37. Regresija (regresijska premica z metodo najmanjših kvadratov)

$y = \alpha + \beta x$ ne poznamo. Recimo, da je približek zanjo premica $y = a + bx$. Določimo jo po *načelu najmanjših kvadratov* z minimizacijo funkcije

$$f(a,b) = \sum_{i=1}^{n} (y_i - (bx_i + a))^2.$$

…izpeljava…

je matrika H pozitivno definitna in zato funkcija f strogo konveksna. Torej je *regresijska premica* enolično določena. Seveda sta parametra a in b odvisna od slučajnega vzorca – torej slučajni spremenljivki. Iz dobljenih zvez za a in b dobimo že znani cenilki za koeficienta α in β

$$B = \frac{C_{xy}}{C_x^2} \quad \text{in} \quad A = \overline{Y} - B\overline{X}.$$

a) Definiraj (Pearsonov) koeficient korelacije za dve številski spremenljivki.

Korelacijski koeficient slučajnih spremenljivk X in Y je definiran z izrazom

$$r(X,Y) = \frac{\mathsf{K}(X,Y)}{\sigma_X\, \sigma_Y} = \frac{\mathsf{E}\Big((X - \mathsf{E}(X))\ (Y - \mathsf{E}(Y))\Big)}{\sigma_X\, \sigma_Y}.$$

(Pearsonov) koeficient korelacije je definiran s formulo

$$r_{XY} = \frac{k(X,Y)}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

b) Kakšne vrednosti lahko zavzame in kaj lahko poveš v primeru, ko doseže največjo oziroma najmanjšo možno vrednost?

V splošnem velja:

$$-1 \leqslant r(X,Y) \leqslant 1.$$

$r(X,Y) = 0$ natanko takrat, ko sta X in Y nekorelirani;
$r(X,Y) = 1$ natanko takrat, ko je $Y = \frac{\sigma_Y}{\sigma_X}\big(X - \mathsf{E}(X)\big) + \mathsf{E}(Y)$ z verjetnostjo 1;
$r(X,Y) = -1$ natanko takrat, ko je $Y = -\frac{\sigma_Y}{\sigma_X}\big(X - \mathsf{E}(X)\big) + \mathsf{E}(Y)$ z verjetnostjo 1.
Torej, če je $|r(X,Y)| = 1$, obstaja med X in Y linearna zveza z verjetnostjo 1.

Koeficient korelacije lahko zavzame vrednosti v intervalu $[-1, 1]$.

Če se z večanjem vrednosti prve spremenljivke
večajo tudi vrednosti druge spremenljivke,
gre za *pozitivno* povezanost.
Tedaj je ρ_{XY} pozitiven in blizu 1.

Če pa se z večanjem vrednosti prve spremenljivke
vrednosti druge spremenljivke manjšajo,
gre za *negativno* povezanost.
Tedaj je ρ_{XY} negativen in blizu -1.

Če ne gre za pozitivno in ne za negativno povezanost,
rečemo da spremenljivki nista povezani in ρ_{XY} je blizu 0.

c) Kaj sta prva in druga regresijska funkcija in kje se sečeta?

Če izračunana parametra vstavimo v regresijsko funkcijo, dobimo:

$$Y = \mu_Y + \frac{\mathsf{K}(X,Y)}{\sigma_X^2}(X - \mu_X).$$

To funkcijo imenujemo tudi **prva** regresijska funkcija.

Podobno bi lahko ocenili linearno regresijsko funkcijo

$$X = a^* + b^* Y.$$

Če z metodo najmanjših kvadratov podobno ocenimo parametra a^* in b^*,
dobimo:

$$X = \mu_X + \frac{\mathsf{K}(X,Y)}{\sigma_Y^2}(Y - \mu_Y).$$

To funkcijo imenujemo **druga** regresijska funkcija.

Regresijski premici se sečeta v točki, določeni z aritmetičnima sredinama spremenljivk X in Y (premislite, kako bi to preverili).

d) Vzemimo spremenljivki X - število ur gledanja TV na teden in Y - število obiskov kinopredstav na mesec. Podatki za 6 oseb so:

X	**10**	**15**	**6**	**7**	**20**	**8**
Y	**2**	**1**	**2**	**4**	**1**	**2**

Z linearno regresijsko funkcijo ocenimo, kolikokrat bo šla oseba v kino na mesec, če gleda 18 ur na teden TV.

e) Kaj so to časovne vrste? Opiši določanje trenda z metodo najmanjših kvadratov.

Časovna vrsta jo niz istovrstnih podatkov, ki se nanašajo na zaporedne časovne razmike ali trenutke.

Osnovni namen analize časovnih vrst je

- opazovati časovni razvoj pojavov,
- iskati njihove zakonitosti in
- predvidevati nadaljni razvoj.

Časovne vrste prikazujejo individualne vrednosti neke spremenljivke v času. Čas lahko interpretiramo kot trenutek ali razdobje; skladno s tem so

Trend lahko obravnavamo kot posebni primer regresijske funkcije, kjer je neodvisna spremenljivka čas (T). Če je trend

$$X_T = f(T),$$

lahko parametre trenda določimo z metoda najmanjših kvadratov

$$\sum_{i=1}^{n}(X_i - X_{iT})^2 = \min.$$

V primeru linearnega trenda

$$X_T = a + bT,$$

$$\sum_{i=1}^{n}(X_i - a - bT_i)^2 = \min.$$

dobimo naslednjo oceno trenda

$$X_T = \overline{X} + \frac{\sum_{i=1}^{n}(X_i - \overline{X})(T_i - \overline{T})}{\sum_{i=1}^{n}(T_i - \overline{T})^2}\,(T - \overline{T}).$$

Ponavadi je čas T transformiran tako, da je $t = 0$. Tedaj je ocena trenda

$$X_T = \overline{X} + \frac{\sum_{i=1}^{n}(X_i - \overline{X}) \cdot t_i}{\sum_{i=1}^{n} t_i^2}\,t.$$

Standardna napaka ocene, ki meri razpršenost točk okoli trenda, je

$$\sigma_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X_{iT})^2},$$

kjer je X_{iT} enak X_T v času t_i.

38. Statistično sklepanje o korelacijski povezanosti in regresija

a) Kaj je cenilka in kdaj je nepristranska?

Točkovna cenilka je pravilo ali formula,
ki nam pove, kako izračunati numerično oceno
parametra populacije na osnovi merjenj vzorca.

Cenilka parametra ζ je vzorčna statistika $C = C(X_1, X_2, \ldots, X_n)$,
katere porazdelitveni zakon *je odvisen* od parametra ζ,
njene vrednosti pa ležijo v prostoru parametrov.

Cenilka je simetrična funkcija:
– njena vrednost je enaka za vse permutacije argumentov.
Seveda je odvisna tudi od velikosti vzorca n .

Cenilka C parametra ζ je **dosledna**, če z rastočim n zaporedje C_n
verjetnostno konvergira k ζ, to je, za vsak $\varepsilon > 0$ velja

$$\lim_{n\to\infty} P(|C_n - \zeta| < \varepsilon) = 1.$$

Izrek. Če za $n \to \infty$ velja $\mathsf{E}(C_n) \to \zeta$ in $\mathsf{D}(C_n) \to 0$,
je C_n dosledna cenilka parametra ζ.

Cenilka C_n parametra ζ je **nepristranska**, če je $\mathsf{E}(C_n) = \zeta$ (za vsak n);
in je **asimptotično nepristranska**, če je $\lim_{n\to\infty} \mathsf{E}(C_n) = \zeta$.

Količino $B(C_n) = \mathsf{E}(C_n) - \zeta$ imenujemo
pristranost (angl. *bias*) cenilke C_n.

b) Definiraj (Pearsonov) koeficient korelacije ρ za dve številski slučajni spremenljivki. Kakšne vrednosti lahko zavzame (morda več zakaj) in kaj se zgodi v primeru, če sta slučajni spremenljivki neodvisni (je to potreben pogoj)?

Če sta slučajni spremenljivki neodvisni, je K(X,Y)=0 -> ρ=0
(Obratno ne velja; če je K(X,Y)=0, sta spremenljivki nekorelirani (in ne nujno tudi neodvisni); nekoreliranost no zadosten pogoj za neodvisnost (obstajajo tudi odvisne spremenljivke, ki so nekorelirane))

c) Kaj lahko poveš v primeru, ko ρ doseže največjo oz. najmanjšo možno vrednost?