

Dobór zmiennych objaśniających (model liniowy).

$X_1, X_2, \dots, X_k$ – zmienne objaśniające,

Y - zmienna objaśniana,

Zmienne objaśniające powinny
charakteryzować się:

- a) wysoką zmiennością (współczynnik zmienności powyżej określonej wartości krytycznej np., $V(X_i) > 0,1$),
- b) silną korelacją z Y ,
- c) słabą korelacją z innymi zmiennymi objaśniającymi.

Przykład.

Mając dane wartości zmiennych

Y	X ₁	X ₂	X ₃	X ₄
5	3	10	7	6
5	4	8	7	5
8	5	7	6	7
8	6	7	7	7
6	5	6	9	6
7	5	5	10	6
10	7	5	12	6
10	7	4	10	7
12	6	4	11	6
12	8	4	12	6

Sprawdź, które zmienne należy wyeliminować jako quasi stałe przyjmując krytyczną wartość współczynnika zmienności równą 0,15?

Współczynniki zmienności dla poszczególnych zmiennych objaśniających są równe:

V(X ₁)	V(X ₂)	V(X ₃)	V(X ₄)
0,255	0,316	0,233	0,097

Zatem należy wyeliminować zmienną X₄ .

Jeśli zmienne X , Y mają pary wartości (x_i, y_i) to współczynnik korelacji Pearsona obliczamy następująco:

$$r_{XY} = \frac{\text{cov}(X, Y)}{S_X \cdot S_Y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} =$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n(\bar{x})^2} \sqrt{\sum y_i^2 - n(\bar{y})^2}}$$

Niech $r_i = r_{X_i Y}$ - współczynniki korelacji między poszczególnymi zmiennymi objaśniającymi a zmienną objaśnianą.

Wektorem korelacji nazywamy wektor

$$R_0 = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix}$$

Niech $r_{ij} = r_{X_i X_j}$ - współczynniki korelacji między poszczególnymi zmiennymi objaśniającymi. Macierzą korelacji nazywamy symetryczną macierz

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

Metoda wyboru zmiennych objaśniających, które są silnie skorelowane ze zmienną objaśniającą i słabo skorelowane z innymi zmiennymi objaśniającymi.

Kolejność postępowania:

- 1) ustalamy krytyczną wartość korelacji r_{kryt} (albo jest narzucona albo obliczamy ze

wzoru $r_{\text{kryt}} = \sqrt{\frac{k_{\alpha}^2}{k_{\alpha}^2 + n - 2}}$ gdzie k_{α} -

odczytujemy z tablicy rozkładu Studenta dla $n - 2$ stopni swobody i poziomu istotności α .

- 2) eliminujemy te zmienne objaśniające dla których:

$$|r_i| \leq r_{\text{kryt}}$$

- 3) spośród pozostałych zmiennych

wybieramy taką zmienną X_s dla której

$$|r_s| = \max \{|r_i|\} \quad (\text{ta zmienna niesie najwięcej informacji})$$

- 4) ze zbioru zmiennych objaśniających **eliminujemy** te dla których

$$|r_{si}| > r_{\text{kryt}}$$

(zmienne silnie skorelowane z wybraną zmienną X_s powielają zawarte w X_s informacje).

Kroki 3) i 4) można ewentualnie powtarzać.

Przykład.

Dla zmiennych X_1, X_2, X_3, X_4 i Y z poprzedniego przykładu wektor korelacji i macierz korelacji są równe:

$$R_0 = \begin{bmatrix} 0,88 \\ -0,82 \\ 0,73 \\ 0,29 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & -0,85 & 0,74 & 0,33 \\ -0,85 & 1 & -0,82 & -0,18 \\ 0,74 & -0,82 & 1 & -0,17 \\ 0,33 & -0,18 & -0,17 & 1 \end{bmatrix}$$

1) Dla poziomu istotności 0,05 i $10 - 2 = 8$ stopni swobody odczytujemy z tablicy rozkładu Studenta $k_\alpha = 2,306$ i wyznaczamy

$$r_{\text{kryt}} = \sqrt{\frac{k_\alpha^2}{k_\alpha^2 + n - 2}} = \sqrt{\frac{2,306^2}{2,306^2 + 10 - 2}} = 0,63$$

2) odrzucamy zmienną X_4 ,

Zredukowany wektor i zredukowana macierz korelacji są równe

$$R_0 = \begin{bmatrix} 0,88 \\ -0,82 \\ 0,73 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & -0,85 & 0,74 \\ -0,85 & 1 & -0,82 \\ 0,74 & -0,82 & 1 \end{bmatrix}$$

3) wybieramy X_1 ,

4) eliminujemy X_2 , X_3 ,

Zatem rozpatrywany model liniowy powinien mieć postać:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Metoda Hellwiga.

Rozpatrujemy wszystkie niepuste podzbiory zbioru zmiennych objaśniających

$\{X_1, X_2, \dots, X_k\}$,

takich podzbiorów jest $L = 2^k - 1$.

Dla każdego podzbioru oblicza się wskaźniki pojemności informacyjnej: indywidualne i integralne (ich wartości należą do przedziału $[0, 1]$).

Indywidualną pojemność informacyjną obliczamy ze wzoru:

$$h_{lj} = \frac{r_j^2}{\sum_{i \in I_l} |r_{ij}|}$$

gdzie $l = 1, 2, \dots, L$ (numer podzbioru - kombinacji),

I_l - zbiór numerów zmiennych wchodzących w skład l - tego podzbioru.

Integralną pojemność informacyjną obliczamy sumując pojemności indywidualne rozpatrywanego podzbioru:

$$H_l = \sum_{j \in I_l} h_{lj}$$

Należy wybrać taki podzbiór zmiennych objaśniających dla którego integralna pojemność informacyjna jest maksymalna.

Przykład.

Dla zmiennych X_1, X_2, Y obliczono

$$R_0 = \begin{bmatrix} 0,8 \\ -0,2 \end{bmatrix} \quad R = \begin{bmatrix} 1 & 0,6 \\ 0,6 & 1 \end{bmatrix}$$

Są 3 podzbiory zbioru $\{X_1, X_2\}$:

$\{X_1\}$,

$\{X_2\}$,

$\{X_1, X_2\}$.

Obliczamy:

$$h_{11} = 0,8^2 = 0,64, \quad H_1 = 0,64,$$

$$h_{22} = (-0,2)^2 = 0,04, \quad H_2 = 0,04,$$

$$h_{31} = 0,8^2/(1 + 0,6) = 0,4,$$

$$h_{32} = (-0,2)^2/(1 + 0,6) = 0,025,$$

$$H_3 = h_{31} + h_{32} = 0,425,$$

Ponieważ największą pojemność informacyjną ma podzbiór $\{X_1\}$, to należy przyjąć, że X_1 jest jedyną zmienną objaśniającą w tym modelu tzn. $Y = \beta_0 + \beta_1 X_1 + \varepsilon$.