Teorie chyb a vyrovnávací počet 1

Téma č. 11: **Regresní a korelační analýza 1. Lineární regrese.**

1. **Regresní a korelační analýza.**
    1. **Princip, výpočet.**
    2. **Funkční vztah, korelační závislost, stochastický (statistický) vztah.**
2. **Lineární regrese**
3. **Korelační koeficient**
    1. **Výpočet z linární regrese**
    2. **Výpočet z kovarianční matice**
    3. **Interpretace koeficientu korelace**
4. **Jiné druhy korelace – pořadová korelace**
    1. **Spermanův koeficient**
    2. **Kendallův koeficient**

1. Regresní a korelační analýza.

V přírodě často probíhají jevy jako funkce jedné nebo více proměnných $y = f(x)$, $z = f(x, y)$, atd., u nichž předem neznáme přesně typ a konstanty (parametry) funkce a teprve je zjišťujeme empiricky.

K empirickému určení analytického typu funkce a číselných hodnot konstant sledujeme průběh jevu měřením hodnot závisle proměnné y při měnících se hodnotách argumentu x. Grafické znázornění průběhu jevu dá vlivem měřických chyb nebo jiných rušivých vlivů nepravidelnou řadu bodů (empirický polygon). Úkolem je najít takovou funkční závislost mezi proměnnými x a y, aby průběh funkce co nejlépe vyjadřoval měřený průběh jevu, tj. aby se vyrovnávací křivka při jednoduchém tvaru funkce dostatečně přimkla empirickému polygonu. Zpravidla k tomu použijeme metodu nejmenších čtverců.

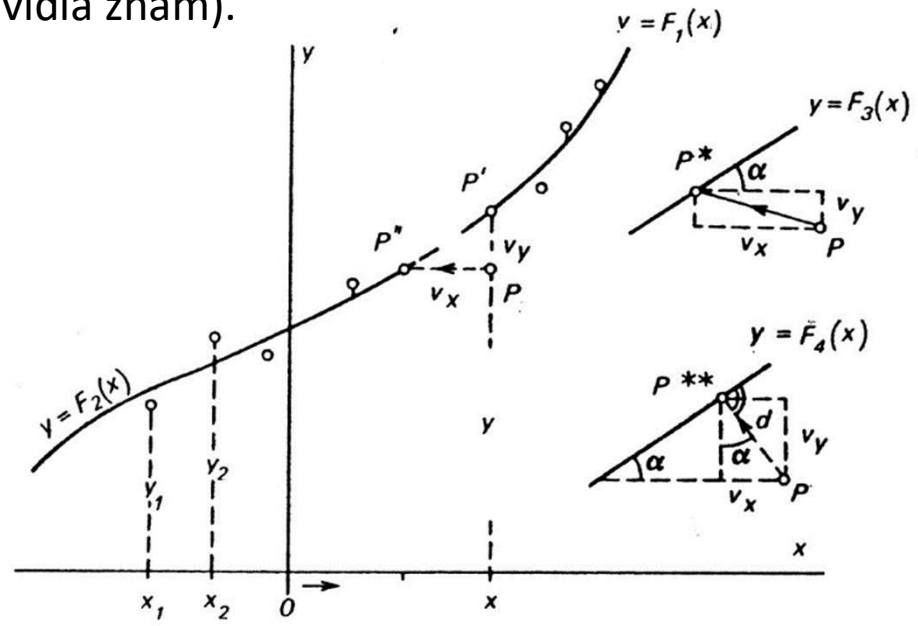
Měřické chyby nebo rušivé vlivy a neznalost přesného analytického typu funkce způsobují, že typ funkce a její konstanty neurčíme s absolutní přesností. Mluvíme proto o aproximaci empirických funkcí a výsledkem bude tzv. regresní křivka.

Druhou, neméně důležitou otázkou, kterou řešíme při hledání aproximačních funkcí, je síla (věrohodnost) jejich platnosti. Regresní analýza sice určí tvar funkce, ale korelační analýza určí věrohodnost platnosti tohoto vztahu v rozmezí absolutní nezávislosti až po funkční vztah dvou veličin.

1. Regresní a korelační analýza.

Měřením n sdružených dvojic x_i, y_i dostaneme k naměřeným argumentům $x_1, \ldots, x_n$, naměřené hodnoty jejich funkce $y_1, \ldots, y_n$. V grafu vlivem měřických chyb nebo jiných rušivých vlivů dostaneme nepravidelnou řadu měřených hodnot y (empirický polygon), kde však pozorujeme určitou celkovou tendenci nebo periodicitu. Ke každému bodu přísluší dvojice proměnných (x_i, y_i).

Vyrovnávací křivka $y = F(x)$ je spojitá a zpravidla prochází mezi body polygonu. Zbylé odstupy bodů od křivky jsou opravy. jsou produktem především měřických chyb a rušivých vlivů a v druhé řadě i použitého nepřesného typu funkce F (teoreticky přesný typ funkce nebývá zpravidla znám).



1. Regresní a korelační analýza.

K určení parametrů hledané funkce měříme zpravidla nadbytečný počet hodnot a zpracování provádíme obvykle podle MNČ vyrovnání měření zprostředkujících. Měřené hodnoty zprostředkujících veličin zde budeme označovat x, y. Při aproximaci funkce podle MNČ hledáme vyrovnané hodnoty (nejspolehlivější odhady) A, B, ... neznámých konstant funkce $\bar{A}$, $\bar{B}$, … . Jsou tři základní varianty řešení.

a) Uvažujeme jen opravy v_y k hodnotě funkce y a vyrovnání provedeme za podmínky $[pvv]_y = min.$ (jako by měřickými chybami byly zatíženy jen měřené hodnoty y a hodnoty x byly bezchybné). Vyrovnávací křivka a příslušné opravy budou mít rovnice a podmínku:

$y_i + v_{y_i} = F_1(x_i)$, $\quad v_{y_i} = F_1(x_i) - y_i$, $\quad v_{x_i} = 0$, $[pvv] = min.$

b) Uvažujeme jen opravy v_x argumentu x a vyrovnání provedeme za podmínky $[pvv]_x = min.$ (jako by chybami byly zatíženy jen měřené argumenty x_i a hodnoty y_i byly bezchybné). Konstanty např. A, B,...vyrovnávací křivky určíme z uvedené podmínky po odvození příslušných normálních rovnic:

$y_i = F_2(x_i + v_{x_i})$, $\quad y_i = F_2(x_i) + \frac{\partial F_2}{\partial(x_i + v_{x_i})} \cdot v_{x_i}$, $v_{y_i} = 0$.

1. Regresní a korelační analýza.

c) Uvažujeme opravy jak hodnot funkce y, tak i hodnot argumentu x a konstanty vyrovnávací křivky najdeme za podmínky

$$[pvv]_x + [pvv]_y = min., \qquad v_x \neq cotg(\alpha) \cdot v_y.$$

Tato varianta je teoreticky nejsprávnější, jestliže rozptyl empirického polygonu kolem vyrovnávací křivky způsobily měřické chyby v obou skupinách zprostředkujících veličin. Rovnice vyrovnávací křivky bude nyní

$$y_i + v_{y_i} = F_3\left(x_i + v_{x_i}\right) = F_3(x_i) + f_{x_i} \cdot v_{x_i} .$$

Opravy v_x, v_y budou nyní složkami vzdálenosti $\overline{PP^*}$ naměřené polohy bodu P od vyrovnané polohy na křivce.

Poznámka: Každá z uvedených tří variant dá jiné číselné hodnoty konstant vyrovnávací křivky (jiný její průběh). Rozdíly budou záviset na velikosti korelačního koeficientu. Všechny uvedené podmínky minima budou blíže vysvětleny na příkladu vyrovnávací přímky.

1. Regresní a korelační analýza.

Funkční vztah

– vztah dvou (či více) veličin x a y lze popsat matematickou funkcí, od které nevykazují žádné odchylky. Pro jednu hodnotu x existuje dle předpisu pouze jedna hodnota y.

Korelační závislost, stochastický (statistický) vztah

– mezi veličinami (měřeními) je vztah, nelze jej však beze zbytku vyjádřit.

Postup vysvětlíme na lineární regresi, která je nejjednodušší a také nejpoužívanější.
zároveň definuje mnohé pojmy používané obecně.

2. Lineární regrese.

Měřené hodnoty y jsou zatíženy chybami:

$$y_i + v_{y_i} = A_y + B_y \cdot x_i \,, \qquad v_{y_i} = A_y + B_y \cdot x_i - y_i \,, \qquad [pvv]_y = min, v_{x_i} = 0.$$

Normální rovnice:

$$\begin{aligned} [p] \cdot A_y \quad &+[px] \cdot B_y \quad -[py] \quad = 0 \\ [px] \cdot A_y \quad &+[pxx] \cdot B_y \quad -[pxy] \quad = 0 \end{aligned}$$

Kovarianční matice:

$$\sigma_0^2 \cdot \begin{pmatrix} \frac{[pxx]}{D} & \frac{[px]}{D} \\ \frac{[px]}{D} & \frac{[p]}{D} \end{pmatrix} \qquad D = [p] \cdot [pxx] - [px]^2$$

Dělíme-li první normální rovnici hodnotou $[p]$ a druhou hodnotou $[px]$, dostaneme:

$A_y + \frac{[px]}{[p]} \cdot B_y - \frac{[py]}{[p]} = 0 \,, \quad A_y + \frac{[pxx]}{[px]} \cdot B_y - \frac{[pxy]}{[px]} = 0 \,.$

takže vyrovnávací přímka prochází těžištěm bodů T a tzv. těžištěm těžkých bodů U:

$x_T = \frac{[px]}{[p]}, y_T = \frac{[py]}{[p]}, \qquad x_U = \frac{[pxx]}{[px]}, y_U = \frac{[pxy]}{[px]}.$

2. Lineární regrese.

Po redukci na těžiště dojde ke zjednodušení:

$x' = x - x_T$, $y' = y - y_T$ a tedy $[px'] = [py'] = 0$

$y'_i + v_{y_i} = B_y \cdot x'_i$, $\quad v_{y_i} = B_y \cdot x'_i - y'_i$, $\quad A'_y = 0$,

$B_y = \frac{[px'y']}{[px'x']}$ $\quad \sigma_{By} = \frac{\sigma_0}{\sqrt{[px'x']}}$

2. Lineární regrese.

Měřené hodnoty x jsou zatíženy chybami ($B = 1/B^*$, $A = -A^*/B^*$):

$$x_i + v_{x_i} = A_x^* + B_y^* \cdot y_i, \qquad [pvv]_x = min, \qquad v_{y_i} = 0.$$
$$v_{x_i} = A_x^* + B_x^* \cdot y_i - x_i$$

Normální rovnice:

$$\begin{aligned} [p] \cdot A_x^* \quad &+[py] \cdot B_x^* \quad -[px] \quad = 0 \\ [py] \cdot A_x^* \quad &+[pyy] \cdot B_x^* \quad -[pxy] \quad = 0 \end{aligned}$$

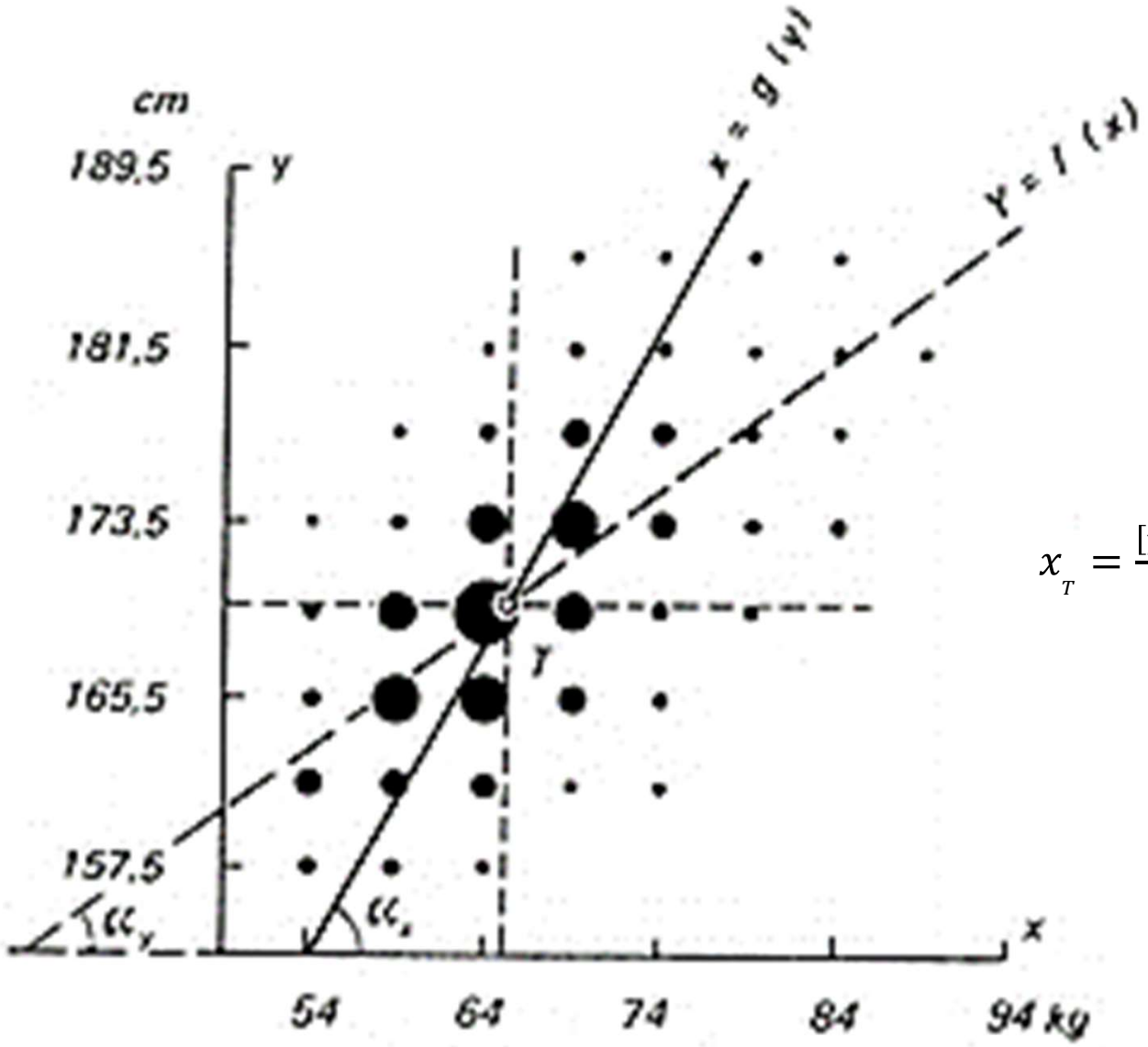
Kovarianční matice:

$$\sigma_0^2 \cdot \begin{pmatrix} \frac{[py\ \ ]}{D} & \frac{[py]}{D} \\ \frac{[py]}{D} & \frac{[p]}{D} \end{pmatrix} \qquad D = [p] \cdot [pyy] - [py]^2$$

Po redukci dtto.

$$B_x^* = \frac{[px'y']}{[py'y']}$$

3. Korelační koeficient.



$$x_T = \frac{[px]}{[p]},\ y_T = \frac{[py]}{[p]}.$$

3. Korelační koeficient.

Výpočet z lineární regrese.

$$B_y = \frac{[px'y']}{[px'x']}, B_x = \frac{[py'y']}{[px'y']}.$$

Koeficient korelace

$$r = \sqrt{\frac{tg(\alpha_y)}{tg(\alpha_x)}} = \sqrt{\frac{B_y}{B_x}} = \sqrt{B_y \cdot B_x^*}$$

(Koeficient korelace je odmocnina z podílu směrnic obou regresních přímek nebo geometrický průměr obou koeficientů regrese.)

$$r = \frac{[px'y']}{\sqrt{[px'x']\cdot[py'y']}}$$

3. Korelační koeficient.

Výpočet z kovarianční matice.

$$\boldsymbol{M} = \begin{pmatrix} \sigma_{x1}^2 & Cov_{12} & Cov_{13} & \dots & Cov_{1n} \\ Cov_{21} & \sigma_{x2}^2 & Cov_{23} & \dots & Cov_{2n} \\ Cov_{31} & Cov_{32} & \sigma_{x3}^2 & \dots & Cov_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov_{1n} & Cov_{2n} & Cov_{3n} & \dots & \sigma_{xn}^2 \end{pmatrix}$$

$$r_{ij} = \frac{Co_{ij}}{\sqrt{\sigma_{xi}^2 \cdot \sigma_{xj}^2}}$$

(X $r = \frac{[px'y']}{\sqrt{[px'x'][py'y']}}$)

3. Korelační koeficient.

Interpretace koeficientu korelace.

a) $r = 0$: mezi oběma proměnnými není lineární vztah (korelace, $E(x' \cdot y') = 0$). Součet $[x' \cdot y'] = 0$. Obě regresní přímky jsou vzájemně kolmé a ztotožňují se s osami souřadnic x', y' (jsou rovnoběžné s osami x, y).

b) $|r| = 1$: stochastický vztah mezi oběma proměnnými přechází v lineární funkční vztah ($[x'y']^2 = [x'x'].[y'y']$) a obě regresní přímky splynou v jedinou.

c) $0 < |r| < 1$: Jestliže se rozptylový obrazec soustřeďuje do I. a III. kvadrantu, procházejí regresní přímky těmito kvadranty, hodnota $[x'y']$ nabývá kladné hodnoty, $0 < r < +1$ a mluvíme o kladné korelaci. Jestliže se naopak rozptylový obrazec soustředí do II. a IV. kvadrantu, budou v nich i regresní přímky, $[x'y']$ budenabývat záporné hodnoty, $0 > r > -1$ a mluvíme o záporné korelaci.

Lze odvodit, že platí

$$r = \sqrt{1 - \frac{[vv]_y}{[y'y']}} = \sqrt{1 - \frac{[vv]_x}{[x'x']}}, \qquad \left([vv]_y = [y'y'] - \frac{[x'y']^2}{[x'x']}\right)$$

3. Korelační koeficient.

Interpretace koeficientu korelace.

a) $r = 0$: mezi oběma proměnnými není lineární vztah (korelace, $E(x' \cdot y') = 0$). Součet $[x' \cdot y'] = 0$. Obě regresní přímky jsou vzájemně kolmé a ztotožňují se s osami souřadnic x', y' (jsou rovnoběžné s osami x, y).

b) $|r| = 1$: stochastický vztah mezi oběma proměnnými přechází v lineární funkční vztah ($[x'y']^2 = [x'x'].[y'y']$) a obě regresní přímky splynou v jedinou.

c) $0 < |r| < 1$: Jestliže se rozptylový obrazec soustřeďuje do I. a III. kvadrantu, procházejí regresní přímky těmito kvadranty, hodnota $[x'y']$ nabývá kladné hodnoty, $0 < r < +1$ a mluvíme o kladné korelaci. Jestliže se naopak rozptylový obrazec soustředí do II. a IV. kvadrantu, budou v nich i regresní přímky, $[x'y']$ budenabývat záporné hodnoty, $0 > r > -1$ a mluvíme o záporné korelaci.

Lze odvodit, že platí

$$r = \sqrt{1 - \frac{[vv]_y}{[y'y']}} = \sqrt{1 - \frac{[vv]_x}{[x'x']}}, \qquad \left([vv]_y = [y'y'] - \frac{[x'y']^2}{[x'x']}\right)$$

4. Jiné druhy korelace – pořadová korelace.

Sledujme na některém jevu dva kvalitativní znaky A, B. Jejich číselné hodnoty x, y nemůžeme registrovat, ale pouze seřadit podle určité zvolené stupnice (např. od nejlepšího k nejhoršímu). Tudíž máme každý jev s odpovídajícím číslem pořadí kvality podle znaku $A(x)$ a $B(y)$. Jevy seřadíme podle pořadí jevu A tj. x bude 1, 2, ..., n. Hodnoty pořadí y budou přeházeny různě. Charakteristiky vazby, které používají pro hodnocení takto seřazená pozorování, se nazývají pořadové (rangové) koeficienty korelace. Uvedeme dva nejužívanější:

a) Spearmanův koeficient korelace τ_c se počítá ze vzorce

$$\tau_c = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n},$$

kde n je rozsah výběru, d_i rozdíl pořadí patřící k jevu i.

b) Kendallův koeficient korelace τ_k se počítá ze vzorce

$$\tau_k = \frac{S}{n \cdot \frac{(n-1)}{2}}, \qquad S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} v_{ij},$$

kde $v_{ij} = -1$, jestliže $y_j < y_i$ a $v_{ij} = +1$, jestliže $y_j > y_i$.

4. Jiné druhy korelace – pořadová korelace.

Příklad:
Provedli jsme 10 pozorování znaků A, B na některém jevu. Výsledky jsme porovnali, určili pořadí a zapsali do tabulky:

i	x_i	y_i	d_i	d^2	ν_i
1	1	2	-1	1	7
2	2	3	-1	1	6
3	3	1	2	4	7
4	4	4	0	0	6
5	5	6	-1	1	3
6	6	5	1	1	4
7	7	9	-2	4	-1
8	8	7	1	1	2
9	9	8	1	1	1
10	10	10	0	0	-
Σ				**14**	**35**

Koeficient τ_c vyšel +0,915, což svědčí o významné kladné pořadové vazbě mezi sledovanými znaky, tj. zvýšení kvality jednoho znaku je zpravidla provázeno zvýšením kvality i druhého znaku.

Pro Kendallův koeficient nejdříve spočítáme pro pevné y_i (např. pro $i = 1$), kolik je hodnot větších (tj. 8 hodnot $y_j > y_i$, tedy pro $j = 2, 4, 5, 6, 7, 8, 9, 10$) a kolik je menších (tj. pouze 1 hodnota $y_j < y_i$, pro $j = 3$).

$S = \sum_{i=1}^{9} \sum_{j=2}^{10} v_{ij} = \sum_{i=1}^{9} v_i = 35$ $\tau_k = \frac{35}{0{,}5 \cdot 10 \cdot 9} = \frac{35}{45} = 0{,}778$, což potvrzuje významnost pořadové korelace mezi oběma znaky.

☺ Konec ☺