ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN — POLONIA
VOL. XXVII, 23 SECTIO H 1993

Zakład Analizy i Metod Rachunku Ekonomicznego Filii UMCS w Rzeszowie

Elżbieta MAKSYMIAK

O pewnej metodzie doboru zmiennych do modelu wielorównaniowego wykorzystującej zmodyfikowany współczynnik determinacji

A Certain Method of Choosing the Variables to a Multiequation Model Making Use of a Modified Coefficient of Determination

W ostatnim okresie w literaturze pojawiło się mnóstwo praca dotyczących różnych etapów budowy modelu ekonometrycznego. Szczególnie dużo miejsca poświęca się tematowi doboru zmiennych objaśniających do modelu ekonometrycznego, tj. jednemu z ważniejszych problemów prawidłowej konstrukcji modelu. Stosowane obecnie metody doboru zmiennych można podzielić na trzy grupy: metody statystyczne, taksonomiczne oraz oparte na analizie czynnikowej[1]. We wszystkich tych metodach preferuje się te zmienne, które są silnie skorelowane ze zmienną endogeniczną a słabo skorelowane między sobą. W pracy Z. Hellwig [9] wykazał na przykładzie prostego modelu, że maksymalizacja współczynników korelacji zmiennych objaśniających ze zmienną endogeniczną i minimalizacja współczynników korelacji między zmiannymi objaśniającymi daje w efekcie zmieniejszenie wariancji resztowej oraz zmniejszanie elementów macierzy wariancji i kowariancji estymatorów paramterów modelu. To zaś jest decydującym czynnikiem wpływającym na poprawność danego modelu, czyli w efekcie na jego praktyczną przydatność.

B. Guzik [6] przedstawil metodę doboru zmiennych do jednorównaniowego liniowego modelu ekonometrycznego wykorzystującą zmodyfikowany wspólczynnik determinacji

$$G = r^2 \det R,$$

[1] Do ważniejszych prac opisujących poszczególne grupy metod należą: — dla pierwszej grupy [2], [3], [4], [6], [8] — dla drugiej grupy [1], [5], [10], [12] — dla trzeciej grupy [7], [11].

gdzie r^2 oznacza kwadrat współczynnika korelacji wielorakiej, natomiast R jest macierzą korelacji dla zmiennych objaśniających. W metodzie tej za zmienne optymalne przyjmuje się taką kombinację zmiennych objaśniających wybraną ze zbioru potencjalnych zmiennych objaśniających, dla której wspólczynnik G przyjmuje wartość maksymalną. Zmienne wybrane do modelu według tej metody są słabo skorelowane między sobą oraz silnie skorelowane ze zmienną endogeniczną.

W niniejszej pracy uogólnimy metodę B. Guzika na liniowy model wielorównaniowy postaci

$$YB + XA + E = 0,$$

gdzie Y_{txm} — macierz obserwacji dokonanych na zmiennej endogenicznej, X_{txm} — macierz obserwacji dokonanych na zmiennych z gory ustalonych, $B = [\beta_{ij}]_{mxm}$ — macierz parametrów strukturalnych związanych ze zmienną endogeniczną, $A = [\alpha_{ij}]_{kxm}$ — macierz paramterów strukturalnych związanych ze zmiennymi z góry ustalonymi, $E = [e_{ij}]_{txm}$ — macierz składników losowych, t — liczba obserwacji, m — liczba zmiennych endogenicznych, k — liczba zmiennych z góry ustalonych.

Będziemy rozważać trzy następujące przypadki:

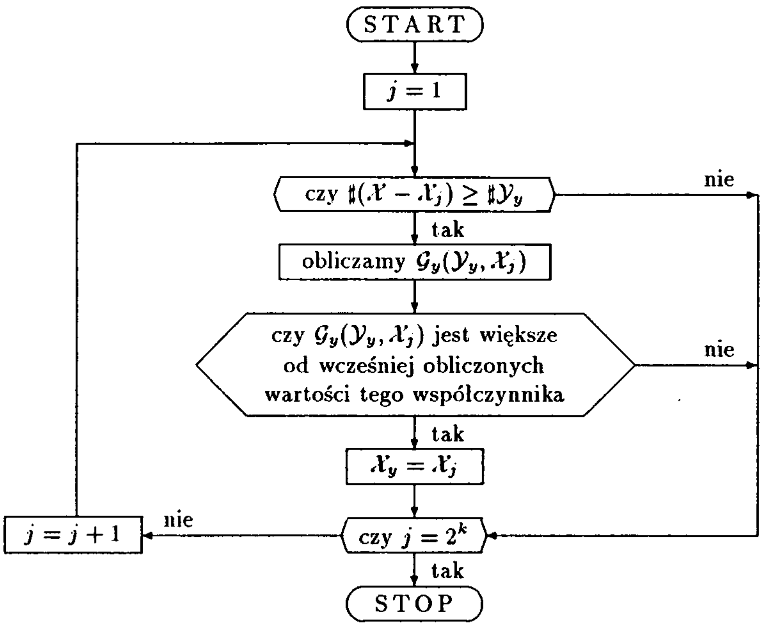
1) macierz B jest określona (tzn. jeżeli wiadomo dla jakich par (i,j) $i,j = 1,2,\ldots,$ m $\beta_{ij} = 0$ oraz macierz A nie jest określona,

2) macierz B nie jest określona i macierz A jest określona,

3) macierz B nie jest określona i macierz A nie jest określona.

Przypadek pierwszy ma miejsce wtedy, gdy teoria ekonomii wyznacza wspólzależność pomiędzy zmiennymi endogenicznymi, czyli gdy dla każdej zmiennej endogenicznej y wyznaczony jest zbiór $\mathcal{Y}_y$ zmiennych endogenicznych opisujących zmienną y. Niech $\mathcal{X}$ będzie zbiorem potencjalnych zmiennych z góry ustalonych. W tym przypadku dla każdej zmiennej y wybieramy podzbiór $\mathcal{X}_y$ zawarty w $\mathcal{X}$ taki, że równanie, w którym zmienną y opisują zmienne ze zbioru $\mathcal{Y}_y$ i $\mathcal{X}_y$ jest identyfikowane o maksymalnym zmodyfikowynym współczynniku determinacji. Z kolei jeżeli macierz A jest określona, tzn. jest wyznaczony zbiór $\mathcal{X}_y \subset \mathcal{X}$, którego elementy opisują zmienną y oraz macierz B nie jest określona, to dla każdej zmiennej endogenicznej y wybieramy zbiór $\mathcal{Y}_y$ taki, że równanie, w którym zmienną y opisują zmienne ze zbioru $\mathcal{Y}_y$ i zmienne ze zbioru $\mathcal{X}_y$, ma największy zmodyfikowany współczynnik determinacji wśród równań identyfikowalnych. W przypadku gdy macierze A i B nie są określone, to dla każdej zmiennej ednogenicznej y wybieramy dwa podzbiory $\mathcal{X}_y$ i $\mathcal{Y}_y$ zawarte odpowiednio w zbiorach $\mathcal{X}$ i $\mathcal{Y}$ takie, by równanie, w którym zmienną y opisują zmienne ze zbiorów $\mathcal{X}_y$ i $\mathcal{Y}_y$

było identyfikowalne o maksymalnym zmodyfikowanym współczynniku determinacji. Identyfikowalność sprawdzamy warunkiem koniecznym wymiarów, tzn. badamy czy liczba elementów zbioru $\mathcal{X} - \mathcal{X}_y$ jest większa lub równa od liczby elementów zbioru $\mathcal{Y}_y$.



Ryc. 1. Algorytm doboru zmiennych, gdy macierz B jest określona oraz macierz A nie jest określona

Algorythm of choice of variables, when matrix B is determined and matrix A is not

Poniżej przedstawimy algorytmy doboru zmiennych w każdym wymienionym wyżej przypadku. Niech $\mathcal{G}_y(\mathcal{Y}_i, \mathcal{X}_j)$ oznacza zmodyfikowany współczynnik determinacji obliczony dla modelu, w którym zmienną y objaśniają zmienne z i-tego podzbioru zbioru $\mathcal{Y}$ oraz z j-tego podzbioru $\mathcal{X}$, gdzie $i = 1, 2, \ldots, 2^m$, $j = 1, 2, \ldots, 2^k$. W przypadku pierwszym, gdy macierz B jest określona oraz macierz A nie jest określona algorytm doboru zmiennych polega na tym, że ze zbioru S wszystkich podzbiorów zbioru $\mathcal{X}$ wybieramy podzbiór S_o, którego elementami są podzbiory $\mathcal{X}_j \subseteq \mathcal{X}$ takie, że równanie objaśniające zmienną y przy pomocy zmiennych ze zbiorów $\mathcal{Y}_y$ i $\mathcal{X}_j$ jest identyfikowalne. Następnie ze zbioru S_o wybieramy zbiór $\mathcal{X}_{jo}$ spełniający warunek:

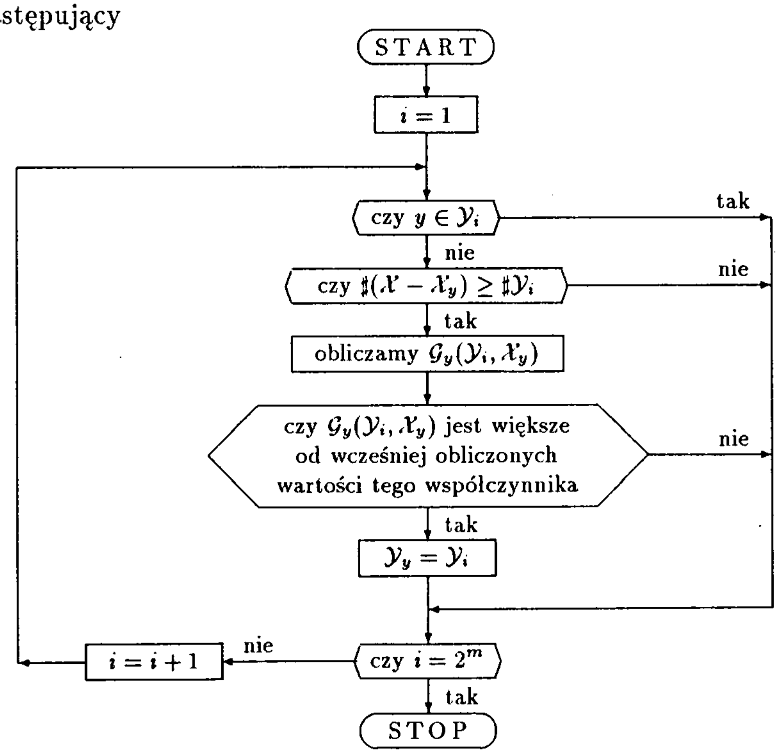
$$\mathcal{G}_y(\mathcal{Y}_y, \mathcal{X}_{jo}) = \max\{\mathcal{G}_y(\mathcal{Y}_y, \mathcal{X}_j) : \mathcal{X}_j \in S_o\}$$

i przyjmujemy, że $\mathcal{X}_y = \mathcal{X}_{jo}$. Aby określić macierz A trzeba zastosować algorytm dla każdej zmiennej endogenicznej $y \in \mathcal{Y}$. Przedstawiony algorytm zilustrujemy przy pomocy schematu blokowego (ryc. 1).

Jeżeli macierz B nie jest określona oraz macierz A jest określona, to algorytm doboru zmiennych polega na tym, że ze zbioru T wszystkich podzbiorów zbioru $\mathcal{Y}$ wybieramy podzbiór T_o tych zbiorów $\mathcal{Y}_i \in \mathcal{Y}$ dla których $y \notin \mathcal{Y}_i$ i równanie objaśniające zmienną y przy pomocy zmiennych ze zbioru $\mathcal{Y}_i$ oraz $\mathcal{X}_y$ jest identyfikowalne. Następnie ze zbioru T_o wybieramy podzbiór $\mathcal{Y}_{io}$ taki, że

$$\mathcal{G}_y(\mathcal{Y}_{io}, \mathcal{X}_y) = \max\{\mathcal{G}_y(\mathcal{Y}_i, \mathcal{X}_y) : \mathcal{Y}_i \in T_o\}$$

i przyjmujemy, że $\mathcal{Y}_y = \mathcal{Y}_{io}$. Algorytm ten stosujemy oddzielnie dla każdej zmiennej endogenicznej $y \in \mathcal{Y}$. Schemat blokowy dla powyższego przypadku jest następujący



Ryc. 2. Algorytm doboru zmiennych, gdy macierz B nie jest określona oraz macierz A jest określona

Algorythm of choice of variables, when matrix B is not determined and matrix A is

W przypadku gdy macierze A i B nie są określone, to ze zbioru U określonego w następujący sposób

$$U = \{(\mathcal{Y}_i, \mathcal{X}_j) : \mathcal{Y}_i \subseteq \mathcal{Y}, \mathcal{Y}_i \neq 0, \mathcal{X}_j \subseteq \mathcal{X}, \ i = 1, 2, \ldots, 2^m - 1; j = 1, 2, \ldots, 2^k\}$$

wybieramy podzbiór U_o tych par, dla których:
a) $y \notin \mathcal{Y}_i$
b) równanie, w którym zmienną y opisują zmienne ze zbiorów $\mathcal{Y}_i$ i $\mathcal{X}_j$ jest identyfikowalne.
Następnie dla każdego elementu $(\mathcal{Y}_i, \mathcal{X}_j)$ ze zbioru U_o obliczamy zmodyfikowany współczynnik determinacji $\mathcal{G}_y(\mathcal{Y}_i, \mathcal{X}_j)$ oraz wyznaczamy

$$\mathcal{G}_y(\mathcal{Y}_{io}, \mathcal{X}_{jo}) = \max\{\mathcal{G}_y(\mathcal{Y}_i, \mathcal{X}_j) : (\mathcal{Y}_i, \mathcal{X}_j) \in U_o\}.$$

W algorytmie tym za $\mathcal{Y}_y$ i $\mathcal{X}_y$ przyjmujemy odpowiednio zbiory $\mathcal{Y}_{io}$ i $\mathcal{X}_{io}$. Aby określić macierze A i B, należy ten algorytm zastosować oddzielnie dla każdej zmiennej endogenicznej $y \in \mathcal{Y}$.

LITERATURA

[1] Aftierowa Z., Jezżewa W.: Zastosowanie teorii grafów w rachunku ekonomicznym, PWE, Warszawa 1974.

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

[3] Forsythe A. B., Engelman L., Jennrid R., May Ph.: Stopping rule for variable selection in multiple regression, "Journal of the American Statistical Association" 1973, vol. 68.

[4] Goldberger A. S.: Teoria ekonometrii, PWE, Warszawa 1972.

[5] Gower J. C., Ros G. J. S.: Minimum spanning tress and single linkage cluster analysis, "Applied Statistics" 1969, vol. 18.

[6] Guzik B.: Propozycja kryterium zmodyfikowanego współczynnika determinacji dla doboru zmiennych objaśniających do modelu ekonometrycznego, „Przegl. Statyst." 1979, z. 1/2.

[7] Grabiński T., Szymanowicz K., Woźniak M., Zeliaś A.: O pewnej metodzie grupowania zmiennych, „Przegl. Statyst." 1976, z. 114.

[8] Hellwig Z.: Problem optymalnego wyboru predykant, „Przegl. Statyst." 1969, nr 3-4.

[9] Hellwig Z.: Rozważania nad istotą modelu ekonometrycznego, „Ekonomista" 1974, nr 2.

[10] Kozłowski S.: O sposobie weryfikacji podziałów przestrzennych w odniesieniu do taksonomicznej metody różnic J. Czekanowskiego, „Wiadom. Statyst." 1972, nr 1.

[11] Maly J.: Prosta metoda wyboru zmiennych objaśniających do modelu ekonometrycznego dla celów predykcji kompleksowej, „Przegl. Statyst." 1974, nr 1.

[12] Piasecki Z.: Nowa metoda taksonomiczna, „Listy Biometryczne" 1971, nr 30-33.

SUMMARY

The paper presents a model of choosing the variables for the econometric model making use of a modified coefficient of determination. The suggested method considers different cases depending on whether there are determined matrixes of structural parameters connected with the variables established in advance and with endogenic variable or whether there are not any.

ANNALES UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA

Druk: Zakład Poligrafii Wydawnictwa UMCS, Radziszewskiego 11, Lublin,
nakład 175 egzemplarzy + 25 nadbitek

A d r e s s e:

UNIWERSYTET MARII CURIE-SKŁODOWSKIEJ
WYDAWNICTWO

Plac Marii
Curie-Skłodowskiej 5 20-031 L U B L I N POLOGNE