ формул (9), (10), (16) показывает, что при построении одномерных регрессий в экспериментах на регрессионную модель аддитивно могут воздействовать независимые случайные величины E_l с различными распределениями, имеющие нулевые математические ожидания и одинаковые дисперсии для фиксированного l. Для разных l дисперсий $\delta^2_{E_l}$ могут быть различными.

Пример 2 (многомерная регрессия).

Исходная модель линии регрессии задается в виде избыточного полинома

$$y=12+11x_1+10x_2+9x_3+8x_1x_2+7x_1^2+6x_1x_3+5x_1x_3^2+4x_2x_3^2+$$
$$+0\cdot x_1^3x_2^2+0\cdot x_1^2x_2+0\cdot x_1x_2x_3+E. \qquad (24)$$

Обозначим $a_0,a_1,a_2,\ldots,a_{11}$ коэффициенты линии регрессии, которые считаются неизвестными. E — случайная величина, имеющая нормальное распределение $ME=0$, $\delta_E=50$.

В этом примере $K_1=\{1;\ 1,2;\ 1,3;\ 1,2,3;\}$; $K_1(1)=1$; $K_1(1,2)=\{1,1;3,2;2,1;\}$; $K_1(1,3)=\{1,1;\}$; $K(1,2,3)=\{1,1,1\}$. Аналогично определяются все K_l, $K_l(i_1,\ldots,i_t)$, $l=\overline{2,3}$.

Для переменной x_1 последовательно фиксируются следующие значения переменных x_2, x_3: $x_2=3{,}39877$, $x_3=9{,}36811$; $x_2=9{,}97516$; $x_3=0{,}137846$; $x_2=-2{,}87215$; $x_3=8{,}77249$; $x_2=9{,}44462$; $x_3=0{,}158521$; $x_2=-4{,}05535$; $x_3=5{,}95574$.

Для каждого набора значений переменных x_2, x_3 восстанавливается одномерная регрессия от переменной x_1, коэффициенты которой позволяют составить такие системы:

- из пяти равенств для нахождения коэффициентов a_1,a_4,a_6,a_7,a_{11} (коэффициенты в (24) при x_1 в первой степени);
- из двух равенств для нахождения коэффициентов a_5,a_{10} (коэффициенты в (24) при x_1 во второй степени);

Таблица 3. Оценки коэффициентов

Количество испытаний, n	Исходные коэффициенты											
	12	11	10	9	8	7	6	5	4	0	0	0
	Оценки коэффициентов											
10	40,9921	14,8957	7,66171	9,51358	8,33979	6,98053	6,29187	4,91568	4,0234	$9{,}4976\cdot10^{-7}$	−0,0036328	−0,0036328
50	12,2543	10,9678	10,0237	8,99369	8,00568	7,00012	6,00292	5,00013	3,99847	$-1{,}10744\cdot10^{-7}$	$-1{,}73976\cdot10^{-5}$	−0,000546877
60	12,6497	11,0074	10,0012	8,99711	7,99925	6,99927	6,00036	4,99981	4,00002	$-6{,}07745\cdot10^{-9}$	−0,000195548	$-9{,}70067\cdot10^{-5}$
70	12,3904	11,0119	10,011	8,99026	7,99915	6,99996	5,99956	4,99947	3,99985	$-2{,}53199\cdot10^{-7}$	$-5{,}09079\cdot10^{-5}$	0,000156464
80	16,0301	10,974	9,49174	9,01362	7,98281	7,00023	6,00584	4,99965	4,0099	$1{,}62543\cdot10^{-6}$	0,00010457	0,00248905
90	12,451	10,8906	10,0072	8,84522	7,9847	7,00009	5,98836	5,00147	3,99982	$-2{,}44712\cdot10^{-7}$	$9{,}15582\cdot10^{-6}$	−0,00280861
100	12,0322	10,989	9,97153	9,00068	8,00138	7,00074	6,00055	5,0004	4,00003	$-4{,}14521\cdot10^{-8}$	0,000120744	0,000133714
110	12,5814	11,0148	9,96272	9,02955	7,99795	7,00002	5,99788	4,99986	4,00069	$-2{,}19275\cdot10^{-8}$	$6{,}60417\cdot10^{-6}$	0,000388347
120	12,0329	10,9907	10,0025	8,99546	8,00007	6,99995	5,99918	5,00037	3,99996	$2{,}15863\cdot10^{-8}$	$-2{,}63362\cdot10^{-7}$	$9{,}69276\cdot10^{-5}$
130	11,1961	10,9948	9,99605	8,97468	7,99934	7,00009	5,99646	4,99949	4,00176	$1{,}42098\cdot10^{-7}$	$6{,}51219\cdot10^{-6}$	0,00025512
140	10,3128	10,9497	10,0357	8,99633	8,003	6,99987	6,00393	5,0031	3,9997	$2{,}22283\cdot10^{-7}$	$1{,}15009\cdot10^{-5}$	−0,00312126
150	11,9329	11,0049	10,013	8,99544	8,00027	7,00014	6,00043	4,99998	3,99983	$-1{,}04791\cdot10^{-5}$	$4{,}08502\cdot10^{-5}$	$2{,}48244\cdot10^{-5}$
160	10,3259	11,0211	9,99883	9,12786	8,00483	6,99923	5,98763	5,00139	4,00048	$1{,}24217\cdot10^{-7}$	−0,000129896	−0,00137244
170	17,7601	10,9917	10,1205	9,00723	7,99704	6,99985	6,00636	5,00079	3,99445	$-3{,}96317\cdot10^{-7}$	$-8{,}84427\cdot10^{-6}$	−0,000230362
180	12,1722	11,0066	9,99337	8,99589	8,00038	6,99985	6,00097	4,99975	4,00019	$-3{,}13437\cdot10^{-7}$	$-3{,}38322\cdot10^{-5}$	$-6{,}88005\cdot10^{-5}$
190	11,658	10,9904	10,0252	9,01683	8,00071	7,00002	5,99723	4,99979	3,99829	$-2{,}6893\cdot10^{-8}$	$-1{,}7055\cdot10^{-5}$	$1{,}57857\cdot10^{-5}$
200	12,545	10,897	9,9477	8,56488	7,98017	7,00054	5,98627	4,99466	4,02602	0,000117637	−0,000981739	0,0113098
210	4,81084	10,9763	9,75834	7,87872	8,01347	6,99998	5,97052	4,98954	4,00278	$-2{,}00435\cdot10^{-7}$	$-2{,}14216\cdot10^{-6}$	0,017214
220	13,0731	10,9937	10,0863	9,03774	8,0098	6,99992	6,00147	5,0003	3,99585	$-1{,}01319\cdot10^{-7}$	$-1{,}65513\cdot10^{-6}$	0,00133661
230	12,9133	10,9997	10,0019	9,03141	7,99923	6,99996	5,99996	4,99892	4,00002	$-6{,}88059\cdot10^{-7}$	$9{,}57041\cdot10^{-6}$	0,000462845
240	15,951	11,0006	10,0061	9,0079	7,99844	7,00017	5,99809	4,99994	3,99547	$1{,}64937\cdot10^{-7}$	$-2{,}37061\cdot10^{-5}$	0,000364645
250	11,9927	11,0092	9,9913	9,00235	8,00143	7,00003	6,00083	4,99956	4,00022	$-5{,}37852\cdot10^{-8}$	$-1{,}22701\cdot10^{-6}$	−0,000401564

- из одного равенства для нахождения a_9 (коэффициент в (24) при x_1 в третьей степени).

Для переменной x_2 фиксируются значения переменных x_1, x_3: $x_1 = 1{,}63987$; $x_3 = 8{,}68112$; $x_1 = -7{,}02188$; $x_3 = -2{,}30255$. Восстанавливаются две одномерные регрессии от переменной x_2. Составляется система из двух равенств для коэффициентов a_2, a_8 (коэффициенты в (24) при x_2 в первой степени). Находятся a_2 и a_8. Для фиксированных переменных $x_1 = 6{,}40309$; $x_2 = 0{,}175851$ строится одномерная регрессия от переменной x_3. Коэффициентом при x_3 в первой степени является a_3. Последним находится коэффициент a_0.

В табл. 3 приведены оценки точных коэффициентов многомерной регрессии, полученные для разного количества экспериментов (n) для каждой одномерной регрессии.

ВЫВОДЫ

Приведен конструктивный метод восстановления многомерной полиномиальной регрессии, представленной избыточным описанием. Показано, что при использовании нормированных ортогональных полиномов Форсайта эту задачу можно свести к последовательности задач восстановления одномерных регрессий и решению систем линейных уравнений с постоянными коэффициентами. На основе анализа проведенных вычислительных экспериментов приведены конкретные практические рекомендации по использованию предложенного метода.

ЛИТЕРАТУРА

1. *Адлер Ю.П., Маркова Е.В., Грановский Ю.В.* Планирование эксперимента при поиске оптимальных условий. — 2-е изд., перераб. и доп. — М.: Наука, 1976. — 280 с.
2. *Айвазян С.А.* Многомерный статистический анализ // Математическая энциклопедия / Под. ред. И.М. Виноградова. — М.: Статистика,1982. — Т. 3. — С. 732–738.
3. *Андерсон Т.* Введение в многомерный статистический анализ / Пер. с англ. Ю.Ф. Кичатова. Под ред. Б.В. Гнеденко. — М.: Физматгиз, 1963. — 500 с.
4. *Ершов А.А.* Стабильные методы оценки параметров: Обзор // Автоматика и телемеханика. — 1978. — № 8. — С. 66–100.
5. *Колмогоров А.Н.* К обоснованию метода наименьших квадратов // Успехи математических наук. — 1946. — Т. 1. — Вып. 1. — С. 57–70.
6. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений / Пер. с англ. Х.Д. Икрамова. — М.: Мир, 1980. — 280 с.
7. *Радченко С.Г.* Устойчивые методы оценивания статистических моделей. — Киев: ПП «Санспарель», 2005. — 504 с.
8. *Дрейпер Норманн Р., Смит Гарри.* Прикладной регрессионный анализ: 3-е изд. / Пер. с англ. — М.: Изд. дом «Вильямс», 2007. — 912 с.
9. *Худсон Д.* Статистика для физиков. — М.: Мир, 1970. — 186 с.
10. *Forsythe G. //* Sos.Ind. Appl. Math. — 1957. — № 5. — С. 74.
11. *Павлов А.А., Чеховский А.В.* Сведение задачи построения многомерной регрессии к последовательности одномерных задач // Вісн. НТУУ «КПІ». Інформатика, управління та обчислювальна техніка. — 2008. — № 48. — С. 18–20.

Поступила 03.06.2008