Ekonometria – wykład do kursu DO i E
na kierunku LIT
W3

Temat: Metoda wskaźników pojemności informacyjnej i jej ilustracja na przykładach.

Założenia: Dany jest MEL, gdzie

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_m X_m + \varepsilon$$

Na podstawie danych z n-krotnej obserwacji $Y, X_1, X_2, \ldots, X_m$, gdzie

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}$$

$$\quad X_1 \qquad X_2 \quad \ldots \quad X_m$$

wyznaczono R_0 – wektor korelacji oraz R – macierz współczynników korelacji.

Cel: Wybrać takie zmienne objaśniające, które są silnie skorelowane ze zmienną objaśnianą, a jednocześnie słabo skorelowane ze sobą.

Metoda: wskaźników pojemności informacyjnej.

Opis metody

Rozpatrujemy wszystkie zestawienia potencjalnych zmiennych objaśniających, a więc wybieramy wszystkie pojedyncze zestawy, podwójne, itd. całości.
Z podstaw kombinatoryki wiadomo, że

$$\# \text{pojedynczych} = \binom{m}{1} = \frac{m!}{1!(m-1)!}$$

$$\# \text{podwójnych} = \binom{m}{2} = \frac{m!}{2!(m-2)!}$$

$$\vdots$$

$$\# \text{wszystkich} = \binom{m}{m} = \frac{m!}{m!(m-m)!}$$

Niech L liczba tych zestawień. Wtedy

$$L = \binom{m}{1} + \binom{m}{2} + \ldots + \binom{m}{m}$$

Ale, z podstaw kombinatoryki wiadomo, iż

$$2^m = \binom{m}{0} + \binom{m}{1} + \ldots + \binom{m}{m} = \binom{m}{0} + L$$

Ponieważ $\binom{m}{0} = \frac{m!}{0!(m-0)!} = \frac{m!}{1 \cdot m!} = 1$,

więc

$$\boxed{L = 2^m - 1}$$

Dla każdego takiego zestawu oblicza się wskaźniki pojemności informacyjnej: indywidualne i integralne. Ich wartości są z przedziału $[0,1]$ oraz przyjmują one tym większe wartości, im zmienne objaśniające są silniej skorelowane ze zmienną objaśnianą oraz im słabiej są skorelowane ze sobą.

Rozwiązaniem problemu jest to zestawienie, dla którego wartość wskaźnika integralnej pojemności informacyjnej jest największa.

Indywidualne wskaźniki pojemności informacyjnej.

Niech l – numer zestawu, $(l = 1, 2, \ldots, L)$

j – numer zmiennej w zestawie

m_l – # zmiennych w tym zestawie

Wtedy

$$(*) \quad h_{lj} \stackrel{df}{=} \frac{r_j^2}{1 + \sum_{\substack{i=1 \\ i \neq j}}^{m_l} |r_{ij}|}$$

Wtedy integralny wskaźnik H_l ma postać

$$(**) \quad H_l = \sum_{j=1}^{m_l} h_{lj} \quad (l = 1, 2, \ldots, L)$$

Niech $H_{l_0} = \max\{H_1, H_2, \ldots, H_L\}$.

Wtedy zestaw l_0 staje się wybranym.

Ze względu na złożoność techniczną metody, szczegóły wyjaśnimy na przykładzie.

Przykład

Wybrać zmienne objaśniające do MEL

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon,$$ gdzie

Y – kształtowanie się produkcji żywca przypadającej na 1 ha użytków rolnych

X_1 – plon kukurydzy przeznaczonej na pasze

X_2 – udział wartości produkcji roślinnej w wartości produkcji globalnej

X_3 – średnia cena skupu żywca

X_4 – zużycie pasz treściwych,

jeśli wiadomo, iż na podstawie danych zebranych za pewien okres R_0 i R ma postać:

$$R_0 = \begin{bmatrix} 0,43 \\ -0,80 \\ 0,18 \\ 0,63 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & -0,64 & 0,14 & 0,41 \\ -0,64 & 1 & -0,13 & -0,55 \\ 0,14 & -0,13 & 1 & -0,03 \\ 0,41 & -0,55 & -0,03 & 1 \end{bmatrix}$$

<u>Krok 1</u> Mamy 4 zmienne objaśniające $(m=4)$, zatem

$$L = 2^4 - 1 = 15$$

<u>Krok 2</u>. Wypisujemy wszystkie zestawy według określonej kolejności:

$Z_1 = (X_1)$

$Z_2 = (X_2)$

$Z_3 = (X_3)$

$Z_4 = (X_4)$

1-el

$Z_5 = (X_1, X_2)$

$Z_6 = (X_1, X_3)$

$Z_7 = (X_1, X_4)$

$Z_8 = (X_2, X_3)$

$Z_9 = (X_2, X_4)$

$Z_{10} = (X_3, X_4)$

2-el. 6

$Z_{11} = (X_1, X_2, X_3)$

$Z_{12} = (X_1, X_2, X_4)$

$Z_{13} = (X_1, X_3, X_4)$

$Z_{14} = (X_2, X_3, X_4)$

3-el.

$Z_{15} = (X_1, X_2, X_3, X_4)$ – 4-el.

Krok 2. Wylicz h_{ij}, H_l, $l = 1, \dots, L$

Zaczynamy od $l = 1, \dots, 4$ – czyli jednoelementowe Z_1, Z_2, Z_3, Z_4. $(m_l = 1)$

Wtedy:

l	j
1	1
2	2
3	3
4	4

Dlatego 2 (*)

$$h_{11} = \frac{r_1^2}{1} = r_1^2 = (0{,}43)^2 = 0{,}185$$

bo $m_l = 1$

$$h_{22} = r_2^2 = (-0{,}80)^2 = 0{,}640$$

$$h_{33} = r_3^2 = (0{,}181)^2 = 0{,}032$$

$$h_{44} = r_4^2 = (0{,}63)^2 = 0{,}370$$

Stąd $H_1 = h_{11}$, $H_2 = h_{22}$, $H_3 = h_{33}$, $H_4 = h_{44}$

Dla ~~Hz~~ $m_c = 2$ mamy zestawy

$$Z_5, Z_6, \ldots, Z_{10}$$

l	j
5	1 ; 2
6	1 ; 3
7	1 ; 4
8	2 ; 3
9	2 ; 4
10	3 ; 4

Dla Z_5: $Z_5 = (X_1, X_2)$

$$h_{51} = \frac{r_1^2}{1 + |r_{12}|} = \frac{(0,43)^2}{1 + 0,64} = 0,113$$

$$h_{52} = \frac{r_2^2}{1 + |r_{21}|} = \frac{(-0,80)^2}{1 + 0,64} = 0,390$$

i dla Z_5 $H_5 = h_{51} + h_{52} = \underline{0,503}$

Zadanie 1

Skorzystać ponownie uzasadnić, że

$h_{61} = 0{,}162 \quad h_{63} = 0{,}028 \Rightarrow M_6 = 0{,}190$

$h_{71} = 0{,}131, \quad h_{7;4} = 0{,}281 \Rightarrow M_7 = 0{,}412$

$h_{82} = 0{,}566 \quad h_{83} = 0{,}029 \Rightarrow M_8 = 0{,}595$

$h_{92} = 0{,}413 \quad h_{94} = 0{,}255 \Rightarrow M_9 = 0{,}668$

$h_{10;3} = 0{,}031 \quad h_{10,4} = 0{,}375 \Rightarrow M_{10} = 0{,}406$

Biorę przypadek $m_l = 3$, czyli $Z_{11}, Z_{12}, Z_{13}, Z_{14}$

l	j
11	1 ; 2 ; 3
12	1 ; 2 ; 4
13	1 ; 3 ; 4
14	2 ; 3 ; 4

Mamy kolejno:

Z_{11} $h_{11,1} = \frac{r_1^2}{1+|r_{12}|+|r_{13}|} = 0{,}104$

$h_{11,2} = \frac{r_2^2}{1+|r_{21}|+|r_{23}|} = 0{,}362$

$h_{11,3} = \frac{r_3^2}{1+|r_{31}|+|r_{32}|} = 0{,}026$

Dlatego $H_{11} = h_{11,1} + h_{11,2} + h_{11,3} = 0{,}492$

Zadanie 2

Stosując powyższe urządzenia, są

$h_{12,1} = 0{,}090$, $h_{12,2} = 0{,}292$, $h_{12,3} = 0{,}202$

$H_{12} = 0{,}583$

$h_{13,1} = 0{,}119$, $h_{13,3} = 0{,}028$, $h_{13,4} = 0{,}276$

$H_{13} = 0{,}423$

$h_{14,2} = 0{,}381$, $h_{14,3} = 0{,}028$, $h_{14,4} = 0{,}251$

$H_{14} = 0{,}660$

Obliczmy M_{15} : $(m_l = 4)$ dla Z_{15}

l	j
15	1 2 3 4

Z (*) dostaję kolejno :

$$h_{15,1} = \frac{r_1^2}{1+|r_{12}|+|r_{13}|+|r_{14}|} = 0,084$$

$$h_{15,2} = \frac{r_2^2}{1+|r_{21}|+|r_{23}|+|r_{24}|} = 0,276$$

$$h_{15,3} = \frac{r_3^2}{1+|r_{31}|+|r_{32}|+|r_{34}|} = 0,025$$

$$h_{15,4} = \frac{r_4^2}{1+|r_{41}|+|r_{42}|+|r_{43}|} = 0,200,$$

skąd $M_{15} = h_{15,1} + \ldots + h_{15,4} = 0,585$

Mamy komplet wskaźników integralnych

$\{H_1, H_2, \dots, H_{15}\}$

Zauważmy, że

$$\max\{H_1, H_2, \dots, H_{15}\} = H_{l^*} = 0{,}668$$

$l^* = 9$, czyli

$Z_9 = (X_2, X_4)$.

WNIOSEK: Dla zmiennych

(Y, X_1, X_2, X_3, X_4)

1^0. X_2, X_4 są najbardziej skorelowane z Y, a jednocześnie

2^0. najmniej ze sobą.

Dlatego MEL

$$\underbrace{Y = \alpha_0 + \alpha_1 X_2 + \alpha_2 X_4 + \varepsilon}$$