

Regresiona analiza

Korelacija je mera povezanosti dve ili više pojava (događaja) ili objekata. Povezanost (zavisnost) znači da je moguće predvideti jedan događaj (pojavu, karakteristiku) na osnovu ostalih. Na primer, ukoliko je prozor na stanu otvoren, a pada kiša, velika je verovatnoća da će stan biti pokvašen. U matematičkom smislu povezanost se odnosi na predviđanje vrednosti jedne zavisno promenljive veličine na osnovu poznavanja vrednosti ostalih nezavisno promenljivih veličina. Procena se uvek vrši uz određenu verovatnoću ishoda.

Jedan od ciljeva u naučnim i stručnim istraživanjima je da se pomoću jednačina opišu veze među pojavama koje su rezultat (ishod) eksperimenatalnog rada. Na primer, može se uspostaviti veza između temperature i pritiska u različitim fazama tehnološkog procesa ili veza između protoka fluida i pada pritiska u određenom aparatu. Oblast matematike (statistike) koja se bavi pronalaženjem jednačina koje povezuju određene rezultate eksperimenata (ili merenja) i procenom kvaliteta dobijenih jednačina se zove regresiona analiza.

Regresiona analiza je skup analitičkih tehnika koje se koriste da bi se procenila međusobna povezanost između pojava ili objekata koji se posmatraju, izraženih u vidu prikupljenih podataka. Kao krajnji rezultat, regresiona analiza treba da ishoduje odgovarajuću jednačinu (korelaciju) i statističku ocenu njene preciznosti. Dobijena jednačina se naziva korelacija (skraćeno od korelaciona jednačina) ili regresija (skr. od regresiona jednačina). *1

Rezultati merenja različitih veličina se dobijaju kao nizovi diskretnih veličina (tabelarna funkcija), koji najčešće ne omogućavaju detaljniju matematičku analizu (diferenciranje, integraljenje, itd.). Nakon uspostavljanja funkcionalne zavisnosti između niza nezavisno promenljivih i zavisno promenljive veličine moguće je koristiti sve pogodnosti matematičke analize¹ za dalje potrebe.

Za inženjerske probleme je od velike važnosti to što je unutar intervala prikupljenih podataka moguće sa (poznatom) tačnošću vršiti proračune (procene, predviđanja) u svrhu projektovanja, konstruisanja, vođenja tehnoloških procesa, itd. Sa nešto manjim poverenjem se dobijene jednačine mogu koristiti i za prognoziranje zavisno

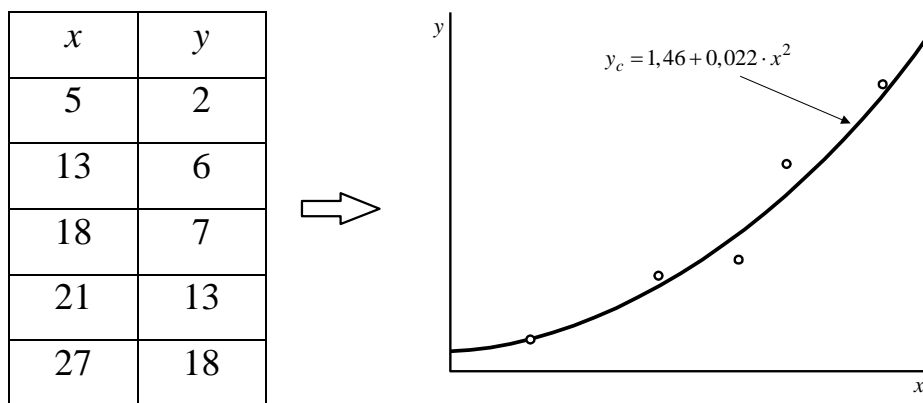
¹ Matematička analiza je oblast matematike koja proučava funkcije i njihove granične vrednosti, diferencijalni i integralni račun, redove itd. Pominje se i pod nazivima viša matematika, infinitezimalni račun, a na engleskom govornom području kao kalkulus (calculus).

promenljivih veličina i van intervala u kojima su obavljeni eksperimenti, kada se proračuni obavljaju u proširenom području korelacija. Naravno, u ovom slučaju je neophodno imati na umu da korišćenje korelacija može da dovede do značajnih grešaka.

Zadatak regresione analize se može matematički formulisati na sledeći način. Postoji skup diskretnih veličina x_1, x_2, \dots, x_k , čiji se uticaj analizira u odnosu na veličinu y . Cilj je da se na osnovu skupova tačaka $(x_{1,i}, x_{2,i}, \dots, x_{k,i}, y_i)$ za $i = 1, 2, 3, \dots, n$ odrediti zavisnost oblika: *2

$$y_c = f(x_1, x_2, \dots, x_k), \quad (1)$$

koja približno predstavlja zadatu tabelarnu funkciju, kao što je predstavljeno na slici 1.



Slika 1 Zadatak regresione analize

Promenljive x_1, x_2, \dots, x_k se nazivaju regresori i mogu biti tačne vrednosti (u matematičkom smislu), ali mogu biti i rezultati merenja različitih pojava ili karakteristika objekta, što znači da mogu biti i slučajne veličine. Zavisno promenljiva y se uvek tretira kao slučajna veličina i naziva se regresand. Sve vrednosti koje potiču iz eksperimentalnog rada su praćene (opterećene) greškama eksperimenta (merenja), pa je utoliko iznalaženje korelacije (1) uvek praćeno i analizom pojedinačnih odstupanja jednačine od samih diskretnih vrednosti zavisno promenljive.

U praksi se često pretpostavlja da su izmereni podaci za regresand i regresore slučajne veličine sa normalnom raspodelom. Ako je ta pretpostavka opravdana, onda se 68,27% vrednosti nalazi u intervalu od $\pm \sigma$ (σ je standardno odstupanje), oko

95,45% vrednosti se nalazi u intervalu od $\pm 2 \cdot \sigma$, a oko 99,73% se nalazi unutar intervala $\pm 3 \cdot \sigma$. Pravilo „68–95–99,7“ se u statistici koristi kao skraćenica pomoću koje se lakše pamte procenti vrednosti koje leže unutar opsega oko srednje vrednosti u normalnoj raspodeli sa širinom od ± 1 , ± 2 i ± 3 standardnih odstupanja. *3

Nakon formiranja korelacije (1) neophodno je izvršiti procenu njene tačnosti u odnosu na tabelarni skup podataka $(x_{1,i}, x_{2,i}, \dots x_{k,i}, y_i)$.

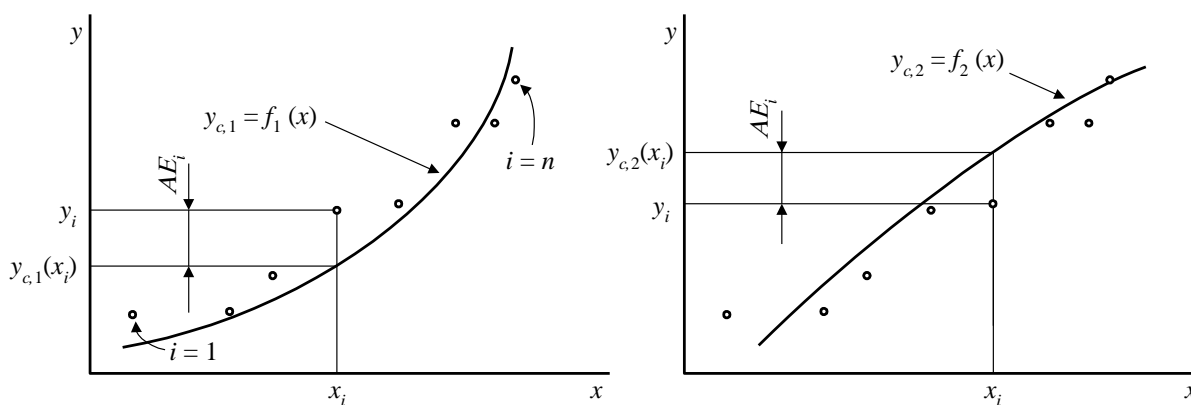
Kada se uspostavi korelacija regresanda i regresora neophodno je jasno i nedvosmisleno izraziti kvalitet dobijene korelacije. U opštem slučaju korelacija:

$$y_{c,1} = f_1(x), \quad (2)$$

ne mora da prođe ni kroz jednu tačku iz skupa od n parova (x_i, y_i) kako je prikazano na levom dijagramu na slici 2. Takođe kroz isti skup tačaka se može postaviti (slika 2 – desni dijagram) i druga korelacija oblika:

$$y_{c,2} = f_2(x). \quad (3)$$

Na osnovu dijagrama sa slike 2 jasno je da se mora uvesti jedan ili više kriterijuma na osnovu kojih se može proceniti kvalitet korelacije između jednačina (2) i (3) i skupa tačaka (x_i, y_i) . Na osnovu tih kriterijuma se može proceniti koja od korelacija bolje predstavlja rezultate merenja.



Slika 2 Korelacije (2) i (3) i skup tačaka (x_i, y_i)

Za svaku pojedinačnu tačku iz skupa merenih veličina (x_i, y_i) može se definisati odstupanje korelacione jednačine u formi apsolutnog odstupanja (greške): *4

$$AE_i = y_i - y_c(x_i), \quad (4)$$

i relativnog odstupanja (greške):

$$RE_i = \frac{|y_i - y_c(x_i)|}{y_i} \quad (5)$$

Na osnovu navedenih pojedinačnih odstupanja može se formirati i sledeće odstupanje za kompletan skup od n tačaka (x_i, y_i) :

- ukupno kvadratno odstupanje na osnovu apsolutne greške:

$$AEK = \sum_{i=1}^n AE_i^2 = \sum_{i=1}^n [y_i - y_c(x_i)]^2. \quad (6)$$

- ukupno kvadratno odstupanje na osnovu relativne greške:

$$REK = \sum_{i=1}^n RE_i^2 = \sum_{i=1}^n \left(\frac{y_i - y_{c,i}}{y_i} \right)^2. \quad (6b)$$

Metod najmanjih kvadrata

Metod najmanjih kvadrata je postupak minimizacije ukupnog kvadratnog odstupanja na osnovu apsolutne greške (zbira kvadrata apsolutnih grešaka) za aproksimaciju tabelarno zadatih funkcija u slučaju kada je zavisna promenljiva y_i slučajna (izmerena) veličina sa normalnom raspodelom, a nezavisna promenljiva x_i je tačna (nema greške merenja). *5

Metoda najmanjih kvadrata za linearnu regresiju koristi kriterijum ukupnog kvadratnog odstupanja na osnovu apsolutne greške.

Linearna korelacija sa jednom nezavisnom promenljivom

Postupak određivanja korelacione jednačine metodom najmanjih kvadrata biće ilustrovan na primeru određivanja linearne zavisnosti oblika: *6

$$y_c = a + b \cdot x. \quad (7)$$

Zbir kvadrata pojedinačnih odstupanja za n parova (x_i, y_i) iznosi:

$$AEK = \sum_{i=1}^n [y_c(x_i) - y_i]^2 = \sum_{i=1}^n (a + b \cdot x_i - y_i)^2 \quad (8)$$

Potreban uslov da bi se zadovoljio zahtev $AEK = AEK_{min}$ je:

$$\frac{\partial AEK}{\partial a} = 2 \cdot \sum_{i=1}^n (a + b \cdot x_i - y_i) = 0, \quad (9)$$

$$\frac{\partial AEK}{\partial b} = 2 \cdot \sum_{i=1}^n (a + b \cdot x_i - y_i) \cdot x_i = 0, \quad (10)$$

odakle se dobija sistem linearnih jednačina: *7

$$a \cdot n + b \cdot \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \quad (11)$$

$$a \cdot \sum_{i=1}^n x_i + b \cdot \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i \cdot y_i. \quad (12)$$

Rešavanjem ovog sistema jednačina po a i b dobija se tražena linearna zavisnost. Da bi se mogli izračunati koeficijenti a i b broj parova (x_i, y_i) u tom slučaju mora biti veći ili jednak 2 ($n \geq 2$).

Linearna korelacija sa dve nezavisno promenljive

Primena metoda najmanjih kvadrata na slučaj linearne regresije će biti takođe ilustrovana na primeru zavisnosti sa dve nezavisno promenljive oblika: *8

$$y_c = a + b \cdot x_1 + c \cdot x_2. \quad (13)$$

Zbir kvadrata pojedinačnih odstupanja za n skupova $(x_{1,i}, x_{2,i}, y_i)$ iznosi:

$$AEK = \sum_{i=1}^n [y_c(x_{1,i}, x_{2,i}) - y_i]^2 = \sum_{i=1}^n (a + b \cdot x_{1,i} + c \cdot x_{2,i} - y_i)^2. \quad (14)$$

Potreban uslov da bi se zadovoljio zahtev na kome je baziran metod najmanjih kvadrata je $AEK = AEK_{min}$, a on sledi iz uslova:

$$\frac{\partial AEK}{\partial a} = 2 \cdot \sum_{i=1}^n (a + b \cdot x_{1,i} + c \cdot x_{2,i} - y_i) = 0, \quad (15)$$

$$\frac{\partial AEK}{\partial b} = 2 \cdot \sum_{i=1}^n (a + b \cdot x_{1,i} + c \cdot x_{2,i} - y_i) \cdot x_{1,i} = 0, \quad (16)$$

$$\frac{\partial AEK}{\partial c} = 2 \cdot \sum_{i=1}^n (a + b \cdot x_{1,i} + c \cdot x_{2,i} - y_i) \cdot x_{2,i} = 0, \quad (17)$$

odakle se dobija sistem linearnih jednačina: *9

$$n \cdot a + b \cdot \sum_{i=1}^n x_{1,i} + c \cdot \sum_{i=1}^n x_{2,i} = \sum_{i=1}^n y_i, \quad (18)$$

$$a \cdot \sum_{i=1}^n x_{1,i} + b \cdot \sum_{i=1}^n x_{1,i}^2 + c \cdot \sum_{i=1}^n x_{1,i} \cdot x_{2,i} = \sum_{i=1}^n x_{1,i} \cdot y_i, \quad (19)$$

$$a \cdot \sum_{i=1}^n x_{2,i} + b \cdot \sum_{i=1}^n x_{1,i} \cdot x_{2,i} + c \cdot \sum_{i=1}^n x_{2,i}^2 = \sum_{i=1}^n x_{2,i} \cdot y_i, \quad (20)$$

na osnovu kojih se mogu dobiti koeficijenti linearne zavisnosti a , b i c .

Sličnim postupkom se može dobiti parabola drugog stepena:

$$y_c = a + b \cdot x + c \cdot x^2, \quad (21)$$

pri čemu se od skupa n tačaka $(x_{1,i}, y_i)$ formira novi skup tačaka $(x_{1,i}, x_{2,i}, y_i)$ uz zamenu:

$$x_{2,i} = x_{1,i}^2. \quad (22)$$

Za izračunavanje koeficijenata a , b i c potrebnih za definisanje parabole drugog stepena, broj parova $(x_{1,i}, y_i)$ mora biti $n \geq 3$.

Linearna korelacija sa k nezavisno promenljivih

Ukoliko postoji n skupova $(x_{1,i}, x_{2,i}, x_{3,i}, \dots, x_{k,i}, y_i)$ onda je moguće dobiti linearnu korelaciju oblika: *10

$$y_c = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + \dots + a_k \cdot x_k, \quad (23)$$

koristeći metod najmanjih kvadrata na sledeći način. *11

Definišu se matrice:

$$X = \begin{bmatrix} x_{01} & x_{11} & \dots & x_{k1} \\ x_{02} & x_{12} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ x_{0n} & x_{1n} & \dots & x_{kn} \end{bmatrix} \quad (24)$$

i

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad (25)$$

Sistem jednačina koji treba rešiti se predstavlja u obliku:

$$X \cdot A = Y, \quad (26)$$

gde je matrica korelacionih koeficijenata:

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_k \end{bmatrix} \quad (27)$$

Matrica A se dobija prema jednačini:

$$A = X^{-1} \cdot Y. \quad (28)$$

Metrike za ocenu kvaliteta regresije

Nakon regresione analize neophodno je izraziti kvalitet dobijene funkcionalne zavisnosti, što se najčešće čini pomoću nekoliko statističkih pokazatelja.

Za skup od n tačaka (x_i, y_i) definicije statističkih pokazatelja su sledeće: *12

- prosečno odstupanje:

$$SD = \sqrt{RENK} = \sqrt{\frac{\sum_{i=1}^n \left(\frac{y_i - y_{c,i}}{y_i} \right)^2}{n}} \quad (29)$$

- korelacioni odnos:

$$CR = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - y_{c,i})^2}{\sum_{i=1}^n (y_i - y_{sr})^2}} \quad (30)$$

pri čemu se srednja vrednost zavisno promenljive za n zadatih parova (x_i, y_i) izračunava korišćenjem izraza:

$$y_{sr} = \frac{\sum_{i=1}^n y_i}{n} \quad (31)$$

Kriterijum *prosečnog odstupanja* je formiran koristeći srednju vrednost ukupnog kvadratnog odstupanja na osnovu relativne greške, a *korelacioni odnos* je nastao razvojem izraza za ukupno kvadratno odstupanje na osnovu apsolutne greške.

Pored navedenih pokazatelja kvaliteta korelacije koji se obavezno koriste, često se navode i prosečna (uprosečena) relativna greška REN i maksimalna relativna greška u dijapazonu promene nezavisno promenljive.

Opšta težnja je da regresiona funkcija ima što manje odstupanje od izmerenih tačaka, pa je utoliko bolja korelacija koja: *13

- ima manje prosečno odstupanje i manje maksimalne relativne greške,
- veći korelacioni odnos (maksimalna vrednost je 1).

Navedeni statistički pokazatelji se primenjuju na funkcionalne zavisnosti bilo kakvog oblika. U specijalnom slučaju kada je korelaciona funkcija linearna, korelacioni odnos se naziva koeficijent korelacije (r_{xy}) i može se izračunati preko izraza *14

$$r_{xy} = CR = \frac{n \cdot \sum_{i=1}^n x_i \cdot y_i - \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{\sqrt{\left[n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \cdot \left[n \cdot \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}} \quad (32)$$

Oblici regresionih linija *15

Najjednostavniji oblik linearne korelacije regresanda je slučaj sa jednim regresorom u obliku

$$y_c = a + b \cdot x. \quad (33)$$

Pored jednačine prave linije u inženjerskoj praksi se koriste i nelinearne funkcije jedne nezavisno promenljive kao što su:

- eksponencijalna korelacija: *16

$$y_c = a \cdot \exp(b \cdot x), \quad (34)$$

transformiše se u opšti oblik linearne korelacije logaritmovanjem kao:

$$\begin{aligned} \ln(y_c) &= \ln(a) + b \cdot x \rightarrow Y = \ln(y_c), A = \ln(a) \\ Y &= A + b \cdot x \end{aligned} \quad (35)$$

- stepena korelacija: *17

$$y_c = a \cdot x^b, \quad (36)$$

transformiše se u opšti oblik linearne korelacije logaritmovanjem kao:

$$\begin{aligned} \ln(y_c) &= \ln(a) + b \cdot \ln(x) \rightarrow Y = \ln(y_c), A = \ln(a), X = \ln(x) \\ Y &= A + b \cdot X \end{aligned} \quad (37)$$

- logaritamska korelacija: *18

$$y_c = a + b \cdot \ln(x), \quad (38)$$

transformiše se u opšti oblik linearne korelacije kao:

$$\begin{aligned} y_c &= a + b \cdot \ln(x) \rightarrow X = \ln(x) \\ y_c &= a + b \cdot X \end{aligned} \quad (39)$$

- polinomska korelacija *19

$$y_c = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \dots + a_k \cdot x^k, \quad (40)$$

transformiše se u opšti oblik linearne korelacije kao:

$$X_1 = x^2, \dots, X_k = x^k \quad y_c = a_0 + a_1 \cdot x + a_2 \cdot X_1 + \dots + a_k \cdot X_k \quad (41)$$

- kao i mnoge druge od kojih se izdvajaju: *20

$$1. \quad y_c = a + b \cdot x^c, \quad (42)$$

transformiše se kao:

$$X = x^c \rightarrow y_c = a + b \cdot X. \quad (43)$$

$$2. \quad y_c = a + b \cdot x + c \cdot x^d, \quad (44)$$

transformiše se kao:

$$X_2 = x^d \rightarrow y_c = a + b \cdot x + c \cdot X_2. \quad (45)$$

$$3. \quad y_c = \frac{1}{a + b \cdot x}, \quad (46)$$

transformiše se kao:

$$z = \frac{1}{y_c} \rightarrow z = a + b \cdot x. \quad (47)$$

$$4. \quad y_c = \frac{x}{a + b \cdot x}, \quad (48)$$

transformiše se kao:

$$z = \frac{x}{y_c} \rightarrow z = a + b \cdot x. \quad (49)$$

$$5. \quad y_c = \frac{1}{a + b \cdot x + c \cdot x^2}, \quad (50)$$

transformiše se kao:

$$z = \frac{1}{y_c}, X_1 = x^2 \rightarrow z = a + b \cdot x + c \cdot X_1. \quad (51)$$

$$6. \quad y_c = \frac{x}{a + b \cdot x + c \cdot x^2}, \quad (52)$$

transformiše se kao:

$$z = \frac{x}{y_c}, X_1 = x^2 \rightarrow z = a + b \cdot x + c \cdot X_1. \quad (53)$$

Pored navedenih oblika u inženjerskoj praksi se koriste i brojni drugi oblici regresionih funkcija koje se načelno mogu podeliti u dve grupe:

1. funkcije koje su formirane na osnovu primene fizičkih ili drugih zakonitosti (teorija sličnosti, itd.) – ovako formirane funkcije imaju opštije značenje;
2. funkcije koje su slobodno formirane radi elementarnog zadovoljenja potrebe da se tabelirani podaci predstave u obliku koji ne mora da ima opštije značenje.

Pitanja:

1. Šta predstavlja regresiona analiza.
2. Formulacija zadatka regresione analize.
3. Pretpostavka raspodele za regresand i regresore i intervali odstupanja.
4. Kriterijumi za procenu kvaliteta korelacije.
5. Šta predstavlja metod najmanjih kvadrata.
6. Linearna korelacija sa jednom slučajnom nezavisnom promenljivom.
7. Sistem jednačina za određivanje koeficijenata linearne korelacije sa jednom nezavisnom promenljivom.
8. Linearna korelacija sa dve nezavisne promenljive.
9. Sistem jednačina za određivanje koeficijenata linearne korelacije sa dve nezavisne promenljive.
10. Linearna korelacija sa više nezavisno promenljivih.
11. Postupak određivanja koeficijenata linearne korelacije sa više nezavisno promenljivih (definisati matrice).
12. Statistički pokazatelji za ocenu kvaliteta regresije (navesti ih i napisati izraze).
13. Opšta težnja regresionih funkcija.
14. Koeficijent korelacije i kada se primenjuje.
15. Oblici regresionih krivih (navesti).
16. Eksponencijalna korelacija (oblik + transformacija).
17. Stepna korelacija (oblik + transformacija).
18. Logaritamska korelacija (oblik + transformacija).
19. Polinomska korelacija (oblik + transformacija).
20. Drugi oblici korelacije (navesti min. 2, oblik + transformacija).