9 Korelacja

Jeśli badaniu statystycznemu podlega jednocześnie wiele cech, to jednym z podstawowych zagadnień staje się analiza zależności pomiędzy nimi.

Do wykrycia zależności pomocne są odpowiednie wykresy, współczynniki mierzące jej siłę oraz testy badające jej istotność.

Metody opisowe:

1. Tabelaryczna - tablice wielodzielcze (korelacyjne).
2. Graficzna - diagramy korelacyjne (wykresy rozrzutu).
3. Statystyki opisowe – współczynniki zależności i korelacji.

Współczynniki zależności i korelacji

Niech $\boldsymbol{x} = (x_1, \ldots, x_n)'$ i $\boldsymbol{y} = (y_1, \ldots, y_n)'$ będą obserwacjami zmiennej X oraz zmiennej Y, odpowiednio.

DEFINICJA

Współczynnikiem korelacji liniowej r-Pearsona nazywamy liczbę:

$$r = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \bar{x})^2 \sum_{k=1}^{n}(y_k - \bar{y})^2}}, \qquad -1 \leq r \leq 1.$$

Dwuwymiarowy model normalny

Zakładamy, że łącznym rozkładem zmiennych X i Y jest dwuwymiarowy rozkład normalny.

FAKT

Zmienne X i Y są nieskorelowane wtedy i tylko wtedy, gdy $\rho = 0$, zatem w tym modelu pojęcia korelacji i niezależności są równoważne.

FAKT

Estymatorem największej wiarogodności współczynnika korelacji ρ jest statystyka:

$$\hat{\rho} = r = \frac{\sum_{k=1}^{n}(X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \bar{X})^2 \sum_{k=1}^{n}(Y_k - \bar{Y})^2}}.$$

Test istotności dla współczynnika korelacji

Założenia: Dwuwymiarowy model normalny.

Hipoteza zerowa: Zmienne XX i YY nie są istotnie (skorelowane) zależne.

$$H_0:\ \rho = 0$$

Hipotezy alternatywne:

$$H_1:\ \rho \neq 0$$

$$H_1:\ \rho > 0$$

$$H_1:\ \rho < 0$$

Statystyka testowa:

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}.$$

Rozkład statystyki testowej:

$$t|_{H_0} \sim t(n - 2).$$

pp-wymiarowy model normalny

Gęstość:

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{p}{2}}\exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})],\ \boldsymbol{x} \in \mathbf{R}^p,$$

gdzie

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \text{- wektor wartości oczekiwanych,}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \text{ - macierz wariancji-kowariancji.}$$

Funkcje związane z analizą zależności i korelacji cech:

cor - estymator macierzy korelacji, estymator współczynnikia korelacji liniowej r-Pearsona,

cor.test - test istotności współczynnika korelacji liniowej.

Redukcja wymiaru

Analiza składowych głównych jest techniką redukcji wymiaru. Jej celem jest znalezienie niewielkiej liczby składowych głównych, które wyjaśniają w maksymalnym stopniu całkowitą wariancję z próby p zmiennych pierwotnych $X_1, \ldots, X_p$, tj. wielkość $\sum_{j=1}^{p} \mathrm{Var}(X_j) = \mathrm{tr}(\mathbf{\Sigma})$, gdzie $\mathbf{\Sigma}$ jest macierzą kowariancji wektora $\boldsymbol{X} = (X_1, \ldots, X_p)'$.

Składowe główne są unormowanymi kombinacjami liniowymi zmiennych pierwotnych:

$$\begin{aligned} Z_1 &= \boldsymbol{a}_1'\boldsymbol{X}, \\ Z_2 &= \boldsymbol{a}_2'\boldsymbol{X}, \\ &\vdots \\ Z_p &= \boldsymbol{a}_p'\boldsymbol{X}. \end{aligned}$$

Przekształcone zmienne (składowe główne) są ortogonalne i nieskorelowane.

Uwaga: Ponieważ macierz $\mathbf{\Sigma}$ nie jest znana, posługujemy się jej oszacowaniem z próby, tj. macierzą $\boldsymbol{S}$.

Algorytm składowych głównych

1. Wyznaczamy współczynniki $\boldsymbol{a}_1 = (a_{11}, \ldots, a_{1p})'$ pierwszej składowej głównej, tak aby
   - zmaksymalizować wariancję zmiennej Z_1: $\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1$,
   - $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$.
2. Wyznaczamy współczynniki $\boldsymbol{a}_2 = (a_{21}, \ldots, a_{2p})'$ drugiej składowej głównej, tak aby
   - zmaksymalizować wariancję zmiennej Z_2: $\boldsymbol{a}_2'\boldsymbol{S}\boldsymbol{a}_2$,
   - $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$,

- składowa Z_2 była nieskorelowana z Z_1: $\boldsymbol{a}_2^{'}\boldsymbol{a}_1 = 0$.

3. Powtarzamy krok 2 (dla następnych składowych głównych), aż do otrzymania współczynników wszystkich p składowych głównych.

Uwaga: Wektor $\boldsymbol{a}_i$ jest wektorem charakterystycznym, odpowiadającym i-tej co do wielkości, wartości własnej λ_i macierzy $\boldsymbol{S}$.

Własności składowych głównych

Mamy

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \mathrm{Var}(Z_j) = \sum_{j=1}^{p} \lambda_j = \mathrm{tr}(\boldsymbol{S}).$$

W analizie składowych głównych oczekujemy, że dla pewnego małego k, suma $\lambda_1 + \lambda_2 + \cdots + \lambda_k$ będzie bliska $\mathrm{tr}(\boldsymbol{S}) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$. Jeśli tak jest, to k pierwszych składowych głównych wyjaśnia dobrze zmienność wektora $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^{'}$ i pozostałe $p - k$ składowe główne wnoszą niewiele, ponieważ mają one małe wariancje z próby.

Wskaźnik

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} 100\%$$

jest procentową miarą wyjaśniania zmienności wektora $\boldsymbol{X}$ przez pierwszych k składowych głównych.

Dobór liczby składowych głównych

1. Jeśli dla pewnego k wskaźnik

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} 100\% \geq \beta,$$

   np. $\beta = 80\%$, to pozostałe $p - k$ składowe główne pomijamy.

2. Pomijamy te składowe główne, których wartości własne są mniejsze od średniej

$$\bar{\lambda} = \frac{1}{p}\sum_{j=1}^{p} \lambda_j.$$

Uwaga: W ustaleniu liczby użytecznych składowych głównych, pomocny jest **wykres osypiska**.

Interpretacja składowych głównych

1. Wartość modułu współczynnika a_{ji}, w j-tej składowej głównej, pokazuje wkład w jej budowę i-tej zmiennej pierwotnej (z uwzględnieniem udziału pozostałych zmiennych pierwotnych).
2. Wartość współczynnika korelacji pomiędzy i-tą zmienna pierwotną, a j-tą składową główną, pokazuje wkład i-tej zmiennej pierwotnej w budowę j-tej składowej głównej (bez uwzględnienia udziału pozostałych zmiennych pierwotnych).

Funkcje związane z analizą składowych głównych:

princomp - analiza składowych głównych, procedura główna.