УДК 519.242.5:519.233.5

С.Г. РАДЧЕНКО

КОНЦЕПЦИЯ ОРТОГОНАЛЬНОСТИ ВО МНОЖЕСТВЕННОМ РЕГРЕССИОННОМ АНАЛИЗЕ

Анотація. *Показано, що одержання регресійних моделей пов'язано з рішенням обернених і некоректних задач і потребує розробки й використання спеціальних методів їх рішення на основі концепції ортогональності. Викладено метод стійкого оцінювання моделей на основі концепції ортогональності ефектів і приведено алгоритм та програмний засіб розв'язання задач.*
Ключові слова: *регресійний аналіз, концепція ортогональності, планування експерименту, некоректно поставлені задачі.*

Аннотация. *Показано, что получение регрессионных моделей связано с решением обратных и некорректно поставленных задач и требует разработки и использования специальных методов их решения на основе концепции ортогональности. Изложен метод устойчивого оценивания моделей на основе концепции ортогональности эффектов и приведены алгоритм и программное средство решения задач.*
Ключевые слова: *регрессионный анализ, концепция ортогональности, планирование эксперимента, некорректно поставленные задачи.*

Abstract. *The paper deals with setting experimental research from general system positions. The requirements to stable (robust) plans of experiments, stable structures of multifactor statistical models and stability of the model coefficients have been formulated for the first time. Examples of a successful use of the developed method of correct solution of multifactor regression problems are presented.*
Keywords: *experimental investigation method, robust experiment design, stable structure, model coefficients, correct solution, regression problems.*

1. Введение

Постановка проблемы

При создании и совершенствовании наукоемких изделий, высоких технологий широко используются многофакторные статистические линейные по параметрам и, в общем случае, нелинейные по факторам модели. Сложность и новизна технических и технологических систем такова, что для получения моделей применяется экспериментально-статистический подход. Получаемые исходные данные являются случайными величинами и представляют результат суммарного влияния управляемых факторов, которые входят в модель, неуправляемых и неконтролируемых факторов.

Решаемая задача относится к классу обратных задач: определение коэффициентов **B** в уравнении **Y=XB+E** по измеренному результату **Y** и условиям наблюдения **X**; **E** – значение случайной ошибки ε. Определение коэффициентов **B** проводится с использованием метода наименьших квадратов. Рассматриваемые обратные задачи в большинстве случаев являются некорректно поставленными задачами. Их решение требует специально разработанных методов.

Анализ публикаций по теме исследований

О. Коши впервые обратил внимание на необходимость использования в методе наименьших квадратов идеи ортогональности.

Р. Фишер заложил основы дисперсионного анализа, теории планирования эксперимента. Идея ортогональности была реализована в планах 2^k, 2^{k-p} и комбинаторных планах экспериментов: латинских, греко-латинских квадратов и других планах. Указанные планы использовались в прикладных исследованиях, в частности, агробиологических.

ISSN 1028-9763. Математичні машини і системи, 2011, № 2

В 50-х годах прошлого века были разработаны ортогональные центральные композиционные планы 2-го порядка.

В планировании эксперимента было предложено эффекты факторов выражать в виде ортогональных контрастов, которые эквивалентны системе ортогональных полиномов Чебышева.

Группа авторов учебного пособия [1, с. 9] отмечает, что «…задача аппроксимации эмпирических зависимостей как на компьютере, так и без него, начиная с ее постановки, во многих отношениях строго не решается, остается неопределенной и неоднозначной».

Классическая теория планирования эксперимента считает, что структура статистической модели задана, после чего строится план эксперимента, соответствующий определенным критериям. В большинстве случаев структура статистической модели исследователю не известна. В работе [2, с. 22, 175–176] отмечается, что «не существует стандартных приемов и методов, которые образовывали бы строгую теоретическую базу для решения этой важнейшей задачи» – «задачи правильного определения структуры модели».

В других публикациях по статистическому моделированию высказываются аналогичные суждения.

Необходимо программно и алгоритмически реализовать «процедуры, помогающие осуществить выбор общего параметрического вида математической модели в задачах регрессии или классификации; различные подходы к получению устойчивых (в определенном смысле) статистических выводов» [3, с. 101].

Практика решения прикладных реальных задач показала, что форма факторного пространства может не соответствовать традиционной – планирование на кубе, сфере, симплексе. В этих задачах факторы коррелированы друг с другом. Необходимо использовать методы устойчивого оценивания статистических моделей в условиях исходной мультиколлинеарности факторов.

Цель статьи

Разработка концепции использования ортогонального планируемого эксперимента и ортогональной структуры статистической модели при условии, что структура модели исследователю не известна. Модель должна соответствовать критериям адекватности, статистической эффективности, устойчивости, семантичности информационной. По одному и тому же плану эксперимента должно быть возможным устойчивое получение различных статистических моделей различных критериев качества систем.

2. Основные подходы ортогонального представления статистических моделей

В основу концепции положено свойство плана полного факторного эксперимента, заключающееся в том, что все эффекты ортогональны друг к другу. Это вытекает из теоремы Бродского В.З. [4, с. 26–29]. Расширенная матрица **X** главных эффектов и взаимодействий содержит столбец фиктивного фактора $x_0 \equiv 1$, столбцы всех главных эффектов и всех возможных взаимодействий главных эффектов. Если эффекты факторов и взаимодействий факторов выразить в виде системы ортогональных нормированных контрастов, т.е.

$$\sum_{u=1}^{N} x_{iu}^{(p)} = 0, \qquad \sum_{u=1}^{N} x_{iu}^{(p)} \times x_{ju}^{(q)} = 0,$$

$$\sum_{u=1}^{N} [x_{iu}^{(p)}]^2 = N, \qquad \sum_{u=1}^{N} [x_{iu}^{(p)} \times x_{ju}^{(q)}]^2 = N,$$

то матрица дисперсий-ковариаций примет вид

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2(\varepsilon) = (1/N)\mathbf{E}\,\sigma^2(\varepsilon),$$

где $x_{iu}^{(p)}$ – значение p-го ортогонального контраста i-го фактора для u-й строки матрицы планирования, $1 \le u \le N$, $1 \le p \le s_i - 1$;

$x_{ju}^{(q)}$ – значение q-го ортогонального контраста j-го фактора для u-й строки матрицы планирования, $1 \le q \le s_j - 1$, $1 \le i < j \le k$;

$\mathbf{X}$ – матрица эффектов полного факторного эксперимента;

$\sigma^2(\varepsilon)$ – теоретическое значение дисперсии воспроизводимости результатов опытов;

N – число опытов в плане эксперимента;

$\mathbf{E}$ – единичная матрица.

По теореме Хоттелинга [5, с. 62–63] выполнение требования ортогональности всех эффектов приводит к минимизации дисперсий коэффициентов статистической модели и получению совместно эффективных оценок.

Автором предложено формализованную структуру многофакторной статистической модели задавать выражением:

$$\prod_{i=1}^{k} (1 + x_i^{(1)} + x_i^{(2)} + \ldots + x_i^{(s_i - 1)}) \to N_{\Pi}, \qquad (1)$$

где 1 – значение фиктивного фактора $x_0 \equiv 1$;

$x_i^{(1)}$, $x_i^{(2)}$, ..., $x_i^{(s_i - 1)}$ – ортогональные контрасты факторов X_i;

s_i – число различных уровней фактора X_i;

k – общее число факторов, $1 \le i \le k$;

(1), (2), …, $(s_i - 1)$ – порядок контрастов фактора X_i;

N_{Π} – число структурных элементов полного факторного эксперимента, равное числу опытов эксперимента.

Условия получения моделей при полном факторном эксперименте наилучшие из всех возможных: все эффекты ортонормированы, а число опытов равно числу определяемых эффектов. Получаемая модель будет адекватна в точках аппроксимации поверхности отклика. Статистическая эффективность также будет наилучшей: модель соответствует критериям *D*-, *A*-, *E*, *G*-, *Q*-оптимальности. Число обусловленности $\mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = 1$, т.е. наилучшее значение.

Однако использование полных факторных экспериментов не всегда возможно из-за значительного числа проводимых экспериментов. Обычно доступное число факторов 3…4. Необходимо применять дробные факторные эксперименты, статистические свойства которых близки к свойствам полного факторного эксперимента. Такими планами являются многофакторные регулярные планы и планы на основе $\mathrm{ЛП}_{\tau}$ равномерно распределенных последовательностей [6, с. 146–169].

В многофакторных регулярных планах все главные эффекты ортогональны друг к другу. Если число необходимых экспериментов $N_{\mathrm{Д}}$ выбирать с учетом эмпирической формулы:

$$N_{\mathrm{Д}} \approx (\mathbf{1,5...2}) \sum_{i=\mathbf{1}}^{k} (s_i - \mathbf{1}),$$

где s_i – число уровней i-го фактора; $1 \le i \le k$;

k – общее число факторов,

то некоторые из взаимодействий факторов будут ортогональны к главным эффектам или коррелированы с ними сравнительно слабо:

$$| r_{ij}(x_{iu}^{(p)}, (x_{iu}^{(p')} \times x_{ju}^{(q)})) | \le 0{,}3...0{,}4,$$

где r_{ij} – коэффициент парной корреляции между эффектами; $1 \le p < p' \le s_i - 1$; $1 \le q \le s_j - 1$.

При использовании дробных факторных экспериментов в регрессионную модель вводятся эффекты (главные и взаимодействия), которые ортогональны или слабо коррелированы со всеми ранее введенными и статистически значимые. Выбор структурных составляющих статистической модели производится из множества структурных элементов модели полного факторного эксперимента (1). Указанное множество элементов необходимо и достаточно для адекватной аппроксимации результатов экспериментов. При выборе структурных составляющих статистической модели используются разработанный алгоритм RASTA3 [6, с. 179–182] и другие алгоритмы.

Для дробных факторных экспериментов число опытов $N_Д$ всегда меньше числа структурных элементов $N_П$ для полного факторного элемента. Условие ортогональности эффектов или слабой коррелированности их между собой позволяет избежать их смешивания и получить раздельные оценки.

Структура эффектов многофакторной статистической модели необходима и достаточна для получения моделей, линейных по параметрам и, в общем случае, нелинейных по факторам. Исключение некоторых эффектов из модели может привести к неадекватности модели. Включение структурных эффектов, не соответствующих модели полного факторного эксперимента, нарушает ортогональность эффектов и ухудшает критерии получаемой модели.

Получение статистических моделей, их проверки по различным критериям осуществляются с использованием программного средства «Планирование, регрессия и анализ моделей» (ПС ПРИАМ) [6, с. 45–47]. Возможна визуализация полученных результатов в виде различных графиков, диаграмм.

Для оценивания устойчивости многофакторных статистических моделей целесообразно использовать меру обусловленности по Нейману-Голдстейну:

$$P(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \lambda_{\max} / \lambda_{\min}$$

и меру обусловленности матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$

$$\mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \| \mathbf{X}^{\mathrm{T}}\mathbf{X} \| \times \| (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \|,$$

где $\mathbf{X}$ – расширенная матрица эффектов уравнения регрессии, имеющая N строк и k' столбцов, т.е. эффектов;

$\lambda_{(\cdot)}$ – собственные числа информационной матрицы Фишера;

$\|\cdot\|$ – обозначение нормы матрицы.

При ортогональности эффектов и их нормировке P будет следующим:

$$|\mathbf{X}^{\mathrm{T}}\mathbf{X} - \lambda \mathbf{E}| = \begin{vmatrix} N-\lambda & & 0 \\ & \ddots & \\ 0 & & N-\lambda \end{vmatrix} = (N-\lambda)^{k'} = 0,$$

$$\lambda_1 = \lambda_2 = ... = \lambda_{k'} = N,\ \lambda_{\max} = \lambda_{\min} = N,$$

$$P(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = N / N = 1.$$

При определении cond($\mathbf{X}^{\mathrm{T}}\mathbf{X}$) предполагается, что матрица $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ не вырождена. Будем использовать следующую норму:

$$\|\mathbf{A}\|_1 = \max_j \sum_{i=1}^{m} \left| a_{ij} \right|,$$

что означает выбор по всем столбцам j максимальной суммы абсолютных значений элементов по строкам i (m – число строк матрицы **A**).

Для произвольной нормы матрицы ||**A**|| выполняется следующее равенство:

$$\|\alpha\mathbf{A}\| = |\alpha| \, \|\mathbf{A}\|,$$

где **A** – любая матрица и α – любое число.

Если все эффекты в расширенной матрице **X** ортогональны друг к другу и нормированы, то

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{Vmatrix} N & & 0 \\ & \ddots & \\ 0 & & N \end{Vmatrix} = N \begin{Vmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{Vmatrix},$$

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{Vmatrix} 1/N & & 0 \\ & \ddots & \\ 0 & & 1/N \end{Vmatrix} = 1/N \begin{Vmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{Vmatrix}.$$

Расширенная матрица **X** имеет размер $\left(N \times k^{'}\right)$, где N – число строк (опытов), $k^{'}$ – число столбцов (эффектов), введенных в модель.

Для матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ размером $\left(k^{'} \times k^{'}\right)$ каждая по столбцам $1 \le j \le k^{'}$ сумма $\sum_{i=1}^{k^{'}} \left| a_{ij} \right| = N$. Для матрицы $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ размером $\left(k^{'} \times k^{'}\right)$ каждая по столбцам $1 \le j \le k^{'}$ сумма $\sum_{i=1}^{k^{'}} \left| a_{ij} \right| = 1/N$.

Число обусловленности $(\mathbf{X}^{\mathrm{T}}\mathbf{X})_1$ будет

$$\mathrm{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = N \times (1/N) = 1.$$

Из анализа приведенных мер обусловленности следует, что ортогональность или близость к ней и нормировка эффектов является необходимым и достаточным условием наилучшей устойчивости статистических моделей.

Предложенная концепция ортогональности во множественном регрессионном анализе создает информационную систему «план эксперимента – структура модели», компоненты которой согласованы между собой.

Новое развитие концепция ортогональности эффектов при построении регрессионных моделей нашла в условиях проведения планируемого эксперимента в областях, не соответствующих стандартным (куб, сфера, симплекс), при начальной мультиколлинеарности факторов. При этом коэффициент парной корреляции эффектов может достигать значений 0,7…0,9 и более [7], что не позволяет найти традиционными методами устойчивое решение построения модели, модель многократно пересчитывают, добиваясь большей точности, структура модели неустойчива. Некорректно поставленную задачу позволяют устойчиво решать разработанные методы топологического отображения хорошо обусловленного факторного пространства (прообраза), в котором строят план

эксперимента с ортогональными эффектами, в плохо обусловленное факторное пространство (образа), заданное условиями эксперимента, в котором эффекты коррелированы [6, с. 183–300; 8]. Область прообраза факторного пространства строится по данным области образа. Области прообраза и образа являются топологически эквивалентными. Решаемая в прообразе задача является корректно поставленной, полученная в нем модель устойчива и соответствует всем критериям качества для статистических моделей [9]. Хорошие свойства оценок коэффициентов регрессионных статистических моделей в области прообраза и их единственность сохраняются при топологическом отображении и в области образа.

3. Выводы и перспективы дальнейших исследований

1. Разработанная концепция ортогональности во множественном регрессионном анализе позволяет создать наилучшие начальные условия для получения статистических моделей и решать корректно поставленные задачи в условиях, когда структура модели исследователю не известна.
2. Приведен общий вид структуры многофакторной статистической модели, позволяющий получить адекватные, семантические (в информационном смысле), устойчивые модели, структуры которых заранее исследователю не известны.
3. В случаях, когда форма факторного пространства не соответствует стандартной, (куб, сфера, симплекс) для корректного решения задачи необходимо использовать топологический метод устойчивого оценивания и его модификации. В области образа устанавливается ортогональная система собственных кодированных координат.
4. Концепция ортогональности в регрессионном анализе была использована при решении более ста прикладных системных задач по техническим, технологическим, измерительным и другим системам и подтвердила высокую эффективность.

Дальнейшее развитие концепции ортогональности в регрессионном анализе проводится с использованием инвариантно-группового подхода в теории планирования эксперимента и отображения прообраза факторного пространства в образ [9].

СПИСОК ЛИТЕРАТУРЫ

1. Компьютерный анализ и интерпретация эмпирических зависимостей: учебник / [С.В. Поршнев, Е.В. Овечкина, М.В. Мащенко и др.]. – М.: ООО «Бином-Пресс», 2009. – 336 с.
2. Айвазян С.А. Прикладная статистика: Исследования зависимостей: справ. изд. / Айвазян С.А., Енюков И.С., Мешалкин Л.Д.; под. ред. С.А. Айвазяна. – М.: Финансы и статистика, 1985. – 487 с.
3. Айвазян С.А. Программное обеспечение персональных ЭВМ по статистическому анализу данных: проблемы, тенденции, перспективы отечественных разработок / С.А. Айвазян // Компьютеры и экономика: экономические проблемы компьютеризации общества. – М., 1991. – С. 91 – 107.
4. Бродский В.З. Введение в факторное планирование эксперимента / Бродский В.З. – М.: Наука, 1976. – 224 с.
5. Себер Дж. Линейный регрессионный анализ / Себер Дж.; пер. с англ. В.П. Носко; под ред. М.Б. Малютова. – М.: Мир, 1980. – 456 с.
6. Радченко С.Г. Устойчивые методы оценивания статистических моделей / Радченко С.Г. – К.: ПП «Санспарель», 2005. – 504 с.
7. Лапач С.М. Проблеми визначення структури рівняння регресії в множинному регресійному аналізі / С.М. Лапач, С.Г. Радченко // Наукові вісті Нац. техн. ун-ту України “Київ. політехн. ін-т”. – 2007. – № 1 (51). – С. 150 – 155.
8. Радченко С.Г. Планирование эксперимента в нестандартных областях факторного пространства / С.Г. Радченко // Вестник Херсонского нац. техн. ун-та. – 2007. – № 2(28). – С. 281 – 285.
9. Лаборатория экспериментально-статистических методов исследований [Электронный ресурс]. – Режим доступа: http://www.n-t.org/sp/lesmi.

Стаття надійшла до редакції 23.12.2010