Inženjerska hidrologija

PRIMENA VEROVATNOĆE I STATISTIKE U HIDROLOGIJI – Korelacija i regresiona analiza

Korelacija i regresiona analiza

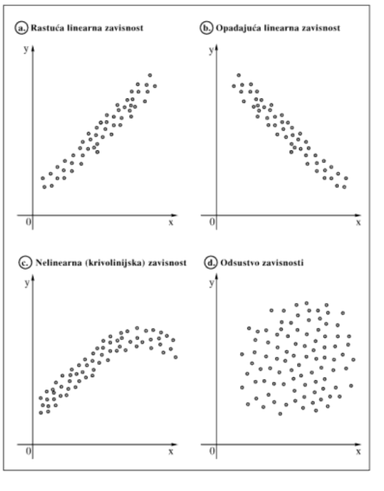
- Povezanost dve ili više promenljivih može biti od velike koristi u hidrologiji
  - najčešće za procenu jedne veličine na osnovu druge koja se meri/osmatra
  - popunjavanje i produžavanje nizova na susednim stanicama
  - veza između oticaja i padavina na jednom slivu
  - veza između oticaja i karakteristika slivova
  - analiza trenda
  - zavisnost padavina od nadmorske visine
  - ...

Korelacija i regresiona analiza

- *Funkcionalna veza* je najjača veza
  - svakoj vrednosti jedne veličine odgovara tačno određena vrednost druge veličine
  - primer 1: površina kruga je u funkcionalnoj vezi sa poluprečnikom kruga
  - primer 2: starost stabla i broj godova
- *Korelaciona ili stohastička veza* je slabija veza
  - između veličina koje su podložne odstupanjima

Korelacija

- Stohastička zavisnost ili korelacija
  - kada se ne može utvrditi funkcionalna zavisnost
  - postoji eksperimentalni skup podataka izmerenih vrednosti parova X i Y
  - grafički prikaz zavisnosti i međuzavisnosti između X i Y je *dijagram rasipanja*
  - primer 1: X = telesna visina i Y = telesna težina
  - primer 2: X = broj sati koje je osoba provela trenirajući i Y = telesna težina
  - ali: ishodi bacanja dve kocke – *nekorelisane veličine*

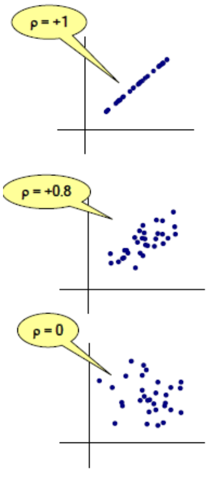


Korelacija

- Mera linearne povezanosti dve promenljive
  - Kovarijansa

$$\operatorname{cov}(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]=\mu_{XY}-\mu_X\cdot\mu_Y$$

Razlikom između μ_{XY} i $\mu_X \cdot \mu_Y$ meri se jačina linearne veze između promenljivih X i Y.



  - Koeficijent korelacije

$$\rho_{XY}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}$$

$$0\le|\rho_{XY}|\le 1$$

Korelacija

- Empirijski koeficijent korelacije dva niza x i y

$$r=\frac{\frac{1}{n-1}\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum(y_i-\bar{y})^2}}=\frac{\frac{1}{n-1}\sum(y_i-\bar{y})(x_i-\bar{x})}{S_xS_y}$$

r = CORREL(nizX, nizY)

Regresiona analiza

- Regresiona analiza: pronalaženje jednačine zavisnosti jedne promenljive od druge ili više njih
- Opšti model

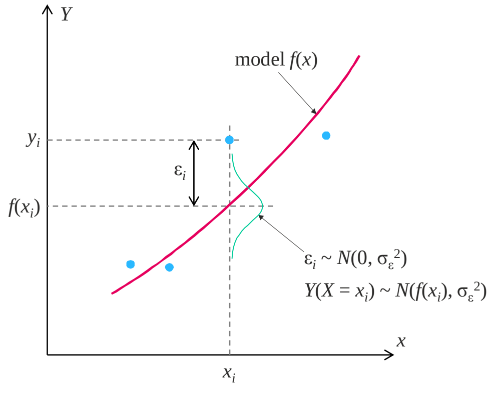
$$\boxed{Y = f(X) + \varepsilon, \quad \varepsilon : N(0, \sigma_\varepsilon^2)}$$

  - X – nezavisna promenljiva (opisna promenljiva)
  - Y – zavisna promenljiva (odzivna promenljiva)
  - ε – slučajna greška (odstupanje, rezidual) sa normalnom raspodelom $N(0, \sigma_\varepsilon)$, tj. sa sredinom 0

Milan Kilibarda

Regresiona analiza

- Regresioni model $f(x)$ je očekivano mesto za vrednosti slučajne promenljive Y
- Realizovana vrednost slučajne promenljive Y odstupa od regresione prave za grešku ε

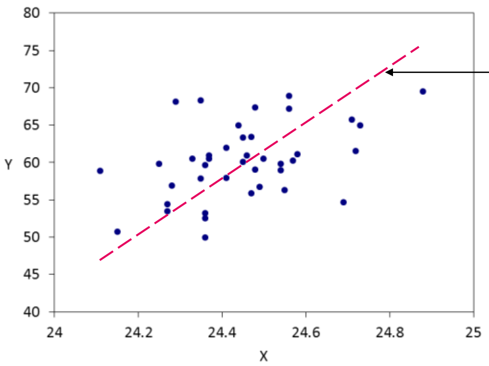


$$\boxed{Y = f(X) + \varepsilon, \quad \varepsilon : N(0, \sigma_\varepsilon^2)}$$

Prosta linearna regresija

- Model proste linearne regresije

$$\boxed{Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon : N(0, \sigma_\varepsilon^2)}$$



$f(x) = \beta_0 + \beta_1 x$

Gde se nalazi očekivano mesto promenljive Y tj. kako odrediti parametre β_1 i β_0?

Prosta linearna regresija

- Određivanje regresionih koficijenata – metoda najmanjih kvadrata
  - cilj: minimizovati zbir kvadrata odstupanja od regresione prave
  - sistem koji se rešava

$$S^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to 0$$

$$\frac{\partial S^2}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S^2}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

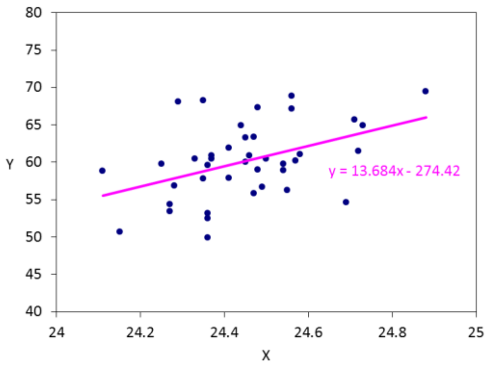
$$\longrightarrow \boxed{\begin{aligned} b_1 &= r \frac{S_y}{S_x} \\ b_0 &= \bar{y} - b_1 \bar{x} \end{aligned}}$$

b1 = SLOPE(nizY, nizX)
b0 = INTERCEPT(nizY, nizX)

Prosta linearna regresija

- Određivanje regresionih koficijenata
  - primer

$$n=40 \quad \bar{x}=24.45 \quad \bar{y}=60.14$$
$$S_x=0.16 \quad S_y=5.01 \quad r=0.436$$



$$b_1=0.436\cdot\frac{5.01}{0.16}=13.68$$
$$b_0=60.14-13.68\cdot 24.45=-274.4$$

$$\boxed{Y_m=f(x)=13.68x-274.4}$$

$$\boxed{Y=f(x)+\varepsilon=13.68x-274.4+\varepsilon}$$

Prosta linearna regresija

- Reziduali

$$\boxed{\varepsilon_i=y_i-(b_1x_i+b_0)}$$

  - srednja vrednost reziduala

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i=\frac{1}{n}\sum_{i=1}^{n}(y_i-b_1x_i-b_0)=\underbrace{\bar{y}-b_1\bar{x}}_{b_0}-b_0=0$$

  - standardna greška regresije = „standardna devijacija“ reziduala – pokazatelj kvaliteta regresionog modela

$$\boxed{S_{Y|X}=\hat{\sigma}_\varepsilon=\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2}=S_Y\sqrt{\frac{n-1}{n-2}(1-r^2)}}$$

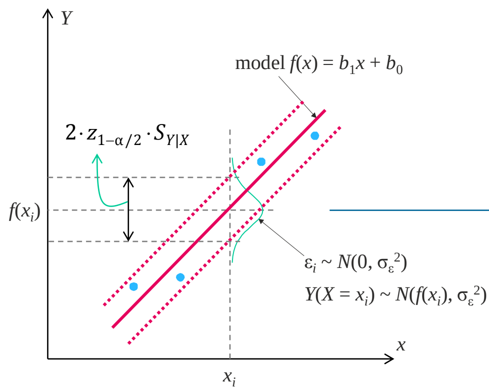
Sy|x = STEYX(nizY, nizX)

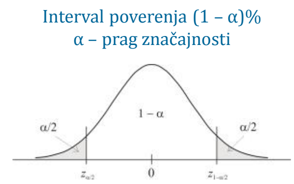
Prosta linearna regresija

- Interval poverenja odstupanja slučajne promenjive Y od regresione prave za fiksirano x

$$\boxed{P\{Y_m(x_0) - z_{\alpha/2}\cdot S_{Y|X} \le Y \mid X = x_0 \le Y_m(x_0) + z_{1-\alpha/2}\cdot S_{Y|X}\} = 1-\alpha}$$

$$Y_m(x_0) = b_0 + b_1 x_0$$

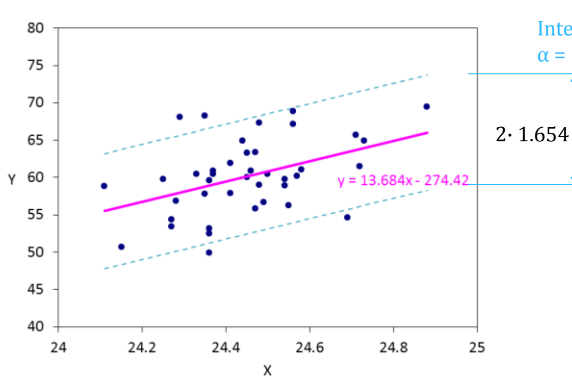


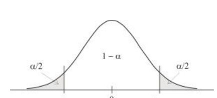


Prosta linearna regresija

- Interval poverenja odstupanja slučajne promenjive Y od regresione prave za fiksirano x
  - primer

$$S_{Y|X} = 5.01\sqrt{\frac{39}{38}(1-0.436^2)} = 4.57$$





$\alpha = 0.1 \quad \rightarrow \quad z_{0.05} = -1.645$

$z_{0.95} = 1.645$

Broj tačaka izvan intervala poverenja 90%: 4 od n = 40

Prosta linearna regresija

- Prognoza slučajne promenjive Y za zadato x_0
  - bez slučajne komponente

$$Y(x_0) = Y_m(x_0) = b_0 + b_1 x_0$$

  - sa slučajnom komponentom

$$Y(x_0) = b_0 + b_1 x_0 + \varepsilon_0, \quad \varepsilon_0 = z_0 \cdot S_{Y|X} \sim N(0, \sigma_\varepsilon^2)$$

gde je z_0 slučajni broj iz normalne raspodele, odnosno kvantil normalne raspodele za slučajan broj p

$$z_0 = \Phi^{-1}(p), \quad 0 < p < 1$$

```
p = RAND()
z0 = NORMSINV(p)
```

Prosta linearna regresija

- Prognoza slučajne promenjive Y za zadato x_0
  - primer
    - bez slučajne komponente

$$x_0 = 24.75$$
$$Y_m(x_0) = 13.68 \cdot 24.75 - 274.4 = 64.18$$

    - sa slučajnom komponentom

$$p = 0.679, \quad z_0 = \Phi^{-1}(0.679) = 0.465$$
$$\varepsilon_0 = 0.465 \cdot 4.57 = 2.12$$
$$Y_m(x_0) = Y_m(x_0) + \varepsilon_0 = 64.18 + 2.12 = 66.30$$