

ЗАМЕТКИ И ПИСЬМА

МЕТОД ИССЛЕДОВАНИЯ КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ НА УСТОЙЧИВОСТЬ ПРИ ВОЗМУЩЕНИИ ИСХОДНОЙ ИНФОРМАЦИИ

В. А. КОРОВОВ

(Москва)

Постановка задачи. В процессе экономико-статистических исследований часто приходится отыскивать формы связи функционального признака с причинными факторами. Наиболее распространенная на практике форма связи — множественная линейная регрессия. В силу относительной несложности и, главное, отработанности математических приемов вычисления и анализа линейной регрессии нелинейные формы связи тем или иным способом (логарифмированием, заменой переменных, тейлоровским разложением и т. д.) обычно приводят к линейному виду. Здесь мы не касаемся вопроса о правомерности выбора той или иной формы связи.

Итак, имеется исходная информация $V_0 = \{v_{ij}^{(0)}\}$, где $i = 0$ — индекс функционального признака; $i = 1, \dots, m$ — индекс причинного фактора; $j = 1, \dots, n$ — индекс наблюдений, m — количество причинных факторов, n — число наблюдений.

По этой информации V_0 на базе метода наименьших квадратов были найдены коэффициенты $b_k^{(0)}$, $k = 0, \dots, m$, множественной линейной регрессии

$$y = b_0^{(0)} + \sum_{k=1}^m b_k^{(0)} X_k, \quad (1)$$

где $b_0^{(0)}$ — свободный член; $b_k^{(0)}$ — регрессионный коэффициент k -го причинного фактора; X_k — величина k -го причинного фактора.

Пусть эти коэффициенты при решении системы нормальных уравнений вычислялись посредством правила Крамера:

$$b_k^{(0)}(V_0) = \frac{A_k(V_0)}{A(V_0)}, \quad k = 0, \dots, m,$$

где $b_k^{(0)}(V_0)$ — выражение k -го коэффициента регрессии как функции от исходной информации $V_0 = \{v_{ij}^{(0)}\}$; $A(V_0)$ — определитель системы нормальных уравнений как функция от V_0 ; $A_k(V_0)$ — определитель как функция от V_0 , численно вычисленный по матрице левой части системы, у которой k -й столбец заменен на столбец свободных членов системы нормальных уравнений. Индекс k здесь характеризует k -й причинный фактор.

Если теперь в информации V_0 произошли какие-либо изменения, то b_k — коэффициенты линейной регрессии — придется вычислять заново. Часто случается, однако, что затраты машинного времени и человеческого труда по повторному вычислению коэффициентов b_k оказываются в целом неэффективными, так как новые $b_k^{(1)}$ практически мало отличаются от первоначальных $b_k^{(0)}$, находясь в допустимых границах своего изменения. Отсюда возникает задача исследования на устойчивость коэффициентов линейной регрессии при «возмущении» информации $V_0 = \{v_{ij}^{(0)}\}$.

Здесь рассматривается самый простейший случай, когда имеется лишь один возмущенный элемент. Для краткости «возмущенным» элементом будем называть тот элемент, величина которого изменилась по сравнению со своим первоначальным значением в исходной информации $V_0 = \{v_{ij}^{(0)}\}$. Задача формулируется следующим образом: в каких пределах может изменяться элемент $v_{ij}^{(0)}$ (прочие элементы информации V_0 остаются неизменными), чтобы найденные регрессионные коэффициенты не выходили из своих априорно заданных интервалов: $\underline{b}_k \leq b_k^{(0)} \leq \bar{b}_k$, $k = 0, \dots, m$.

Задача заключается, следовательно, в вычислении верхней $\tilde{x}_{ij}$ и нижней x_{ij} границ возмущенного элемента x_{ij} . Теперь возмущенный элемент $v_{ij}^{(0)}$ является неизвестной величиной x_{ij} .

Данная постановка отличается от задачи, решаемой в [1, стр. 274].

В нашем случае имеется непосредственная связь между информацией — элементами системы нормальных уравнений — коэффициентами линейной регрессии. Можно найти некоторую аналогию между этими задачами, хотя речь и идет о разных вещах. В [1] исследуется вторая стадия — устойчивость решения при возмущении элементов матрицы системы. Рассмотрев еще и первую стадию — зависимость элементов системы от исходной информации, можно было бы получить обобщенную задачу об устойчивости решения при варьировании исходной информации. Такая задача представляет наибольший теоретический интерес.

Тем не менее интерес представляет и определение границ допустимого возмущения информации, при которых решение остается в своих заданных границах, т. е. будет устойчивым. Конечно, решение системы нормальных уравнений является всегда единственным и вполне определенным независимо от возмущения исходной информации (подразумевается, что информация состоит из действительных чисел). Никаких «разрывов», комплексных областей и прочих тонкостей! Поэтому под «устойчивостью» здесь понимается не «выход» решения из заданных границ, а колебание его вокруг определенного уровня (первоначально полученного решения).

Данная постановка задачи отличается и от чисто классического исследования устойчивости (например, по Ляпунову) тем, что она идет от зафиксированных границ изменения функции к искомым границам возмущения аргумента. Смена «ролей» между аргументом и функцией с целью сведения задач к классической является в данном случае весьма трудоемкой и не столь очевидной, как кажется на первый взгляд. Возмущение информации влияет на коэффициенты регрессии не «прямо», а через систему нормальных уравнений. Пришлось бы вначале решать очень сложную и самостоятельную задачу по представлению возмущенного элемента в информации как функции от коэффициентов регрессии и затем уже перейти к классическому рассмотрению проблемы устойчивости. Наша постановка задачи и особенно метод ее решения являются упрощенными и предназначены для практики.

Метод определения границ «возмущения» элементов исходной информации. Необходимыми характеристиками устойчивости являются величины

$$b_{kij} = \frac{\partial b_k(V_0)}{\partial x_{ij}} = \left(\frac{\partial A_k(V_0)}{\partial x_{ij}} A_{m+1}(V_0) - \frac{\partial A_{m+1}(V_0)}{\partial x_{ij}} A_k(V_0) \right) / A_{m+1}^2(V_0), \quad (2)$$

так как эти частные производные показывают меру изменения соответствующего коэффициента при изменении на единицу элемента x_{ij} информации V_0 и неизменных величинах остальных ее элементов.

Дифференцирование определителей по элементу x_{ij} информации V_0 легче всего производить по правилу Лягуилля [2]. Например,

$$\frac{\partial A_k(V_0)}{\partial x_{ij}} = \sum_{w=0}^m A_k^{(w)}(V_0), \quad (3)$$

где $A_k^{(w)}(V_0)$ — определитель, матрица которого отличается от матрицы определителя $A_k(V_0)$ лишь w -й строкой, элементы которой представляют собой продифференцированные по x_{ij} соответствующие элементы $A_k(V_0)$.

По формуле (3) вычисляются необходимые $\partial A_k(V_0) / \partial x_{ij}$ и $\partial A(V_0) / \partial x_{ij}$.

Новые значения $b_k^{(1)}$ каждого регрессионного коэффициента можно рассматривать как сумму его первоначального значения, найденного по первичной информации, и его приращения, связанного с изменением элемента x_{ij} , т. е.

$$b_k^{(1)} = b_k^{(0)} + \Delta b_k \approx b_k^{(0)} + b_{kij}^{(0)} \Delta x_{ij}, \quad (4)$$

где $b_k^{(0)}$ — первоначально найденные регрессионные коэффициенты, $k = 0, \dots, m$; Δb_k , Δx_{ij} — приращение соответствующих величин; $b_{kij}^{(0)}$ — частные производные соответствующих коэффициентов, численно вычисляемые, как и $b_k^{(0)}$, $k = 0, \dots, m$, по первоначальной информации $V_0 = \{v_{ij}^{(0)}\}$.

Как видно, в самом приращении регрессионных коэффициентов Δb_k , связанном с возмущением элемента x_{ij} , взята основная его часть, т. е. дифференциал. Учет последующих членов ряда Тейлора не представляет принципиальной трудности для метода, но значительно усложняет процесс вычисления. Как показали расчеты на ЭВМ,

в данном случае дифференциал практически полностью совпадает с приращением функции.

Если возмущение элемента x_{ij} произвольно, то понятно, что новые величины $b_k^{(1)}$ могут выйти из своих заданных верхних $\bar{b}_k$ или нижних b_k границ. Цель заключается в нахождении таких возможных $\bar{x}_{ij}$, x_{ij} , чтобы ни один из коэффициентов не вышел из своих допустимых границ. Возьмем один какой-либо коэффициент, например b_1 , и найдем для него такое максимальное приращение элемента x_{ij} (обозначим его через $\Delta x_{ij}^{(b)}$), которое удовлетворяло бы только одному его b_1 . Очевидно,

$$\bar{b}_1 = b_1^{(0)} + b_{1ij}^{(0)} \Delta x_{ij}^{(b_1)},$$

откуда

$$\Delta x_{ij}^{(b_1)} = (\bar{b}_1 - b_1^{(0)}) / b_{1ij}^{(0)}. \quad (5)$$

Соответственно, исходя из $\underline{b}_1 = b_1^{(0)} + b_{1ij}^{(0)} \Delta x_{ij}^{(b_1)}$, находим

$$\Delta x_{ij}^{(b_1)} = (\underline{b}_1 - b_1^{(0)}) / b_{1ij}^{(0)}. \quad (6)$$

Аналогично вычисляется для каждого коэффициента максимально возможное приращение «возмущенного» элемента. Здесь за основу берется предположение о приближенной пропорциональности частных производных регрессионных коэффициентов приращению «возмущенного» элемента на всем заданном интервале допустимого колебания коэффициентов. Априорно заданные границы допустимого колебания коэффициентов b_k должны быть, таким образом, достаточно «тесными», чтобы требуемая пропорциональность соблюдалась на всем интервале изменения «возмущенного» элемента. По формулам (5) и (6) получаем два набора, подобных $\Delta x_{ij}^{(b_1)}$ и $\Delta x_{ij}^{(b_1)}$ величин, т. е.

$$\bar{M}_{ij} = \{\Delta x_{ij}^{(b_k)}\} \text{ и } \underline{M}_{ij} = \{\Delta x_{ij}^{(b_k)}\}, k = 0, \dots, m.$$

Каждый набор может иметь как положительные, так и отрицательные значения своих элементов, что связано со знаком частных производных $b_{kij}^{(0)}$, т. е. верхний (нижний) предел какого-либо коэффициента может быть достигнут не увеличением (уменьшением) возмущенного элемента, а, наоборот, его уменьшением (увеличением), если частная производная коэффициента отрицательная. Поэтому из всех элементов обоих наборов $\bar{M}_{ij}$ и $\underline{M}_{ij}$ выберем отдельно для положительных и отрицательных элементов (это разделение существенно) минимальное по абсолютной величине приращение элемента Δx_{ij} . Получим набор N , состоящий из двух элементов, т. е.

$$N = \{\tilde{\Delta x}_{ij}, -\Delta x_{ij}\}, \quad (7)$$

где

$$\begin{aligned} \tilde{\Delta x}_{ij} &= \min_{(\Delta x_{ij}^{(b_k)} > 0)} \{\bar{M}_{ij}, \underline{M}_{ij}\}; \\ -\Delta x_{ij} &= \min_{(|\Delta x_{ij}^{(b_k)}| < 0)} \{\bar{M}_{ij}, \underline{M}_{ij}\}, \end{aligned}$$

подразумевая под стоящими в фигурных скобках $\bar{M}_{ij}$ и $\underline{M}_{ij}$ элементы этих наборов, обозначенных кратко через $|\Delta x_{ij}^{(b_k)}| < 0$, означает абсолютную величину отрицательных $\Delta x_{ij}^{(b_k)}$.

Набор N является искомым. Действительно, любое приращение элемента Δx_{ij} исходного x_{ij} вызывает увеличение одних и уменьшение других коэффициентов. Это связано со знаками $b_{kij}^{(0)}$. Приращение Δx_{ij} представляет собой минимальное увеличение элемента x_{ij} из всех его максимальных положительных верхних и нижних приращений для каждого параметра из набора $\bar{M}_{ij}$ и $\underline{M}_{ij}$.

Очевидно, все регрессионные коэффициенты будут находиться в допустимых верхних и нижних пределах. Аналогично $-\Delta x_{ij}$ является минимальным уменьшением элемента x_{ij} , также гарантирующее всем коэффициентам остаться в заданных пределах. С помощью элементов набора N определяем искомые верхние $\bar{x}_{ij}$ и нижние $\underline{x}_{ij}$

границы изменения элемента исходной информации, т. е.

$$\bar{x}_{ij} = v_{ij}^{(0)} + \tilde{\Delta} x_{ij}; \quad x_{ij} = v_{ij}^{(0)} - \underline{\Delta} x_{ij},$$

где $v_{ij}^{(0)}$ — первоначальная величина b информации $V_0 = \{v_{ij}^{(0)}\}$ до своего «возмущения».

Устойчивость коэффициентов регрессии при возмущении исходной информации. Вышеизложенным методом определяются для *всех* элементов информации $V_0 = \{v_{ij}^{(0)}\}$ искомые верхние $\bar{x}_{ij}$ и нижние x_{ij} границы возмущения каждого элемента в отдельности.

Получаем еще два «информационных» множества:

$$\bar{V} = \{\bar{v}_{ij}\} \text{ и } \underline{V} = \{v_{ij}\},$$

где $\bar{v}_{ij}, v_{ij}$ — найденные искомые $\bar{x}_{ij}, x_{ij}$.

Одновременно с процессом определения искомых $\bar{x}_{ij}$ и x_{ij} надо было вычислить частные производные $b_{kij}^{(0)}$. Оформляем их в таблицы. Получаем $(m+1)$ таблиц (для каждого k -го коэффициента регрессии, $k = 0, \dots, m$, состоящих из $(m+1)n$ величин частных производных, $(m+1)n$ — размерность информации $V_0 = \{v_{ij}^{(0)}\}$.

Если теперь в информации $V_0 = \{v_{ij}^{(0)}\}$ изменился какой-либо элемент $v_{ij}^{(0)}$ на $v_{ij}^{(1)}$, то если $v_{ij} \leq v_{ij}^{(1)} \leq \bar{v}_{ij}$, тогда для всех коэффициентов регрессии: $\underline{b}_k \leq b_k^{(1)} \leq \bar{b}_k$. Иначе говоря, решение системы нормальных уравнений остается устойчивым. В противном случае, когда $v_{ij}^{(1)}$ не лежит в интервале $[\bar{v}_{ij}, v_{ij}]$, то какой-то коэффициент регрессии выходит из своего интервала $[\bar{b}_k, \underline{b}_k]$, т. е. решение уже неустойчиво. Не решая задачу заново, можно получить необходимые $b_k^{(1)}$, исходя из корректировки

$$b_k^{(1)} = b_k^{(0)} + \Delta b_k \approx b_k^{(0)} + \sum_{i,j} b_{kij}^{(0)} \Delta v_{ij},$$

где $\Delta v_{ij} = v_{ij}^{(1)} - v_{ij}^{(0)}$ — приращение (мера «возмущения») возмущенного элемента.

Если в информации V_0 возмущилось несколько элементов, например $v_{01}^{(0)}, v_{11}^{(0)}, v_{23}^{(0)}$.

то новое решение $\pi_1 = \{b_k^{(1)}\}$ можно получить корректировкой

$$b_k^{(1)} = b_k^{(0)} + \Delta B_k \approx b_k^{(0)} + D_k \approx b_k^{(0)} + b_{k01}^{(0)} \Delta v_{01} + b_{k11}^{(0)} \Delta v_{11} + b_{k23}^{(0)} \Delta v_{23},$$

где ΔB_k — полное приращение для k -го коэффициента регрессии; D_k — полный дифференциал для каждого коэффициента регрессии; $b_{kij}^{(0)}$ — частные производные для k -го коэффициента регрессии по соответствующему возмущенному элементу, вычисленное по информации $V_0 = \{v_{ij}^{(0)}\}$.

Это приближительное соотношение базируется на учете только главной, линейной части полного приращения ΔB_k . В свою очередь полный дифференциал D_k представляется как сумма частных дифференциалов.

Сравнивая далее решение $\pi_1 = \{b_k^{(1)}\}$, полученное таким образом, с заданными интервалами колебаний для каждого k -го коэффициента регрессии, т. е. с $[\bar{b}_k, \underline{b}_k]$, можно судить о том, является ли $\pi_1 = \{b_k^{(1)}\}$ устойчивым или нет при данных возмущениях в информации $V_0 = \{v_{ij}^{(0)}\}$.

В принципе можно было бы уменьшить разницу между ΔB_k и D_k учетом и нелинейной части приращения. Однако это влечет за собой резкое усложнение вычислительного процесса. И вряд ли целесообразно добиваться точного учета величины ΔB_k , если вспомнить, что само по себе уравнение регрессии характеризует корреляционную, а не функциональную зависимость.

На первый взгляд кажется, что вычисление необходимых таблиц слишком громоздко и легче перебить перфокарты и решить задачу на ЭВМ заново, чем воспользоваться данным методом. В единичных случаях — да. Но чем больше число таких единичных случаев, тем преимущество повторного решения быстрее сокращается. Как показывает практика, данный метод по затратам машинного времени на ЭВМ (экономию человеческого труда в пользу предлагаемого метода даже без учета работников,

обслуживающих ЭВМ, не вызывает сомнения) равносильно 5—6 единичным решениям (при одних и тех же размерах задачи). Дело в том, что получение $\pi_1 = \{b_k^{(1)}\}$ прямым путем связано с обработкой исходной информации $V_0 = \{v_{ij}^{(0)}\}$, ее транспонированием, формированием системы нормальных уравнений, вычислением определителей, обращением матрицы, решением системы нормальных уравнений. Это занимает больше машинного времени, чем вычисление одной частной производной b_{kij} — самого трудоемкого момента в данном методе. Но для получения полной картины приходится вычислять все $(m+1)^2 n$ частных производных. Отсюда значительный проигрыш нашего метода в затратах машинного времени по сравнению с обыкновенным решением «в лоб». Но зато он позволяет, получив один раз искомые таблицы с частными производными $\bar{V} = \{\bar{v}_{ij}\}$ и $V = \{v_{ij}\}$, больше уже не выходить на ЭВМ и получать $\pi_1 = \{b_k^{(1)}\}$ при изменении в информации V_0 посредством корректировки ранее найденного плана $\pi_0 = \{b_k^{(0)}\}$.

Выводы. Решение задач об устойчивости коэффициентов множественной линейной регрессии позволяет:

- 1) определить, насколько допустимы изменения каждого информационного элемента в отдельности при условии сохранения каждым регрессионным коэффициентом своих пределов заданной вариации;
- 2) предъявлять особенно строгие требования к точности как измерения, так и определения той части информации, которая имеет наиболее узкие границы «возмущения»;
- 3) корректировать значения регрессионных коэффициентов при «возмущении» исходной информации;
- 4) исследовать на устойчивость коэффициенты не только множественной линейной регрессии, но и любой криволинейной функции, которая при обращении к методу наименьших квадратов тем или иным способом (логарифмированием, например) приводится к линейному виду; в том числе определить устойчивость производственной функции, параметры которой определяются на базе метода наименьших квадратов;
- 5) получить определенное решение единожды и корректировать все последующие решения (при варьировании информации) на основе этого первоначального решения и таблиц с частными производными и искомыми верхними и нижними границами допустимого возмущения элементов информации, что даст гораздо большую экономию дорогостоящего машинного времени, чем традиционные повторные решения.

ЛИТЕРАТУРА

1. Д. К. Фаддеев и В. Н. Фаддеева. Вычислительные методы линейной алгебры. М.—Л., Физматгиз, 1963.
2. Л. С. Понтрягин. Обыкновенные дифференциальные уравнения. М., Физматгиз, 1964.

Поступила в редакцию
7 IV 1969

К ПРОГНОЗУ МАКРОСПРОСА НА ТОВАРЫ НАРОДНОГО ПОТРЕБЛЕНИЯ

У. Г. ЧЕРНЯВСКИЙ

(Москва)

При построении моделей макроспроса (по укрупненным группам товаров) в государственной и кооперативной торговле для корреляционного анализа используют выборочные обследования бюджетов рабочих и колхозников либо данные о товарной структуре розничного оборота*. В обоих случаях возникают серьезные затруднения — недостаток данных. В самом деле, бюджетные обследования разрабатываются лишь по 10-12 группам в зависимости от размера дохода на члена семьи. Можно провести исследование по динамическому ряду о розничном товарообороте и других показателях по СССР в целом или по отдельной зоне, допустим за период с 1950 по 1969 г., увеличив, таким образом, количество данных до двадцати. Однако использование информации за годы, далеко отстоящие от прогнозируемого, снижает ее значение

* Розничный товарооборот этой торговли учитывался по 78 группам товаров. С 1967 г. номенклатура учета стала еще шире — 91 группа.