Korelacijska analiza

Pojem korelacije

- Korelacijska analiza proučuje soodvisnost (povezanost, usklajenost) med dvema (ali več) statističnima spremenljivkama.
- Soodvisnost med spremenljivkama lahko povzroča tudi kakšna tretja spremenljivka.

Korelacijska analiza – predpostavke

- X, Y – statistični spremenljivki.
- Podatki nastopajo v parih.
- Vzorec velikosti n: $(x_1, y_1), \ldots, (x_n, y_n)$.
- Podatke prikažemo z razsevnim diagramom.

1. PRIMER: X meri telesno višino štorklje.

   Y meri telesno težo štorklje.

   Ugotovitve:
   - Podatki so razpršeni v oblaku.
   - Oblak je nagnjen – se dviguje.
   - Intuitivno: višje štorklje so težje.
   - Pozitivna linearna povezanost oz. korelacija.

2. PRIMER: X meri prisotnost NaCl v zemlji.

   Y meri oddaljenost zemlje od ceste.

   Ugotovitve:
   - Podatki so razpršeni v oblaku.
   - Oblak je nagnjen – pada.
   - Intuitivno: dlje od ceste, manj NaCl.
   - Negativna linearna povezanost oz. korelacija.
- Jakost linearne korelacije: majhna, velika, močna.
- Opomba: Obstajajo tudi nelinearne korelacije.

Kovarianca

- Kovarianca – mera za medsebojno povezanost dveh spremenljivk. Kovarianca spremenljivk X in Y:
  $K(X,Y) = E((X - E(X))(Y - E(Y)))$.
- Cenilka za populacijsko kovarianco je vzorčna kovarianca:
  $S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$; $\overline{X}$ in $\overline{Y}$ sta vzorčni povprečji.
  - Če je $K(X,Y) = 0$, sta X in Y nekorelirani (nepovezani).
  - $K(X,X) = D(X)$.
  - $K(X,Y) = K(Y,X)$.
  - Pomen pozitivne in negativne kovariance.
- Če sta X in Y neodvisni, potem sta tudi nekorelirani. Obrat v splošnem ne velja – velja ob določenih predpostavkah (npr., če sta X in Y normalno porazdeljeni).

Pearsonov korelacijski koeficient

- Jakost linearne povezave med sprem. X in Y merimo s Pearsonovim korelacijskim koeficientom: $\rho = \frac{K(X,Y)}{\sigma(X)\sigma(Y)}$.
- Uporaba Pearsonovega korelacijskega koeficienta kot mere za jakost linearne povezave, je utemeljena le za normalno porazdeljene spremenljivke.
- Cenilka za populacijski korelacijski koeficient je vzorčni korelacijski koeficient: $r = \frac{S_{XY}}{S_X S_Y}$, kjer sta
  $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ in $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$
  vzorčni disperziji.

Lastnosti Pearsonovega korelacijskega koeficienta

- $\rho \in [-1, 1]$.
- Če je $\rho = 1$ ali $\rho = -1$, tedaj obstaja med X in Y linearna funkcijska zveza, $Y = a + bX$.
- Če je $\rho = 0$, sta X in Y nepovezani, nekorelirani. Če sta normalno porazdeljeni, sta tudi neodvisni.
- Če $\rho \neq -1, 0, 1$, potem med X in Y ni le linearne povezanosti.
  - $\rho < 0$: negativna povezava.
  - $\rho > 0$: pozitivna povezava.
  - $0 - 0,2$: neznatna linearna povezanost.
  - $0,2 - 0,4$: nizka linearna povezanost.
  - $0,4 - 0,7$: zmerna linearna povezanost.
  - $0,7 - 0,9$: visoka linearna povezanost.
  - $0,9 - 1$: zelo visoka linearna povezanost.

Ocenjevanje in testiranje korelacijskega koeficienta ρ

- Točkovna cenilka: vzorčni korelacijski koeficient $r = \frac{S_{XY}}{S_X S_Y}$.
- Testiranje nekoreliranosti sprem. X in Y (oz. neodvisnosti pri normalno porazdeljenih spremenljivkah).
  Na stopnji značilnosti α testiramo:
  $H_0(\rho = 0) : H_1(\rho \neq 0)$ (parametrični preizkus).
  H_0: X in Y sta nekorelirani.
  H_1: X in Y sta korelirani.
- Testna statistika: $T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim S(n-2)$.
- Kritično območje dvostranskega testa:
  - Izberemo tak t_α, da velja $P(|T| < t_\alpha) = 1 - \alpha$ (tabela B).
  - Izračunamo vrednost testne statistike T na vzorcu: T_e.
  - Če je $|T_e| \geq t_\alpha$, hipotezo H_0 zavrnemo in potrdimo H_1.
  - Če je $|T_e| < t_\alpha$, o hipotezi H_0 ne odločimo.
  - Če je $p \leq \alpha$, H_0 zavrnemo (p je signifikanca vzorca).

Ostali koeficienti

- Če X in Y ne zadoščata pogojem za uporabo Pearsonovega korelacijskega koeficienta ρ, lahko uporabimo (izbira je odvisna od vrste statističnih spremenljivk):
  - Spearmanov korelacijski koeficient.
  - Biserialni korelacijski koeficient.
  - Točkovni biserialni korelacijski koeficient.
  - Fi-koeficient.
  - Koeficient eta-kvadrat.

Spearmanov korelacijski koeficient

- Spearmanov korelacijski koeficient meri jakost monotone povezave med spremenljivkama X in Y.
- Uporabimo ga, če X in Y ne zadoščata pogojem za uporabo Pearsonovega korelacijskega koeficienta ρ.
- Naj bosta spremenljivki X in Y vsaj ordinalni.
- Vzorec velikosti n: $(x_1, y_1), \ldots, (x_n, y_n)$.
  Spearmanov korelacijski koeficient ρ_S je pravzaprav Pearsonov korelacijski koeficient, izračunan na podlagi rangov podatkov.
- Rangiramo podatke za X: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$.
- Rangiramo podatke za Y: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n-1)} \leq y_{(n)}$.
  Pri tem po potrebi izračunamo povprečni rang.
- Označimo:
  - I_k = rang podatka x_k.
  - J_k = rang podatka y_k.
  - $D_k = I_k - J_k$ (razlika rangov).

Vzorčni Spearmanov koeficient korelacije

- Vzorčni Spearmanov koeficient korelacije je definiran kot:
  $$r_S = 1 - \frac{6}{n^3 - n} \sum_{k=1}^{n} D_k^2.$$
  Lastnosti:
- $r_S \in [-1, 1]$.
- Če je $r_S = 1$, velja $D_k = 0$ za vsak k, torej je vedno $I_k = J_k$.
  x_k ima isti rang kot y_k. Popolna pozitivna povezanost.
- Če je $r_S = -1$, sta ranžirni vrsti „narobe obrnjeni“: $(1, n)$, $(2, n-1), \ldots$ (popolna negativna povezanost).
- Če je $r_S = 0$, so rangi nekorelirani.
- Interpretacija r_S kot pri Pearsonovem koeficientu.

Neparametrično testiranje nekoreliranosti

- Naj bosta spremenljivki X in Y vsaj ordinalni.
- Vzorec velikosti n.
- Na stopnji značilnost α testiramo:
  $H_0(\rho_S = 0) : H_1(\rho_S \neq 0)$.
  H_0: X in Y sta nekorelirani.
  H_1: X in Y sta korelirani.
- Izračunamo r_S.
- Za $n > 30$ je testna statistika $Z = r_S\sqrt{(n-1)} \sim N(0, 1)$.