

## regrese

konečnými momenty. Hledejme  $X_1, \dots, X_n$ . To znamená,

$X_1, \dots, X_n$

pomocí střední čtvercové

distí

$\text{cov}(X, Y)$

$E(X)$

pak pro ni platí  $EZ^2 =$   
odtud

$(X)$

$-\beta'X) = 0$ , tj. pro  $\alpha =$

$(Y, X)\beta + \beta'V\beta$

$-V^{-1}\text{cov}(X, Y)$

$(X, Y)$

$(X, Y)$

$(Y)$ .  $\square$

$+\beta'X$ , kde  $\alpha$  a  $\beta$  jsou

$\text{cov}(X, Y)$

celkový rozptyl získá vzorec

(2.7)

## 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*

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

Někdy místo  $\rho$  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  $-1 \leq \rho_{X,Y} \leq 1$ . 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

$$|\text{E}(X - EX)(Y - EY)| \leq \sqrt{\text{E}(X - EX)^2 \text{E}(Y - EY)^2}$$

má v našem označení podobu  $|\text{cov}(X, Y)| \leq \sqrt{(\text{var } X)(\text{var } Y)}$ . Z toho plyne, že  $-1 \leq \rho_{X,Y} \leq 1$ . Rovnosti ve Schwarzově nerovnosti je dosahováno právě tehdy, platí-li buď  $X - EX = 0$  nebo  $Y - EY = 0$  skoro jistě (což v našem případě nepřichází v úvahu vzhledem k předpokladu  $\text{var } X > 0$ ,  $\text{var } Y > 0$ ), nebo když platí  $Y - EY = b(X - EX)$  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  $b < 0$ .  $\square$

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  $X - EX$  a  $Y - EY$  rovny skoro všude. Definujme na třídách ekvivalence skalární součin předpisem  $(X, Y) = \text{cov}(X, Y)$  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  $\rho$  veličin  $X \neq 0$  a  $Y \neq 0$  vyjádřit ve tvaru

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

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  $X - EX$  a  $Y - EY$ .

Mějme náhodné vektory  $\mathbf{X} = (X_1, \dots, X_n)'$  a  $\mathbf{Y} = (Y_1, \dots, Y_m)'$  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  $\mathbf{X}$  a  $\mathbf{Y}$*  a označíme ji  $\text{cor}(\mathbf{X}, \mathbf{Y})$ . Matice  $\text{cor}(\mathbf{X}, \mathbf{X})$  se nazývá *korelační matice vektoru  $\mathbf{X}$*  a značí se  $\text{cor} \mathbf{X}$ .

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

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

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

Důkaz je zřejmý.  $\square$

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

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

$$\begin{aligned} (\mathbf{b}' \mathbf{V} \mathbf{c})^2 &= (\mathbf{b}' \mathbf{V}^{\frac{1}{2}} \mathbf{V}^{\frac{1}{2}} \mathbf{c})^2 = [(\mathbf{V}^{\frac{1}{2}} \mathbf{b})' (\mathbf{V}^{\frac{1}{2}} \mathbf{c})]^2 \\ &\leq [(\mathbf{V}^{\frac{1}{2}} \mathbf{b})' (\mathbf{V}^{\frac{1}{2}} \mathbf{b})][(\mathbf{V}^{\frac{1}{2}} \mathbf{c})' (\mathbf{V}^{\frac{1}{2}} \mathbf{c})] = (\mathbf{b}' \mathbf{V} \mathbf{b})(\mathbf{c}' \mathbf{V} \mathbf{c}). \end{aligned} \quad \square$$

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

$$\text{cov}(Y, \beta' \mathbf{X}) = \text{cov}(Y, \mathbf{X})\beta = \text{cov}(Y, \mathbf{X})\mathbf{V}^{-1} \text{cov}(\mathbf{X}, Y) \geq 0,$$

je koeficient mnohonásobné korelace vždy nezáporný.

**Věta 2.19** *Označme  $\mathbf{P} = \text{cor } \mathbf{X}$ . Pak platí*

$$\rho_{Y, \mathbf{X}}^2 = \text{cor}(Y, \mathbf{X})\mathbf{P}^{-1} \text{cor}(\mathbf{X}, Y). \quad (2.8)$$

*Důkaz.* Pro  $\beta = \mathbf{0}$  je tvrzení  $\rho_{Y, \alpha + \beta' \mathbf{X}} = \rho_{Y, \beta' \mathbf{X}}$ . Protože

Dosadíme-li  $\text{cov}(Y, \mathbf{X}) = \beta'$

Užitím lemmatu 2.17 dostaneme

$$\beta = \mathbf{V}^{-1} \text{cov}(\mathbf{X}, Y) =$$

a proto

$$\begin{aligned} \beta' \mathbf{V} \beta &= \sigma_Y^2 \\ &= \sigma_{Y, \mathbf{X}}^2 \end{aligned}$$

Dosazením do (2.9) se dostaneme

Koeficient mnohonásobné korelace

Ze vzorců (2.7) a (2.10) dostaneme

$$\sigma_{Y, \mathbf{X}}^2 = \sigma_Y^2$$

Často se počítá koeficient mnohonásobné korelace  $\rho_{Y, (X_1, X_2)'}$ , který má jen jeden hodnotu. Koeficient mezi  $X_i$  a  $X_j$  ( $i \neq j$ ) je pak stejný, a to vždy v případě z věty 2.19 plyne.

Korelační matice  $\mathbf{P}$  vektoru  $\mathbf{X}$  je diagonální a determinanta je nezáporná.

$$|\mathbf{P}| = \begin{vmatrix} 1 & \rho_{01} \\ \rho_{01} & 1 \\ \rho_{02} & \rho_{12} \end{vmatrix}$$

Předpokládejme nyní, že  $\rho_{01} < 1$ . Vidíme, že v případě  $|\rho_{02}| < 1$  je  $|\rho_{02}|$  blízké 1.

Další zdůvodnění koeficientu mnohonásobné korelace je věta 2.19.

## Teoretické základy korelace

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

$$\rho_{Y,X}^2 = \frac{[\text{cov}(Y, X)\beta]^2}{\text{var } Y \cdot \beta'V\beta}. \quad (2.9)$$

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

$$\rho_{Y,X}^2 = \frac{\beta'V\beta}{\text{var } Y}. \quad (2.10)$$

Užitím lemmatu 2.17 dostaneme

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

a proto

$$\begin{aligned} \beta'V\beta &= \sigma_Y^2 \text{cor}(Y, X)P^{-1}D^{-1}VD^{-1}P^{-1}\text{cor}(X, Y) \\ &= \sigma_Y^2 \text{cor}(Y, X)P^{-1}\text{cor}(X, Y). \end{aligned}$$

Dosazením do (2.9) se dostane tvrzení věty.  $\square$

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

$$\sigma_{Y,X}^2 = \sigma_Y^2(1 - \rho_{Y,X}^2), \quad \rho_{Y,X}^2 = 1 - \frac{\sigma_{Y,X}^2}{\sigma_Y^2}.$$

Často se počítá koeficient mnohonásobné korelace mezi  $Y$  a vektorem  $X = (X_1, X_2)'$ , který má jen dvě složky. Položme  $X_0 = Y$  a označme  $\rho_{ij}$  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

$$\rho_{0,12}^2 = \frac{\rho_{01}^2 + \rho_{02}^2 - 2\rho_{01}\rho_{02}\rho_{12}}{1 - \rho_{12}^2}.$$

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

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

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

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

(2.8)

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

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

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

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

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

$$\begin{aligned} \rho_{Y,d'\mathbf{X}}^2 &= \frac{[\text{cov}(Y, d'\mathbf{X})]^2}{\text{var } Y \cdot \text{var } d'\mathbf{X}} = \frac{(\beta'\mathbf{V}d)^2}{\text{var } Y \cdot d'\mathbf{V}d} \leq \frac{(\beta'\mathbf{V}\beta)(d'\mathbf{V}d)}{\text{var } Y \cdot d'\mathbf{V}d} \\ &= \frac{(\beta'\mathbf{V}\beta)^2}{\text{var } Y \cdot \beta'\mathbf{V}\beta} = \frac{[\text{cov}(Y, \beta'\mathbf{X})]^2}{\text{var } Y \cdot \text{var } \beta'\mathbf{X}} = \rho_{Y,\beta'\mathbf{X}}^2. \quad \square \end{aligned}$$

Z věty 2.20 např. plyne, že koeficient mnohonásobné korelace  $\rho_{Y,X_1,\dots,X_n}$  není nikdy menší než absolutní hodnota kteréhokoli obyčejného korelačního koeficientu  $\rho_{Y,X_i}$ ,  $i = 1, \dots, n$ . 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

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

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

$$(x, y)\|z\| \geq |(x, z)| \cdot \|y\|. \quad (2.12)$$

Jelikož  $(x, z) = ((x - y) + y, z) = (x - y, z) + (y, z) = (y, z)$ , 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  $|(x, z)| \leq \|y\| \cdot \|z\|$ .  $\square$

Hodnotu  $\frac{(x, z)}{\|x\| \cdot \|z\|}$  lze interpretovat jako kosinus úhlu mezi vektory  $x$  a  $z$ . Tento úhel je tedy nejmenší úhel, pod kterým se vektory  $x$  a  $z$  svírají různoběžka s danou rovinou.

Mějme nyní dvě náhodné vektory  $X_1, \dots, X_n$ . Předpokládejme, že vektor  $\mathbf{X} = (X_1, \dots, X_n)'$  má korelace mezi  $Y$  a  $Z$ , jestliže je možné spočítat v tom, že by se vektor  $\mathbf{X}$  nemění. To se však v matematickém prostředku nemůže dít. Matematické vektory  $\mathbf{X}$  a  $Y$  pomoci vektoru  $\mathbf{V}$  mohou mít korelace mezi  $Y$  a  $Z$ .

$$\beta_1 = \mathbf{V}^{-1}\mathbf{X}$$

Tu část veličiny  $Y$ , kterou vektorem  $\beta_1 = \mathbf{V}^{-1}\mathbf{X}$  vystihujeme, nazýváme  $\hat{Y}$ . Stejná úvaha platí pro  $\beta_2 = \mathbf{V}^{-1}\mathbf{Z}$ , kde

$$\beta_2 = \mathbf{V}^{-1}\mathbf{Z}$$

To vede k následující definici. Definujeme *parciální korelace* mezi  $Y$  a  $Z$  jako korelační koeficient  $\rho_{Y-\hat{Y}, Z-\hat{Y}}$ , kde  $\hat{Y}$  je vektor, který nezávisí na  $\mathbf{X}$ . Platí rovněž

**Věta 2.22** Nechť  $P = \text{cor } \mathbf{X}$

$$\rho_{Y,Z,\mathbf{X}} = \frac{\text{cov}(Y, Z)}{\sqrt{[1 - \text{cor}(Y, \mathbf{X})][1 - \text{cor}(Z, \mathbf{X})]}}$$

*Důkaz.* Nejprve vypočteme

$$\text{cov}(Y - \beta_1'\mathbf{X}, Z - \beta_2'\mathbf{X})$$

Dosadíme-li odtud do vzorce (2.14), získáme

$$\rho_{Y,Z,\mathbf{X}} = \frac{\text{cov}(Y, Z)}{\sqrt{[1 - \text{cor}(Y, \mathbf{X})][1 - \text{cor}(Z, \mathbf{X})]}}$$

kde  $\sigma_Y^2 = \text{var } Y$  a  $\sigma_Z^2 = \text{var } Z$ , a to je výsledek věty 2.19. Tím je dokázána věta 2.22.

Je-li  $\mathbf{X}$  jednorozměrný vektor, pak výsledek věty 2.22 je ekvivalentní s výsledkem věty 2.21.

$$\rho_{Y,Z,\mathbf{X}} = \frac{\text{cov}(Y, Z)}{\sqrt{[1 - \text{cor}(Y, \mathbf{X})][1 - \text{cor}(Z, \mathbf{X})]}}$$