M5VM05 Statistické modelování

4. Základy regresní a korelační analýzy

Jan Koláček (kolacek@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

Motivace

V předchozím jsme zkoumali jednotlivé jevy (statistické znaky) izolovaně; zabývali jsme se tzv. jednorozměrnými soubory, tj. soubory popisujícími pouze jeden statistický znak a nezajímaly nás jeho vazby a vztahy k jiným jevům. V reálném světě (v přírodě, společnosti, ekonomice,...) se ovšem jevy nacházejí ve více nebo méně složitých vzájemných vztazích – navzájem na sobě závisí a podmiňují se. Proto se statistická analýza nemůže omezit pouze na zkoumání izolovaných jevů, ale musí se také zabývat analýzou jejich vzájemných vztahů. Tato analýza se dá obecně rozdělit na dvě části: regresní a korelační.

Úloha regresní analýzy

Hlavní úlohou regresní analýzy je provést **predikci** nějaké závisle proměnné náhodné veličiny Y na základě informace, kterou poskytují měření nějakých jiných náhodných veličin, řekněme $X_1, \ldots, X_k$. Veličinám $X_1, \ldots, X_k$ se potom říká **nezávisle proměnné** nebo též **doprovodné proměnné**, nebo také **kovariáty**. Měření nezávislých proměnných jsou pro experimentátora snáze dostupné než měření závisle proměnné Y.
Predikce spočívá v nalezení nějaké funkce $g(X_1, \ldots, X_k)$, která vhodně aproximuje závisle proměnnou Y. Kvalita predikce se obvykle posuzuje pomocí tzv. **střední kvadratické chyby predikce** $E[Y - g(X_1, \ldots, X_k)]^2$. Za optimální se považuje volba takové predikční funkce g, která uvedenou střední kvadratickou chybu **minimalizuje**.

Úloha korelační analýzy

Vedle průběhu sledované závislosti Y na $X_1, \ldots, X_k$ dané funkcí g je také třeba se zaměřit na měření **těsnosti** tohoto vztahu, tedy je nutné zavést nějaké míry velikosti statistické vazby (závislosti) závisle proměnné Y na nezávisle proměnných $X_1, \ldots, X_k$ s ohledem na vybranou funkci g a případně také s ohledem na závislosti mezi náhodnými veličinami $X_1, \ldots, X_k$. Tato problematika je hlavní úlohou korelační analýzy. K tomuto účelu byly zkonstruovány různé koeficienty, které nabývají hodnot od 0 do 1 (resp. od -1 do 1). Čím je takový koeficient bližší 1 (resp. -1), tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.
Korelační analýza většinou přirozeně navazuje na regresní analýzu. Nejprve pomocí regresní analýzy najdeme nějaký model závislosti v datech. Poté pomocí regresní analýzy zkoumáme vhodnost tohoto modelu.

Optimální volba predikční funkce g

Věta 1

Nechť $Y, X_1, \ldots, X_k$ jsou náhodné veličiny. Označme $\mathbf{X} = (X_1, \ldots, X_k)'$ a nechť $EY^2 < \infty$. Pak pro každou měřitelnou funkci

$$g : \mathbb{R}^k \to \mathbb{R}$$

platí

$$E(Y - g(\mathbf{X}))^2 \geq E[Y - E(Y|\mathbf{X})]^2$$

a rovnost v uvedené nerovnosti nastává právě když

$$P(g(\mathbf{X}) = E(Y|\mathbf{X})) = 1.$$

Podmíněná střední hodnota

$\mathbf{Z} = (Y, X)'$ ... sdruž. hustota $f(y,x)$; X a Y ... margin. hustoty $f_X(x)$, $f_Y(y)$.
Označme $M_X = \{x \in \mathbb{R} : f_X(x) > 0\}$, $M_Y = \{y \in \mathbb{R} : f_Y(y) > 0\}$.
Pak **podmíněná distribuční funkce** je v tomto případě definována vztahem

$$F(y|x) = \begin{cases} \int_{-\infty}^{y} \frac{f(t,x)}{f_X(x)} dt & \text{pro } x \in M_X, \\ 0 & \text{pro } x \in \mathbb{R} \setminus M_X \end{cases}$$

a **podmíněná hustota** je rovna

$$f(y|x) = \begin{cases} \frac{f(y,x)}{f_X(x)} & \text{pro } x \in M_X, \\ 0 & \text{pro } x \in \mathbb{R} \setminus M_X. \end{cases}$$

Položme

$$h(x) = E(Y|X = x) = \int_{\mathbb{R}} y dF(y|x) = \int_{\mathbb{R}} y \frac{f(y,x)}{f_X(x)} dy, \quad \text{pro} \quad \forall x \in M_X.$$

Pak náhodnou veličinu

$$E(Y|X) = h(X)$$

nazveme **podmíněnou střední hodnotou**.

Vlastnosti

- Nechť Y_1, Y_2, X jsou náhodné veličiny a a_0, a_1, a_2 jsou reálné konstanty, pak pokud střední hodnoty EY_1, EY_2 existují, platí

$$E(a_0 + a_1 Y_1 + a_2 Y_2 | X) = a_0 + a_1 E(Y_1|X) + a_2 E(Y_2|X).$$

- Nechť X, Y jsou náhodné veličiny a střední hodnota EY existuje, pak

$$E\left[E(Y|X)\right] = EY.$$

Podmíněný rozptyl

Definujeme také **podmíněný rozptyl** náhodné veličiny Y při daném X vztahem

$$D(Y|X) = E\left\{[Y - E(Y|X)]^2 \,|X\right\}.$$

Platí

$$DY = E\left[D(Y|X)\right] + D\left[E(Y|X)\right]. \tag{1}$$

Korelační koeficient

Definice 2

Pearsonův koeficient korelace náhodných veličin X, Y (které jsou aspoň intervalového charakteru) je definován vztahem

$$R(X,Y) = \begin{cases} \frac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} & \text{pro } \sqrt{D(X)},\ \sqrt{D(Y)} > 0, \\ 0 & \text{jinak,} \end{cases}$$

kde $C(X,Y) = E[(X - EX)(Y - EY)]$ je **kovariance** náhodných veličin X a Y.

Připomeneme jeho vlastnosti:

- $R(X,X) = 1$
- $R(X,Y) = R(Y,X)$
- $R(a + bX, c + dY) = \operatorname{sgn}(bd) R(X,Y)$
- $-1 \leq R(X,Y) \leq 1$ a rovnosti je dosaženo tehdy a jen tehdy, když existují reálné konstanty a, b, kde $b \neq 0$ tak, že $P(Y = a + bX) = 1$, přičemž $R(X,Y) = 1$ pro $b > 0$ a $R(X,Y) = -1$ pro $b < 0$.

Z těchto vlastností plyne, že $R(X,Y)$ je vhodnou mírou těsnosti **lineárního** vztahu náhodných veličin X, Y.

Vlastnosti

Věta 3

Mějme náhodnou veličinu Y s konečným a nenulovým rozptylem a náhodný vektor $\mathbf{X} = (X_1, \ldots, X_k)'$. Potom pro libovolnou měřitelnou funkci

$$g : \mathbb{R}^k \to \mathbb{R}$$

takovou, že existuje korelační koeficient $R(Y, g(\mathbf{X}))$ platí

$$|R(Y, g(\mathbf{X}))| \leq R(Y, E(Y|\mathbf{X})) = \sqrt{\frac{D[E(Y|\mathbf{X})]}{DY}}$$

a rovnost nastává v případě, že $D[E(Y|\mathbf{X})] \neq 0$ právě když $g(\mathbf{X})$ je lineární funkcí $E(Y|\mathbf{X})$ skoro všude vzhledem k P. V případě, že $D[E(Y|\mathbf{X})] = 0$ nastává rovnost při libovolné volbě funkce g.

Vlastnosti

Výsledky uvedené v předchozích dvou větách ukazují velký význam podmíněné střední hodnoty $E(Y|\mathbf{X})$ v **regresní** a **korelační analýze**.

(1) Z první věty plyne, že nejlepší predikci náhodné veličiny Y pomocí náhodných veličin $X_1, \ldots, X_k$, která minimalizuje střední kvadratickou chybu $E(Y - g(\mathbf{X}))^2$, dostaneme, když položíme

$$g(\mathbf{X}) = E(Y|\mathbf{X}).$$

V této souvislosti potom nejlepší prediktor $g(\mathbf{X}) = E(Y|\mathbf{X})$ nazýváme **regresní funkcí** náhodné veličiny Y na náhodných veličinách $X_1, \ldots, X_k$.

(2) Z druhé věty plyne, že regresní funkce $E(Y|\mathbf{X})$ je prediktor, který má ze všech možných prediktorů $g(\mathbf{X})$ největší korelační koeficient s predikovanou náhodnou veličinou Y. To znamená, že regresní funkce $E(Y|\mathbf{X})$ je optimálním prediktorem v tom smyslu, že má maximální statistickou vazbu (měřenou korelačním koeficientem) s predikovanou náhodnou veličinou Y.

Korelační poměr

Definice 4

Mějme náhodnou veličinu Y s konečným a nenulovým rozptylem a náhodný vektor $\mathbf{X} = (X_1, \ldots, X_k)'$. Potom číslo

$$\eta^2_{Y|\mathbf{X}} = \frac{D[E(Y|\mathbf{X}))]}{DY}$$

nazýváme **korelačním poměrem** náhodné veličiny Y na náhodném vektoru $\mathbf{X} = (X_1, \ldots, X_k)'$, nebo též **korelačním poměrem** náhodné veličiny Y na náhodných veličinách $X_1, \ldots, X_k$ a pak jej též značíme $\eta^2_{Y|X_1,\ldots,X_k}$.

Poznámky

(1) Z předchozích vět plyne, že

$$\eta^2_{Y|\mathbf{X}} = [R(Y, E(Y|\mathbf{X}))]^2$$

a tedy pro korelační poměr platí nerovnost

$$0 \leq \eta^2_{Y|\mathbf{X}} \leq 1.$$

(2) Po vydělení rovnosti (1) rozptylem DY a jednoduché úpravě dostaneme

$$1 = \frac{E(Y - E(Y|\mathbf{X}))^2}{DY} + \eta^2_{Y|\mathbf{X}}.$$

Označme symbolem $\sigma^2_{Y|\mathbf{X}}$ **střední kvadratickou chybu** predikce, když prediktorem je regresní funkce $E(Y|\mathbf{X})$, tj.

$$\sigma^2_{Y|\mathbf{X}} = E(Y - E(Y|\mathbf{X}))^2,$$

pak díky předchozímu máme

$$\eta^2_{Y|\mathbf{X}} = 1 - \frac{\sigma^2_{Y|\mathbf{X}}}{DY}.$$

Interpretace

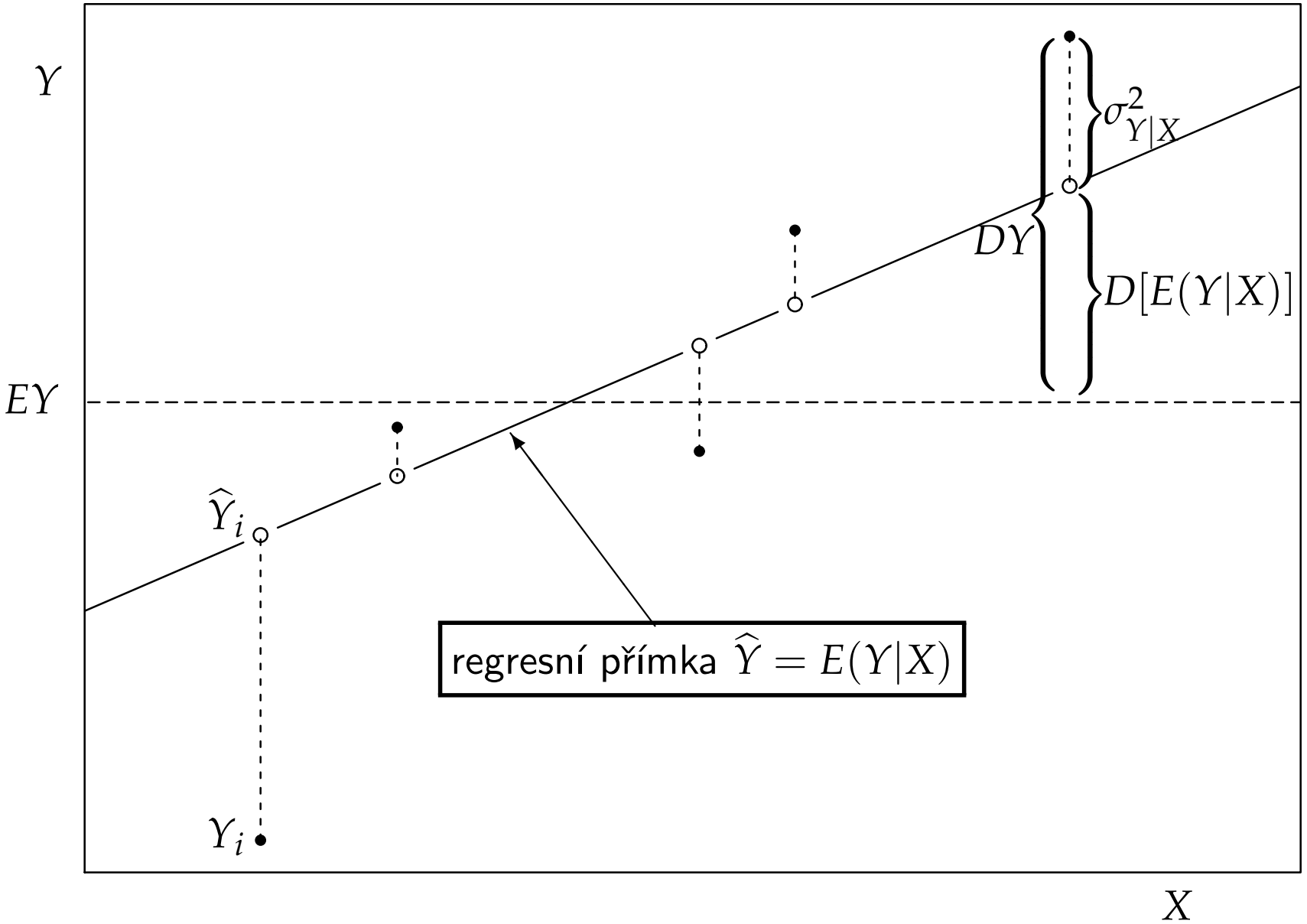
Z tohoto vztahu plyne velice názorná **interpretace** korelačním poměru $\eta^2_{Y|\mathbf{X}}$.

(a) Je-li střední kvadratická chyba predikce $\sigma^2_{Y|\mathbf{X}} = 0$, tedy v případě **ideální predikce**, je korelační poměr $\eta^2_{Y|\mathbf{X}} = 1$.

(b) V druhém krajním případě, když střední kvadratická chyba predikce je rovna DY, tj. $\sigma^2_{Y|\mathbf{X}} = DY$, pak je $\eta^2_{Y|\mathbf{X}} = 0$ a využití informace, kterou o náhodné veličině Y poskytuje náhodný vektor $\mathbf{X}$, nepřináší žádné zmenšení chyby predikce.

Tedy korelační poměr $\eta^2_{Y|\mathbf{X}}$ poskytuje **míru přesnosti predikce** a je velice užitečný při srovnávání různých vektorů doprovodných proměnných.

Graficky


Y
EY
$\widehat{Y}_i$
Y_i
X
$\sigma^2_{Y|X}$
DY
$D[E(Y|X)]$
regresní přímka $\widehat{Y} = E(Y|X)$

Praktický výpočet

Návod 5

Při praktických výpočtech se příslušné rozptyly odhadují výběrovými rozptyly. Odhadnutý korelační poměr $\eta^2_{Y|\mathbf{X}}$ *se pak nazývá* **index determinace**.

Nechť tedy máme realizace $y_1, \ldots, y_n$ *a jejich predikované hodnoty* $\hat{y}_1, \ldots, \hat{y}_n$. *Koeficient determinace má tvar*

$$ID = \frac{s^2_{\widehat{Y}}}{s^2_Y} = 1 - \frac{s^2_{Y\widehat{Y}}}{s^2_Y},$$

kde

$$s^2_{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \quad s^2_{Y\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad s^2_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Příklad

Příklad 1

Při laboratorním pokusu bylo získáno následujících 8 výsledků měření

	1	2	3	4	5	6	7	8
x_i	2,2840	2,8170	2,8367	3,5288	4,1031	4,4262	4,5211	4,9446
y_i	4,3046	6,3235	3,7082	7,6835	7,0239	8,7973	10,2961	8,4979

Zvolený model nám predikoval tyto hodnoty

$$\hat{\mathbf{y}} = (4,2614; 5,3352; 5,3750; 6,7694; 7,9264; 8,5774; 8,7685; 9,6217).$$

Určete index determinace a interpretujte ho.

Řešení

Řešení Ukážeme oba způsoby výpočtu. Vypočteme nejprve příslušné výběrové rozptyly: $\bar{y} = 7,079$, $s^2_{\widehat{Y}} = \frac{1}{8}\sum_{i=1}^{8}(\hat{y}_i - 7,079)^2 = 3,283$, $s^2_{Y\widehat{Y}} = \frac{1}{8}\sum_{i=1}^{8}(y_i - \hat{y}_i)^2 = 1,131$, $s^2_Y = \frac{1}{8}\sum_{i=1}^{8}(y_i - 7,079)^2 = 4,414$.

Podle definice je

$$ID = \frac{s^2_{\widehat{Y}}}{s^2_Y} = \frac{3,283}{4,414} = 0,7438$$

nebo

$$ID = 1 - \frac{s^2_{Y\widehat{Y}}}{s^2_Y} = 1 - \frac{1,131}{4,414} = 0,7438.$$

Výsledek lze interpretovat tak, že 74,38% celkové variability je vysvětleno zvoleným modelem.

Analýza závislosti

Výpočet podmíněné střední hodnoty $E(Y|\mathbf{X})$ vyžaduje znalost sdruženého rozdělení náhodného vektoru $\mathbf{Z} = (Y, X_1, \ldots, X_k)'$, což činí hlavní potíž, neboť v praktických situacích nebývá sdružené rozdělení vektoru $\mathbf{Z} = (Y, X_1, \ldots, X_k)'$ známé. Proto se, pokud to praktická situace dovolí, uvažují pouze **lineární modely** typu

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = \beta_0 + \boldsymbol{\beta}'\mathbf{X},$$

jestliže označíme $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$. Úloha predikce se pak redukuje na nalezení neznámých koeficientů $\beta_0, \ldots, \beta_k$, které minimalizují střední kvadratickou chybu této predikce, tj.

$$(\beta_0, \ldots, \beta_k)' = \underset{(c_0,\ldots,c_k)' \in \mathbb{R}^{k+1}}{\arg\min} \; E(Y - c_0 - c_1 X_1 - \cdots - c_k X_k)^2$$

Označme $\widehat{Y} = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$ nejlepší lineární predikci náhodné veličiny Y. Střední kvadratickou chybu nejlepší **lineární** predikce označíme tentokrát

$$\sigma^2_{Y \cdot \mathbf{X}} = E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{X})^2$$

Koeficient mnohonásobné korelace

Definice 6

Pearsonův korelační koeficient $R(Y, \widehat{Y})$ označíme $\rho_{Y \cdot \mathbf{X}}$ a budeme jej nazývat **koeficientem mnohonásobné korelace** náhodné veličiny Y na náhodném vektoru $\mathbf{X} = (X_1, \ldots, X_k)'$ (nebo též na náhodných veličinách $X_1, \ldots, X_k$ a pak budeme podrobněji psát $\rho_{Y \cdot (X_1, \ldots, X_k)}$).

Definice 7 (Korelační matice)

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ a $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ jsou náhodné vektory. Potom matici

$$R(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} R(X_1, Y_1) & \cdots & R(X_1, Y_m) \\ \vdots & \ddots & \vdots \\ R(X_n, Y_1) & \cdots & R(X_n, Y_m) \end{pmatrix} = \big(R(X_i, Y_j)\big)_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}}$$

nazýváme **korelační maticí náhodných vektorů $\mathbf{X}$ a $\mathbf{Y}$**.
Dále matici $R(\mathbf{X}, \mathbf{X})$ budeme značit $R(\mathbf{X})$ a budeme ji nazývat **korelační maticí náhodného vektoru $\mathbf{X}$**.

Vlastnosti

Věta 8

Koeficient mnohonásobné korelace $\rho_{Y\cdot\mathbf{X}}$ má následující vlastnosti

(1) *Koeficient mnohonásobné korelace $\rho_{Y\cdot\mathbf{X}}$ je vždy nezáporný.*

(2) *Pomocí regresních koeficientů $\beta_0, \beta_1, \ldots, \beta_k$ jej lze vyjádřit ve tvaru*

$$\rho^2_{Y\cdot\mathbf{X}} = \frac{\boldsymbol{\beta}' D\mathbf{X}\boldsymbol{\beta}}{DY}.$$

(3) *Pomocí korelačních matic jej lze vyjádřit ve tvaru*

$$\rho^2_{Y\cdot\mathbf{X}} = R(Y,\mathbf{X})(R(\mathbf{X}))^{-1}R(\mathbf{X},Y)$$

(4) *Pomocí reziduálního rozptylu lineární predikce jej lze vyjádřit ve tvaru*

$$\rho^2_{Y\cdot\mathbf{X}} = 1 - \frac{\sigma^2_{Y\cdot\mathbf{X}}}{DY}$$

Poznámka

1. Vzorec $\boxed{\rho^2_{Y\cdot\mathbf{X}} = \frac{\boldsymbol{\beta}' D\mathbf{X}\boldsymbol{\beta}}{DY}}$ je vhodný pro výpočet koeficientu mnohonásobné korelace v případě, že je k dispozici vektor regresních koeficientů $(\beta_0, \beta_1, \ldots, \beta_k)'$.
2. Vzorec $\boxed{\rho^2_{Y\cdot\mathbf{X}} = R(Y,\mathbf{X})(R(X))^{-1}R(\mathbf{X},Y)}$ se využívá v případě, že jsou k dispozici korelační koeficienty mezi náhodnými veličinami $Y, X_1, \ldots, X_k$.
3. Identity $\boxed{\rho^2_{Y\cdot\mathbf{X}} = 1 - \frac{\sigma^2_{Y\cdot\mathbf{X}}}{DY}}$ a $\boxed{\eta^2_{Y|\mathbf{X}} = 1 - \frac{\sigma^2_{Y|\mathbf{X}}}{DY}}$ ukazují, že korelační poměr $\eta^2_{Y|\mathbf{X}}$ je roven kvadrátu koeficientu mnohonásobné korelace $\rho^2_{Y\cdot\mathbf{X}}$ v případě, že teoretická regresní funkce $g(X) = E(Y|\mathbf{X})$ je lineární funkcí proměnných $X_1, \ldots, X_k$. Dále je z tohoto vzorce patrné, že pokud se omezíme na lineární predikce, je interpretace koeficientu mnohonásobné korelace stejná jako je interpretace korelačního poměru v obecném případě.

Pokračování

4. Podle uváděných vzorců lze koeficient mnohonásobné korelace $\rho_{Y\cdot\mathbf{X}}$ počítat i v případě, kdy podmíněná střední hodnota $E(Y|\mathbf{X})$ není lineární. V tomto případě potom díky vztahu (dokázaném ve Větě 1)

$$\underbrace{E(Y-\beta_0-\boldsymbol{\beta}'\mathbf{X})^2}_{=\sigma^2_{Y\cdot\mathbf{X}}} \geq \underbrace{E[Y-E(Y|\mathbf{X})]^2}_{=\sigma^2_{Y|\mathbf{X}}}$$

snadno vidíme, že

$$0 \leq \rho^2_{Y\cdot\mathbf{X}} \leq \eta^2_{Y|\mathbf{X}} \leq 1$$

Vlastnosti

Věta 9

Pro libovolný nenulový vektor $\mathbf{c} = (c_1, \ldots, c_k)' \in \mathbb{R}^k$ *a* $c_0 \in \mathbb{R}$ *platí*

$$\rho^2_{Y \cdot \mathbf{X}} \geq R^2(Y, c_0 + \mathbf{c}'\mathbf{X}),$$

tj. koeficient mnohonásobné korelace je maximální korelační koeficient mezi náhodnou veličinou Y *a libovolnou lineární funkcí* $c_0 + \mathbf{c}'\mathbf{X}$ *náhodného vektoru* $\mathbf{X}$.

Důsledek 10

Pro libovolné $j = 1, \ldots, k$ *platí*

$$\rho^2_{Y \cdot \mathbf{X}} \geq R^2(Y, X_j),$$

tj. absolutní hodnota libovolného korelačního koeficientu mezi náhodnou veličinou Y *a libovolnou z náhodných veličin* $X_1, \ldots, X_k$ *je nejvýše rovna koeficientu mnohonásobné korelace mezi náhodnou veličinou* Y *a náhodným vektorem* $\mathbf{X} = (X_1, \ldots, X_k)'$.

Prakticky

Definice 11

Mějme náhodný výběr rozsahu n s vektory $\mathbf{X}_1 = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Z}_1 \end{pmatrix}, \ldots, \mathbf{X}_n = \begin{pmatrix} \mathbf{Y}_n \\ \mathbf{Z}_n \end{pmatrix}$, kde pro $i = 1, \ldots, n$ jsou náhodné vektory $\mathbf{Y}_i$ typu $p \times 1$ a $\mathbf{Z}_i$ typu $q \times 1$, přičemž $p + q = k$.
Definujme **výběrové kovarianční matice**

$$\mathbf{S}_{YZ} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{Y}_i - \overline{\mathbf{Y}})(\mathbf{Z}_i - \overline{\mathbf{Z}})' = (S_{ij}) \qquad (\text{typu } p \times q),$$

kde

$$\overline{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i = \begin{pmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_p \end{pmatrix} \qquad \text{a} \qquad \overline{\mathbf{Z}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i = \begin{pmatrix} \overline{Z}_1 \\ \vdots \\ \overline{Z}_q \end{pmatrix},$$

a **výběrovou korelační matici**

$$\mathbf{R}_{ZY} = (r_{ij}) = \left(\frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}} \right).$$

Prakticky

Definice 12

Mějme náhodné vektory $\begin{pmatrix} Y_1 \\ \mathbf{X}_1 \end{pmatrix}, \cdots, \begin{pmatrix} Y_n \\ \mathbf{X}_n \end{pmatrix}$, kde Y_i jsou náhodné veličiny a $\mathbf{X}_i \quad (i = 1, \cdots, n)$ jsou náhodné vektory typu $p \times 1$.
Jestliže matice $\mathbf{R}_{XX}$ je regulární, pak **výběrový koeficient mnohonásobné korelace** je definován vztahem:

$$r^2_{Y \cdot \mathbf{X}} = \mathbf{R}_{YX} \mathbf{R}^{-1}_{XX} \mathbf{R}_{XY}.$$

Návod 13 (praktický výpočet)

V praxi se většinou výběrový koeficient mnohonásobné korelace počítá pomocí nějakého software. Hledání inverzní matice $\mathbf{R}^{-1}_{XX}$ může být obecně složitý proces, proto ještě uvedeme alternativní výpočet. Položme $\mathbf{Z} = (Y, \mathbf{X})$ a $\mathbf{R} = \mathbf{R}_{ZZ}$. Pak

$$r^2_{Y \cdot \mathbf{X}} = 1 - \frac{\det(\mathbf{R})}{\det(\mathbf{R}_{XX})}.$$

Příklad

Příklad 2

Zjišťujeme závislost koncentrace ozónu[a] (proměnná Y) ve spodních vrstvách atmosféry na meteorologických podmínkách, které jsou popsány intenzitou slunečního záření (X_1), rychlosti větru (X_2) a teplotě vzduchu (X_3). Naměřená data udává následující tabulka.

i	Y	X_1	X_2	X_3
1	23	148	8,00	82
2	21	191	14,90	77
3	37	284	20,70	72
4	20	37	9,20	65
5	12	120	11,50	73
6	13	137	10,30	76
7	135	269	4,10	84
8	49	248	9,20	85
9	32	236	9,20	81
10	64	175	4,60	83

Vypočtěte výběrový koeficient mnohonásobné korelace.

[a] část datového souboru *airquality* implementovaného v jazyce R

Řešení

Řešení $\mathbf{R}_{YX} = (0,55; -0,51; 0,54)$.

$$\mathbf{R}_{XX} = \begin{pmatrix} 1,00 & 0,19 & 0,60 \\ 0,19 & 1,00 & -0,52 \\ 0,60 & -0,52 & 1,00 \end{pmatrix}$$

Její inverze je tvaru

$$\mathbf{R}_{XX}^{-1} = \begin{pmatrix} 3,29 & -2,25 & -3,13 \\ -2,25 & 2,91 & 2,85 \\ -3,13 & 2,85 & 4,34 \end{pmatrix}$$

a celkově dostáváme $r^2_{Y\cdot\mathbf{X}} = \mathbf{R}_{YX}\mathbf{R}_{XX}^{-1}\mathbf{R}_{XY} = 0,8557$.

Řešení

Pokud bychom použili druhý způsob uvedený v Návodu 18, je třeba vypočítat matici $\mathbf{R}$, kterou lze z předešlého vyjádřit $\mathbf{R} = \begin{pmatrix} 1 & \mathbf{R}_{YX} \\ \mathbf{R}'_{YX} & \mathbf{R}_{XX} \end{pmatrix}$, tj.

$$\mathbf{R} = \begin{pmatrix} 1,00 & 0,55 & -0,51 & 0,54 \\ 0,55 & 1,00 & 0,19 & 0,60 \\ -0,51 & 0,19 & 1,00 & -0,52 \\ 0,54 & 0,60 & -0,52 & 1,00 \end{pmatrix}.$$

Pak

$$r^2_{Y\cdot\mathbf{X}} = 1 - \frac{\det(\mathbf{R})}{\det(\mathbf{R}_{XX})} = 1 - \frac{0,032}{0,22} = 0,8557.$$

Hodnota tohoto koeficientu poukazuje na do jisté míry velkou lineární závislost proměnné Y na ostatních proměnných. Tato hodnota je však značně ovlivněna také korelacemi proměnných X_1, X_2 a X_3 mezi sebou. Při pohledu na prvky matice $\mathbf{R}_{XX}$ vidíme, že je např. významná korelace mezi intenzitou slunečního záření (X_1) a teplotou vzduchu (X_3). Pro vyloučení těchto vlivů je třeba spočítat parciální korelační koeficienty – viz dále.

Parciální korelační koeficient

Budeme uvažovat náhodné veličiny

$$Y, Z, X_1, \ldots, X_k.$$

Motivací k zavedení tohoto korelačního koeficientu je fakt, že korelační koeficient $R(Y, Z)$ mezi náhodnou veličinou Y a Z může být dosti vysoký proto, že obě náhodné veličiny jsou silně závislé na náhodném vektoru $\mathbf{X} = (X_1, \ldots, X_k)'$.
Zajímá nás proto, jaká by byla korelace mezi Y a Z při vyloučení vlivu, který je způsoben náhodným vektorem $\mathbf{X}$.
Toto odstranění vlivu náhodného vektoru $\mathbf{X}$ lze uskutečnit tak, že se sleduje korelace mezi Y a Z při pevných hodnotách náhodného vektoru $\mathbf{X}$.
Protože v praktických situacích není možné uspořádání experimentu takovým způsobem, aby byla provedena eliminace vlivu náhodného vektoru $\mathbf{X}$, je třeba ji provést pomocí vhodného matematického modelu. Obdobně jako v případě koeficientu mnohonásobné korelace se omezíme pouze na lineární vztahy.

Definice

Označme $\widehat{Y}$ a $\widehat{Z}$ nejlepší lineární predikce náhodných veličin Y a Z pomocí náhodného vektoru $\mathbf{X}$. Korelaci očištěnou od vlivu náhodného vektoru $\mathbf{X}$ dostaneme, budeme-li počítat korelaci $R(Y - \widehat{Y}, Z - \widehat{Z})$.

Definice 14

Nechť existuje korelační koeficient $R(Y - \widehat{Y}, Z - \widehat{Z})$. Potom jej budeme nazývat **parciálním korelačním koeficientem** náhodných veličin Y a Z při pevném $\mathbf{X}$ a budeme jej značit

$$\rho_{Y,Z \cdot \mathbf{X}} = R(Y - \widehat{Y}, Z - \widehat{Z}).$$

Vlastnosti

Věta 15

Pro parciální korelační koeficient náhodných veličin Y a Z při pevném $\mathbf{X}$ platí

$$\rho_{Y,Z\cdot\mathbf{X}} = \left[R(Y,Z) - R(Y,\mathbf{X})(R(\mathbf{X}))^{-1}R(\mathbf{X},Z)\right]\left[\left(1-\rho^2_{Y\cdot\mathbf{X}}\right)\left(1-\rho^2_{Z\cdot\mathbf{X}}\right)\right]^{-\frac{1}{2}}$$

Z hodnoty korelačního koeficientu $R(Y,Z)$ nelze usuzovat na velikost parciálního korelačního koeficientu $\rho_{Y,Z\cdot\mathbf{X}}$. Tyto dva koeficienty se od sebe mohou dosti odlišovat, mohou mít i různé znaménko a v případě, že jeden z nich je roven nule, může být druhý různý od nuly a podobně. Jejich vztah je tedy odlišný od vztahu $R(Y,X_j)$ a $\rho_{Y\cdot\mathbf{X}}$, který dává Důsledek 15.

Prakticky

Definice 16

Mějme náhodné vektory $\begin{pmatrix} Y_1 \\ Z_1 \\ \mathbf{X}_1 \end{pmatrix}, \cdots, \begin{pmatrix} Y_n \\ Z_n \\ \mathbf{X}_n \end{pmatrix}$, kde Y_i, Z_i jsou náhodné veličiny a $\mathbf{X}_i \quad (i = 1, \cdots, n)$ jsou náhodné vektory typu $p \times 1$.
Pak **výběrový parciální korelační koeficient** je definován vztahem

$$r_{Y,Z\cdot\mathbf{X}} = \frac{r^2_{YZ} - r^2_{Y\cdot\mathbf{X}} r^2_{Z\cdot\mathbf{X}}}{\sqrt{\left(1 - r^2_{Y\cdot\mathbf{X}}\right)\left(1 - r^2_{Z\cdot\mathbf{X}}\right)}},$$

kde r^2_{YZ} je výběrový koeficient korelace náhodných veličin Y, Z a $r^2_{Y\cdot\mathbf{X}}$, $r^2_{Z\cdot\mathbf{X}}$ jsou příslušné výběrové koeficienty mnohonásobné korelace.

Prakticky

Definice 17

Mějme náhodné vektory $\begin{pmatrix} Y_1 \\ Z_1 \\ \mathbf{X}_1 \end{pmatrix}, \cdots, \begin{pmatrix} Y_n \\ Z_n \\ \mathbf{X}_n \end{pmatrix}$, kde Y_i, Z_i jsou náhodné veličiny a $\mathbf{X}_i \quad (i = 1, \cdots, n)$ jsou náhodné vektory typu $p \times 1$.
Pak **výběrový parciální korelační koeficient** je definován vztahem

$$r_{Y,Z\cdot\mathbf{X}} = \frac{r^2_{YZ} - r^2_{Y\cdot\mathbf{X}} r^2_{Z\cdot\mathbf{X}}}{\sqrt{\left(1 - r^2_{Y\cdot\mathbf{X}}\right)\left(1 - r^2_{Z\cdot\mathbf{X}}\right)}},$$

kde r^2_{YZ} je výběrový koeficient korelace náhodných veličin Y, Z a $r^2_{Y\cdot\mathbf{X}}$, $r^2_{Z\cdot\mathbf{X}}$ jsou příslušné výběrové koeficienty mnohonásobné korelace.

Prakticky

Návod 18

V praxi se pro výpočet parciálního korelačního koeficientu používá následujícího postupu. Položme $\mathbf{W} = (Y, Z, \mathbf{X})$ *a* $\mathbf{R} = \mathbf{R}_{WW}$. *Pak*

$$r_{Y,Z\cdot\mathbf{X}} = \frac{\det(\mathbf{R}_{(12)})}{\sqrt{\det(\mathbf{R}_{(11)})\det(\mathbf{R}_{(22)})}},$$

kde $\mathbf{R}_{(ij)}$ *je submatice, která vznikne z* $\mathbf{R}$ *vynecháním* i*-tého řádku a* j*-tého sloupce.*

Příklad

Příklad 3

Na datech z Příkladu 2 vypočtěte parciální korelační koeficient $r_{Y,X_1\cdot(X_2,X_3)}$.

Řešení Připomeňme matici $\mathbf{R}$, která byla tvaru

$$\mathbf{R} = \begin{pmatrix} 1,00 & 0,55 & -0,51 & 0,54 \\ 0,55 & 1,00 & 0,19 & 0,60 \\ -0,51 & 0,19 & 1,00 & -0,52 \\ 0,54 & 0,60 & -0,52 & 1,00 \end{pmatrix}.$$

Řešení

Příslušné submatice jsou

$$\mathbf{R}_{(11)} = \begin{pmatrix} 1,00 & 0,19 & 0,60 \\ 0,19 & 1,00 & -0,52 \\ 0,60 & -0,52 & 1,00 \end{pmatrix},$$

$$\mathbf{R}_{(12)} = \begin{pmatrix} 0,55 & 0,19 & 0,60 \\ -0,51 & 1,00 & -0,52 \\ 0,54 & -0,52 & 1,00 \end{pmatrix},$$

$$\mathbf{R}_{(22)} = \begin{pmatrix} 1,00 & -0,51 & 0,54 \\ -0,51 & 1,00 & -0,52 \\ 0,54 & -0,52 & 1,00 \end{pmatrix}.$$

Po dosazení dostáváme

$$r_{Y,X_1\cdot(X_2,X_3)} = \frac{0,2827}{\sqrt{0,2220 \cdot 0,4654}} = 0,8795.$$

Výsledek lze interpretovat jako velikost lineární závislosti ozónu na intenzitě slunečního záření s vyloučením vlivu rychlosti větru a teploty vzduchu. Podobně by šlo zkoumat ostatní vazby mezi proměnnými.

Úlohy k procvičení

Příklad 4.1

V tabulce jsou uvedeny výsledky měření (x_i, y_i) *a predikované hodnoty* $\hat{y}_i$, $i = 1, \ldots, 10$

i	1	2	3	4	5	6	7	8	9	10
x_i	1,60	1,86	2,21	2,29	3,38	3,42	3,62	3,65	3,76	4,27
y_i	3,24	3,12	3,81	5,12	6,28	7,15	7,33	7,81	8,08	8,43
$\hat{y}_i$	2,98	3,54	4,31	4,48	6,85	6,94	7,37	7,44	7,68	8,79

Určete index determinace a interpretujte ho.

[$ID = 0.95532$*]*

Úlohy k procvičení

Příklad 4.2

Během 14-ti dní byla měřena polední teplota vzduchu. K predikci teploty byly použity dva modely – model A a model B. Naměřené hodnoty a predikované hodnoty obou modelů jsou uvedeny v následující tabulce.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_i	0,35	-1,54	0,47	-0,50	-1,99	-2,17	-1,86	-1,37	-1,88	-2,30	-2,13	-2,12	-1,76	-1,06
$\hat{y}_i^A$	-0,62	-0,75	-0,87	-0,99	-1,11	-1,24	-1,36	-1,48	-1,60	-1,73	-1,85	-1,97	-2,09	-2,22
$\hat{y}_i^B$	-0,17	-0,35	-0,52	-0,70	-0,87	-1,05	-1,22	-1,39	-1,57	-1,74	-1,92	-2,09	-2,27	-2,44

Na základě indexu determinace rozhodněte, který z modelů je lepší.

$[ID_A = 0,31;\ ID_B = 0,24]$

Příklad 4.3

Na datech ze Cvičení 4.2 byla predikována hodnota polední teploty vzduchu v 15. den. Model A tuto hodnotu odhadl $\hat{y}_{15}^A = -2,34$*, predikce pomocí modelu B byla* $\hat{y}_{15}^B = -2,61$*. Ve skutečnosti byla naměřena hodnota* $y_{15} = -1,34$*. Na nových datech opět porovnejte oba modely pomocí indexu determinace.*

$[ID_A = 0,22;\ ID_B = 0,09]$

Úlohy k procvičení

Příklad 4.4

Zjišťujeme závislost spotřeby paliva osobních automobilů[a] *(proměnná* Y*, počet mil/galon) na vlastnostech motoru, které jsou popsány objemem válců (*X_1*, kubické palce), výkonem (*X_2*, počet koní), hmotností vozidla (*X_3*, kilolibry) a zrychlením (*X_4*, počet sekund na 1/4 míle). Naměřená data udává tabulka na další straně.*

Vypočtěte závislost spotřeby paliva osobních automobilů na objemu válců, výkonu, hmotnosti a zrychlením vozidla.

$$[r^2_{Y\cdot\mathbf{X}} = 0,934]$$

[a] část datového souboru *mtcars* implementovaného v jazyce R

Model (r.v. 1974)	Y	X_1	X_2	X_3	X_4
Mazda RX4 Wag	21,00	160,00	110,00	2,88	17,02
Datsun 710	22,80	108,00	93,00	2,32	18,61
Hornet 4 Drive	21,40	258,00	110,00	3,21	19,44
Valiant	18,10	225,00	105,00	3,46	20,22
Merc 280C	17,80	167,60	123,00	3,44	18,90
Cadillac Fleetwood	10,40	472,00	205,00	5,25	17,98
AMC Javelin	15,20	304,00	150,00	3,44	17,30
Fiat X1-9	27,30	79,00	66,00	1,94	18,90
Porsche 914-2	26,00	120,30	91,00	2,14	16,70
Ford Pantera L	15,80	351,00	264,00	3,17	14,50

Úlohy k procvičení

Příklad 4.5

V rámci biometrického výzkumu byl na jednotlivých stromech zjišťován vztah mezi veličinami objem (Y, m^3), výčetní tloušťka (X_1, cm), výška (X_2, m) a délka zelené koruny (X_3, m). Naměřené hodnoty jsou uvedeny v tabulce na další straně. Vyšetřete korelační závislost objemu na tloušťce, výšce a délce zelené koruny.

$[r^2_{Y\cdot\mathbf{X}} = 0,9634]$

Strom	Y	X_1	X_2	X_3
1	0,013	8	9,8	3,6
2	0,021	8	10,2	3,6
3	0,012	7	9,4	3,0
4	0,009	7	7,8	1,4
5	0,065	12	11,2	4,6
6	0,071	12	12,0	5,1
7	0,102	13	13,5	6,9
8	0,048	10	12,1	4,6
9	0,049	11	10,8	4,3
10	0,011	7	8,9	3,9
11	0,017	8	9,3	3,5
12	0,059	11	12,0	4,8

Úlohy k procvičení

Příklad 4.6

Na datech ze Cvičení 4.4 vypočtěte parciální korelační koeficienty $r_{Y,X_1\cdot(X_2,X_3,X_4)}$, $r_{Y,X_2\cdot(X_1,X_3,X_4)}$, $r_{Y,X_3\cdot(X_1,X_2,X_4)}$, $r_{X_1,X_4\cdot(X_1,X_2,X_3)}$.

$[r_{Y,X_1\cdot(X_2,X_3,X_4)} = 0,2319;\ r_{Y,X_2\cdot(X_1,X_3,X_4)} = -0,5219;\ r_{Y,X_3\cdot(X_1,X_2,X_4)} = -0,7405;$
$r_{X_1,X_4\cdot(X_1,X_2,X_3)} = -0,0736.]$

Příklad 4.7

Na datech ze Cvičení 4.5 vypočtěte všechny parciální korelační koeficienty.

$[r_{Y,X_1\cdot(X_2,X_3)} = 0,8558;\ r_{Y,X_2\cdot(X_1,X_3)} = 0,1938;\ r_{Y,X_3\cdot(X_1,X_2)} = 0,2974;$
$r_{X_1,X_2\cdot(Y,X_3)} = 0,1248;\ r_{X_1,X_3\cdot(Y,X_2)} = -0,22;\ r_{X_2,X_3\cdot(Y,X_1)} = 0,6161.]$