

8. Regresní a korelační analýza

Problém: hledání, zkoumání a hodnocení souvislostí, závislostí mezi dvěma a více statistickými znaky (veličinami).

Typy závislostí: *pevné* a *volné*

Pevná závislost – každé hodnotě jedné veličiny odpovídá jedna a jen jedna hodnota jiných veličin

– (většinou v teoretické oblasti)

Volná závislost – hodnotám jedné veličiny odpovídají různé hodnoty jiné veličiny

– při změnách hodnot těchto veličin se projevuje určitá obecná tendence

– (v praktických situacích)

Statistická závislost – volná závislost mezi kvantitativními veličinami

Metody *regresní a korelační analýzy* – slouží k poznání, matematickému popisu stat. závislostí a k hodnocení závěrů o vztahu zkoumaných veličin.

Jednostranné závislosti – *regresní analýza*

– zkoumání obecné tendence ve změnách závislé veličiny vzhledem ke změnám nezávislých vel.

Vzájemné závislosti – *korelační analýza*

– důraz na sílu vzájemného vztahu mezi vel.

- **Lineární rovnice s jednou nezávislou proměnnou**

Obecný tvar *lineární rovnice s jednou nezávislou proměnnou*

$$y = b_0 + b_1x$$

b_0 a b_1 – konstanty

x – nezávislá veličina, y - závislá veličina

Graf lineární rovnice s 1 nezávislou proměnnou –
přímka; každá přímka, která není kolmá na osu x

Geometrická interpretace b_0, b_1

b_0 – *y-úsek (intercept)*

b_1 – *směrnice (slope)*: indikuje změnu y -hodnoty, která je způsobena změnou x -hodnoty o jednu jednotku

8.1 Regresní přímka

Předpoklady:

X – nezávislá (vysvětlující) veličina (proměnná),
regresor

Y – závislá (vysvětlovaná) veličina (proměnná)
náhodná veličina

P1. Teoretická regresní přímka:

$$\exists \text{ přímka } y = \beta_0 + \beta_1 x : \quad \forall x \quad E(Y|X = x) = \beta_0 + \beta_1 x$$

P2. Shodné směrodatné odchylky:

$$\sigma(Y|X = x) = \sigma(Y) \quad \forall x$$

P3. Normalita:

$$\forall x \quad Y \sim N\text{-rozdělení}$$

If $\exists \beta_0, \beta_1$ a σ : $\forall x \quad Y \sim N(\beta_0 + \beta_1 x, \sigma^2) \implies$

P1–P3 splněny

Předpoklady P1, P2, P3 – *model regresní přímky*

Symbolické vyjádření:

$$Y = \beta_0 + \beta_1 X + \epsilon = \eta + \epsilon$$

$$\epsilon \sim N[0; \sigma^2]$$

β_0, β_1 – *parametry (koeficienty) regresní přímky*

• Výběrová (empirická) regresní přímka

x_1, x_2, \dots, x_n – pozorované hodnoty veličiny X

y_1, y_2, \dots, y_n – pozorované hodnoty i.i.d. náh. v.

$Y_1, Y_2, \dots, Y_n, \quad Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \eta_i = E(Y_i | X = x_i) = \beta_0 + \beta_1 x_i$

β_0, β_1, σ – obecně neznámé

Cíl: Odhadnout β_0, β_1, σ na základě dvojic dat $(x_i, y_i), i = 1, 2, \dots, n$

b_0 a b_1 - bodové odhady parametrů β_0 a β_1

$\hat{y} = b_0 + b_1 x$ – *výběrová (empirická) regresní přímka*
– odhad teoretické regresní přímky

Reziduum:

$$e_i = y_i - \hat{y}_i \quad \sum_{i=1}^n e_i = 0$$

e_i – odhad hodnoty náhodné veličiny ϵ_i

Reziduální součet čtverců:

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

• Bodové odhady parametrů β_0 a β_1

Kritérium: Minimalizace součtu čtverců S_R

$$S_R = S(\beta_0, \beta_1) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Nutná podmínka pro minimum ryze konvexní funkce $S(\beta_0, \beta_1)$ dvou proměnných β_0, β_1 :

$$\frac{\partial S}{\partial \beta_0} \Big|_{\beta_0=b_0, \beta_1=b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} \Big|_{\beta_0=b_0, \beta_1=b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i = 0$$

System normálních rovnic

$$\begin{aligned} n b_0 + b_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \end{aligned}$$

Řešení:

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{s_{xy}}{s_x^2} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Výběrová regresní přímka:

$$\hat{y} = b_0 + b_1 x = \bar{y} - b_1 \bar{x} + b_1 x = \bar{y} - b_1 (x - \bar{x}) = \bar{y} - b_{yx} (x - \bar{x})$$

- Bodový odhad rozptylu σ^2

Předpoklady: P1–P3 pro model regresní přímky

Bodový odhad σ : s_R^2

$$s_R^2 = \frac{S_R}{n - 2}, \quad S_R = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Směrodatná chyba odhadu rozptylu σ (reziduální směrodatná odchylka) (standard error of estimate)

$$s_R = \sqrt{\frac{S_R}{n - 2}}$$

Interpretace s_R : vyjadřuje jak se v průměru hodnota \hat{y} veličiny Y liší od pozorované hodnoty y

• Rozdělení odhadů b_0 , b_1 a \hat{y}

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t[n - 2]; \quad \frac{b_1 - \beta_1}{s_{b_1}} \sim t[n - 2]; \quad \frac{\hat{y}_i - \eta_i}{s_{\hat{y}}} \sim t[n - 2]$$

s_{b_0}, s_{b_1} – *směrodatné chyby odhadů b_0 a b_1*

$$s_{b_0}^2 = s_R^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = s_R^2 \left(\frac{\overline{x^2}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2$$

$$s_{b_1}^2 = s_R^2 \left(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$s_{\hat{y}}$ – *chyba regresní přímky pro i -té pozorování y_i*

$$s_{\hat{y}} = s_R^2 \left(\frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

Pro $n > 30$ lze použít aproximaci $N[0; 1]$ -rozdělením

$$\frac{b_0 - \beta_0}{s_{b_0}} \approx N[0; 1]; \quad \frac{b_1 - \beta_1}{s_{b_1}} \approx N[0; 1]; \quad \frac{\hat{y}_i - \eta_i}{s_{\hat{y}}} \approx N[0; 1]$$

- **Intervaly spolehlivosti pro β_0, β_1, η_i**

Předpoklady: P1–P3 pro model regresní přímky

- ◇ **Koeficient spolehlivosti: $(1 - \alpha)$**

- ◇ **Bodové odhady β_0, β_1, η_i : b_0, b_1, \hat{y}_i**

- ◇ **Krajní body $100(1 - \alpha)\%$ intervalu spolehlivosti:**

$$b_0 \pm t_{1-\frac{\alpha}{2}}(n-2) s_{b_0}$$

$$b_1 \pm t_{1-\frac{\alpha}{2}}(n-2) s_{b_1}$$

$$\hat{y}_i \pm t_{1-\frac{\alpha}{2}}(n-2) s_{\hat{y}_i}$$

$$i = 1, \dots, n$$

s_{b_i} – směrodatná (standardní) chyba odhadu
 $b_i, i = 0, 1$

$s_{\hat{y}_i}$ – chyba regresní přímky pro i -té pozorování y_i

$t_{1-\frac{\alpha}{2}}(n-2)$ – $100(1 - \alpha/2)\%$ kvantil Studentova
 t -rozdělení s $(n - 2)$ stupni volnosti

- **Testy hypotéz o parametrech β_0, β_1**

Individuální t-test :

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0 \quad i = 0, 1$$

- **Testovací statistika:**

$$T_i = \frac{b_i}{s_{b_i}} \sim t[n - 2]$$

- **Kritický obor:**

$$|T_i| > t_{1-\frac{\alpha}{2}}(n - 2)$$

s_{b_i} – směrodatná (standardní) chyba odhadu b_i , $i = 0, 1$

$t_{1-\frac{\alpha}{2}}(n - 2)$ – 100(1 - $\alpha/2$)% kvantil Studentova t -rozdělení s $(n - 2)$ stupni volnosti

• Odhad a předpověď (predikce)

Využití výběrové regresní přímky:

- pro odhad střední hodnoty závislé veličiny Y odpovídající určité hodnotě nezávislé veličiny X
- pro předpověď individuální hodnoty veličiny Y odpovídající určité hodnotě nezávislé veličiny X

x_P – určitá hodnota nezávislé veličiny X

$\hat{y}_P = b_0 + b_1 x_P$ – předpověď hodnoty y_P veličiny Y
pro $X = x_P$

$E(Y | X = x_P)$ – střední hodnota Y na úrovni x_P

Bodový odhad $E(Y | X = x_P) : b_0 + b_1 x_P$

Bodový odhad střední hodnoty Y na úrovni x_P je shodný s předpovědí individuální hodnoty y_P .

- **Interval spolehlivosti pro $E(Y \mid X = x_P)$**

t -rozdělení pro IS v regresi

$$T = \frac{\hat{Y}_P - (\beta_0 + \beta_1 x_P)}{s_R \sqrt{\frac{1}{n} + \frac{(x_P - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t[n - 2]$$

Předpoklady: P1–P3 pro model regresní přímky

- **Koeficient spolehlivosti:** $(1 - \alpha)$
- **Bodový odhad $E(Y \mid X = x_P)$:** $b_0 + b_1 x_P$
- **Krajní body IS pro $E(Y \mid X = x_P)$:**

$$\hat{y}_p \pm t_{1-\frac{\alpha}{2}}(n - 2) s_R \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$s_R^2 = \frac{S_R}{n-2}$ – **reziduální rozptyl**

$t_{1-\frac{\alpha}{2}}(n - 2)$ – **100(1 - $\alpha/2$)% kvantil Studentova t -rozdělení s $(n - 2)$ stupni volnosti**

- **Interval spolehlivosti pro y_P**

**IS pro y_P – interval předpovědi (predikce) pro y_P
(IP)**

t -rozdělení pro IP v regresi

$$T = \frac{Y_P - \hat{y}_P}{s_R \sqrt{1 + \frac{1}{n} + \frac{(x_P - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t[n - 2]$$

Předpoklady: P1–P3 pro model regresní přímky

- **Koeficient spolehlivosti: $(1 - \alpha)$**
- **Předpověď hodnoty veličiny Y pro hodnotu x_P veličiny X :**

$$\hat{y}_p = b_0 + b_1 x_p$$

- **Krajní body IS pro hodnotu y_P na úrovni x_P :**

$$\hat{y}_P \pm t_{1-\frac{\alpha}{2}}(n - 2) s_R \sqrt{1 + \frac{1}{n} + \frac{(x_P - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

IP je širší než IS

8.2 Kvalita regresní přímky a intenzita závislostí

- Korelace a regrese

Korelační model: Y a X náhodné veličiny

Regresní model: Y náhodná

Regresní model – širší uplatnění

- *Korelační koeficient* r_{xy} (výběrový):
 - popisná míra síly lineárního (přímkového) vztahu mezi dvěma proměnnými

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{yx}}{s_y s_x} = r_{yx} \in \langle -1, 1 \rangle$$

- Regresní parametr b_1 a korelační koeficient r_{yx}

$$b_1 = \frac{s_{xy}}{s_x^2}, \quad r_{yx} = \frac{s_{xy}}{s_x s_y} \implies b_1 = r_{yx} \frac{s_y}{s_x}$$

$$b_1 = 0 \iff r_{yx} = 0$$

Pro teoretické hodnoty platí:

$$\beta_1 = 0 \iff \rho_{yx} = 0$$

- **Vysvětlený a nevysvětlený součet čtverců**

$(x_i, y_i), i = 1, \dots, n$ – pozorované hodnoty X a Y

- Pro odchylky platí:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

celková = vysvětlená + nevysvětlená
odchylka odchylka odchylka

- Pro součet čtverců platí:

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

celkový = vysvětlený + nevysvětlený
součet součet součet
čtverců čtverců čtverců

- Vysvětlený součet čtverců je vysvětlen regresorem (veličinou X):

$$\sum_{i=1}^n (y_i - \bar{y})^2 = b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

celkový = součet čtverců + nevysvětlený
součet vysvětlený (reziduální)
čtverců z X součet čtverců

• Regresní identita

Celkový součet čtverců:

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

Reziduální součet čtverců:

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

Regresní součet čtverců:

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

Regresní identita:

$$S_y = S_R + S_T$$

Výpočetní vzorce pro součty čtverců

Celkový součet čtverců: $S_y = \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2/n$

Regresní součet čtverců:

$$S_T = \frac{[\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)/n]^2}{[\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n]}$$

Reziduální součet čtverců: $S_R = S_y - S_T$

- Regresní t -test $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$

Test významnosti parametru β_1 – rozhodování o užitečnosti X pro Y

Jestliže $\beta_1 = 0 \implies \eta = E(Y) = \beta_0, \quad D(Y) = \sigma^2$

$\eta, D(Y)$ nezávisí na $X \implies$

X neposkytuje žádnou informaci o rozdělení $Y \implies$
neexistuje lineární vztah mezi X a Y

Regresní t -test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$

- Testovací statistika:

$$T = \frac{b_1}{s_{b_1}} \sim t[n - 2]$$

- Kritický obor:

$$|T| > t_{1-\frac{\alpha}{2}}(n - 2)$$

s_{b_1} – směrodatná (standardní) chyba odhadu

$t_{1-\frac{\alpha}{2}}(n - 2)$ – 100(1 - $\alpha/2$)% kvantil Studentova t -rozdělení s $(n - 2)$ stupni volnosti

• Analýza rozptylu – regresní přímka

Zdroj variability	SS	Df	MS	F
Vysvětlený (regresí)	S_T	1	$\frac{S_T}{1}$	$\frac{S_T}{\frac{S_R}{n-2}}$
Nevysvětlený	S_R	$n - 2$	$\frac{S_R}{n-2}$	
Celkový	S_y	$n - 1$		

$$F = \frac{\text{rozptyl vysvětlený regresí}}{\text{nevysvětlený rozptyl}} = \frac{b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{s_R^2}$$

F-test analýzy rozptylu

– alternativní způsob testování hypotézy

$H_0 : \beta_1 = 0$ (X nemá žádný vztah k Y)

$H_1 : \beta_1 \neq 0$

○ Hladina významnosti: α

○ Testovací statistika:

$$F = \frac{\frac{S_T}{1}}{\frac{S_R}{n-2}} \sim F[1; n - 2]$$

○ Kritický obor:

$$F > F_{1-\alpha}(1; n - 2)$$

$F_{1-\alpha}(1; n - 2)$ – 100(1 - α)% kvantil Fisherova-Snedecorova rozdělení s 1 a $(n - 2)$ stupni volnosti

Rovnocenné způsoby testování hypotézy:

$$\underline{H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0}$$

- ◇ Regresní t -test
- ◇ F -test analýzy rozptylu
- ◇ Test nulovosti korelačního koeficientu $\rho = 0$

Rovnocennost regresního t -testu a F -testu:

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

- Vztah mezi F -statistikou a T -statistikou:

$$F = \frac{b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{s_R^2} = \frac{b_1^2}{s_R^2 / \sum_{i=1}^n (x_i - \bar{x})^2} = \left(\frac{b_1}{s_R} \right)^2 = T^2$$

- Vztah mezi kvantily $F[\nu_1; \nu_2]$ a $t[\nu_2]$ -rozdělení:
Pro $\nu_1 = 1$, ν_2 libovolné, α platí:

$$F_{1-\alpha}(1, \nu_2) = t_{1-\alpha/2}^2(\nu_2)$$

Výhoda t -testu – možnost sestavit IS pro β_1

• Koeficient (index) determinace

Koeficient determinace (v regresní analýze také název *index determinace* I^2)(Coefficient of determination, R-squared):

$$R^2 = \frac{\text{vysvětlený součet čtverců}}{\text{celkový součet čtverců}} = \frac{S_T}{S_y} = 1 - \frac{S_R}{S_y} \in \langle 0, 1 \rangle$$

$R = \sqrt{R^2}$ – index korelace

R^2 – charakteristika kvality regresního modelu:

- udává jakou část celkové variability lze vysvětlit zvoleným regresním modelem
- poměrné snížení celkového součtu čtverců chyb, kterého docílíme použitím regresní rovnice místo aritmetického průměru

Interpretace:

- R^2 blízké 0 naznačuje, že zvolená regresní funkce není příliš vhodná pro popis vztahu X a Y
- R^2 blízké 1 naznačuje, že regresní přímka velice dobře vystihuje vztah X a Y

8.3 Obecný regresní model

X_1, X_2, \dots, X_k – nezávislé (vysvětlující) proměnné

Y – závislá (vysvětlovaná) veličina

Regresní funkce:

$$\eta \equiv E(Y) = f(x_1, x_2, \dots, x_k; \beta_0, \beta_1, \dots, \beta_p)$$

x_1, x_2, \dots, x_k – naměřené (dané) hodnoty
veličin X_1, X_2, \dots, X_k

$\beta_0, \beta_1, \dots, \beta_p$ – *regresní parametry*

$$Y = f(x_1, x_2, \dots, x_k; \beta_0, \beta_1, \dots, \beta_p) + \epsilon = \eta + \epsilon$$

η – deterministická složka

ϵ – náhodná složka: $\epsilon \sim N[0; \sigma]$

Funkce f :

– zpravidla známá funkce

– nebo se předpokládá znalost tvaru fce

$\beta_0, \beta_1, \dots, \beta_p, \sigma$ – neznámé parametry

Dva základní typy regrese:

- *Jednoduchá regrese* - jedna nezávislá veličina ($k = 1$)

$$\eta = f(x, \beta_0, \beta_1, \dots, \beta_p)$$

- *Vícenásobná regrese* - více nezávislých veličin ($k \geq 2$)

$$\eta = f(x_1, x_2, \dots, x_k; \beta_0, \beta_1, \dots, \beta_p)$$

• Jednoduchá regrese

$$\eta = f(x, \beta_0, \beta_1, \dots, \beta_p)$$

- Lineární regresní funkce – lineární z hlediska parametrů

$$\eta = \beta_0 + \beta_1 f_1(x) + \dots + \beta_p f_p(x)$$

$\beta_0, \beta_1, \dots, \beta_p$ – neznámé regresní parametry

f_1, f_2, \dots, f_p – známé funkce nezávislé veličiny X

- ◇ Speciální případ: Modely lineární z hlediska parametrů i z hlediska vysvětlujících proměnných

$$\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

Příklady

(a) *přímková regrese*: $k = 1$, $f_1(x) = x$

$$\eta = \beta_0 + \beta_1 x$$

(b) *parabolická regrese*: $f_1(x) = x$, $f_2(x) = x^2$

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2$$

(c) *polynomická regrese p -tého stupně*:

$$f_i(x) = x^i, \forall i = 1, 2, \dots, p$$

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p$$

(d) *hyperbolická regrese*: $f_1(x) = x^{-1}$

$$\eta = \beta_0 + \frac{\beta_1}{x}$$

(e) *hyperbolická regrese p-tého stupně*:

$$f_i(x) = x^{-i}, \forall i = 1, 2, \dots, p$$

$$\eta = \beta_0 + \frac{\beta_1}{x} + \frac{\beta_2}{x^2} + \dots + \frac{\beta_p}{x^p}$$

(e) *logaritmická regrese*: $k = 1$, $f_1(x) = \log x$

$$\eta = \beta_0 + \beta_1 \log x$$

- Nelineární regresní funkce – nelineární z hlediska parametrů

Příklady

(α) *exponenciální regrese p-tého stupně*

$$\eta = \beta_0 \beta_1^{f_1(x)} \beta_2^{f_2(x)} \dots \beta_p^{f_p(x)}$$

(β) *exponenciální regrese prvního stupně*:

$$p = 1, f_1(x) = x$$

$$\eta = \beta_0 \beta_1^x$$

(γ) *mocninná regrese*

$$\eta = \beta_0 x^{\beta_1}$$

• Bodové odhady regresních parametrů

y_1, y_2, \dots, y_n – n nezávislých pozorování veličiny Y

$x_{1j}, x_{2j}, \dots, x_{nj}$ – dané hodnoty X_j , $j = 1, 2, \dots, k$.

Metoda nejmenších čtverců:

$$\min_{\beta_0, \dots, \beta_p} \sum_{i=1}^n [y_i - f(x_{1i}, x_{2i}, \dots, x_{ki}; \beta_0, \beta_1, \dots, \beta_p)]^2$$

$$b_0 = \hat{\beta}_0, b_1 = \hat{\beta}_1, \dots, b_p = \hat{\beta}_p$$

Řešení:

– v případě regresních funkcí, které nejsou lineární z hlediska parametrů – MNČ vede na soustavu nelineárních rovnic – iterační algoritmy

– použití vhodné transformace

Příklad: převedení pomocí logaritmické transformace

$$Y = \beta_0 \beta_1^{f_1(x)} \beta_2^{f_2(x)} \dots \beta_p^{f_p(x)}$$

na

$$Y = \beta_0 + \beta_1 f_1(x) + \dots + \beta_p f_p(x)$$

8.4 Vícenásobná regrese a korelace

Vícenásobná lineární regrese

- **Klasický lineární regresní model**

K1. Tvar regresní funkce:

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon = \eta + \epsilon$$

K2. X_1, X_2, \dots, X_p – nenáhodné, neexistuje mezi nimi lineární funkční vztah

$x_{j1}, x_{j2}, \dots, x_{jn}$ – dané hodnoty proměnné X_j ,
 $j = 1, 2, \dots, p$

K3. Rozdělení náhodné složky: $\epsilon \sim N[0; \sigma^2]$

K4. y_1, y_2, \dots, y_n – pozorované hodnoty náh. veličin

Y_1, Y_2, \dots, Y_n

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i$$

$$\epsilon_i \sim N[0; \sigma^2]$$

$$\text{cov}(\epsilon_i \epsilon_j) = 0 \quad \forall i \neq j, \quad i, j = 1, 2, \dots, n$$

Odhadnutá regresní funkce:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 \dots + b_p x_p$$

nebo

$$\hat{y} = b_0 + b_{yx_1 \cdot x_2 \dots x_p} x_1 + b_{yx_2 \cdot x_1 \dots x_p} x_2 \dots + b_{yx_p \cdot x_1 x_2 \dots x_{p-1}} x_p$$

Parciální (dílčí) regresní koeficienty:

$$b_{yx_1 \cdot x_2 \dots x_p}, b_{yx_2 \cdot x_1 \dots x_p}, \dots, b_{yx_p \cdot x_1 x_2 \dots x_{p-1}}$$

Interpretace parciálních regresních koeficientů:

- charakteristiky k posouzení individuálního vlivu jednotlivých vysvětlujících proměnných na závislou proměnnou
- udávají odhad toho, jak se změnila v průměru závislá proměnná Y při jednotkové změně nezávisle proměnné před tečkou, za předpokladu konstantní úrovně proměnných uvedených za tečkou.

• Regresní rovina (p=2) - (dvojnásobná r.)

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

y_1, y_2, \dots, y_n – n nezávislých pozorování veličiny Y

$x_{1j}, x_{2j}, \dots, x_{nj}$ – dané hodnoty X_j , $j = 1, 2$

Metoda nejmenších čtverců:

$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=1}^n [y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}]^2$$

Ze soustavy normálních rovnic dostaneme:

$$\bar{y} = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2$$

$$b_0 = \hat{\beta}_0, b_1 = \hat{\beta}_1, b_2 = \hat{\beta}_2$$

Odhadnutá regresní funkce:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

nebo

$$\hat{y} = b_0 + b_{yx_1 \cdot x_2} x_1 + b_{yx_2 \cdot x_1} x_2$$

$$\hat{y} = \bar{y} + b_{yx_1 \cdot x_2} (x_1 - \bar{x}_1) + b_{yx_2 \cdot x_1} (x_2 - \bar{x}_2)$$

- **B-koeficienty** (*Beta Coefficients*)

- normalizované regresní koeficienty (bezrozměrné charakteristiky)

Důvod zavedení: hodnoty parciálních korelačních koeficientů závisí na jednotkách, v jakých jsou vyjádřeny jednotlivé proměnné.

Použití: pro srovnání a posouzení individuálního vlivu jednotlivých regresorů na závisle proměnnou.

- Transformace:

$$\hat{y}'_i = \frac{\hat{y}_i - \bar{y}}{s_y}, \quad x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}}, \quad i = 1, 2, \dots, n, \quad j = 1, 2$$

- Odhadnutá regresní funkce pro $p = 2$:

$$\hat{y}' = B_{yx_1 \cdot x_2} x'_1 + B_{yx_2 \cdot x_1} x'_2$$

$B_{yx_1 \cdot x_2}, B_{yx_2 \cdot x_1}$ – **B-koeficienty**

- Odhady B-koeficientů:

- ◇ MNČ

- ◇ Výpočet z dílčích regresních koeficientů

$$B_{yx_1 \cdot x_2} = \frac{s_{x_1}}{s_y} b_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}$$

$$B_{yx_2 \cdot x_1} = \frac{s_{x_2}}{s_y} b_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}$$

Kvalita a intenzita vícenásobné lineární závislosti

Míry těsnosti závislosti Y na X_1, X_2, \dots, X_p

- *Koeficient dílčí korelace (výběrový)* $r_{yx_1 \cdot x_2 \dots x_p}$
– míra intensity lineární závislosti y na x_1 při konstantních x_2, \dots, x_p

$r_{yx_1 \cdot x_2 \dots x_p}$ – odhad $\rho_{yx_1 \cdot x_2 \dots x_p}$

$p = 2$:

$$|r_{yx_1 \cdot x_2}| = \sqrt{b_{yx_1 \cdot x_2} b_{yx_2 \cdot x_1}}$$

Rekurentní vzorce pro výpočet $r_{yx_1 \cdot x_2}$ a $r_{yx_2 \cdot x_1}$:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1 x_2}^2)}}$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{x_1 x_2}^2)}}$$

$p \geq 2$:

$$r_{yx_1 \cdot x_2 \dots x_p} = \frac{r_{yx_1 \dots x_{p-1}} - r_{yx_p \cdot x_2 x_3 \dots x_{p-1}} r_{x_1 x_p \cdot x_2 x_3 \dots x_{p-1}}}{\sqrt{(1 - r_{yx_p \cdot x_2 x_3 \dots x_{p-1}}^2)(1 - r_{x_1 x_p \cdot x_2 x_3 \dots x_{p-1}}^2)}}$$

- *Koeficient vícenásobné korelace (výběrový)*

$$r_{y \cdot x_1 x_2 \dots x_p}$$

– míra těsnosti lineární závislosti y na všech x_1, x_2, \dots, x_p dohromady

$p = 2$:

$$r_{y \cdot x_1 x_2} = \sqrt{\frac{r_{yx_1}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2} + r_{yx_2}^2}{1 - r_{x_1x_2}^2}}$$

Platí:

$$\diamond 0 \leq r_{y \cdot x_1 x_2 \dots x_p} \leq 1$$

$$\diamond r_{y \cdot x_1 x_2 \dots x_p} > \max_{j=1,2,\dots,p} r_{yx_j}$$

$r_{y \cdot x_1 x_2 \dots x_p}$ – odhad teoretického koef. vícenásobné korelace $\rho_{y \cdot x_1 x_2 \dots x_p}$

- *Koeficient vícenásobné determinace:*

$$R^2 = \frac{\text{vysvětlený součet čtverců}}{\text{celkový součet čtverců}} = \frac{S_T}{S_y} = 1 - \frac{S_R}{S_y} \in \langle 0, 1 \rangle$$

- *Upravený (korigovaný) koeficient determinace (Adjusted R-squared):*

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

(bere v úvahu počet parametrů p a rozsah n)

Intervaly spolehlivosti a testy hypotéz v regresi a korelaci

- **IS pro regresní parametry**

- **Koeficient spolehlivosti:** $(1 - \alpha)$

- **Bodové odhady β_j :** $b_j, j = 0, 1, \dots, p$

- **Krajní body $100(1 - \alpha)\%$ IS:**

$$b_j \pm t_{1-\frac{\alpha}{2}}(n - p - 1) s_{b_j}$$

s_{b_j} – směrodatná (standardní) chyba odhadu

$t_{1-\frac{\alpha}{2}}(n - p - 1)$ – $100(1 - \frac{\alpha}{2})\%$ kvantil t -rozdělení o $(n - p - 1)$ stupních volnosti.

- **IS pro koeficienty korelace**

- *Párový korelační koeficient* ρ_{yx} :

- (a) ρ_{yx} se málo liší od nuly, $n > 100$

- ◊ **Koeficient spolehlivosti:** $(1 - \alpha)$

- ◊ **Bodový odhad** ρ_{yx} : r_{yx}

- ◊ **Krajní body** $100(1 - \alpha)\%$ **IS:**

$$r_{yx} \pm u_{1-\frac{\alpha}{2}} \frac{1 - r_{yx}^2}{\sqrt{n}}$$

$u_{1-\frac{\alpha}{2}}$ – $100(1 - \alpha/2)\%$ kvantil $\mathcal{N}(0; 1)$

- (b) $\rho_{yx} > 0,5$ n malé

- ◊ **Fisherova transformace:**

$$z_r = \frac{1}{2} \ln \frac{(1 + r_{yx})}{(1 - r_{yx})}, \quad Z_r \approx \mathcal{N}(E(Z_r), D(Z_r))$$

$$E(Z_r) = \frac{1}{2} \ln \frac{(1 + \rho_{yx})}{(1 - \rho_{yx})} + \frac{\rho_{yx}}{2(n - 1)}, \quad D(Z_r) = \frac{1}{n - 3}$$

- ◊ **Krajní body IS** ($\rho_{yx}/[2(n - 1)]$ lze zanedbat):

$$z_r \pm u_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{n - 3}}$$

- *Parciální koeficient korelace:* **IS** nemají praktické využití

• Testy hypotéz o regresních parametrech

Test: $H_0 : \beta_j = \beta_{0j} \quad \forall j = 0, 1, \dots, p$ versus

a) $H_1 : \beta_j \neq \beta_{0j}$ b) $H_1 : \beta_j > \beta_{0j}$ c) $H_1 : \beta_j < \beta_{0j}$

○ Testovací statistika:

$$\diamond T = \frac{b_j - \beta_{0j}}{s_{b_j}} \sim t[n - p - 1]$$

$$\diamond U = \frac{b_j - \beta_{0j}}{s_{b_j}} \approx \mathcal{N}(0; 1) \quad n - p > 30$$

○ Kritické obory:

$$\begin{array}{ll} \text{a)} & |T| > t_{1-\frac{\alpha}{2}}(n - p - 1) & |U| > u_{1-\frac{\alpha}{2}} \\ \text{b)} & T > t_{1-\alpha}(n - p - 1) & U > u_{1-\alpha} \\ \text{c)} & T < t_{\alpha}(n - p - 1) & U < u_{\alpha} \end{array}$$

• Celkový F-test o modelu

zdroj variability	SS	DF	MS	F
regresní	S_T	p	$\frac{S_T}{p}$	$\frac{S_T/p}{S_R/(n-p-1)}$
reziduální	S_R	$n - p - 1$	$\frac{S_R}{n-p-1}$	
celkový	S_y	$n - 1$		

$p + 1$ – počet regresních parametrů

p – počet vysvětlujících veličin

Celkový F-test:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1 : \text{alespoň jeden regresní parametr } \beta_j \neq 0$$

◇ Hladina významnosti: α

◇ Testovací statistika:

$$F = \frac{\frac{S_T}{p}}{\frac{S_R}{(n-p-1)}} \sim F[p; n - p - 1]$$

◇ Kritický obor:

$$F > F_{1-\alpha}(p; n - p - 1)$$

$F_{1-\alpha}(p; n - p - 1)$ – 100(1 - α)% kvantil Fisherova-Snedecorova rozdělení s p a $(n-p-1)$ stupni volnosti.

• Testy hypotéz o korelačních koeficientech

◦ Párový korelační koeficient ρ_{yx} :

Y a X lineárně nezávislé: $\rho_{yx} = 0$

Test: $H_0 : \rho_{yx} = 0$ versus

a) $H_1 : \rho_{yx} \neq 0$ b) $H_1 : \rho_{yx} > 0$ c) $H_1 : \rho_{yx} < 0$

◊ Testovací statistika:

$$\triangleleft T = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \sqrt{n-2} \sim t[n-2]$$

$$\triangleleft U = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \sqrt{n-2} \approx \mathcal{N}(0; 1) \quad n > 30$$

◊ Kritické obory:

$$\text{a) } |T| > t_{1-\frac{\alpha}{2}}(n-2) \quad |U| > u_{1-\frac{\alpha}{2}}$$

$$\text{b) } T > t_{1-\alpha}(n-2) \quad U > u_{1-\alpha}$$

$$\text{c) } T < t_{\alpha}(n-2) \quad U < u_{\alpha}$$

Test: $H_0 : \rho_{yx} = \rho_0$ ($\rho_0 \in (-1, 1)$) versus

a) $H_1 : \rho_{yx} \neq \rho_0$ b) $H_1 : \rho_{yx} > \rho_0$ c) $H_1 : \rho_{yx} < \rho_0$

◊ Testovací statistika:

$$U = |Z_r - z_{\rho_0}| \sqrt{n-3} \sim \mathcal{N}(0; 1)$$

◊ Kritické obory:

$$\text{a) } |U| > u_{1-\frac{\alpha}{2}} \quad \text{b) } U > u_{1-\alpha} \quad \text{c) } U < u_{\alpha}$$

- Koeficient dílčí korelace $\rho_{yx_1 \cdot x_2 \dots x_p}$:

Test:

$$H_0 : \rho_{yx_1 \cdot x_2 \dots x_p} = 0 \text{ versus } \text{non}H_0$$

- ◇ Testovací statistika:

$$T = \frac{r_{yx_1 \cdot x_2 \dots x_p} \sqrt{n - p - 1}}{\sqrt{1 - r_{yx_1 \cdot x_2 \dots x_p}^2}} \sim t[n - p - 1]$$

- ◇ Kritický obor:

$$|T| > t_{1-\alpha/2}(n - p - 1)$$

- Koeficient vícenásobné korelace $\rho_{y \cdot x_1 x_2 \dots x_p}$:

Test:

$$H_0 : \rho_{y \cdot x_1 x_2 \dots x_p} = 0 \text{ versus } H_1 : \rho_{y \cdot x_1 x_2 \dots x_p} > 0$$

- ◇ Testovací statistika:

$$F = \frac{r_{y \cdot x_1 x_2 \dots x_p}^2 (n - p - 1)}{(1 - r_{y \cdot x_1 x_2 \dots x_p}^2) p} \sim F[p; n - p - 1]$$

- ◇ Kritický obor:

$$F > F_{1-\alpha}(p; n - p - 1)$$

Korelační analýza a regresní model

- Výběr nezávislých veličin v regresním modelu

Multikolinearita – závislost mezi nezávislými (vysvětlujícími proměnnými, regresory)

Matice párových korelačních koeficientů \mathbf{R} :

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ r_{31} & r_{32} & \dots & r_{3p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

$$r_{ij} \equiv r_{x_i x_j} \quad i, j = 1, 2, \dots, p$$

Indikátor multikolinearity: det R

- neexistuje multikolinearita – v praxi vzácné

$$r_{ij} = 0 \quad \forall i \neq j, \quad i, j = 1, 2, \dots, p \implies \det \mathbf{R} = 1$$

- multikolinearita:

$$r_{ij} \neq 0 \quad \forall i \neq j, \quad i, j = 1, 2, \dots, p \implies 0 \leq \det \mathbf{R} < 1$$

- ◇ úplná multikolinearita – v praxi vzácné

$$\det \mathbf{R} = 0$$

$\det \mathbf{R} = 0 \implies$ alespoň jeden $r_{ij} = 1$

◁ Neexistuje řešení MNČ

◁ Interpretace: alespoň jeden $r_{ij} = 1 \implies$ všechny hodnoty jedné z vysvětlujících proměnných jsou stejným nenulovým násobkem některé jiné vysvětlující proměnné

◁ Důsledek: přidávání dalších vysvětlujících proměnných do modelu není účelné

- ◇ Multikolinearitu považujeme za vysokou:

$$|r_{ij}| > 0,75$$

alespoň pro jeden korelační koeficient

- Určení nejlepší podmnožiny regresorů v regresním modelu

Zařazujeme pouze regresory, které výrazně zlepší odhad modelu tak, aby model nebyl zbytečně složitý.

Sekvenční F -test

– ověření správnosti přidání $(k + 1)$ -ní nezávislé proměnné (regresoru) do modelu

○ Sekvenční F -test:

$$H_0 : \beta_{k+1} = 0$$

$(x_{k+1}$ nepřispívá k vysvětlení variability y)

$$H_1 : \beta_{k+1} \neq 0$$

$(x_{k+1}$ přispívá k vysvětlení variability y)

◇ Testovací statistika:

$$F = \frac{S_{T(k+1)} - S_{T(k)}}{\frac{S_R}{n-k-2}} \sim F[1; n - k - 2]$$

$S_{T(k+1)} - S_{T(k)}$ – přírůstek regresního součtu čtverců S_T po přidání $(k + 1)$ -ní proměnné do modelu

S_R – reziduální součet čtverců v modelu s $(k + 1)$ regresory

◇ Kritický obor:

$$F > F_{1-\alpha}(1; n - k - 2)$$

○ Metoda Stepwise (krokovací metoda)

1. metoda dopředná (forward)

- postupné přidávání přínosných regresorů do modelu

2. metoda zpětná (backward)

- postupné odstraňování nepřínosných regresorů z modelu

Maticový přístup k lineární regresi

Regresní model lineární v parametrech i v nezávislých proměnných

Předpoklady:

M1. (Y_1, Y_2, \dots, Y_n) – náhodné veličiny

M2. \mathbf{X} – matice daných čísel $(n \times (p + 1))$, $p + 1 < n$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

M3. Pro náhodný vektor $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ platí:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

$\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ – vektor neznámých parametrů

$\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ – vektor náhodných veličin:

$$E(\epsilon) = \mathbf{0}, \quad \Sigma_\epsilon = \sigma^2 \mathbf{I}$$

○ **Odhady regresních parametrů β**

$\mathbf{X}\beta$ – nenáhodný vektor

Z M3. $\implies E(\mathbf{Y}) = \mathbf{X}\beta, \quad \Sigma_Y = \sigma^2\mathbf{I}$

$\mathbf{b} = (b_0, b_1, \dots, b_p)^\top$ – odhad $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$

Předpoklady:

○ $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$ – pozorovaná hodnota \mathbf{Y}

○ $h(\mathbf{X}) = p + 1 \implies \mathbf{X}^\top\mathbf{X}$ – regulární matice

Metoda nejmenších čtverců: minimalizace

$$S(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

Řešení:

$$\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

$E(\mathbf{b}) = \beta$ – \mathbf{b} nestranný odhad β

$\Sigma_b = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ – kovarianční matice β

Příklad (vícenásobná lineární regrese)

Byly sledovány *výdaje* Y (v tisících) za potraviny a nápoje u jednotlivých domácností v závislosti na *počtu členů* domácnosti X_1 a na *celkovém čistém příjmu* domácnosti X_2 (v tisících).

V tabulce jsou uvedeny údaje o 7 náhodně vybraných domácnostech.

Výdaje (Y)(v tisících)	12	9	12	3	18	12	15
Počet členů (X_1)	4	2	4	1	5	3	4
Čistý příjem (X_2) (v tisících)	30	24	36	9	45	24	39

- Určete regresní rovnici závislosti výdajů za potraviny a nápoje na 2 uvažovaných regresorech.*
- Který regresor má větší vliv na výdaje za potraviny a nápoje?*
- Vypočítejte parciální korelační koeficient mezi výdaji za potraviny a nápoje a čistými příjmy domácností při konstantním počtu členů domácnosti.*
- Vypočítejte parciální korelační koeficient mezi výdaji za potraviny a nápoje a počtem členů domácnosti při konstantní výši čistého příjmu domácnosti.*
- Pomocí metody stepwise-forward vyberte vhodnou podmnožinu regresorů (nezávislých proměnných).*
- Pomocí metody stepwise-backward vyberte vhodnou podmnožinu regresorů (nezávislých proměnných).*