ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1980

Seria: GÓRNICTWO z. 106 Nr kol. 646

Czesław POTOCKI
Jerzy CHOWANIEC

ALGORYTM POSTĘPOWANIA W ZASTOSOWANIU REGRESJI KROKOWEJ PRZY ROZWIĄZYWANIU ZAGADNIEŃ INŻYNIERSKO-ORGANIZACYJNYCH

Streszczenie. W artykule podano szczegółową metodykę postępowania przy zastosowaniu regresji krokowej. Metoda ta jest przydatna tam, gdzie należy rozpatrywać dużą liczbę zmiennych w różnym stopniu ze sobą skorelowanych i gdzie należy ograniczyć się do zmiennych najistotniejszych.

1. SFORMUŁOWANIE PROBLEMU

Załóżmy, że dokonujemy obserwacji pewnej zmiennej Y, która jest zależna od k zmiennych objaśniających X i w wyniku tej obserwacji otrzymujemy -n-elementową próbę. Modele matematyczne, opisujące zależność pomiędzy zmienną objaśniającą Y, a zmiennymi objaśniającymi $X_i (i=1,2,.....k)$, można podzielić na dwie grupy:

- funkcje addytywne,
- funkcje multyplikatywne.

Pierwszą grupę stanowią funkcje zakładające, że zmienna objaśniana jest sumą wpływów działających na nią czynników.

$$Y = \sum_{i=1}^{k} \alpha_i X_i + \alpha_o + \xi, \tag{1}$$

gdzie α_o, α_i - parametry strukturalne modelu.

Drugą grupę modeli stanowią funkcje zakładające, że zmienna objaśniana jest iloczynem wpływów działających na nią zmiennych objaśniających. Spośród nich najczęściej stosowana jest funkcja Cobb-Douglasa o wzorze analitycznym:

$$Y = \alpha_o \prod_{i=1}^{k} \alpha_i X_i e^{\xi}. \tag{2}$$

Jest to model nieliniowo zależny od parametrów strukturalnych α_i, lecz łatwo sprowadzalny do postaci modelu liniowo-zależnego od α_i drogą prostych przekształceń nieliniowych.

Po ustaleniu postaci analitycznej funkcji występującej w problemie można przystąpić do oceny przydatności zmiennych w regresji wielokrotnej bezwarunkową metodą najmniejszej sumy kwadratów. Istnieje kilka procedur statystycznych, które pozwalają dokonać wyboru najlepszego równania regresji. Przy wyborze najlepszego równania regresji należy się kierować następującymi zasadami, które z natury rzeczy są antagonistyczne:

- do równania regresji wprowadza się możliwie najwięcej zmiennych w celu zapewnienia wiarygodności wyznaczonych ocen,
- ze względu na pracochłonność i koszty uzyskania informacji o dużej liczbie zmiennych należałoby uwzględnić jak najmniejszą liczbę nieskorelowanych ze sobą zmiennych najbardziej istotnie wpływających na analizowane zjawiska.

Spełnienie tych dwóch przeciwstawnych postulatów napotyka w praktyce na trudności, stąd należy się posługiwać określonymi procedurami doboru zespołu zmiennych kształtujących zmienną objaśnianą Y.

2. PROCEDURA REGRESJI KROKOWEJ

W celu właściwego doboru zmiennych objaśniających do modelu opisującego zmienność badanego zjawiska zastosowano procedurę, w której badania istotności regresji podlegają na każdym etapie zmiennej wprowadzonej do równania w poprzednich etapach.

Dokonuje się tego badania ze względu na to, że zmienna objaśniająca, która mogła być najlepszą pojedynczą zmienną do wprowadzenia w poprzedzającym etapie, może w etapie późniejszym być zbyteczna ze względu na swoją zależność od innych zmiennych objaśniających, mimo faktu, że jej współczynnik korelacji ze zmienną objaśnianą był większy od współczynnika korelacji zmiennych wprowadzonych później do funkcji regresji. Każda zmienna, która nie wnosi istotnego wkładu do wyjaśnienia zmienności badanego zjawiska jest usuwana z modelu.

Przedstawiamy teraz kilka kroków metody obliczeniowej, służącej do znalezienia najlepszego równania predykcji (1).

Krok 1. Procedura regresji krokowej rozpoczyna się od wyznaczenia symetrycznej macierzy korelacji R stopnia k + 1.

$$R = \begin{array}{c} \\ X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_k \\ Y \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & \cdots & X_k & Y \end{array} \\ \begin{bmatrix} 1 & v_{12} & \cdots & v_{1k} & v_{1k+1} \\ & 1 & \cdots & v_{2k} & v_{2k+1} \\ & & \cdots & & \cdots \\ & & \cdots & & \cdots \\ & & \cdots & & \cdots \\ & & & 1 & v_{k\,k+1} \\ & & & & 1 \end{bmatrix} \end{array}$$

Następnie rozszerza się daną macierz korelacji w następujący sposób:

$$A = [a_{ij}] = \begin{bmatrix} R\ (k \times k) & T'\ (k \times 1) & I\ (k \times k) \\ T\ (1 \times k) & S\ (1 \times 1) & 0\ (1 \times k) \\ -I\ (k \times k) & 0\ (k \times 1) & 0\ (k \times k) \end{bmatrix},$$

gdzie:

R (k x k) - macierz korelacji cząstkowej dla k zmiennych objaśniających,

T (1 x k) - wektor korelacji k zmiennych objaśniających ze zmienną objaśnianą Y,

T' (k x 1) - macierz transponowana macierzy T.

S (1 x 1) - macierz jednoelementowa (korelacja własna zmiennej objaśnianej),

I (k x k) - macierz jednostkowa,

-I (k x k) - ujemna macierz jednostkowa.

Dla wprowadzenia zmiennych objaśniających do równania regresji stosuje się ciąg statystyk:

$$V_i = \frac{v_{iy}\ \ y_{yi}}{v_{ii}} = \frac{a_{i.k+1}\ \ a_{k+1.i}}{a_{ii}}, \quad \text{dla } i = 1,2, \ldots k.$$

Wyboru pierwszej zmiennej dla wprowadzenia do regresji dokonuje się na podstawie warunku:

$$\max V = \max \left\{ V_1, V_2, \ldots V_k \right\}.$$

Warunek ten wyznacza zmienną najbardziej skorelowaną ze zmienną objaśnianą Y. Macierz $A = [a_{ij}]$ musi być dostosowana do wprowadzenia wybranej zmiennej do regresji. W tym celu wiersz macierzy $A = [a_{ij}]$, odpowiadający wprowadzonej zmiennej, należy podzielić przez element diagonalny w tym wierszu i wstawić otrzymany w ten sposób wiersz do macierzy $B = [b_{ij}]$. Pozostałe elementy macierzy $B = [b_{ij}]$ uzyskuje się przez zastosowanie następującego wzoru

$$b_{ij} = a_{ij} - \frac{a_{ii}\, a_{ij}}{a_{ii}}, \quad i = 1,2, \ldots, k,$$

gdzie i - numer zmiennej wprowadzonej do regresji w danym kroku.

Następnie przy użyciu typowego testu F sprawdza się, czy w ogóle wybrana zmienna powinna być wprowadzona do regresji. W tym celu oblicza się wartość funkcji testowej

$$F_{obl} = \frac{\hat{R}^2}{1 - \hat{R}^2} \cdot \frac{n - k - 1}{k}.$$

Kwadrat współczynnika korelacji wielokrotnej $\hat{R}^2$ jest określony wzorem

$$R^2 = \frac{\text{suma kwadratów w regresji}}{\text{centrowana suma kwadratów}}.$$

Jeżeli model regresyjny jest istotny, to w pierwszym a także w każdym następnym kroku zestawia się tablicę analizy wariancji oraz oblicza się duże miary struktury stochastycznej:
- kwadrat współczynnika korelacji wielokrotnej,
- odchylenie standardowe reszt.

Z tablic rozkładu F Snedecora wyznacza się dla poziomu α (np. α=0,05) przy $k = 1$ i $n - k - 1 = n - 2$ stopniach swobody wartość krytyczną $F(1; n-2; 0,05)$. Jeżeli $F_{obl} > F\ (1; n-2; 0,05)$, to odrzuca się hipotezę o nieistotności równania regresji, tzn. wprowadza się zmienną X_{ij} do równania regresji. W przypadku przeciwnym, gdy $F_{obl} \leq F\ (1; n-2; 0,05)$ nie wprowadza się do równania żadnej zmiennej. Wartość krytyczna F ustalona jest w każdym kroku oddzielnie zarówno dla wprowadzenia, jak i dla usuwania zmiennych objaśniających. W późniejszych etapach stosuje się sekwencyjny test F, służący do testowania hipotezy, czy ostatnia zmienna wprowadzona do regresji ma istotny udział w zmniejszeniu niewyjaśnionej zmienności ciągu danych empirycznych.

T e s t e l i m i n a c j i zmiennej występującej w regresji

Wartość krytyczna testu F dla wprowadzenia zmiennej jest nie mniejsza od wartości krytycznej F dla eliminacji zmiennej. Zazwyczaj obie wartości są równe. Oczywiście na tym etapie test eliminacji zmiennej występującej w regresji nie jest przeprowadzony, ponieważ w równaniu wystę-

puje tylko jedna zmienna. Test dotyczący eliminacji zmiennych występujących już w regresji zostanie omówiony w kroku 2.

Krok 2. Na tym etapie przeprowadza się wybór drugiej zmiennej dla wprowadzenia do regresji. Posługując się macierzą $B = [b_{ij}]$ określa się ciąg statystyk $\{V_s\}$ dla zmiennych nie występujących w regresji:

$$V_s = \frac{b_{s,k+1}\, b_{k+1,s}}{b_{ss}}.$$

Wyboru drugiej zmiennej dla wprowadzenia do regresji dokonuje się na podstawie tego samego warunku:

$$\max V = \max \{V_1, V_2, \ldots V_s\}.$$

Następnie oblicza się wartość F_{wp} testu sekwencyjnego dla wprowadzenia drugiej zmiennej. Łatwo stwierdzić, że w dowolnym etapie wielkość testowana, dotycząca pozycji następnej zmiennej, ma postać:

$$F_{wp} = \frac{q \max V}{\hat{s} - \max V},$$

gdzie:

q - liczba stopni swobody zmiennej resztowej,
$\hat{s}$ - suma kwadratów zmiennej resztowej.

Jeśli spełniona jest nierówność $F_{wp} > F(2; n-3; 0{,}05)$, to przyjmuje się drugą zmienną do równania regresji. W przypadku przeciwnym, gdy $F_{wp} \leq F(2; n-3; 0{,}05)$ drugiej zmiennej, a tym bardziej pozostałych zmiennych nie należy wprowadzać do modelu.

T e s t e l i m i n a c j i zmiennych występujących w regresji.

Na tym etapie eliminuje się zmienną wprowadzoną do regresji, na pierwszym kroku za pomocą częściowego testu F_{cz}. Częściowy test F jest wygodnym kryterium dla usuwania zmiennych z modelu. Wpływ pewnej zmiennej X_q na wyjaśnienie zmiennej objaśniającej Y może być duży, jeżeli równanie regresji zawiera tylko zmienną X_q. Jeśli jednak taka zmienna wchodzi do równania wraz z innymi zmiennymi, może ona oddziaływać bardzo mało na zmienną Y, ze względu na to, że X_q jest silnie skorelowana ze zmiennymi już występującymi w równaniu regresji. Częściowy test F można przeprowadzić dla wszystkich zmiennych występujących w regresji, tak jak gdyby była ostatnią zmienną wprowadzaną do równania, a więc stwierdzić względne oddziaływanie każdej zmiennej w stosunku do innych. Można jednak nie rozpatrywać częściowego testu F dla ostatniej zmiennej, gdyż wartość krytyczna F dla wprowadzenia zmiennej jest zawsze większa lub równa wartości krytycznej F dla usuwania zmiennej. Dla usunięcia zmiennej z regresji tworzy się nową macierz $C = [c_{ij}]$. W tym celu elementy wiersza

macierzy $B = [b_{ij}]$ odpowiadające wprowadzonej zmiennej dzieli się przez pierwszy element diagonalny w tym wierszu i wstawia się otrzymany w ten sposób wiersz do macierzy $C = [c_{ij}]$.
Wszystkie pozostałe elementy macierzy $C = [c_{ij}]$ wyznacza się ze wzoru:

$$c_{ij} = b_{ij} - \frac{b_{il}\ b_{lj}}{b_{ll}},$$

gdzie l - jest aktualnie podawaną zmienną (wynik ten jest ważny dla każdego elementu, z wyjątkiem znajdujących się w wierszu odpowiadającym właśnie wprowadzonej zmiennej).

Wartość F testu częściowego dla eliminacji zmiennej wprowadzonej do równania w poprzednim kroku oblicza się według wzoru

$$F_{cz} = \frac{q(c_{i,k+1})^2}{(c_{k+1,\ k+1})\ (c_{i+k+1,\ i+k+1})}.$$

W przypadku gdy $F_{cz} > F\ (2;\ n-3;\ 0{,}05)$ nie ma podstaw do wyeliminowania pierwszej zmiennej z regresji, w przypadku przeciwnym $[F_{cz} \leq F\ (2;\ n-3;\ 0{,}05)]$ usuwa się zmienną z regresji. Dla ustalenia, jak dalece funkcja będzie regresji przydatna jako predyktor podsumowuje się otrzymaną informację po wprowadzeniu każdej zmiennej. W tym celu zestawia się tablicę analizy wariancji oraz wyznacza się nieobciążone estymatory parametrów strukturalnych modelu, standardowe błędy ocen parametrów strukturalnych, kwadrat współczynnika korelacji wielokrotnej i odchylenie standardowe reszt.

Krok 3. Podobnie jak w kroku poprzednim do określenia zmiennej dla wprowadzenia do regresji tworzy się macierz $D = [d_{ij}]$, oblicza się ciąg statystyk $\{V_1\}$ dla zmiennych nie występujących w regresji i wybiera się zmienną, której odpowiada największa wartość V_i. Następnie oblicza się wartość F testu dla wprowadzenia oraz wartość F testów dla eliminacji zmiennych wprowadzonych do regresji w dwóch poprzednich krokach. Proces ten trwa tak długo, aż żądana ze zmiennych objaśniających nie będzie mogła być już wprowadzona do równania regresji i żadna odrzucona. Przyjęta metoda doboru zmiennych objaśniających do równania regresji zabezpiecza przed wprowadzeniem do modelu zmiennych przypadkowych, nieistotnie wpływających na badane zjawisko. Metoda regresji krokowej jest przydatna tam, gdzie należy rozpatrywać dużą liczbę zmiennych w różnym stopniu ze sobą skorelowanych i gdzie należy ograniczyć się do zmiennych najistotniejszych.

Na podstawie wyznaczonych w ten sposób równań można oszacować wartości oczekiwane zmiennej Y i odwrotnie. Wartości zmiennej Y odpowiadające określonym wartościom zmiennych objaśniających można oczasować również za pomocą przedziałów ufności.

LITERATURA

[1] Chajkim W., Najdzienow W., Crałzow S.: Korelacja i modelowanie statystyczne w rachunku ekonomicznym. PWN, Warszawa 1968.

[2] Draper N.R., Smith H.: Analiza regresji stosowana. PWN, Warszawa 1973.

АЛГОРИТМ ПРОВЕДЕНИЯ В ПРИМЕНЕНИИ ШАГОВОЙ РЕГРЕССИИ
ПРИ РЕШЕНИИ ИНЖЕНЕРНО-ОРГАНИЗАЦИОННЫХ ВОПРОСОВ

Р е з ю м е

В статье приводится подробная методика поведения при применении шаговой регресии. Этот метод пригоден там, где следует рассматривать большое число переменных в разной степени коррелированных друг с другом и где следует ограничиться к самым существенным переменным.

THE ALGORITHM OF PROCEDING IN THE APPLICATION OF STEPWISE REGRESSION
IN SOLVING ENGINEERING-ORGANISATIONAL PROBLEMS

S u m m a r y

The paper presents a detailed method of operations undertaken in applying the stepwise regression. The method is especially useful in the case of analysing a big number of variables correlated to different degree and where it is necessary to constraint to the most essential variables.