PRZEGLĄD STATYSTYCZNY
R. XXXIV – zeszyt 1 – 1987

MAREK WALESIAK

ZMODYFIKOWANE KRYTERIUM DOBORU ZMIENNYCH OBJAŚNIAJĄCYCH DO LINIOWEGO MODELU EKONOMETRYCZNEGO

Celem prezentowanego artykułu jest zaproponowanie zmodyfikowanego kryterium doboru zmiennych objaśniających do liniowego modelu ekonometrycznego.(1) Idea tego kryterium wywodzi się ze wskaźnika pojemności integralnej nośników informacji Z. Hellwiga [5], [6] oraz ze zmodyfikowanego współczynnika determinacji B. Guzika [4].

1. KRYTERIUM DOBORU ZMIENNYCH

Rozważany jest liniowy model dla zmiennej Y, w którym X_i $(i=1, \ldots, k)$ są potencjalnymi zmiennymi objaśniającymi. Na zmiennej Y oraz każdej zmiennej X_i $(i=1, \ldots, k)$ dokonano T łącznych pomiarów $(y_t; x_{t1}, \ldots, x_{tk})$ $(t=1, \ldots, T; T>k)$. Niech $\mathbf{y}=[y_t]$ oznacza wektor wyników obserwacji zmiennej Y, a $\mathbf{x}_i=[x_{ti}]$ wektor wyników obserwacji zmiennej X_i $(i=1, \ldots, k)$.

Symbolem r_i oznaczany będzie współczynnik korelacji między wektorami $\mathbf{y}$, $\mathbf{x}_i$, a r_{ij} – współczynnik korelacji między wektorami $\mathbf{x}_i, \mathbf{x}_j$ $(i, j=1, \ldots, k)$. Ponadto symbolem $\mathbf{R}_0$ oznaczany będzie k-wymiarowy wektor kolumnowy o składowych r_i, natomiast symbolem $\mathbf{R}$ macierz kwadratowa stopnia k o elementach r_{ij}.

Dany jest zbiór K kombinacji potencjalnych zmiennych objaśniających. Spośród kombinacji $K_m \in K$ $(m=1, \ldots, L)$ wybiera się do liniowego modelu dla zmiennej Y taką, dla której pewien wskaźnik „jakości" $M(K_m)$ osiąga wartość optymalną.(2)

W tym miejscu należy zdefiniować pożądane własności wskaźnika $M(K_m)$.

1° $M(K_m)=1$, gdy wszystkie wektory $\mathbf{x}_i$ $(i \in K_m)$ są wzajemnie nieskorelowane i gdy równocześnie

$$\sum_{i \in K_m} r_i^2 = 1 .$$

2° $M(K_m)=0$, gdy ma miejsce jedna z poniższych sytuacji:

---

(1) Obszerny przegląd metod doboru zmiennych objaśniających do liniowego modelu ekonometrycznego zawarty jest między innymi w pracach [1], [2], [8], [9].

(2) Kombinacja zmiennych objaśniających K_m będzie dalej utożsamiana ze zbiorem numerów tych zmiennych.

a) w zbiorze zawierającym wektory $\mathbf{x}_i$ ($i \in K_m$) i T-elementowy wektor jedynek przynajmniej jeden wektor jest liniową kombinacją innych wektorów z tego zbioru, tzn. nie jest spełnione założenie

$$r(\mathbf{X}_m)=k_m+1, \tag{1}$$

gdzie: $\mathbf{X}_m$ – macierz wyników obserwacji zmiennych X_i ($i \in K_m$) rozszerzona o wektor jedynek, k_m – liczba zmiennych X_i w kombinacji K_m, $r(\mathbf{X})$ – rząd macierzy $\mathbf{X}$;
b) wektory $\mathbf{x}_i$ ($i \in K_m$) nie są skorelowane z wektorem $\mathbf{y}$;
c) zachodzi a i b łącznie.

3° Jeśli dla dwóch równolicznych kombinacji o numerach m i m'

$$\mathbf{R}_m=\mathbf{R}_{m'} \quad \wedge \quad |\mathbf{R}_{om}| \leqslant |\mathbf{R}_{om'}| \tag{2}$$

gdzie: $\mathbf{R}_m$, $\mathbf{R}_{m'}$ – macierz współczynników korelacji wektorów $\mathbf{x}_i$, $\mathbf{x}_j$ dla $i, j \in K_m$ oraz dla $i, j \in K_{m'}$, $\mathbf{R}_{om}$, $\mathbf{R}_{om'}$ – wektory współczynników korelacji między wektorami $\mathbf{y}$, $\mathbf{x}_i$ dla $i \in K_m$ oraz dla $i \in K_{m'}$, $|\mathbf{R}|$ – moduł wektora $\mathbf{R}$; to $M(K_m) \leqslant M(K_{m'})$.

4° Powinien być niemianowany.

5° Koncepcja oraz postać analityczna wskaźnika powinna być wystarczająco prosta.

W punktach 1° i 2° przedstawiono dwie sytuacje krańcowe (ważne z punktu widzenia teoretycznych rozważań), które w praktyce w czystej postaci rzadko występują.(3) Postulaty 1° i 2° w aspekcie poprawności budowanego liniowego modelu ekonometrycznego są oczywiste (potwierdzenie słuszności takiego postawienia problemu można znaleźć m.in. w pracach [4], [6], [7]).

Prawidłowo skonstruowany wskaźnik $M(K_m)$ powinien właściwie reagować nie tylko na przypadki krańcowe, ale również na przypadki „przeciętne". Realizacji tego celu ma służyć właśnie postulat 3°.

Ze względu na liczne walory, największe uznanie wśród ekonometryków prowadzących badania empiryczne uzyskała metoda Z. Hellwiga (por. prace [2], [5], [6], [8], [9]), w której kryterium wyboru kombinacji optymalnej jest maksymalizacja wskaźnika pojemności integralnej nośników informacji:

$$M(K_m)=H_m=\sum_{i \in K_m} \frac{r_i^2}{\sum\limits_{j \in K_m} |r_{ij}|}, \tag{3}$$

gdzie: H_m – wskaźnik pojemności informacji m-tej kombinacji zmiennych objaśniających.

Wskaźnik ten spełnia postulaty 2°b, 2°c, 3° 4° i 5°, natomiast nie spełnia postulatów 1° i 2°a. Postulat 3° jest przez wskaźnik (3) spełniony, ponieważ – przy przyjętych założeniach – sumy w mianownikach H_m i $H_{m'}$, są identyczne oraz

$$\sum_{i \in K_m} r_i^2 \leqslant \sum_{i \in K_{m'}} r_i^2,$$

co powoduje iż $H_m \leqslant H_{m'}$.

(3) Wielkości ekonomiczne są na ogół silnie skorelowane ze sobą.

H_m może przyjąć wartość jeden nie tylko w sytuacji przedstawionej w 1°, ale również wtedy gdy między wektorami $\mathbf{x}_i$ $(i \in K_m)$, a także między $\mathbf{y}$ oraz $\mathbf{x}_i$ $(i \in K_m)$ występuje ścisła zależność korelacyjna, tzn. gdy

$$\bigwedge_{i,j \in K_m} |r_{ij}| = 1 \quad \wedge \quad \bigwedge_{i \in K_m} |r_i| = 1 . \tag{4}$$

Ponadto, kierując się wskaźnikiem (3) można uznać za optymalną kombinację zmiennych objaśniających, utworzoną przez zmienne, których obserwacje, wektory $\mathbf{x}_i$ $(i \in K_m)$, są współliniowe (por. przykład 1).(4)

Inne kryterium wyboru optymalnej kombinacji podaje B. Guzik w pracy [4]. Odpowiedni wskaźnik jest iloczynem współczynnika determinacji i tzw. współczynnika rozsiewu (mierzącego wewnętrzne skorelowanie wektorów $\mathbf{x}_i$),

$$M(K_m) = G_m = R_m^2 (\det \mathbf{R}_m)^{0,5}, \tag{5}$$

gdzie: G_m – zmodyfikowany współczynnik determinacji dla m-tej kombinacji, R_m^2 – współczynnik determinacji dla m-tej kombinacji, $\det \mathbf{R}_m$ – wyznacznik macierzy $\mathbf{R}_m$ ($\det \mathbf{R}_m \in \langle 0; 1 \rangle$; $\det \mathbf{R}_m = 1$, gdy wszystkie wektory $\mathbf{x}_i$, $\mathbf{x}_j$ $(i, j \in K_m, i \neq j)$ są wzajemnie nieskorelowane; $\det \mathbf{R}_m = 0$, gdy przynajmniej jeden wektor $\mathbf{x}_i$ $(i \in K_m)$ jest liniową kombinacją innych wektorów z tego zbioru, tzn. gdy $r(\mathbf{X}_m) < k_m + 1$). Wskaźnik (5) nie spełnia tylko postulatu 3°, a wynika to stąd, że współczynnik R_m^2 przyjmuje dużą wartość nie tylko w przypadku silnej korelacji pomiędzy wektorami $\mathbf{y}$ oraz $\mathbf{x}_i$ $(i \in K_m)$, lecz również w przypadku silnego skorelowania między wektorami $\mathbf{x}_i$, $\mathbf{x}_j$ $(i, j \in K_m)$ (por. przykład 2).

Wszystkie przedstawione wcześniej postulaty 1° - 5° spełnia zmodyfikowany w następujący sposób wskaźnik pojemności integralnych nośników informacji:(5)

$$M(K_m) = H'_m = H_m (\det \mathbf{R}_m)^{0,5}. \tag{6}$$

Kombinacją optymalną w sensie kryterium (6) jest ta, dla której H'_m jest maksymalne.

Przedstawione w tym miejscu dwa przykłady pokazują niektóre własności wskaźników H_m i G_m.

PRZYKŁAD 1. (B. Guzik [4], s. 73). Dana jest macierz $\mathbf{R}$ i wektor $\mathbf{R}_0$

$$\mathbf{R} = \begin{bmatrix} 1 & 0{,}2835 & 1 \\ 0{,}2835 & 1 & 0{,}2835 \\ 1 & 0{,}2835 & 1 \end{bmatrix}, \quad \mathbf{R}_0 = \begin{bmatrix} 0{,}75 \\ 0{,}4725 \\ 0{,}75 \end{bmatrix}.$$

Dwie kombinacje są najlepsze w sensie G_m i H'_m, tzn. $\{X_1, X_2\}$ i $\{X_2, X_3\}$, natomiast w sensie H_m najlepsza jest kombinacja $\{X_1, X_2, X_3\}$. Wektory $\mathbf{x}_1$ i $\mathbf{x}_3$ są współliniowe, a mimo to wskaźnik H_m preferuje kombinację zawierającą te zmienne.

---

(4) W pracy [3] B. Guzik podaje pewną modyfikację metody Hellwiga dla przypadku współliniowości par zmiennych objaśniających.

(5) Wskaźnik (6) spełnia postulat 3° z uwagi na to, że $\det \mathbf{R}_m = \det \mathbf{R}_{m'}$ oraz $H_m \leqslant H_{m'}$. Ogólnie, wskaźnik (6) jest szczególnym przypadkiem wskaźnika $H'_m = H_m^p (\det \mathbf{R}_m)^v$ (p, v – liczby nieujemne), w którym $p = 1$, $v = 0{,}5$.

PRZYKŁAD 2. Dana jest macierz $\mathbf{R}$ i wektor $\mathbf{R}_0$

$$\mathbf{R}=\begin{bmatrix}1 & 0{,}875 & 0{,}875\\ 0{,}875 & 1 & 0{,}8125\\ 0{,}875 & 0{,}8125 & 1\end{bmatrix}, \quad \mathbf{R}_0=\begin{bmatrix}0{,}4677\\ 0{,}0668\\ 0{,}4009\end{bmatrix}.$$

Wartości miar H_m, G_m, H'_m dla kombinacji $\{X_1, X_2\}$ i $\{X_1, X_3\}$ są następujące

	H_m	G_m	H'_m
$\{X_1, X_2\}$	0,1190	0,3481	0,0576
$\{X_1, X_3\}$	0,2024	0,1060	0,0980

W sensie wskaźnika G_m lepsza jest kombinacja o niższych wartościach współczynników korelacji między wektorami $\mathbf{x}_i$ oraz wektorem $\mathbf{y}$, przy takiej samej macierzy współczynników korelacji między wektorami $\mathbf{x}_i$, $\mathbf{x}_j$.

Między miarą H'_m a miarami H_m i G_m istnieją następujące zależności:

$$H'_m \leqslant H_m, \quad H'_m \leqslant G_m.$$

Nierówności te są oczywiste. Pierwsza wynika stąd, że $(\det \mathbf{R}_m)^{0,5}$ zawiera się w przedziale $\langle 0, 1\rangle$, natomiast druga stąd, że – jak udowodnił Z. Hellwig w pracy [5] – $H_m \leqslant R_m^2$.

Można wskazać kilka przypadków, w których miary H_m, G_m i H'_m są sobie równe (por. [4]):

1. W przypadku jednoelementowej kombinacji zmiennych objaśniających

$$H_m = G_m = H'_m = r_i^2 \quad (i \in K_m).$$

2. Gdy wektory $\mathbf{x}_i$ $(i \in K_m)$ są wzajemnie nieskorelowane

$$H_m = G_m = H'_m = \sum_{i \in K_m} r_i^2.$$

3. Gdy wektory $\mathbf{x}_i$ $(i \in K_m)$ są nieskorelowane z wektorem $\mathbf{y}$

$$H_m = G_m = H'_m = 0.$$

2. PRZYKŁAD EMPIRYCZNY

Dla zilustrowania zaproponowanego kryterium doboru zmiennych (do liniowego modelu ekonometrycznego) przedstawimy przykład doboru zmiennych do modelu kształtowania się plonów 4 zbóż w q/ha w Polsce (Y_t) w latach 1970 - 1984. Potencjalnymi zmiennymi objaśniającymi są:

X_{1t} – zużycie nawozów sztucznych (w przeliczeniu na czysty składnik) na 1 ha użytków rolnych w kg;

X_{2t} – zużycie nawozów wapniowych na 1 ha użytków rolnych w kg;

X_{3t} – powierzchnia użytków rolnych na 1 ciągnik w ha;

X_{4t} – zmienna zero-jedynkowa, przyjmująca wartość jeden w latach urodzaju (1971 - 1974, 1976, 1978, 1982 - 1984) i wartość zero w latach nieurodzaju (lata pozostałe),

X_{5t} – dostawy pestycydów na zaopatrzenie rolnictwa w tys. ton.

Na podstawie danych statystycznych z lat 1970 - 1984 zaczerpniętych z Roczników Statystycznych obliczono wektor korelacji $\mathbf{R}_0$ oraz macierz korelacji $\mathbf{R}$:

$$\mathbf{R}_0 = \begin{bmatrix} 0{,}422 \\ 0{,}572 \\ -0{,}512 \\ 0{,}692 \\ 0{,}267 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1 & 0{,}725 & -0{,}855 & -0{,}171 & -0{,}071 \\ 0{,}725 & 1 & -0{,}819 & 0{,}084 & -0{,}274 \\ -0{,}855 & -0{,}819 & 1 & 0{,}050 & 0{,}140 \\ -0{,}171 & 0{,}084 & 0{,}050 & 1 & 0{,}205 \\ -0{,}071 & -0{,}274 & 0{,}140 & 0{,}205 & 1 \end{bmatrix}$$

Tablica 1 zawiera zestawienie wartości miar H_m, H'_m i G_m oraz ich rang (w nawiasach) dla poszczególnych kombinacji zmiennych objaśniających.

Tablica 1

m	Kombinacja	H_m	H'_m	G_m
1	X_1	0,1780 (30)	0,1780 (25)	0,1780 (27)
2	X_2	0,3271 (18)	0,3271 (13)	0,3271 (18)
3	X_3	0,2621 (27)	0,2621 (18)	0,2621 (21)
4	X_4	0,4788 (15)	0,4788 (7)	0,4788 (11)
5	X_5	0,0712 (31)	0,0712 (31)	0,0712 (31)
6	X_1, X_2	0,2928 (24)	0,2016 (21)	0,2253 (22)
7	X_1, X_3	0,2372 (28)	0,1230 (28)	0,1364 (29)
8	X_1, X_4	0,5608 (13)	0,5525 (5)	0,7686 (5)
9	X_1, X_5	0,2326 (29)	0,2320 (19)	0,2662 (20)
10	X_2, X_3	0,3239 (20)	0,1858 (22)	0,1909 (26)
11	X_2, X_4	0,7434 (1)	0,7407 (1)	0,7420 (6)
12	X_2, X_5	0,3126 (22)	0,3006 (15)	0,5014 (8)
13	X_3, X_4	0,7056 (3)	0,7047 (2)	0,7777 (3)
14	X_3, X_5	0,2923 (25)	0,2894 (16)	0,3752 (16)
15	X_4, X_5	0,4564 (16)	0,4467 (8)	0,4851 (10)
16	X_1, X_2, X_3	0,2955 (23)	0,0876 (30)	0,0995 (30)
17	X_1, X_2, X_4	0,6562 (6)	0,4212 (9)	0,5163 (7)
18	X_1, X_2, X_5	0,3156 (21)	0,2051 (20)	0,3425 (17)
19	X_2, X_3, X_4	0,7343 (2)	0,4106 (10)	0,4425 (12)
20	X_2, X_3, X_5	0,3404 (17)	0,1856 (23)	0,2843 (19)
21	X_3, X_4, X_5	0,6547 (7)	0,6343 (3)	0,7941 (1)
22	X_1, X_3, X_4	0,6175 (11)	0,3099 (14)	0,4024 (14)
23	X_1, X_3, X_5	0,2825 (26)	0,1444 (27)	0,1957 (25)
24	X_1, X_4, X_5	0,5470 (14)	0,5271 (6)	0,7730 (4)
25	X_2, X_4, X_5	0,6604 (5)	0,6213 (4)	0,7829 (2)
26	X_1, X_2, X_3, X_4	0,6522 (8)	0,1813 (24)	0,2252 (23)
27	X_1, X_2, X_3, X_5	0,3243 (19)	0,0907 (29)	0,1490 (28)
28	X_1, X_3, X_4, X_5	0,5991 (12)	0,2882 (17)	0,4016 (15)
29	X_2, X_3, X_4, X_5	0,6822 (4)	0,3482 (12)	0,4371 (13)
30	X_1, X_2, X_4, X_5	0,6214 (10)	0,3578 (11)	0,4944 (9)
31	X_1, X_2, X_3, X_4, X_5	0,6265 (9)	0,1552 (26)	0,2139 (24)

Źródło: obliczenia własne.

Kombinacją najlepszą w sensie miary H_m i H'_m jest $\{X_2, X_4\}$, natomiast w sensie miary G_m – kombinacja $\{X_3, X_4, X_5\}$. Chociaż kryteria H_m i H'_m wykazują tę samą kombinację jako najlepszą, to zasadnicze różnice ujawniają się na dalszych pozycjach. Skrajnym przypadkiem jest kombinacja zawierająca wszystkie zmienne objaśniające. Według miary H_m znajduje się na 9 pozycji, a według miary H'_m dopiero na 26. Daleka pozycja tej kombinacji wynika stąd, że wektory $\mathbf{x}_1$, $\mathbf{x}_2$ i $\mathbf{x}_3$ są silnie skorelowane między sobą.

Akademia Ekonomiczna we Wrocławiu

LITERATURA

[1] Draper N. R., Smith H., *Analiza regresji stosowana*, PWN, Warszawa 1973.

[2] Grabiński T., Wydymus S., Zeliaś A., *Metody doboru zmiennych w modelach ekonometrycznych*, PWN, Warszawa 1982.

[3] Guzik B., *Metoda Hellwiga w warunkach współliniowości par zmiennych objaśniających*, Przegląd Statystyczny, 1 (1985), s. 33 - 39.

[4] Guzik B., *Propozycja kryterium zmodyfikowanego współczynnika determinacji dla doboru zmiennych objaśniających do modelu ekonometrycznego*, Przegląd Statystyczny, 1 - 2 (1979), s. 67 - 78.

[5] Hellwig Z., *Efekt katalizy w modelu ekonometrycznym, jego wykrywanie i usuwanie*, Przegląd Statystyczny, 2 (1977), s. 179 - 191.

[6] Hellwig Z., *O jakości modelu ekonometrycznego*, Przegląd Statystyczny, 1 (1985), s. 3 - 23.

[7] Hellwig Z., *Rozważania nad istotą modelu ekonometrycznego*, Ekonomista, 2 (1974), s. 305 - 324.

[8] Nowak E., *Problemy doboru zmiennych do modelu ekonometrycznego*, PWN, Warszawa 1984.

[9] Strahl D., *Modelowanie zjawisk złożonych. Modele infrastruktury społecznej*, Prace Naukowe AE we Wrocławiu, nr 158, Wrocław 1980.

Praca wpłynęła do Redakcji w lutym 1986 r.
Wersja ostateczna w grudniu 1986 r.

МОДИФИЦИРОВАННЫЙ КРИТЕРИЙ ПОДБОРА ОБЪЯСНЯЮЩИХ ПЕРЕМЕННЫХ К ЛИНЕЙНОЙ ЭКОНОМЕТРИЧЕСКОЙ МОДЕЛИ

Резюме

В статье исследуются свойства трех критериев подбора переменных (к линейной эконометрической модели) с учётом некоторых дополнительных постулатов.

A MODIFIED CRITERION OF EXPLANATORY VARIABLES' SELECTION FOR LINEAR ECONOMETRIC MODEL

Summary

In the paper, properties of three criteria of explanatory variables' selection (for linear econometric model) are examined from the point of view of predetermined postulates.