2.6 Teoretické základy korelace

Nechť X a Y jsou náhodné veličiny s konečnými druhými momenty a s kladnými rozptyly. Závislost těchto veličin na sobě se často měří pomocí *korelačního koeficientu*

$$\rho = \frac{\mathsf{cov}(X,Y)}{\sqrt{(\mathsf{var}\,X)(\mathsf{var}\,Y)}}.$$

Někdy místo ρ píšeme $\rho_{X,Y}$, abychom vyznačili, o které veličiny se jedná. Je zřejmé, že $\rho_{X,Y} = \rho_{Y,X}$. Nechť dále a, b, c, d jsou taková reálná čísla, že $ac \neq 0$. Snadno se dá dokázat, že platí

$$\rho_{aX+b,cY+d} = \begin{cases} \rho_{X,Y} & \text{pro } ac > 0, \\ -\rho_{X,Y} & \text{pro } ac < 0. \end{cases}$$

Při lineární transformaci se tedy korelační koeficient buď nezmění vůbec, nebo pouze změní znaménko.

Věta 2.16 *Pro korelační koeficient platí $-1 \le \rho_{X,Y} \le 1$. Rovnost $\rho_{X,Y} = 1$ platí právě tehdy, je-li $Y = a + bX$ s pravděpodobností 1, přičemž $b > 0$. Analogicky rovnost $\rho_{X,Y} = -1$ platí právě tehdy, je-li $Y = a + bX$ s pravděpodobností 1, přičemž $b < 0$.*

Důkaz. Schwarzova nerovnost

$$|\mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y)| \le \sqrt{\mathsf{E}(X - \mathsf{E}X)^2\mathsf{E}(Y - \mathsf{E}Y)^2}$$

má v našem označení podobu $|\mathsf{cov}(X,Y)| \le \sqrt{(\mathsf{var}\,X)(\mathsf{var}\,Y)}$. Z toho plyne, že $-1 \le \rho_{X,Y} \le 1$. Rovnosti ve Schwarzově nerovnosti je dosahováno právě tehdy, platí-li buď $X - \mathsf{E}X = 0$ nebo $Y - \mathsf{E}Y = 0$ skoro jistě (což v našem případě nepřichází v úvahu vzhledem k předpokladu $\mathsf{var}\,X > 0$, $\mathsf{var}\,Y > 0$), nebo když platí $Y - \mathsf{E}Y = b(X - \mathsf{E}X)$ skoro jistě pro nějaké $b \neq 0$. Výpočtem se ověří, že v tomto posledním případě je $\rho_{X,Y} = 1$ pro $b > 0$ a $\rho_{X,Y} = -1$ pro $b < 0$. □

Při práci s korelačním koeficientem bývá užitečná následující interpretace. Uvažujme prostor náhodných veličin s konečnými druhými momenty, přičemž tyto veličiny jsou definovány na stejném pravděpodobnostním prostoru. Vytvořme třídy ekvivalence těchto veličin tak, že X a Y prohlásíme za ekvivalentní, když si jsou veličiny $X - \mathsf{E}X$ a $Y - \mathsf{E}Y$ rovny skoro všude. Definujme na třídách ekvivalence skalární součin předpisem $(X,Y) = \mathsf{cov}(X,Y)$ pro libovolné reprezentanty těchto tříd, jak je obvyklé ve funkcionální analýze. Prostor těchto tříd je Hilbertův prostor a označíme ho $\mathcal{H}$. Norma veličiny X je $\|X\| = \sqrt{(X,X)}$. V tomto označení lze korelační koeficient ρ veličin $X \neq 0$ a $Y \neq 0$ vyjádřit ve tvaru

$$\rho_{X,Y} = \frac{(X,Y)}{\|X\|\,.\,\|Y\|}.$$

Na základě analogie se vzorcem známým z geometrie můžeme tedy korelační koeficient interpretovat jako kosinus úhlu mezi veličinami X a Y, nebo názorněji jako kosinus úhlu mezi veličinami $X - \mathsf{E}X$ a $Y - \mathsf{E}Y$.

Mějme náhodné vektory $\boldsymbol{X} = (X_1, \dots, X_n)'$ a $\boldsymbol{Y} = (Y_1, \dots, Y_m)'$ s konečnými druhými momenty. Nechť všechny složky těchto dvou vektorů mají kladné rozptyly. Matici typu $n \times m$, jejíž (i,j)-tý prvek je roven korelačnímu koeficientu veličin X_i a Y_j, nazveme *korelační maticí vektorů* $\boldsymbol{X}$ a $\boldsymbol{Y}$ a označíme ji $\mathsf{cor}(\boldsymbol{X}, \boldsymbol{Y})$. Matice $\mathsf{cor}(\boldsymbol{X}, \boldsymbol{X})$ se nazývá *korelační matice vektoru* $\boldsymbol{X}$ a značí se $\mathsf{cor}\,\boldsymbol{X}$.

Lemma 2.17 *Nechť $\boldsymbol{X} = (X_1, \dots, X_n)'$ je náhodný vektor s konečnými druhými momenty, jehož všechny složky mají kladné rozptyly. Označme $\boldsymbol{V} = \mathsf{var}\,\boldsymbol{X}$, $\boldsymbol{P} = \mathsf{cor}\,\boldsymbol{X}$, $\sigma_i = \sqrt{\mathsf{var}\,X_i}$ pro $i = 1, \dots, n$ a*

$$\boldsymbol{D} = \mathsf{Diag}\{\sigma_1, \dots, \sigma_n\} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n \end{pmatrix}.$$

Pak $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}$.

Důkaz je zřejmý. □

Lemma 2.18 *Budiž $\boldsymbol{V}$ symetrická pozitivně definitní matice typu $n \times n$. Pak pro libovolné n-rozměrné vektory $\boldsymbol{b}$ a $\boldsymbol{c}$ platí $(\boldsymbol{b}'\boldsymbol{V}\boldsymbol{c})^2 \le (\boldsymbol{b}'\boldsymbol{V}\boldsymbol{b})(\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c})$.*

Důkaz. Je známo, že za uvedených předpokladů existuje taková symetrická pozitivně semidefinitní matice $\boldsymbol{V}^{\frac{1}{2}}$, pro kterou platí $\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{V}^{\frac{1}{2}} = \boldsymbol{V}$. Tato matice $\boldsymbol{V}^{\frac{1}{2}}$ se nazývá *odmocninová*. Ze Schwarzovy nerovnosti plyne

$$\begin{aligned}(\boldsymbol{b}'\boldsymbol{V}\boldsymbol{c})^2 &= (\boldsymbol{b}'\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{c})^2 = [(\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{b})'(\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{c})]^2 \\ &\le [(\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{b})'(\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{b})][(\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{c})'(\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{c})] = (\boldsymbol{b}'\boldsymbol{V}\boldsymbol{b})(\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}). \qquad \square\end{aligned}$$

Mějme náhodnou veličinu Y a náhodný vektor $\boldsymbol{X} = (X_1, \dots, X_n)'$ s konečnými druhými momenty podobně jako v případě lineární regrese. Nechť $\mathsf{var}\,Y > 0$ a nechť $\boldsymbol{V} = \mathsf{var}\,\boldsymbol{X}$ je regulární matice. Závislost mezi Y a celým vektorem $\boldsymbol{X}$ měříme pomocí *koeficientu mnohonásobné korelace* $\rho_{Y,\boldsymbol{X}}$, což je korelační koeficient mezi veličinou Y a její nejlepší lineární aproximací $\hat{Y} = \alpha + \boldsymbol{\beta}'\boldsymbol{X}$, kde α a $\boldsymbol{\beta}$ jsou uvedeny ve větě 2.15. Je-li $\boldsymbol{\beta} = \boldsymbol{0}$, definuje se $\rho_{Y,\boldsymbol{X}} = 0$. Někdy se místo $\rho_{Y,\boldsymbol{X}}$ píše podrobněji $\rho_{Y.X_1,\dots,X_n}$. Protože platí $\rho_{Y,\boldsymbol{X}} = \rho_{Y,\alpha+\boldsymbol{\beta}'\boldsymbol{X}} = \rho_{Y,\boldsymbol{\beta}'\boldsymbol{X}}$ a

$$\mathsf{cov}(Y, \boldsymbol{\beta}'\boldsymbol{X}) = \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{\beta} = \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y) \ge 0,$$

je koeficient mnohonásobné korelace vždy nezáporný.

Věta 2.19 *Označme $\boldsymbol{P} = \mathsf{cor}\,\boldsymbol{X}$. Pak platí*

$$\rho^2_{Y,\boldsymbol{X}} = \mathsf{cor}(Y, \boldsymbol{X})\boldsymbol{P}^{-1}\mathsf{cor}(\boldsymbol{X}, Y). \tag{2.8}$$

Důkaz. Pro $\boldsymbol{\beta} = 0$ je tvrzení zřejmé. Nechť tedy $\boldsymbol{\beta} \neq \mathbf{0}$. Označme $\boldsymbol{V} = \mathsf{var}\,\boldsymbol{X}$. Máme $\rho_{Y,\alpha+\boldsymbol{\beta}'\boldsymbol{X}} = \rho_{Y,\boldsymbol{\beta}'\boldsymbol{X}}$. Protože $\mathsf{cov}(Y, \boldsymbol{\beta}'\boldsymbol{X}) = \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{\beta}$ a $\mathsf{var}\,\boldsymbol{\beta}'\boldsymbol{X} = \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}$, je

$$\rho^2_{Y,\boldsymbol{X}} = \frac{[\mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{\beta}]^2}{\mathsf{var}\,Y \,.\, \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}}. \tag{2.9}$$

Dosadíme-li $\mathsf{cov}(Y, \boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{V}$, dostaneme odtud

$$\rho^2_{Y,\boldsymbol{X}} = \frac{\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}}{\mathsf{var}\,Y}. \tag{2.10}$$

Užitím lemmatu 2.17 dostaneme

$$\boldsymbol{\beta} = \boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y) = \boldsymbol{D}^{-1}\boldsymbol{P}^{-1}\boldsymbol{D}^{-1}\mathsf{cov}(\boldsymbol{X}, Y) = \boldsymbol{D}^{-1}\boldsymbol{P}^{-1}\mathsf{cor}(\boldsymbol{X}, Y)\sigma_Y,$$

a proto

$$\begin{aligned} \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta} &= \sigma^2_Y\,\mathsf{cor}(Y, \boldsymbol{X})\boldsymbol{P}^{-1}\boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{P}^{-1}\mathsf{cor}(\boldsymbol{X}, Y) \\ &= \sigma^2_Y\,\mathsf{cor}(Y, \boldsymbol{X})\boldsymbol{P}^{-1}\mathsf{cor}(\boldsymbol{X}, Y). \end{aligned}$$

Dosazením do (2.9) se dostane tvrzení věty. □

Koeficient mnohonásobné korelace těsně souvisí s *reziduálním rozptylem* $\sigma^2_{Y.\boldsymbol{X}}$. Ze vzorců (2.7) a (2.10) dostáváme

$$\sigma^2_{Y.\boldsymbol{X}} = \sigma^2_Y(1 - \rho^2_{Y,\boldsymbol{X}}), \qquad \rho^2_{Y,\boldsymbol{X}} = 1 - \frac{\sigma^2_{Y.\boldsymbol{X}}}{\sigma^2_Y}.$$

Často se počítá koeficient mnohonásobné korelace mezi Y a vektorem $\boldsymbol{X} = (X_1, X_2)'$, který má jen dvě složky. Položme $X_0 = Y$ a označme ρ_{ij} korelační koeficient mezi X_i a X_j $(i, j = 0, 1, 2)$. Dále označme $\rho_{0.12} = \rho_{Y.X_1,X_2}$. V tomto případě z věty 2.19 plyne

$$\rho^2_{0.12} = \frac{\rho^2_{01} + \rho^2_{02} - 2\rho_{01}\rho_{02}\rho_{12}}{1 - \rho^2_{12}}.$$

Korelační matice $\boldsymbol{P}$ vektoru $(X_0, X_1, X_2)'$ je pozitivně semidefinitní, a proto její determinant je nezáporný. Odtud dostáváme

$$|\boldsymbol{P}| = \begin{vmatrix} 1 & \rho_{01} & \rho_{02} \\ \rho_{01} & 1 & \rho_{12} \\ \rho_{02} & \rho_{12} & 1 \end{vmatrix} = 1 + 2\rho_{01}\rho_{02}\rho_{12} - \rho^2_{01} - \rho^2_{02} - \rho^2_{12} \geq 0.$$

Předpokládejme nyní, že $\rho_{01} = 0$. Pak $|\boldsymbol{P}| = 1 - \rho^2_{02} - \rho^2_{12} \geq 0$, takže $\rho^2_{02}/(1 - \rho^2_{12}) \leq 1$. Vidíme, že v případě $\rho_{01} = 0$ může být $\rho^2_{0.12} = \rho^2_{02}/(1 - \rho^2_{12})$ blízké 1, pokud je $|\rho_{02}|$ blízké 1.

Další zdůvodnění koeficientu mnohonásobné korelace jako míry závislosti přináší následující tvrzení.

Věta 2.20 *Koeficient mnohonásobné korelace $\rho_{Y,\boldsymbol{X}}$ je největší korelační koeficient mezi náhodnou veličinou Y a libovolnou lineární funkcí vektoru $\boldsymbol{X}$ v tom smyslu, že*

$$\rho_{Y,\boldsymbol{X}} = \max_{\substack{c\in\mathbb{R}\\ \boldsymbol{0}\neq \boldsymbol{d}\in\mathbb{R}_n}} |\rho_{Y,c+\boldsymbol{d}'\boldsymbol{X}}|.$$

Důkaz. Nechť α a $\boldsymbol{\beta}$ jsou koeficienty optimální lineární aproximace $\hat{Y}$ veličiny Y uvedené ve větě 2.15. V případě $\boldsymbol{\beta} = \boldsymbol{0}$ je tvrzení zřejmé. Pak totiž $\mathsf{cov}(Y, \boldsymbol{X}) = \boldsymbol{0}$, takže pro libovolné c a $\boldsymbol{d}$ platí $\mathsf{cov}(Y, c + \boldsymbol{d}'\boldsymbol{X}) = \boldsymbol{0}$. Je-li $\boldsymbol{\beta} \neq \boldsymbol{0}$, je třeba ukázat, že $\rho^2_{Y,\alpha+\boldsymbol{\beta}'\boldsymbol{X}} \geq \rho^2_{Y,c+\boldsymbol{d}'\boldsymbol{X}}$. Protože čísla α a c žádný z korelačních koeficientů neovlivňují, jde o ověření nerovnosti $\rho^2_{Y,\boldsymbol{\beta}'\boldsymbol{X}} \geq \rho^2_{Y,\boldsymbol{d}'\boldsymbol{X}}$. Pro libovolný vektor $\boldsymbol{d} \in \mathbb{R}_n$ platí

$$\mathsf{cov}(Y, \boldsymbol{d}'\boldsymbol{X}) = \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{d} = \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{d} = \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{d}.$$

Z lemmatu 2.18 pro $\boldsymbol{d} \neq \boldsymbol{0}$ tudíž dostáváme

$$\rho^2_{Y,\boldsymbol{d}'\boldsymbol{X}} = \frac{[\mathsf{cov}(Y, \boldsymbol{d}'\boldsymbol{X})]^2}{\mathsf{var}\,Y \,.\, \mathsf{var}\,\boldsymbol{d}'\boldsymbol{X}} = \frac{(\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{d})^2}{\mathsf{var}\,Y \,.\, \boldsymbol{d}'\boldsymbol{V}\boldsymbol{d}} \leq \frac{(\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta})(\boldsymbol{d}'\boldsymbol{V}\boldsymbol{d})}{\mathsf{var}\,Y \,.\, \boldsymbol{d}'\boldsymbol{V}\boldsymbol{d}}$$

$$= \frac{(\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta})^2}{\mathsf{var}\,Y \,.\, \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}} = \frac{[\mathsf{cov}(Y, \boldsymbol{\beta}'\boldsymbol{X})]^2}{\mathsf{var}\,Y \,.\, \mathsf{var}\,\boldsymbol{\beta}'\boldsymbol{X}} = \rho^2_{Y,\boldsymbol{\beta}'\boldsymbol{X}}. \qquad \square$$

Z věty 2.20 např. plyne, že koeficient mnohonásobné korelace $\rho_{Y.X_1,\ldots,X_n}$ není nikdy menší než absolutní hodnota kteréhokoli obyčejného korelačního koeficientu ρ_{Y,X_i}, $i = 1, \ldots, n$. To bývá užitečná kontrola při výpočtech.

Vlastnost koeficientu mnohonásobné korelace vyjádřená větou 2.20 odpovídá následujícímu obecnějšímu tvrzení.

Věta 2.21 *Budiž $\mathcal{H}$ Hilbertův prostor a $\mathcal{H}_1$ jeho podprostor. Nechť $x \in \mathcal{H}$ a nechť y je projekce prvku x na $\mathcal{H}_1$. Je-li $y \neq 0$, pak $x \neq 0$ a pro každý nenulový prvek $z \in \mathcal{H}_1$ platí*

$$\frac{(x, y)}{\|x\| \cdot \|y\|} \geq \frac{|(x, z)|}{\|x\| \cdot \|z\|}. \tag{2.11}$$

Důkaz. Implikace $[y \neq 0] \Rightarrow [x \neq 0]$ je zřejmá. Nerovnost (2.11) je ekvivalentní s

$$(x, y)\|z\| \geq |(x, z)| \cdot \|y\|. \tag{2.12}$$

Jelikož $(x, z) = \big((x - y) + y, z\big) = (x - y, z) + (y, z) = (y, z)$, plyne z toho speciálně $(x, y) = (y, y)$. Proto (2.12) je ekvivalentní s nerovností

$$(y, y)\|z\| \geq |(x, z)| \cdot \|y\|,$$

tj. s $\|y\|^2\|z\| \geq |(x, z)| \cdot \|y\|$. Podle předpokladu $\|y\| \neq 0$, a tak je náš vztah ekvivalentní se Schwarzovou nerovností $|(x, z)| \leq \|y\| \cdot \|z\|$. $\square$