Korelačná analýza

Korelačná analýza

skúma **intenzitu** stochastickej - náhodnej závislosti medzi štatistickými premennými

- **jednoduchá** korelácia – medzi dvomi premennými Y a X

- **viacnásobná** korelácia – medzi premennou Y a dvoma alebo viacerými premennými X_1, X_2,..., X_n

Miery intenzity závislosti

- **korelačný pomer** a pomer determinácie - ak nepoznáme riešenie regresnej úlohy
- **koeficient korelácie** a **koeficient determinácie** – ak riešením korelačnej úlohy je regresná funkcia lineárna v parametroch
- **Spearmanov koeficient poradovej korelácie** – pre ordinálne premenné
- pri viacnásobnej korelácii – **koeficient mnohoásobnej korelácie**

Pomer determinácie η_{xy}^2

- meria intenzitu závislosti v prípade, že nepoznáme regresný model
- výpočet sa odvodzuje z rozkladu rozptylu na zložky (1. rozklad rozptylu):

  celkový rozptyl = vnútroskupinový rozptyl + medziskupinový rozptyl

$$s_y^2 = \overline{s_i^2} + s_{\bar{y}}^2$$

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 n_i}{n} = \frac{\sum_{i=1}^{n} s_i^2 n_i}{n} + \frac{\sum_{i=1}^{n}(\bar{y}_i - \bar{y})^2 n_i}{n}$$

Ak rovnicu vydelíme celkovým rozptylom s_y^2, dostaneme

$$1 = \frac{\sum_{i=1}^{n} s_i^2 n_i}{\sum_{i=1}^{n} (y_i - \bar{y})^2 n_i} + \frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2 . n_i}{\sum_{i=1}^{n} (y_i - \bar{y})^2 n_i}$$

kde druhý sčítanec je **pomer determinácie**

$$\eta_{yx}^2 = \frac{\sum_{i=1}^{m} (\bar{y}_i - \bar{y})^2 . n_i}{\sum_{j=1}^{l} (y_j - \bar{y})^2 n_j}$$

Pomer determinácie

nadobúda hodnoty z intervalu < 0;1 >

- interpretácia: koľko percent variability premennej Y je vysvetlených variabilitou premennej X

Korelačný pomer je druhou odmocninou pomeru determinácie

- interpretácia: čím má vyššiu hodnotu, tým je závislosť medzi premennými tesnejšia

Koeficient korelácie a koeficient determinácie

- meranie intenzity závislosti medzi X a Y, ak parametre regresnej funkcie sú odhadnuté metódou najmenších štvorcov

- výpočet vychádza z rozkladu rozptylu na zložky (2. rozklad variability):

celková variabilita = (teoretická) variabilita spôsobená regresiou + reziduálna variabilita

$$CSS = TSS + RSS$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$1 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} + \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Koeficient determinácie:

$$R^2 = \frac{\sum_{i=1}^{m}(\hat{y}_i - \bar{y})^2 n_i}{\sum_{i=1}^{m}(y_i - \bar{y})^2 n_i} = \frac{TSS}{CSS} = 1 - \frac{RSS}{CSS}$$

Koeficient korelácie:

$$R = \sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_i - \bar{y})^2 n_i}{\sum_{i=1}^{m}(y_i - \bar{y})^2 n_i}}$$

- hodnoty z intervalu < 0;1 >

Upravený koeficient determinácie

- použitie: porovnávanie viacerých zvolených regresných modelov

$$R^2_{adj} = 1 - \frac{n-1}{n-p}\left(1 - R^2\right)$$

- môže nadobudnúť aj záporné hodnoty, ale interpretuje sa len v intervale od 0 do 1

Párový koeficient korelácie

- meranie intenzity závislosti, ak je medzi X a Y lineárna závislosť

$$\rho_{yx} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{D[X]D[Y]}}, \quad \text{ak} \quad 0 < D[X] < \infty,\ 0 < D[Y] < \infty$$

Pearsonov koeficient korelácie:

$$r_{yx} = \frac{\mathrm{cov}(x,y)}{s_x s_y}$$

je špeciálnym prípadom koeficientu korelácie R

- je realizáciou odhadu párového koeficienta korelácie
- hodnoty nadobúda z intervalu <-1;1 >,
- záporná hodnota – nepriama závislosť, kladná hodnota – priama závislosť
- $r_{xy} = r_{yx}$; $r_{xx} = 1$; r_{xy} a koeficient b priamky majú rovnaké znamienko

Spearmanov koeficient poradovej korelácie

- zisťovanie závislosti medzi poradovými znakmi

$$r_{i_x i_y} = 1 - \frac{6\sum_{i=1}^{n}\left(i_x - i_y\right)^2}{n\left(n^2 - 1\right)}$$

- kde i_x a i_y sú poradia hodnôt v usporiadanom rade

Test hypotézy – párový korelačný koeficient

1^0 $H_0: \rho = \rho_0$ $H_1: \rho \neq \rho_0$

2^0 α

3^0 testovacia štatistika $$U = \frac{r_z - \rho_{0z}}{\sqrt{\frac{1}{n-3}}}$$

kde $$r_z = \frac{1}{2} \ln \frac{1+r}{1-r}$$ je Fisherova z-transformácia výberového korelačného koeficientu r a ρ_{0z} koeficientu ρ_0

- testovacia štatistika U má normované normálne rozdelenie

Interval spoľahlivosti pre korelačný koeficient

- získame z intervalu pre ρ_z

$$\left(r_z - u_{1-\alpha/2} \cdot \sqrt{\frac{1}{n-3}}; r_z + u_{1-\alpha/2} \cdot \sqrt{\frac{1}{n-3}} \right)$$

- vypočítané hranice intervalu je potrebné odtransformovať

Koeficient viacnásobnej korelácie

- meria intenzitu závislosti medzi závislou premennou Y a nezávislými premennými x_1, x_2,..., x_k
- hodnoty nadobúda z intervalu < 0;1 >

$$r_{y.x_1x_2\ldots} = \sqrt{\left(1 - \frac{|R|}{|R_{yy}|}\right)} = \sqrt{1 - \frac{RSS}{TSS}}$$

kde $|R|$ je determinant korelačnej matice

$$R = \begin{pmatrix} r_{yy} & r_{yx_1} & \cdots & r_{yx_k} \\ \vdots & \vdots & & \vdots \\ r_{x_k y} & r_{x_k x_1} & \cdots & r_{x_k x_k} \end{pmatrix}$$

- $|R_{yy}|$ je determinant submatice korelačnej matice (vznikne vynechaním riadku y a stĺpca y z matice R)

- **koeficient viacnásobnej determinácie** je

$$r^2_{y.x_1x_2...x_k}$$

- má hodnoty z intervalu < 0;1 >

- interpretácia:

aké percento variability závislej premennej je vysvetlené variabilitou všetkých nezávislých premenných