Regresija i korelacija

- Podaci iz uzorka se mogu koristiti za utvrđivanje međusobne zavisnosti obeležja. Neka obeležja su više, a neka manje povezana.
- Na osnovu pojma uslovnog matematičkog očekivanja se rešava problem određivanja funkcionalne zavisnosti dve slučajne promenljive (obeležja).
- U *teoriji verovatnoće* moguće je odrediti raspodelu Y ($Y = f(X)$) ako je poznata raspodela X

$$F_Y(y) = P(Y<y) = P(f(X)<y) = P(X<f^{-1}(y)) \quad g_Y(y) = \frac{\mathrm{d}F_Y}{\mathrm{d}y}$$

ako funkcija f ima inverznu,
i ako je ona rastuća

Uslovno matematičko očekivanje

- Ako je (X, Y) diskretna 2D slučajna promenljiva sa vrednostima (x_i, y_j), $i=1,..,r$, $j=1,..,s$ i zakonom raspodele $p_{ij}=P[X=x_i, Y=y_j]$, $i=1,...,r$, $j=1,...,s$, tada je

$$E(Y/X=x_i)=\sum_{j=1}^{s} y_j P[Y=y_j/X=x_i]=\sum_{j=1}^{s} y_j \frac{P[Y=y_j, X=x_i]}{P[X=x_i]}$$

- Ako je (X, Y) neprekidna 2D slučajna promenljiva sa gustinom raspodele $g(y, x)$ i ako je $g_X(x)$ gustina raspodele slučajne promenljive X, tada je

$$E(Y/X=x)=\int_{-\infty}^{\infty} y \frac{g(x,y)}{g_X(x)} \mathrm{d}y$$

Regresija

- Slučajna promenljiva $E(Y/X) = R(X)$ se naziva **regresija**.
- To je funkcija koja najbolje opisuje zavisnost Y od X u smislu da je srednje kvadratno odstupanje minimalno

$$E(Y-h(X))^2$$

$h(X)= E(Y/X) = R(X)$.

- Da bismo odredili $R(X)$, potrebno je poznavanje zajedničke raspodele slučajnih promenljivih X i Y.

Metoda najmanjih kvadrata

- Ako zajednička raspodela nije poznata, moguće je na osnovu podataka iz uzorka (x_j, y_j), $j=1,..,n$ odrediti zavisnost Y od X u obliku $Y= f(X, a_1, a_2, \ldots, a_m)$.
- Funkciju f zovemo regresija, a parametre $a_1,\ldots,a_m$ određujemo iz uslova da suma kvadrata odstupanja bude minimalna

$$S(a_1,...,a_m) = \sum_{j=1}^{n} (y_j - f(x_j, a_1,...,a_m))^2$$

- Rešava se sistem jednačina (normalni sistem)

$$\frac{\partial S}{\partial a_1} = 0 ,..., \frac{\partial S}{\partial a_m} = 0$$

Određivanje parametara

- Neka su $(\hat{a}_1, \hat{a}_2, ..., \hat{a}_m)$ jedno rešenje normalnog sistema. Proveravamo da li je za te vrednosti postignut lokalni minimum f-je

$$\mathrm{d}^2 S = \frac{\partial^2 S}{\partial a_1^{\,2}} \mathrm{d}a_1^{\,2} + ... + \frac{\partial^2 S}{\partial a_m^{\,2}} \mathrm{d}a_m^{\,2} + 2\frac{\partial^2 S}{\partial a_1 \partial a_2} \mathrm{d}a_1 \mathrm{d}a_2 + ... + 2\frac{\partial^2 S}{\partial a_{m-1} \partial a_m} \mathrm{d}a_{m-1} \mathrm{d}a_m$$

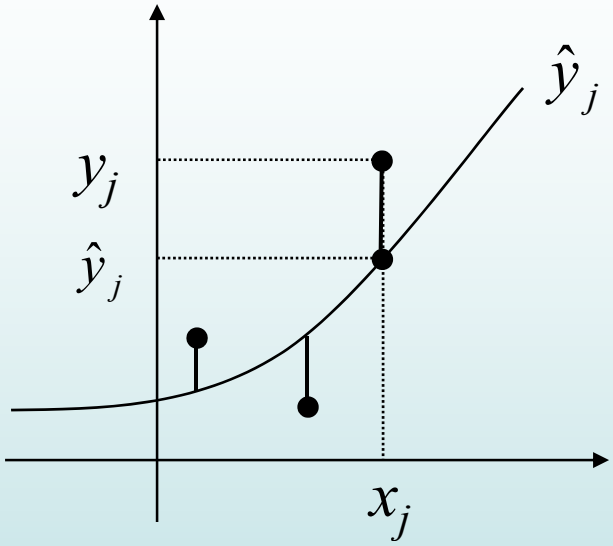
$$\mathrm{d}^2 S(\hat{a}_1, \hat{a}_2, ..., \hat{a}_m) > 0$$

tu je minimum funkcije

- Ako ima više tačaka u kojima se postiže lokalni minimum, među svima treba odrediti onu koja ima najmanju vrednost.

Suma kvadrata odstupanja

- Provera adekvatnosti izabranog modela – pomoću sume kvadrata odsupanja (SKO)



$$\hat{y}_j = f(x_j, \hat{a}_1, ..., \hat{a}_m)$$

$$\text{SKO} = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$$

- Indeks krivolinijske korelacije $R^2 = 1 - \dfrac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{\sum_{j=1}^{n} (y_j - \bar{y}_n)^2}$ $\quad \bar{y}_n = \dfrac{1}{n} \sum_{j=1}^{n} y_j$

$$R^2 \in [0,1]$$

- Bolji izbor f što je R^2 bliže 1.

Linearna regresija

- Linearna zavisnost veličina X i Y je najjednostavniji oblik zavisnosti ($Y=a\,X+b$).
- Treba odrediti koeficijente a i b.
- Ako raspodela za (X, Y) nije poznata, tada se a i b mogu odrediti iz uslova da

$$E(Y-(aX+b))^2$$

bude minimalno. Iz sistema jednačina

$$\frac{\partial S(a,b)}{\partial a}=0 \qquad \frac{\partial S(a,b)}{\partial b}=0$$

dobijaju se koeficijenti

$$a=\frac{E(XY)-E(X)E(Y)}{D(X)} \qquad b=E(Y)-aE(X)$$

Linearna regresija

- Dat je uzorak (x_j, y_j), $j=1,..,n$. Neka je zavisnost oblika $f(x,a,b) = a\,x + b$. Koeficijenti a i b se nalaze iz uslova

$$\min_{a,b\in R} S(a,b) = \min_{a,b\in R} \sum_{j=1}^{n} (y_j - (ax_j + b))^2$$

$$\frac{\partial S(a,b)}{\partial a} = 0 \qquad \frac{\partial S(a,b)}{\partial b} = 0$$

Sistem normalnih jednačina

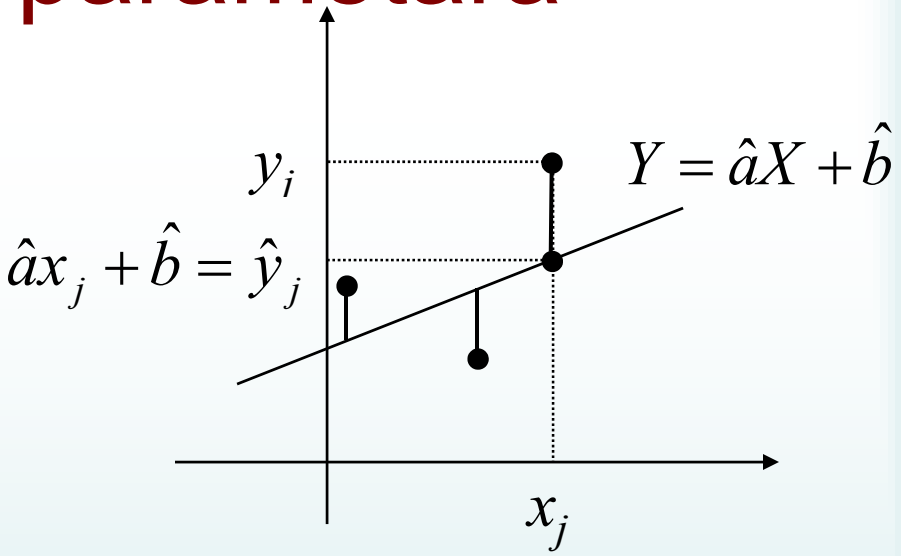
$$a\sum_{j=1}^{n} x_j + nb = \sum_{j=1}^{n} y_j$$

$$a\sum_{j=1}^{n} x_j^{\,2} + b\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_j y_j$$

$\hat{a}$ i $\hat{b}$

rešenje sistema

Određivanje parametara



$$\Delta = (\sum_{j=1}^{n} x_j)^2 - n\sum_{j=1}^{n} x_j^{\,2}$$

$$\hat{a} = \frac{\sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j - n\sum_{j=1}^{n} x_j y_j}{\Delta}$$

$$\hat{b} = \frac{\sum_{j=1}^{n} x_j \sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j^{\,2}}{\Delta}$$

$$\hat{b} = \frac{1}{n}(\sum_{j=1}^{n} y_j - \hat{a}\sum_{j=1}^{n} x_j)$$

$$\bar{y}_n = \frac{1}{n}\sum_{j=1}^{n} y_j \qquad \bar{x}_n = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Veza parametara, $\rho_{X,Y}$ i uzor. disp.

- Koeficijent korelacije

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

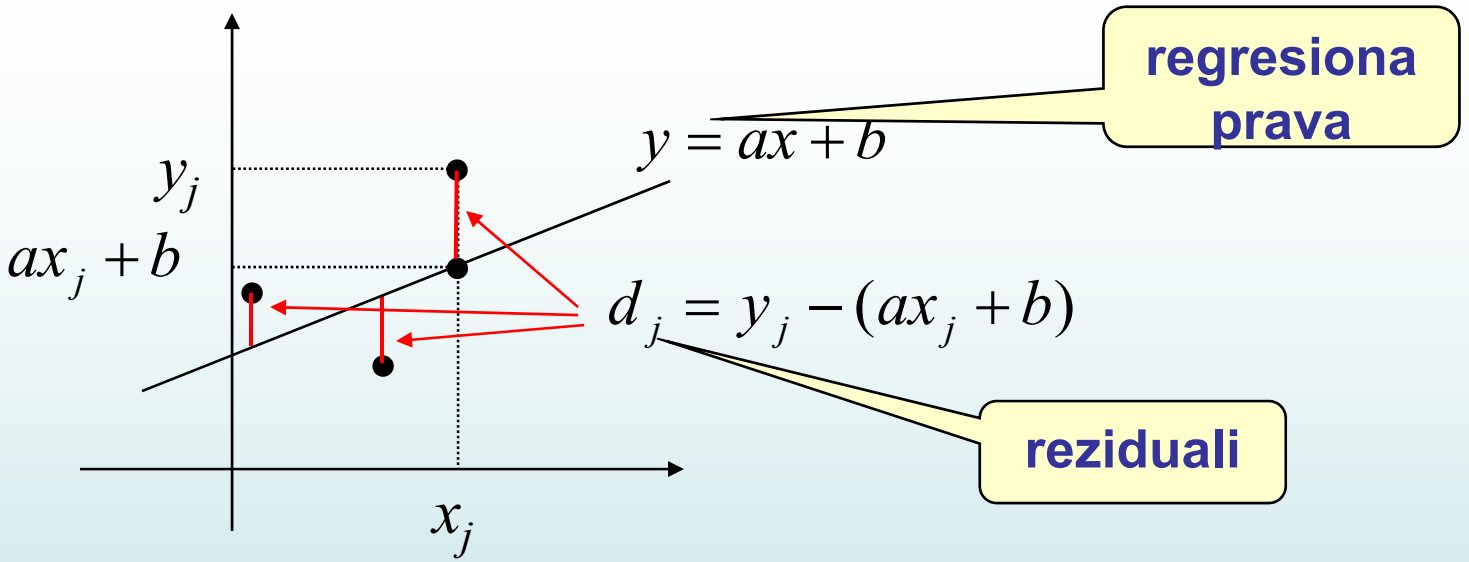
- Uzorački koeficijent korelacije

$$E(X) \to \bar{x}_n \qquad E(Y) \to \bar{y}_n \qquad E(X,Y) \to \overline{x_n y_n} = \frac{1}{n}\sum_{j=1}^{n} x_j y_j$$

$$r_{X,Y} = \frac{\frac{1}{n}\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\sum_{j=1}^{n} x_j \frac{1}{n}\sum_{j=1}^{n} y_j}{\sqrt{\left[\frac{1}{n}\sum_{j=1}^{n} x_j^{\,2} - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2\right]\left[\frac{1}{n}\sum_{j=1}^{n} y_j^{\,2} - \left(\frac{1}{n}\sum_{j=1}^{n} y_j\right)^2\right]}}$$

$$\hat{a} = \frac{\sqrt{\overline{S}^2_{nY}}}{\sqrt{\overline{S}^2_{nX}}} r_{XY}$$

Reziduali, regresija



- Od svih pravih $y=ax+b$ biramo onu za koju je zbir kvadrata odsečaka minimalan.