СТАТИСТИЧЕСКИЕ МЕТОДЫ И ТЕОРИЯ ВЕРОЯТНОСТЕЙ

ОПТИМИЗАЦИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Баландин В. С.

(Саратов)

В статье предлагается подход к формированию и использованию оптимизационных статистических моделей, т. е. задач математического программирования, построенных из регрессионных функций. Все относящееся к данной проблеме излагается применительно к моделированию экономической эффективности качества проектируемой техники*.

Пусть $X = (x_1, \dots, x_r)$ — вектор независимых (внутренних) показателей качества проектируемой техники, $Y = (y_1, \dots, y_n)$ — вектор зависимых качественных и экономических показателей. Считаем, что y_i — приведенные затраты на производство и применение техники. В общем случае зависимость между выходными и входными показателями будет стохастической

$$y_i = f_i(X, A_i, \varepsilon_i), \quad i=1, \dots, n, \quad (1)$$

где A_i — $1 \times r$ -вектор детерминированных параметров; ε_i — вектор случайных параметров той же размерности.

Содержательный анализ параметров проектируемой техники обычно позволяет сделать выводы о том, какие из них принять за входные, какие — за выходные, влиянием каких следует пренебречь. Но вопрос определения вида зависимости f_i , количественных оценок для параметров модели A_i , для законов распределения случайных величин ε_i остается открытым. В настоящее время его решение возможно только с использованием аппарата математической статистики, который по уже имеющейся информации о значениях входных и выходных показателей позволяет установить существующую зависимость f_i и детерминированные параметры A_i . Кроме того, этот аппарат дает возможность оценить отклонения фактических значений выходных показателей от полученных по (1).

Процесс построения статистических моделей по результатам наблюдений, представленным в табл. 1, начинается с поиска необходимой формы связи между входными и выходными показателями. На этом этапе совокупность исходных данных исследуется на однородность. Если в ходе исследования окажется, что однородность имеет место, то необходимо строить линейную регрессионную модель. Исходная совокупность считается однородной, если многомерное распределение зависимой переменной y_i и независимых показателей $x_1, \dots, x_r$ будет нормальным или близким к нему (из теории вероятностей известно, что если многомерные случайные величины $y_i, x_1, \dots, x_r$ распределены нормально, то теоретическая линия регрессии любой из них относительно всех остальных является линейной функцией). Поэтому анализ совокупности данных на однородность сводится на практике к проверке гипотезы о нормальности распределения зависимой и независимых случайных величин.

* Обзор известных оптимизационных статистических моделей см. в [1].

Предположим, что можно построить линейную зависимость между входными и выходными показателями вида

$$y_i = \alpha_{i0} + \sum_{j=1}^r \alpha_{ij} x_j + \varepsilon_i, \quad i=1, \dots, n, \quad (2)$$

где α_{i0} , α_{ij} — детерминированные параметры линейной регрессионной модели; ε_i — случайный параметр, показывающий отклонение от линейной взаимосвязи зависимой переменной y_i и системы независимых переменных $(x_1, \dots, x_r)$, причем математическое ожидание $\varepsilon_i=0$, дисперсия постоянна, а последовательные значения не зависят друг от друга.

Таблица 1

Данные статистических наблюдений

Номер наблюдения	Значения показателей					
	независимых			зависимых		
	x_1	...	x_r	y_1	...	y_n
1	x_{11}	...	x_{1r}	y_{11}	...	y_{1n}
...	...	...	...	...	...	...
N	x_{N1}	...	x_{Nr}	y_{N1}	...	y_{Nn}

Рассчитав математическое ожидание для обеих частей (2), получим уравнения

$$M[y_i] = \bar{y}_i = \alpha_{i0} + \sum_{j=1}^r \alpha_{ij} x_j, \quad i=1, \dots, n, \quad (3)$$

характеризующие среднее значение выходного показателя y_i для данного комплекса входных переменных $(x_1, \dots, x_r)$. На практике коэффициенты α_{ij} определяются методом наименьших квадратов на основании таблицы. Допустим, что эти коэффициенты найдены. Тогда (3) примет вид

$$\bar{y}_i = a_{i0} + \sum_{j=1}^r a_{ij} x_j, \quad i=1, \dots, n, \quad (4)$$

где a_{i0} , a_{ij} — оценки параметров α_{i0} , α_{ij} , полученные после решения соответствующей системы нормальных уравнений.

Исходя из введенных обозначений можно сформулировать модель оптимизации

$$y_i = a_{i0} + \sum_{j=1}^r a_{ij} x_j \rightarrow \min, \quad (5)$$

$$y_i^{\min} \leq a_{i0} + \sum_{j=1}^r a_{ij} x_j \leq y_i^{\max}, \quad i=2, \dots, n, \quad (6)$$

$$y_i \leq y_i^{\max} \quad (7)$$

$$x_i^{\min} \leq x_i \leq x_i^{\max}, \quad i=1, \dots, r, \quad (8)$$

где

$$y_i^{\min} = \min_{1 \leq j \leq N} y_{ji}; \quad y_i^{\max} = \max_{1 \leq j \leq N} y_{ji}; \quad x_i^{\min} = \min_{1 \leq j \leq N} x_{ji}; \quad x_i^{\max} = \max_{1 \leq j \leq N} x_{ji};$$

N — число наблюдений.

В итоге решения задачи (5)–(8) может оказаться, что оптимальный план: 1) полностью совпадает с одним из результатов наблюдений, представленных в таблице; 2) его компоненты принадлежат различным наблюдениям; 3) его компоненты не содержатся вообще в информации, которой мы располагаем.

Прежде чем дать содержательную интерпретацию каждой из трех возможных ситуаций, заметим, что в экономическом анализе проектных вариантов новой техники статистические модели используются уже давно, они применялись и в оптимизационных расчетах сравнительной экономической эффективности в качестве источника необходимых данных. Решение одной и той же оптимизационной задачи может осуществляться с помощью разных методов получения информации. Если же все исходные показатели определяются в результате применения статистических моделей, то такая задача фактически эквивалентна (5)–(8). Однако вопрос об исследовании результатов решения со статистической точки зрения долгое время не ставился. Проведение статистического анализа оптимального решения следует считать развитием методов оптимизации.

В дальнейшем используем понятия: структура выборочной совокупности и структура оптимизационной модели. Под первой будем понимать количественное соотношение между x_i , между y_i , а также между показателями x_i и y_j . Структуру выборочной совокупности можно охарактеризовать заданием нижней и верхней границы пар x_i/x_j , y_i/y_j , y_i/x_i . Второе понятие определяется структурой входящих в оптимизационную модель регрессионных зависимостей. Количественная определенность структуры модели выражается в коэффициентах регрессий, изменение значений которых означает изменение самой оптимизационной модели.

Если оптимальный план полностью совпадает с одним из результатов наблюдений в таблице, на основе которой и была построена задача (5)–(8), то процесс оптимизации следует считать законченным. А если компоненты оптимального плана принадлежат различным наблюдениям или вообще не содержатся в располагаемой нами информации, то полученный оптимальный план может оказаться статистически недопустимым. В этом случае и структура оптимизационной модели, построенной на выборке фактических значений показателей производимой техники, не соответствует оптимальному плану. Это означает, что закономерности, действующие в выборке, не распространяются на оптимальный план и, следовательно, структура ее не годится для предсказания оптимальных (будущих) вариантов проектируемой техники и требует изменения.

Для проверки принадлежности оптимального плана к исходной выборке в нее добавляют этот план и методами многомерной классификации и распознавания образов производят классификацию новой, увеличенной выборки [1]. Так как выборка однородна и представляет единый класс (мы рассматриваем именно такую ситуацию, поскольку речь идет о линейных регрессиях), то принадлежность оптимального плана к исходной выборке говорит о конце процесса оптимизации статистической модели. Если оптимальный план образует отдельный класс, то строят новую оптимизационную модель, соответствующую уже увеличенной выборке. Затем проверяют новый оптимальный план на принадлежность выборке и т. д. Может оказаться, что на каждом шаге в выборку добавляются элементы,

резко отличающиеся от фактических наблюдений, и тогда число таких элементов будет все увеличиваться и станет намного больше объема исходной выборки. Поэтому от шага к шагу однородность выборки повышается, значения случайных величин выборки все более приближаются к рассчитываемым по регрессионным уравнениям. Отсюда и конечность рассмотренного процесса поиска оптимального статистически допустимого решения.

Отметим, что исследование оптимального решения на статистическую допустимость приведет к его соответствию структуре той совокупности, на основе которой была построена оптимизационная модель. Однако это уже соответствие структуре не исходной, а новой совокупности, причем такое изменение может оказаться настолько сильным, что полученная совокупность будет иметь совсем другую структуру (дело в том, что число шагов оптимизации статистической модели иногда бывает большим). Что же касается исходной совокупности, то оптимальный план, составленный на последнем шаге оптимизации, может так и не соответствовать ей. Практическое использование такого результата оптимизации статистической модели недопустимо с точки зрения закономерностей, действующих в исходной выборке фактических величин входных и выходных показателей проектируемой техники. В подобных случаях значение расширенной выборки в том, что она показывает одно из направлений достижения определенного конечного результата. Это, по сути дела, дополнительная информация, рекомендации проектировщикам новой техники, указывающие некоторый путь ее создания.

На наш взгляд, сохранить структуру оптимального плана (если это необходимо) можно следующим образом: нужно не корректировать исходную совокупность, а изменить оптимизационную модель, сделать ее адекватной структуре исходной совокупности. Для этого к (5)–(8) требуется добавить условия сохранения этой структуры, например, такие структурные ограничения

$$g_{ij}^x \left(\frac{\bar{x}_i}{\bar{x}_j} \right) \leq \frac{x_i}{x_j} \leq f_{ij}^x \left(\frac{\bar{x}_i}{\bar{x}_j} \right), \quad i, j=1, \dots, r, \quad i \neq j, \quad (9)$$

$$g_{ij}^y \left(\frac{\bar{y}_i}{\bar{y}_j} \right) \leq \frac{y_i}{y_j} \leq f_{ij}^y \left(\frac{\bar{y}_i}{\bar{y}_j} \right), \quad i, j=1, \dots, n, \quad i \neq j, \quad (10)$$

$$g_{ij}^{xy} \left(\frac{\bar{y}_i}{\bar{x}_j} \right) \leq \frac{y_i}{x_j} \leq f_{ij}^{xy} \left(\frac{\bar{y}_i}{\bar{x}_j} \right), \quad i=1, \dots, n, \quad j=1, \dots, r, \quad (11)$$

где

$$\bar{x}_i = \frac{1}{N} \sum_{j=1}^N x_{ji}, \quad \bar{y}_i = \frac{1}{N} \sum_{j=1}^N y_{ji}.$$

Функции g_{ij} и f_{ij} отражают изменение структуры показателей (как входных, так и выходных) при проектировании новой техники. Так, если

$$f_{ij}^x \left(\frac{\bar{x}_i}{\bar{x}_j} \right) = \frac{\bar{x}_i}{\bar{x}_j}, \quad (12)$$

то соотношение между указанными показателями должно оставаться постоянным. В общем случае структура их меняется, и это должно быть обязательно учтено при оптимизации соответствующих статистических моделей. Различные пары показателей будут иметь свои функции g_{ij} и f_{ij} . Они могут быть как линейными (однородными или неоднородными), так и нелинейными; общих рекомендаций по их построению дать невозможно.

Однако при решении конкретных инженерных задач эти функции можно получить с учетом опыта проектировщиков новой техники или требований потребителей. Результаты проектирования зубообрабатывающих станков в Саратовском станкообъединении говорят о том, что в ряде случаев целесообразно использовать структурные ограничения

$$(1-\alpha_{ij}^x) \frac{\bar{x}_i}{\bar{x}_j} \leq \frac{x_i}{x_j} \leq (1+\alpha_{ij}^x) \frac{\bar{x}_i}{\bar{x}_j}, \quad (13)$$

$$(1-\alpha_{ij}^y) \frac{\bar{y}_i}{\bar{y}_j} \leq \frac{y_i}{y_j} \leq (1+\alpha_{ij}^y) \frac{\bar{y}_i}{\bar{y}_j}, \quad (14)$$

$$(1-\alpha_{ij}^{xy}) \frac{\bar{y}_i}{\bar{x}_j} \leq \frac{y_i}{x_j} \leq (1+\alpha_{ij}^{xy}) \frac{\bar{y}_i}{\bar{x}_j}, \quad (15)$$

где α_{ij} — коэффициенты соотношения между i -м и j -м показателями, $0 < \alpha_{ij} < 1$.

Если предположить, что изменение структуры показателей проектируемой техники должно подчиняться неравенствам (13)–(15), то в модель (5)–(8) нужно их добавить. В общем случае в (5)–(8) добавляются условия (9)–(11). Задача (5)–(11) позволит получить статистически допустимое оптимальное решение за один шаг. Это серьезное преимущество предлагаемого способа уточнения статистической оптимизационной модели перед изложенной выше процедурой изменения структуры исходной совокупности. Кроме того, условия (9)–(11) обеспечат действительно статистически допустимое решение, а структура оптимального решения будет соответствовать фактически сложившейся структуре показателей техники аналогичного назначения.

Однако линейная статистическая модель не всегда адекватна описанию зависимости входных и выходных показателей проектируемой техники и поэтому годится лишь для однородной выборки значений случайных переменных. Если исследование выборки показало, что она неоднородна (распределение случайных величин не является нормальным или близким к нему), то линейные регрессии не будут отражать реальную связь входных и выходных переменных, и для связи их может быть использована нелинейная модель. Наиболее распространена на практике модель вида

$$y_i = a_{i0} \prod_{j=1}^r x_j^{a_{ij}}, \quad i=1, \dots, n. \quad (16)$$

На базе зависимостей типа (16) аналогично можно построить оптимизационную модель. После перехода в ней к логарифмическим переменным получим задачу линейного программирования, решение которой эквивалентно решению исходной задачи. Применимы и другие нелинейные модели, но в таких случаях затруднительно как их построение, так и решение соответствующих оптимизационных моделей.

Ограничения вида (13)–(15) путем задания различных значений параметра α_{ij} позволяют проводить исследование при выборе оптимальных направлений создания новой техники. Тем не менее опыт решения задач с такими ограничениями свидетельствует о том, что их введение иногда может привести к несовместности системы ограничений. В таких ситуациях можно использовать тот же подход, что и при записи ограничений вида (8)

$$\min_k \left(\frac{x_{ki}}{x_{kj}} \right) \leq \frac{x_i}{x_j} \leq \max_k \left(\frac{x_{ki}}{x_{kj}} \right), \quad k=1, \dots, N, \quad i, j=1, \dots, r, \quad i \neq j, \quad (17)$$

$$\min_k \left(\frac{y_{ki}}{y_{kj}} \right) \leq \frac{y_i}{y_j} \leq \max_k \left(\frac{y_{ki}}{y_{kj}} \right), \quad k=1, \dots, N, \quad i, j=1, \dots, n, \quad i \neq j, \quad (18)$$

где x_{ki} , x_{kj} , y_{ki} , y_{kj} берутся из таблицы.

Рассмотрим пример построения и применения статистической оптимизационной модели со структурными ограничениями.

Таблица 2

Значение экономических и качественных показателей зубострогальных станков разных моделей

Модель станка	Экономические (зависимые) показатели станка			Качественные (независимые) показатели станка		
	y_1	y_2	y_3	x_1	x_2	x_3
5A250П	31,3	6,8	2	8,7	4,5	6437
5C276П	51,2	13,3	2,1	9	11,5	8579
5C286П	54,9	13,8	28,7	15	13,6	8161
5A26	10,4	5,3	3	6,5	4	5127
5283	26,7	15,5	5,4	18	10,1	5744
5A250	21,2	6,8	3,8	8,7	4,5	6579
5282	24,7	6,1	13,4	12,5	9,3	6985
5A283	26,6	15,5	5,6	19,2	13,6	6854
5A283С1	80,2	26,7	8,1	19,2	13,6	6927
52ТМ-2	36,5	45	6,6	43	27,3	6615
5Е283	80	26,7	8,5	19	12	6320
5284	21,2	28,1	5,1	25	7,1	5722

На основе данных по важнейшим показателям зубострогальных станков разных типов (эти данные были оформлены в виде приведенной таблицы) имеем следующие регрессионные зависимости

$$y_1 = -15,7 - 0,496x_1 + 1,66x_2 + 0,00671x_3, \quad (19)$$

$$y_2 = -10,4 + 1,30x_1 - 0,292x_2 + 0,00136x_3, \quad (20)$$

$$y_3 = -15,5 + 0,0305x_1 + 0,0908x_2 + 0,00325x_3, \quad (21)$$

где y_1 — цена станка, тыс. руб.; y_2 — трудоемкость производства станка, тыс. нормо-часов; y_3 — себестоимость обработки годового объема деталей на станке, тыс. руб.; x_1 — вес станка, т; x_2 — мощность электродвигателей, установленных на станке, квт; x_3 — общее количество деталей в станке, шт. Коэффициенты множественной корреляции для каждой из зависимостей соответственно равны: 0,945; 0,798; 0,785, коэффициенты детерминации: 0,893; 0,637; 0,616.

Полученные регрессионные зависимости (19)–(21) были использованы при построении статистической оптимизационной модели

$$y = y_3 + E_n y_1 \rightarrow \min, \quad (22)$$

$$y_i^{\min} \leq y_i \leq y_i^{\max}, \quad i=1, 2, 3, \quad (23)$$

$$x_i^{\min} \leq x_i \leq x_i^{\max}, \quad i=1, 2, 3. \quad (24)$$

Здесь E_n — нормативный коэффициент эффективности капитальных вложений, равный 0,15. Минимальные и максимальные значения x_i и y_i , т. е. показатели $x_i^{\min}$, $x_i^{\max}$, $y_i^{\min}$, $y_i^{\max}$ находятся из соответствующих столбцов табл. 2. Очевидно, что целевая функция (22) содержательно показывает приведенные затраты на годовое производство деталей на

одном станке. С математической точки зрения, модель (22)–(24) – задача линейного программирования. Ее решение, полученное по известной программе «Мелена» на ЭВМ «Минск-32», имеет вид: $x_1=28,4$; $x_2=4$; $x_3=5127$; $y=4,1$. Значения y_1, y_2, y_3 , соответствующие этому решению, оказались равными соответственно 10,9; 31,9; 2,4.

Расчеты показали, что полученное оптимальное решение не укладывается ни в рамки ограничений (13)–(14) при любом α , ни (17)–(18). Лишь соотношение x_2 и x_3 соответствует исходной выборке. Следовательно, полученное решение является статистически не допустимым. Это сделало необходимым введение в модель (22)–(24) структурных ограничений. В данном примере использовались ограничения типа (17)–(18).

После включения в модель (22)–(24) условий (17)–(18) это решение изменилось: $x_1=14,2$; $x_2=4,02$; $x_3=5140$; $y_1=18,5$; $y_2=13,6$; $y_3=2$; $y=4,8$. Отсюда видно, что значение целевой функции несколько ухудшилось в результате введения структурных ограничений (17) и (18).

Полученное решение показывает оптимальное направление создания новых зубострогальных станков. Оно применяется как некоторый ориентир проектировщикам таких станков.

Независимо от вида регрессионных зависимостей оптимизационная модель должна обязательно включать в себя структурные ограничения. Это позволит получать статистически допустимое оптимальное решение. На наш взгляд, такого типа ограничения должны найти место также в оптимизационных моделях текущего и перспективного, отраслевого и внутривзаводского планирования. И не только в моделях, построенных на базе регрессионных зависимостей, а в любых других типах указанных моделей. Если, например, решается задача оптимизации годового плана производства промышленного предприятия, то в ограничения соответствующей модели должны включаться и условия вида

$$a_{ij} \leq \frac{x_i}{x_j} \leq b_{ij}, \quad (25)$$

где x_i, x_j – объемы производства соответственно виду продуктов i и j . Установление нижней и верхней границ отношения a_{ij} и b_{ij} должно быть в компетенции работников плановых органов либо предприятия, либо вышестоящих органов. Их следует устанавливать с учетом сложившейся ассортиментной структуры производства на данном предприятии.

ЛИТЕРАТУРА

1. Кулешов В. В., Лукацкая М. Л., Ягольницер М. А. Проблемы статистического моделирования и оптимизация отраслевых планов. Новосибирск.: Наука, 1977.

Поступила в редакцию
16 III 1978