7

KORELACE

Pro vyjádření intenzity vztahů mezi složkami $\xi_1, ..., \xi_m$ m-rozměrného náhodného vektoru $\boldsymbol{\xi}$ se používá *korelačních koeficientů*. Data tvoří *náhodný výběr* z m-rozměrného rozdělení náhodného vektoru $\boldsymbol{\xi}$. Neuvažuje se obyčejně a priori, která složka ξ_j náhodného vektoru $\boldsymbol{\xi}$ je *vysvětlovaná* (u lineárního regresního modelu označovaná jako výstupní závisle proměnná) a které složky vektoru ξ jsou *vysvětlující* (u lineárního regresního modelu označované jako vstupní nezávisle proměnné). Náhodný výběr $\{x_{ij}\}$, $i = 1, ..., n$, $j = 1, ..., m$, velikosti n je tvořen ($n \times m$) rozměr-ným polem dat

$$\begin{bmatrix} x_{11} & . & . & . & x_{12} & . & . & . & . & x_{1m} \\ x_{21} & . & . & . & x_{22} & . & . & . & . & x_{2m} \\ . & & & & & & & & & \\ . & & & & & & & & & \\ . & & & & & & & & & \\ x_{n1} & . & . & . & x_{n2} & . & . & . & . & x_{nm} \end{bmatrix} .$$

Platí, že

a) počet řádků n (tj. počet m-rozměrných "bodů" $\boldsymbol{x}_i$) je výrazně větší, než počet sloupců m (tj. počet "proměnných" čili složek vektoru $\boldsymbol{x}$).

b) Všechny složky vektoru $\boldsymbol{x}_i$ jsou *náhodné* a předem *neovlivnitelné* experimen-tátorem.

c) Mezi složkami jsou pouze lineární vazby.

7.1 Druhy korelačních koeficientů

7.1.1 Párový korelační koeficient

Korelační koeficienty slouží jako míry pro vyjádření "těsnosti lineární stochastické vazby" mezi složkami náhodného vektoru **>**. *Pearsonův párový korelační koeficient* $\rho(\xi_i, \xi_j) = r_{ij}$ vyjadřuje míru lineární stochastické vazby mezi náhodnou veličinou ξ_i a ξ_j. Označme *populační párový korelační koeficient* ρ a *výběrový párový korelační koeficient* r. Nahradíme

střední hodnoty μ_1 a μ_2 aritmetickými průměry $\bar{x}_1$ a $\bar{x}_2$, dále rozptyly σ_1^2 a σ_2^2 výběrovými rozptyly s_1^2 a s_2^2. Pro *výběrový korelační koeficient* platí výraz

$$r = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2}} .$$

K interpretaci korelačních koeficientů je třeba přistupovat velmi obezřetně. Platí pravidlo, že *významná párová korelace není důkazem příčinné souvislosti*. Někdy vznikají falešné korelace, kdy jak ξ_1, tak i ξ_2 silně korelují s neuvažovanou náhodnou veličinou ξ_3 a vysoká hodnota $\rho(\xi_1, \xi_2)$ je důsledek vysokých hodnot $\rho(\xi_1, \xi_3)$ a $\rho(\xi_2, \xi_3)$. Při interpretaci korelačních koeficientů je pak vhodné užít i parciální korelační koeficienty.

Při konstrukci testů významnosti se využívá testační statistiky

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

která má pro případ $\rho = 0$ Studentovo rozdělení s $(n - 2)$ stupni volnosti. Toho lze využít k testování nekorelovanosti, resp. lineární nezávislosti dvojice náhodných veličin. Je-li jejich rozdělení dvourozměrné normální, je nekorelovanost totožná s nezávislostí. Testuje se hypotéza H_0: $\rho = 0$ proti různým alternativám H_A. Vyjde-li $|t|$ větší než odpovídající kvantil Studentova rozdělení, zamítá se H_0 a náhodné veličiny nejsou nekorelované. Uvedený test je silně nerobustní a platí pouze v případě dvourozměrné normality ξ_1, ξ_2. Pro urychlení konvergence $f(r)$ k normálnímu rozdělení se používá různých transformací. Jednoduchá *Rubenova transformace* má tvar

$$R(r) = \frac{\sqrt{n - 2.5}\, r}{\sqrt{1 - 0.5\, r^2}} .$$

Náhodná veličina $R(r)$ již má i pro menší výběry normované normální rozdělení N(0, 1).

7.1.2 Parciální korelační koeficient

V řadě případů je účelné sledovat vztah mezi dvěma složkami ξ_i a ξ_j náhodného vektoru při zkonstantnění dalších složek vektoru $\boldsymbol{\xi}$. Pro vyjádření intenzity tohoto vztahu se používají parciální korelační koeficienty různých řádů. Nejjednodušší jsou parciální korelační koeficienty nultého řádu, které odpovídají párovým korelačním koeficientům.

Parciální korelační koeficienty prvního řádu $r_{1,3(2)}$ odpovídají párovému korelačnímu

koeficientu mezi rezidui $\varepsilon_2 = \xi_1 - E(\xi_1/x_2)$

a rezidui $\eta_2 = \xi_3 - E(\xi_3/x_2)$

a mají tvar $r_{1,3(2)} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$.

Analogicky lze definovat i další parciální korelační koeficienty $r_{1i(j)}$ prvního řádu jako párové korelační koeficienty mezi rezidui $g_j = >_1 - E(>_1/x_j)$

a rezidui $i_j = >_i - E(>_i/x_j)$,

pro které platí tvar $r_{1,i(j)} = \frac{r_{1i} - r_{1j} r_{ij}}{\sqrt{(1 - r_{1i}^2)(1 - r_{ij}^2)}}$.

Parciální korelační koeficienty druhého řádu $r_{1i(j,k)}$ jsou vlastně párové korelační koeficienty reziduí $g_{j,k} = >_1 - E(>_1/(x_j, x_k))$

a reziduí $i_{j,k} = >_i - E(>_i/(x_j, x_k))$

a mají tvar $r_{1i(j,k)} = \frac{r_{1i(j)} - r_{1j(k)} r_{ij(k)}}{\sqrt{(1 - r_{1j(k)}^2)(1 - r_{ij(k)}^2)}}$.

Parciální korelační koeficient (m - 1). řádu $r_{1i(2, 3,\ldots,m)}$ odpovídá jednoduchému korelačnímu koeficientu mezi rezidui $g_{2,\ldots,m} = >_1 - E(>_1/\boldsymbol{x}^*)$

a rezidui $i_{2,\ldots,m} = >_i - E(>_i/\boldsymbol{x}^*)$,

kde vektor $\boldsymbol{x}^*$ obsahuje složky $x_2, x_3, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m$.

Obecně se počítají *parciální korelační koeficienty vyšších řádů* podle rekurentní formule

$$r_{1,j(2,3,\ldots,j-1)} = \frac{A - B C}{\sqrt{(1 - B^2)(1 - C^2)}} ,$$

kde $A = r_{1,j(2,3,\ldots,j-2)}$ $B = r_{1,j-1(2,3,\ldots,j-2)}$ $C = r_{j,j-1(2,3,\ldots,j-2)}$.

Pro statistické testování a konstrukci intervalů spolehlivosti se využívá pravidlo, že rozdělení parciálního korelačního koeficientu řádu (m - 1) je stejné jako rozdělení párového korelačního koeficientu pro rozsah výběru (n - m + 1).

7.1.3 Vícenásobný korelační koeficient

Vícenásobný korelační koeficient $R_{1(2,\ldots,m)}$ definuje *míru lineární stochastické závislosti* mezi náhodnou veličinou ξ_1 a nejlepší lineární kombinací složek $\xi_2, \ldots, \xi_m$ náhodného vektoru. Pro tento korelační koeficient platí, že

$$R_{1(2,\ldots,m)} = \sqrt{1 - \frac{\det(\boldsymbol{R})}{\det(\boldsymbol{R}_{11})}},$$

kde det(.) označuje determinant a $\boldsymbol{R}_{ij}$ je matice vzniklá vypuštěním *i*-tého řádku a *j*-tého sloupce korelační matice $\boldsymbol{R}$.

Mezi základní vlastnosti vícenásobného korelačního koeficientu patří:

1. Platí nerovnost $0 \leq R_{1(2,\ldots,m)} \leq 1$.

2. Pokud je $R_{1(2,\ldots,m)} = 1$, znamená to, že náhodná veličina ξ_1 je přesně lineární kombinací veličin $\xi_2, \ldots, \xi_m$.

3. Pokud je $R_{1(2,\ldots,m)} = 0$, jsou také všechny odpovídající párové korelační koeficienty rovny nule $\rho(\xi_1, \xi_j) = 0, j = 2, \ldots, m$.

4. Pro případ jedné vysvětlující proměnné je $R_{1(2)} = |\rho(\xi_1, \xi_2)|$, tj. vícenásobný korelační koeficient je totožný s absolutní hodnotou párového korelačního koeficientu.

5. Platí, že s růstem počtu vysvětlujících proměnných vícenásobný korelační koeficient nikdy neklesá

$$R^2_{1(2)} \leq R^2_{1(2,3)} \leq R^2_{1(2,3,4)} \leq \ldots \leq R^2_{1(2,\ldots,m)} .$$

Při znalosti jednotlivých parciálních korelačních koeficientů všech řádů je možné vyčíslit také vícenásobný korelační koeficient ze vztahu

$$R^2_{1(2,\ldots,m)} = 1 - (1 - R^2_{1,2})(1 - R^2_{1,3(2)})(1 - R^2_{1,4(2,3)}) \ldots$$

$$\ldots (1 - R^2_{1,m(2,3,\ldots,m-1)}).$$

Pro výpočet parciálních korelačních koeficientů je výhodné využít vztah

$$R_{1i(2,3,\ldots,m)} = \frac{(-1)^i \det(\boldsymbol{R}_{1,i})}{\sqrt{\det(\boldsymbol{R}_{11})\ \det(\boldsymbol{R}_{i,i})}} ,$$

kde $\boldsymbol{R}$ je korelační matice odpovídající vektoru $\boldsymbol{x}$ a $\boldsymbol{R}_{i,j}$ je matice vzniklá vynecháním *i*-tého řádku a *j*-tého sloupce matice $\boldsymbol{R}$.

7.2 Pořadový korelační koeficient

V některých případech je výhodné nahradit klasický párový korelační koeficient pořadovým (neparametrickým) korelačním koeficientem podle Spearmana, který je málo citlivý na přítomnost vybočujících hodnot. Pořadí *i*-tého prvku výběru je rovno indexu odpovídající pořádkové statistiky. Označme pořadí prvků výběru vzhledem k proměnné ξ_1 jako x_{1si} a pořadí prvků výběru vzhledem k proměnné ξ_2 jako x_{2si}.

Pro *Spearmanův pořadový korelační koeficient* pak platí

$$D_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (x_{1si} - x_{2si})^2.$$

Rozdělení veličiny D_s je symetrické se střední hodnotou $E(D_s) = 0$ a rozptylem $D(D_s) = 1/(n - 1)$. Pro $n > 10$ se často využívá toho, že veličina

$$t_s = \frac{|D_s| \sqrt{n-2}}{\sqrt{1-D_s^2}}$$

má asymptoticky Studentovo rozdělení s (n - 2) stupni volnosti, pokud teoretický koeficient $\rho_s = 0$.

V praxi se stává, že pro několik prvků výběru vychází stejné pořadí. Pak se všem přiřadí průměr z pořadí, které by měly, pokud by nabývaly různých hodnot, a *Spearmanův korelační koeficient* se počítá dle upravené formule

$$D_s = \frac{\frac{n(n^2-1)}{6} - \sum_{i=1}^{n} (x_{1si} - x_{2si})^2 - a - b}{\sqrt{\left(\frac{n(n^2-1)}{6} - 2a\right)\left(\frac{n(n^2-1)}{6} - 2b\right)}},$$

kde a, b jsou opravné koeficienty na pořadí

$$a = \frac{1}{12} \sum_{(j)} (a_j^3 - a_j),$$

$$b = \sum_{(k)} (b_k^3 - b_k),$$

kde j označují čísla shluků stejných pořadí pro x_1 a a_j je počet hodnot se stejným pořadím v j-tém shluku. Analogicky je definováno také k a b_k.

Spearmanův pořadový korelační koeficient ρ_s leží v intervalu $-1 \leq \rho_s \leq 1$. Pokud výběr pochází z dvourozměrného normálního rozdělení a $n \geq 30$, platí vztah, že

$$D(\varphi_1, \varphi_2) = 2 \sin\left(\frac{\pi}{6} D_s\right).$$

Při použití pořadových korelačních koeficientů je třeba mít stále na paměti, že při přechodu z dat x_{1i}, x_{2i} na pořadí x_{1si}, x_{2si} dochází vždy ke ztrátě informace. Na druhé straně je však docíleno zrobustnění a snížení citlivosti na odchylky od normality.

7.3 Cronbachův korelační koeficient γ spolehlivosti výsledku

Spolehlivost výsledku, měření může být rozdělena na dvě kategorie: správnost a přesnost (viz 1. kapitola). *Správnost* se týká důkazu, zda naměřená hodnota je správná. *Přesnost* se týká důkazu, zda naměřené hodnoty jsou stejné při svém opakovaní. Přístroj může být správný při měření jedné veličiny, ale nemusí být správný při měření jiné. Bylo navrženo několik metod na prokázání spolehlivosti přístroje. Zaměříme se nyní na ověření *vnitřní jednotnosti výsledku (konzistentnosti)*.

Cronbachův korelační koeficient γ: představuje nejrozšířenější kritérium posou-zení vnitřní jednotnosti výsledku a vypočte se dle vzorce

$$\gamma = \frac{m}{m-1}\left[1 - \frac{\sum_{i=1}^{m}\sigma_{ii}}{\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}}\right],$$

kde m je počet proměnných a σ_{ij} je vypočtená kovariance mezi proměnnou i a j, σ_{ii} je rozptyl proměnné i. Jsou-li data předem standardizována (odečtením průměru a podělením směrodatnou odchylkou položky), dostaneme standardizovanou verzi Cronbachova koeficientu

$$\gamma = \frac{m\,\bar{r}}{1 + \bar{r}\,(m-1)},$$

kde $\bar{r}$ je průměr všech korelačních koeficientů mezi všemi m proměnnými.

Cronbachův koeficient γ má několik interpretací: rovná se průměru všech Cronbachových koeficientů, získaných pro všechny možné kombinace rozdělením $2m$ proměnných do dvou skupin, každé o m proměnných, a vypočtením dvou polovičních testů. Dále odhaduje očekávanou korelaci jednoho přístroje s alternativní formou jiného, obsahujícího stejný počet měřených proměnných. Může odhadovat také očekávanou korelaci mezi aktuálním testem a hypotetickým testem, který nikdy nebyl popsán. Protože jde o korelační koeficient, je Cronbachův koeficient γ definován v intervalu -1 až +1. Ve většině případů jde o kladné číslo. Existuje pravidlo, že γ by mělo pro většinu přístrojů dosáhnout hodnoty alespoň 0.8. Koeficient γ lze zlepšit či zvýšit zvětšením počtu měření nebo zvýšením průměrné korelace mezi proměnnými.

Postup analýzy korelace

1. Návrh modelu: zařadíme obvykle i absolutní člen β_0 a nejprve budeme uvažovat lineární regresní model ve tvaru $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$. Polohu a proměnlivost proměnných y, x_1, x_2, x_3 přináší *průměr* a *směrodatná odchylka* hodnot každé proměnné. Zatímco *Pearsonův vícenásobný korelační koeficient r* ukazuje, do jaké míry je navržený lineární regresní model statisticky významný, hodnota *koeficientu determinace* $D = r^2$ vyjadřuje kolik procent bodů dobře koresponduje s modelem. *Predikovaný koeficient determinace* D_p má podobný význam jako *koeficient determinace D*, je však vyčíslen jinak, místo sumy čtverců odchylek *RSC* se ve vztahu užije střední kvadratická chyba predikce *MEP*.

2. Korelační matice Pearsonovy a Spearmanovy pořadové: výpočet umožňuje likvidaci děravých cel párovým nebo řádkovým způsobem. Korelace jsou však silně ovlivněny odlehlými hodnotami, heteroskedasticitou, nenormalitou rozdělení a nelinearitami. Vhodným doplňkem Pearsonova korelačního koeficientu je Spear-manův pořadový korelační koeficient. Pořadová korelace se vyčíslí Pearsonovým korelačním vzorcem, aplikovaným na pořadové číslo dat ne na numerické hodnoty dat samotných. V případě odlehlých hodnot se bude velice lišit parametrická a neparametrická míra korelace, tj. Pearsonův korelační koeficient a Spearmanův pořadový korelační koeficient. V případě kolinearity jsou vysoké hodnoty párových korelací první indikací kolinearity.

3. Matice rozdílů: Aby se umožnilo porovnat tyto dva typy korelačních matic, vypočte se také matice rozdílů. Tím se ukáže, která dvojice proměnných si žádá hlubšího vyšetření.