С. Г. ФЕДОРЧЕНКО, А. Г. ВАСИЛЕВСКИЙ

Молдова, г. Тирасполь, Приднестровский гос. ун-т им. Т. Г. Шевченко

Дата поступления в редакцию
10.01—14.05 1999 г.
Оппонент *д. т. н. Р. А. ВОРОБЕЛЬ*

МЕТОД ПРОВЕРКИ ЗНАЧИМОСТИ ОЦЕНОК КОЭФФИЦИЕНТОВ МОДЕЛИ

Предложен метод выделения значимых коэффициентов регрессионной модели, основанный на вычислении числа обусловленности матрицы нормальных уравнений.

Для нахождения математической модели исследуемого объекта широко применяются методы планирования эксперимента [1], в частности, полный факторный эксперимент (ПФЭ), разработанный Дж. Боксом в 1956 г. Описание методов планирования эксперимента можно найти в обширном перечне книг и статей. При этом, по нашему мнению, основное внимание исследователи уделяли построению плана эксперимента, удовлетворяющего различным требованиям ($D-A-E-F$-оптимальные планы и т. д.). Методика же обработки экспериментальных данных, полученных в ходе эксперимента, была заимствована из теории регрессионного анализа и оставлена без изменения. Вследствие этого для обработки результатов эксперимента, в частности ПФЭ, используется метод наименьших квадратов (МНК), который, в свою очередь, предъявляет к экспериментальным данным свои требования, а именно:

— выходная величина должна подчиняться нормальному закону;

— дисперсии частных выборок выходной величины во всех точках факторного пространства должны быть одинаковы, и др.

Для проверки выполнения этих требований необходимо набрать некоторую статистику, для чего требуется время и средства, что затрудняет, а в ряде случаев делает невозможным эту работу. В связи с этим возникает вопрос об устойчивости (робастности) методов планирования эксперимента к нарушению предпосылок регрессионного анализа, а также о поиске таких способов обработки экспериментальных данных, которые не требовали бы соблюдения всех или части вышеуказанных условий.

В качестве альтернативы МНК можно предложить использовать оценки коэффициентов модели с учетом поправки на гетероскедастичность (т. е. неравенство дисперсий частных выборок) [2] или так называемый взвешенный МНК [3, с. 121]. В этом случае результаты эксперимента можно обработать даже если дисперсии частных выборок выходной величины различны.

Однако требование нормальности распределения выходной величины сохраняется.

Строго говоря, реализовать план эксперимента можно независимо от того, выполняется это требование или нет. МНК-оценки значений коэффициентов модели также могут быть найдены независимо от выполнения этого требования [3, с. 323]. Однако при этом возникает проблема проверки значимости оценок коэффициентов модели, т. к. использование для этих целей t — критерия Стьюдента становится, строго говоря, неправомерным.

Целью данной работы является поиск метода проверки значимости оценок коэффициентов модели, который не был бы основан на конкретном виде закона распределения оцениваемого параметра.

Основные положения

При обработке результатов ПФЭ нахождение оценок коэффициентов модели сводится к решению системы нормальных уравнений вида

$$\boldsymbol{X}\cdot\overline{b}=\overline{Y}, \qquad (1)$$

где $\boldsymbol{X}$ — матрица значений факторов;
$\overline{b}$ — вектор неизвестных оценок коэффициентов модели;
$\overline{Y}$ — вектор значений выходной величины.

Тогда

$$\boldsymbol{X}^{\text{т}}\cdot\boldsymbol{X}\cdot\overline{b}=\boldsymbol{X}^{\text{т}}\cdot\overline{Y}, \qquad (2)$$

где $\boldsymbol{X}^{\text{т}}$ — транспонированная матрица $\boldsymbol{X}$.

Введем вектор $\overline{f}=\boldsymbol{X}^{\text{т}}\cdot\overline{Y}$. В силу плана ПФЭ матрица $\boldsymbol{X}^{\text{т}}\cdot\boldsymbol{X}$ является диагональной. Поделим каждую строку матрицы $\boldsymbol{X}^{\text{т}}\cdot\boldsymbol{X}$ на соответствующий ей элемент вектора $\overline{f}$ и обозначим матрицу коэффициентов при неизвестных в новом уравнении как $\boldsymbol{X}^{\text{н}}$. В этом случае уравнение (2) примет вид

$$\boldsymbol{X}^{\text{н}}\cdot\overline{b}=\overline{1}, \qquad (3)$$

где $\overline{1}$ — единичный вектор-столбец.

Теперь вся информация о величине оценок коэффициентов модели сосредоточена в элементах матрицы $\boldsymbol{X}^{\text{н}}$, которая также является диагональной.

Назовем полученную систему уравнений, коэффициентами при неизвестных которой являются элементы матрицы $\boldsymbol{X}^{\text{н}}$, нормированной системой нормальных уравнений. Точность решения системы линейных уравнений, имеющих матрицу коэффициентов при неизвестных X, характеризуется числом обусловленности матрицы $\boldsymbol{M}$ [4], которое для диагональной матрицы определяется как отношение максимального элемента к минимальному.

Будем искать зависимость между числом обусловленности матрицы $\boldsymbol{X}^{\text{н}}$, вычисленном по значимым коэффициентам, и параметрами ПФЭ.

Имитационное моделирование на ЭВМ

Исследование проводилось методом имитационного моделирования. С помощью программы-имитатора имитировался объект, математическая модель которого была задана в виде

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x+ +b_{123}x_1x_2x_3 + \varepsilon, \qquad (4)$$

где b — постоянные, заданные оператором коэффициенты;
x — относительные значения факторов, поступающих на вход объекта;
ε — нормально распределенная случайная величина с нулевым средним.

Устанавливая значения факторов в соответствии с планом эксперимента 2^3 и находя для каждой комбинации факторов соответствующие им значения отклика по формуле (4), имитируем проведение ПФЭ.

На первом этапе работы, используя критерий Стьюдента (уровень значимости — 5 %) были отсеяны незначимые оценки коэффициентов модели и по матрице $\boldsymbol{X}^{н}$, из которой были удалены строки и столбцы, соответствующие незначимым факторам, найдено число обусловленности матрицы. Варьируя значения коэффициентов b_{ij} (оценка коэффициента модели для взаимодействий i-го и j-го факторов) около границы значимости, искали соответствующее ей значение числа обусловленности, которое в дальнейшем будем обозначать как $M_{пор}$.

Расчеты проводились для различного числа параллельных опытов m в каждой точке факторного пространства и различных видов математической модели — полной, когда в модели присутствуют как линейные члены, так и взаимодействия факторов; линейной, когда присутствуют только линейные члены модели; нелинейной, когда присутствуют только нелинейные члены модели. Результаты приведены в **табл. 1**.

Таблица 1

Значения $M_{пор}$ для различных видов модели и числа параллельных опытов m

m	Линейная модель	Нелинейная модель	Полная модель
3	15,9	9,1	9,0
4	14,7	11,6	10,2
5	18,6	11,5	8,8
6	16,7	12,2	8,6
7	18,1	9,0	9,8

Как видно из таблицы, величина $M_{пор}$ зависит от вида модели и количества параллельных опытов в каждой точке факторного пространства. Величина $M_{пор}$ для больших значений m не была вычислена, т. к. начиная с некоторого значения m, которое будем обозначать как $m_{кр}$, не выполнялась одна из вышеперечисленных предпосылок регрессионного анализа, а именно — равенство дисперсий частных выборок выходной величины, и использование МНК-оценок коэффициентов модели становилось невозможным [5]. Для линейной модели $m_{кр}$ равно 15, для чисто нелинейной модели 12, для полной модели 8.

Назовем критерий, основанный на использовании числа обусловленности матрицы нормированных нормальных уравнений, M-критерием.

Предлагается следующая методика использования M-критерия для отсеивания незначимых факторов.

Строится матрица $\boldsymbol{X}^{н}$. Вычисляется величина M. Если $M>M_{пор}$, то отбрасывается (признается незначимым) коэффициент, которому соответствует наибольший диагональный элемент матрицы $\boldsymbol{X}^{н}$. Данная процедура продолжается до тех пор, пока вычисленное значение M не будет меньше $M_{пор}$. Первоначально имеет смысл предполагать полный вид модели и использовать значения $M_{пор}$ из соответствующего столбца табл. 1. После того, как будут отсеяны незначимые коэффициенты модели и будет ясен вид модели, можно уточнить значимость ряда коэффициентов, попавших в зону неопределенности, используя данные соответствующего столбца табл. 1.

Производственный пример

Проиллюстрируем наши рассуждения на производственном примере. В [5] приведены результаты ПФЭ. План эксперимента и результаты его реализации приведены в **табл. 2**. Количество параллельных измерений в каждой точке факторного пространства m=3.

Таблица 2

Результаты реализации плана ПФЭ 2^3

x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$	Y_i
+	-	-	-	+	+	+	+	2,45
+	+	-	-	-	-	+	+	1,61
+	-	+	-	-	+	-	+	1,85
+	+	+	-	+	-	-	+	2,05
+	-	-	+	+	-	-	-	2,14
+	+	-	+	-	+	-	-	1,35
+	-	+	+	-	-	+	-	1,55
+	+	+	+	+	+	+	-	1,92

Первые восемь столбцов табл. 2 представляют собой план эксперимента, последний столбец — среднее значение отклика для каждой строки плана. Матрица $\boldsymbol{X}^{т}\boldsymbol{X}$, в силу ортогональности плана ПФЭ, имеет диагональную структуру:

$$\boldsymbol{X}^{т}\boldsymbol{X} = \mathrm{diag}(\, 8, 8, 8, 8, 8, 8, 8, 8 \,),$$

а вектор

$$\bar{f}=\boldsymbol{X}^{т}\bar{Y}=\mathrm{col}(\, 377{,}6,\ 52,\ 33{,}6,\ -44,\ 0,\ 0,\ 0,\ 0 \,).$$

Поделив каждую строку матрицы $\boldsymbol{X}^{т}\boldsymbol{X}$ на соответствующий ей элемент вектора $\bar{f}$, получим нормированную матрицу системы нормальных уравнений:

$$\boldsymbol{X}^{н} = \mathrm{diag}(0{,}0212,\ 0{,}1538,\ 0{,}2381,\ -0{,}1818,\ \infty,\ \infty,\ \infty,\ \infty).$$

Число обусловленности этой матрицы

$$M=\infty / 0{,}0212=\infty.$$

Отбрасывая столбцы и строки матрицы $\boldsymbol{X}^{н}$, содержащие наибольшие члены (в нашем случае их величина бесконечно велика), получим усеченную матрицу $\boldsymbol{X}^{н*}$ следующего вида:

$$\boldsymbol{X}^{н*}=\mathrm{diag}(\, 0{,}0212,\ 0{,}1538,\ 0{,}2381,\ -0{,}1818 \,).$$

Число обусловленности матрицы $\boldsymbol{X}^{н*}$ будет равно M=0,2381/0,0212=11,23. Факторы и взаимодействия, соответствующие удаленным столбцам матрицы $\boldsymbol{X}^{н}$, признаем незначимыми.

Поскольку все оставшиеся элементы матрицы $\boldsymbol{X}^{н*}$ соответствуют линейным членам модели, восполь-

Таблица 3

Значения $M_{пор}$ для различных видов модели ($\chi^2_{табл}$=11,1)

m	Линейная модель			Нелинейная модель			Полная модель		
	χ^2=12,6	χ^2=19,8	χ^2=20,6	χ^2=12,6	χ^2=19,8	χ^2=20,6	χ^2=12,6	χ^2=19,8	χ^2=20,6
3	14,8	15,3	6,8	10,3	11,9	7,3	8,4	8,9	6,1
4	16,0	15,6	6,1	11,1	10,8	6,8	7,9	8,1	6,6
5	15,4	15,5	8,8	12,3	12,9	7,7	8,6	11,5	5,7
6	16,2	15,6	9,3	12,7	12,5	10,5	8,9	9,1	8,5
7	20,3	18,3	8,0	11,8	14,4	8,0	9,6	10,2	8,0
8	21,0	19,0	10,6	11,9	13,2	9,0			
9	26,6	19,7	10,4	15,5	14,5	11,0			

зуемся значением $M_{пор}$ из табл. 1 для линейной модели при m=3. $M_{пор}$=15,9.

Так как $M<M_{пор}$, признаем коэффициенты при x_0, x_1, x_2, x_3 значимыми. Анализ на значимость оценок коэффициентов модели с помощью критерия Стьюдента привел к такому же результату.

Случай выходной величины, не распределенной по нормальному закону

Рассмотрим возможность применения M-критерия для случая распределения выходного параметра, отличного от нормального. Для этого проведем имитационный эксперимент и определим величины $M_{пор}$ для такого выходного параметра. Результаты расчетов приведены в **табл. 3**.

Для оценки степени отличия от нормального распределения используем статистику χ^2. Как известно, для проверки гипотезы о нормальности закона распределения выходной величины используется, в частности, критерий Пирсона [1], который сводится к сравнению величины меры отклонения закона распределения выходной величины от нормального (χ^2) с табличным значением, которое будем обозначать как $\chi^2_{табл}$. Если $\chi^2<\chi^2_{табл}$, то принимается гипотеза о нормальном распределении выходной величины.

Как видно из табл. 3, $M_{пор}$ демонстрирует ярко выраженную зависимость от количества параллельных опытов m и меры отличия от нормальности закона распределения выходной величины χ^2. В табл. 3 приведены результаты расчетов для трех значений меры χ^2 при одном и том же значении $\chi^2_{табл}$. Апробация данной таблицы для отсеивания незначимых факторов показала ее эффективность.

Обобщение полученных результатов

Так как предлагаемая методика использования M-критерия выглядит слишком громоздкой, попытаемся несколько упростить ее.

Рассмотрим уравнение (2) и перепишем его в следующем виде:

$$\boldsymbol{Z}\cdot\overline{b}=\overline{f},$$

где матрица $\boldsymbol{Z}=\boldsymbol{X}\cdot\boldsymbol{X}^n$, а вектор $\overline{f}=\boldsymbol{X}^n\cdot\overline{Y}$.

Матрица $\boldsymbol{Z}$ в случае ПФЭ может быть записана следующим образом:

$$\boldsymbol{Z}=2^n\cdot\boldsymbol{I},$$

где $\boldsymbol{I}$ — единичная матрица, n — количество управляемых факторов.

То есть все элементы матрицы **Z**, отличные от нуля, расположены на главной диагонали и равны 2^n. Следовательно, вся информация о величинах оценок коэффициентов модели сосредоточена в векторе $\overline{f}$ и $\hat{b}_i=f_i/2^n$. Здесь $\hat{b}_i$ — оценка i-го коэффициента модели, f_i — i-й элемент вектора $\overline{f}$. Тогда величина M может быть найдена как отношение $\max\{f_i\}/\min\{f_i\}$.

Введем понятие динамического диапазона метода и определим его как отношение максимальной значимой оценки коэффициента модели к минимальной значимой оценке. Это отношение численно равно M и может считаться оценкой динамического диапазона ПФЭ. Таким образом, предлагаемая методика отсева незначимых коэффициентов модели основана на таблицах значений динамического диапазона ПФЭ и в силу этого не требует нормального закона распределения выходного параметра.

Данный подход является достаточно общим и может быть использован не только в случае МНК-оценок коэффициентов модели.

***

Предлагаемая методика отсева незначимых оценок коэффициентов модели показала свою эффективность как для случая нормального, так и для случаев (в рассмотренном диапазоне) отличия от нормального распределения выходного параметра.

Необходима дальнейшая проработка возможности использования данной методики в случае сильного отличия от нормального распределения выходного параметра, а также в случае его неунимодального распределения.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Хартман К., Шефер В., Лецкий Э. и др. Планирование эксперимента в исследовании технологических процессов.— М. : Мир, 1977.

2. Долгов Ю. А. Модифицированный метод случайного баланса // Электрон. моделирование.— 1987.— № 4.— С. 79—84.

3. Линник Ю. В. Метод наименьших квадратов и основы теории обработки наблюдений.— М. : Физматгиз, 1962.

4. Федорченко С. Г. Проведение эксперимента в условиях значительной нестабильности управляемых факторов // Мат-лы МНТК "Современные проблемы устройств телекоммуникации, компьютерной инженерии и подготовки специалистов".— 23—28 фев. 1998 г., г. Львов.— С. 32—34.

5. Долгов Ю. А., Олейник Т. В., Панасенко Е. П., Цуркан К. В. Связь скорости напыления резисторов с коэффициентом термостабилизации / В сб.: Вопросы электроники. Полупроводниковые приборы и электронная аппаратура.— Кишинев: Штиинца, 1976.— С. 86—89.