KVANTITATIVNE METODE U GRAĐEVINSKOM MENADŽMENTU

predavanja 2017/18

KORELACIJA I LINEARNA REGRESIJA

1. Teorija korelacije; kovarijansa, koeficijent korelacije
2. Regresija; linearna regresija

P9

ODNOS SLUČAJNIH PROMJENLJIVIH

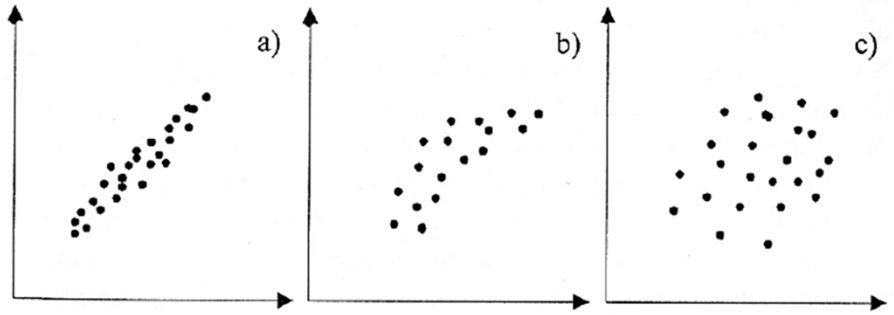
- proučavanje pojava okarakterisanih sa dvije ili više slučajnih promenljivih ne može se svesti na proučavanje svake slučajne promenljive posebno, već se moraju ispitati i međusobne zavisnosti[1]
- povezanost između dvije slučajne promjenljive (npr. dva obilježja jednog statističkog skupa) može biti:
  - prema jačini veze:
    - funkcionalna (deterministička) povezanost- svakoj vrijednosti jednog obilježja odgovara tačno određena vrijednost drugog obilježja
    - korelativna (ili stohastička) veza –za datu vrijednost x slučajne promjenljive X ne može se predskazati vrijednost slučajne promjenljive Y, već je vrijednost jedne promjenljive moguće sa određenom vjerojatnoćom predvidjeti na osnovu saznanja o vrijednosti druge promjenljive
    - ako veza ne postoji, onda se radi o nekorelativnim (nezavisnim) slučajnim promjenljivim- (broj tačaka koji se pojavljuje na dvije kockice koje sa bacaju)
  - prema smjeru:
    - pozitivni smjer– porastom vrijednosti jednog obilježja raste i drugo sa njim povezano
    - negativni smjer - porastom vrijednosti jednog obilježja drugo sa njim povezano opada
  - prema obliku povezanosti:
    - linearna (pravolinijska)
    - nelinearna (krivolinijska)
- pokazatelji korelacionih veza (ne otrkivaju da li između pojava postoji uzročno-posledična veza, u smislu da je jedna pojava uzrok, a druga posledica!):
  - koeficijent korelacije - pokazuje da li postoji linearna veza (neki koeficijenti se računaju da bi se provjerilo da li postoji nelinearna veza)
  - jednačine regresije –određuje model koji najbolje opisuje uočene veze između promjenljivih i služi da se predvide vrijednosti „zavisne“ promjenljive

1) Paunović, R., Omorjan, R. Osnovi inženjerske statistike, Tehnološki fakultet u Novom Sadu, http://www.tf.uns.ac.rs/~omorr/radovan_omorjan_003_is/Osnovi inzenjerske statistike.pdf

„Pretpostavimo da, na osnovu podataka uzetih od velikog broja ljudi, proučavamo raspodelu dvodimenzionalne slučajne promenljive, gde prva promenljiva X predstavlja visinu neke osobe, a druga promenljiva, Y njenu masu. Marginalna raspodela prve promenljive daje raspodelu visina, a druge, raspodelu masa ljudi, ali se iz njih ne može dobiti nikakva informacija o neospornoj vezi između visine i mase čoveka. Da bi analizirali tu vezu, neophodno je posmatrati raspodele masa osobe, Y pri različitim, odabranim visinama, X. Te raspodele, dobijene iz dvodimenzionalne raspodele, zovemo uslovne, jer su izvedene pod uslovom da je vrednost druge promenljive zadata.“

TEORIJA KORELACIJE

- **Teorija korelacije** - skup statističkih metoda kojima se proučavaju uzajamne veze statističkih obilježja i pojava, pri čemu se opisuju jačina, smjer, oblik ovih veza
- Za statističku analizu korelacije dvije slučajne promenljive (obilježja) X i Y, neophodno je raspolagati parovima (odgovarajućih) vrijednosti promjenljivih, tzv. vezanim uzorkom: (x_i,y_i), i=1,2...n
- **Dijagram rasipanja (raspršenosti=rasturanja)** (scatter diagram). – dijagram u x0y sistemu na kojem su predstavljeni uređeni parovi iz vezanog uzorka (visina i težina studenta kao uređeni par). Na apscisnu osu nanose se jedinice pojave koju smo označili nezavisnom (kontrolisanom = prediktornom= objašnjavajućom=regresor) promjenljivom X, a na ordinatnu osu jedinice zavisne promjenljive Y(regresand). Ucrtavanjem svih empirijskih parova podataka može se dobiti važna slika o eventualnom postojanju, obliku, smeru i jačini veze između posmatranih pojava
- Na osnovu rasporeda tačaka u dijagramu, može se grubo procijeniti:
  - da li postoji stohastička zavisnost promjenljivih (korelacija), (na sl. a) i b) korelacija postoji)
  - ako postoji korelacija, da li je ona linearna ili nelinearna, (a-linearna, b-nelinearna)
  - ako postoji korelacija, da li je ona slaba ili jaka (a- jaka, b-slaba)
- **linija regresije (regresiona kriva)–** linija (kriva) koja reprezentuje povezanost obilježja X i Y



Paunović, R., Omorjan, R. Osnovi inženjerske statistike, Tehnološki fakultet u Novom Sadu, http://www.tf.uns.ac.rs/~omorr/radovan_omorjan_003_is/Osnovi inzenjerske statistike.pdf

8 Analiza korelacije

Predmet ove glave je analiza međuzavisnosti (korelacije) dve neprekidne slučajne promenljive, na bazi paralelnog praćenja njihovih vrednosti. Najočigledniji primer međuzavisnosti dve slučajne veličine su visina i masa čoveka i kao što smo u Pogl. 3.6 konstatovali, u pitanju je ne funkcionalna, već stohastička veza između ta dva obeležja. Drugi primer je uticaj sadržaja neke komponente u složenom materijalu, recimo građevinskom, na neko svojstvo tog materijala, recimo čvrstinu. Želimo metodama statistike da, na osnovu merenja, dođemo do zaključka da li posmatrani sadržaj komponente utiče na čvrstinu građevinskog materijala i uz to, • koliko je ta korelacija izražena (jaka), • da li je ona pozitivna ili negativna, tj. da li sa porastom sadržaja posmatrane komponente čvrstina građevinskog materijala raste ili opada.

KOVARIJANSA I KOEFICIJENT KORELACIJE SLUČAJNIH PROMJENLJIVIH

- **Kovarijansa (korelacioni moment)** za slučajne promjenljive X i Y je matematičko očekivanje proizvoda odstupanja vrijednosti slulajnih promjenljivih od njihovih matematičkih očekivanja:

$$\boldsymbol{Cov(X,Y) = \mu_{XY} = E[(X - E(X)) \cdot (Y - E(Y))] = E(X \cdot Y) - E(X) \cdot E(Y)}$$

- Osobine kovarijanse:
  - Ako su X i Y nezavisne slučajne promjenljive, onda je Cov(X,Y)=0. Obrnuto ne mora da važi
  - Cov(X,Y)=Cov(Y,X)
  - Cov(X,X)=D(X)
  - Cov(aX,bY)=abCov(X,Y), gdje su a i b realni brojevi
  - Cov(X+a,Y+b)=Cov(X,Y), gdje su a i b realni brojevi
  - Može uzimati pozitivne i negativne vrijednosti i zavisna je o mjernim jeinicama promjenljivih X i Y, pa slući za zaključivanje o postojanju i smjeru veze, ali ne i o stepenu veze

- **Koeficijent korelacije osnovnog skupa (Pirsonov koeficijent)** je mjera linearne zavisnosti između slučajnih promjenljivih X i Y na nivou osnovnog skupa a računa se prema

$$\rho(X,Y) = \rho_{xy} = \frac{Cov(X,Y)}{\sqrt{D(X)} \cdot \sqrt{D(Y)}}$$

- Osobine korelacije:
  - $-1 \leq \rho \leq +1$
  - $\rho = \pm 1$. ako i samo ako je $P(Y = aX + b) = 1$, $a \neq 0$, $b \epsilon R$
  - Ako je $\rho = 0$ ne postoji linearna veza između X i Y, ali one mogu biti povezane nekom drugom nelinearnom vezom

https://dlscrib.com/download/statistika_5899a18e6454a74548b1e91c_pdf

KORELACIONA ANALIZA I UZORAČKI KOEFICIJENT KORELACIJE

- **Cilj korelacione analize** je da se ispita da li između varijacija posmatranih pojava postoji kvantitativno slaganje i, ako postoji, u kom stepenu. Sprovodi se na osnovu stvarnih vrijednost pojava (promjenljivih) u uzorku. Svejedno je koja se promjenljiva klasifikuje kao nezavisna, a koja jao zavisna promjenljiva
- **Uzorački koeficijent korelacije**: saglasna, nepristasna i asimptotski efikasna ocjena koeficijenta korelacije iz osnovnog skupa (populacije). Ova ocjena zavisnosti dva obilježja dobija se iz uzorka na osnovu parova njihovih vrijednosti (x_i, y_i), i=1,2,..n, odnosno na osnovu nepristrasnih (centriranih) ocjena kovarijanse i disperzija iz uzorka

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

- Uzorački koeficijent korelacije ima smisla računati samo kada ima indikacija (teoretska znanja, dijagram rasipanja) da je veza između posmatranih promjenljivih linearna ili približno linearna
- Ako je veza nelinearna, uzorački koeficijent korelacije r_{xy} nije mjerilo jačine korelacije i može biti i blizak nuli, uprkos jakoj vezi. U tim slučajevima treba sračunati neke druge koeficijente (Spearmanov i sl.)

podsjecanje: Grupe ocjena:

Saglasna (stabilna) je ocjena koja konvergira u vjerovatnoći ka parametru osnovne populacije kada n raste, odnosno ako je ispunjeno:

$[\![\lim P]\!]_{\top}(n\to\infty)(|U-Q|<\varepsilon)=1$, gdje je ε proizvoljno mali pozitivan broj

Centrirana ili nepristrasna je ocjena ako je njeno matematičko očekivanje jednako parametru osnovne populacije koji se procjenjuje, odnosno ako je ispunjeno: M(U)=Q, a asimptotski centrirana je ona kod koje je ispunjeno:

$[\![[\![lim]\!]_{\top}(n\to\infty)\ M]\!](U)=Q$

najefikasnija je ocjena koja je saglasna i centrirana i koja ima najmanju disperziju. asimptotski najefikasnija je ona kod koje je ispunjeno: $[\![[\![lim]\!]_{\top}(n\to\infty)\ e]\!](U)=1$

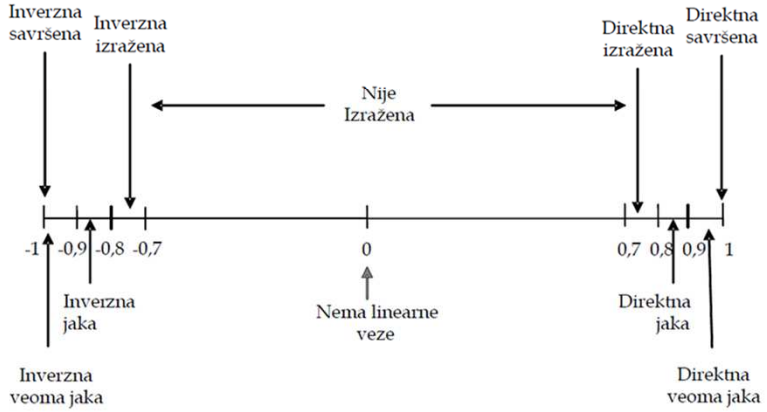
https://dlscrib.com/download/statistika_5899a18e6454a74548b1e91c_pdf

„8.4 TUMAČENJE KOEFICIJENATA KORELACIJE

S obzirom na smisao teoretskog koeficijenta korelacije ρxy, njegovu procenu rxy, ima smisla računati samo kada ima indikacija (teoretska znanja, dijagram rasipanja) da je veza između posmatranih promenljivih linearna ili približno linearna. Ako je veza nelinearna, uzorački koeficijent korelacije rxy nije merilo jačine korelacije i može biti i blizak nuli, uprkos jakoj vezi. Takođe je važno imati u vidu da statistički značajna vrednost koeficijenta korelacije nije dokaz da između posmatranih promenljivih postoji kauzalna (suštinska) veza. Tako, visoka vrednost rxy može biti rezultat delovanja treće promenljive, koja se menja u toku eksperimenata, a koja je prouzrokovala istovremene promene posmatranih promenljivih i privid njihove međuzavisnosti. Instruktivan i duhovit primer daju Boks i sar. [Box G., Hunter W i Hunter S, 1978]. U periodu od 7 godina, na kraju svake godine, je određivan broj stanovnika Oldenburga i broj roda i zapažena je jaka linearna korelacija između te dve veličine. Da li iz toga treba zaključiti da je porast nataliteta prouzrokovan porastom broja roda (rode donose decu?)? U ovom primeru, treća promenljiva, sa kojom su rasle posmatrane dve jeste vreme. U laboratorijskim i pogonskim merenjima, primer "treće" ili "nekontrolisane" promenljive je temperatura, koja deluje na veliki broj fizičko-hemijskih parametara i ako se ne kontroliše (drži konstantnom) u toku praćenja neke dve veličine, može stvoriti privid kauzalne veze izmedju njih. Tako, da bi se utvrdila suštinska povezanost između dve promenljive, neophodno je dobro poznavati njihovu fizičko-hemijsku prirodu s jedne strane, i vrlo pažljivo kontrolisati eksperimente, s druge strane. „

JAČINA LINEARNE KORELACIJE

- Neka su X i Y slučajne promjenljive i neka je ρ (X, Y) njihov koeficijent korelacije; X i Y su:
  - nekorelisane ako je ρ(X, Y) = 0;
  - pozitivno korelisane ako je ρ(X, Y) > 0;
  - negativno korelisane ako je ρ(X, Y) < 0:
- Kao empirijsko pravilo prihvataju se sljedeće granice Pirsonovog koeficijenta ***r*** (ove granice zavise od citirane literature):



- Kad se sračuna koeficijent ***r,*** potrebno je testirati koliko je on reprezentativan za osnovni skup. To se radi testiranjem parametarskih hipoteza, ali time se nećemo baviti

https://sh.wikipedia.org/wiki/Korelacija

„Rezultati korelacije imaju brojne praktičke primjene, ali se ni u kojem slučaju ne bi smjeli samo na osnovu rezultata utvrđene korelacije donositi zaključci o uzročno-poljedičnoj vezi. Korelacija se ne bi trebala koristiti za donošenje zaključaka o uzročno-posljedičnoj vezi između dvije varijable pošto je velika vjerojatnost da će zaključak biti kriv. Čest slučaj je da se promatra odnos između dvije varijable koje su u korelaciji visokog stupnja. Međutim, postoji i skrivena treća varijabla koju bi također trebalo staviti u odnos sa promatrane dvije, kako bi se ispravno protumačio uzročno-posljedični odnos.

Jedan od klasičnih, u literaturi često spominjanih primjera, je pojava uočena u Kopenhagenu nekoliko godina poslije završetka Drugog svjetskog rata. Zamijećena je korelacija između povećanja broja novorođene djece i broja roda koje su se gnijezdile u gradu. Ako bi se korelacija bez razmišljanja protumačila kao uzročno-posljedični odnos, moglo bi se zaključiti da rode donose djecu. Pravi uzrok leži u tome što se po završetku rata velik dio stanovništva sa sela preselio u grad, što je uzrokovalo povećanje broja stanovnika u gradu, a samim tim i povećanje broja novorođene djece. Istovremeno, za nove stanovnike grada izgradile su se nove kuće, tako da su i rode dobile veći broj dimnjaka za svoja gnijezda. Tu je dakle, postojala skrivena varijabla - broj stanovnika, koju je prilikom donošenje zaključka o uzročno-posljedičnoj vezi trebalo uzeti u obzir.

Naravno, ima i suprotnih primjera kada ne postoji skrivena varijabla. Vrlo rano je ustanovljena korelacija između pušenja i vjerojatnosti da će osoba oboljeti od raka. Duhanska industrija branila je svoju tezu da se ne može uspostaviti uzročno-posljedična veza između pušenja i vjerojatnosti dobivanja raka. Oni su tezu obrazlagali time da su pušači vrlo često nervozne osobe, koje zbog toga što su nervozne počinju pušiti. Istovremeno postoji korelacija između toga da je osoba nervozna i vjerojatnosti da će takva osoba dobiti rak. S druge strane, liječnici su tvrdili da postoji izravna uzročno-posljedična veza između pušenja i vjerojatnosti da će osoba dobiti rak, što je kasnije i potvrđeno.

Na osnovu utvrđene korelacije ne možemo sa sigurnošću utvrditi uzročno-posljedičnu vezu između dviju varijable. Unatoč tome korelacija nam daje informaciju o tome da su te dvije varijable na određeni način povezane. Iako ne shvaćamo u potpunosti mehanizam te povezanosti, znamo da povezanost postoji i prilikom opisa varijabli to možemo uzeti u obzir. Npr. poznato nam je da je povećana tjelesna težina u korelaciji sa povećanom smrtnošću i možemo reći da su te dvije varijable u međusobnom odnosu. Korelacija se najčešće koristi za predviđanje vrijednosti jedne varijable ovisno o promjeni vrijednosti druge varijable, u slučaju ako su te dvije varijable u korelaciji. Saznanje o korelaciji između dvije varijable pomaže nam da s većom sigurnošću predvidimo na koji način će se mijenjati vrijednost druge varijable. Npr. poznato nam je da su količina unesene soli u organizam i visina krvnog tlaka osoba određenog spola i dobi u korelacijskom odnosu i taj odnos nam je poznat. Na osnovu tih informacija možemo dozirati unos potrebne količine soli u organizam kako bi krvni tlak ostao unutar

granica normale, a organizam bi primio dovoljnu količinu soli za normalno funkcioniranje.“

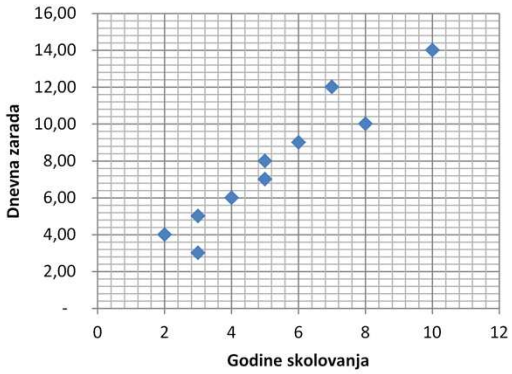
Primjer 1. Za promjenljive čije su vrijednosti date u tabeli:
a) sračunati koeficijent korelacije r_{xy} i zaključiti kakva je korelacija između X i Y
b) nacrtati dijagram rasturanja (rasipanja)

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

Redni br.	Godine školovanja X	Dnevna zarada Y
1	8	10,00
2	10	14,00
3	3	3,00
4	3	5,00
5	2	4,00
6	7	12,00
7	5	8,00
8	6	9,00
9	5	7,00
10	4	6,00

Redni br.	X	Y	XY	X^2	Y^2
1	8,00	10,00	80,00	64,00	100,00
2	10,00	14,00	140,00	100,00	196,00
3	3,00	3,00	9,00	9,00	9,00
4	3,00	5,00	15,00	9,00	25,00
5	2,00	4,00	8,00	4,00	16,00
6	7,00	12,00	84,00	49,00	144,00
7	5,00	8,00	40,00	25,00	64,00
8	6,00	9,00	54,00	36,00	81,00
9	5,00	7,00	35,00	25,00	49,00
10	4,00	6,00	24,00	16,00	36,00
SUMA	53,00	78,00	489,00	337,00	720,00

$$\frac{756,00}{23,69 \cdot 33,4} = 0,955$$



Na osnovu r=0,955 zaključujemo da se postoji izražena direktna korelacija između X i Y

http://people.dmi.uns.ac.rs/~zlc/fajlovi/Regresija.pdf

Utvrđivanjem korelacije između vrijednosti dvije varijable može se dobiti prva informacija o njihovoj međusobnoj povezanosti. Nakon toga se utvrđena povezanost može detaljnije istražiti drugim statističkim metodama. Npr. korelacijom se utvrdi da postoji veza između korištenje nekog kemijskog sredstva i pojave određene bolesti. Nakon toga se može u eksperimentalnim uvjetima, na laboratorijskim životinjama utvrditi da li stvarno postoji uzročno-posljedična veza između tih varijabli. Korelacija je tu odigrala ulogu da izolira varijable koje međusobno na neki način utječu jedna na drugu, a nakon toga druge metode, koje to mogu, potvrđuju ili odbacuju odgovarajuću uzročno-posljedičnu hipotezu. Korelacija se često koristi za provjeru rezultata testiranja. Nakon provednog testiranja utvrđuje se odgovarajuća korelacija između testiranja i dobivenih rezultata. Nakon što se testiranje ponovi, ponovno se utvrđuje korelacija između novih i prethodno dobivenih rezultata. U slučaju da korelacija ne postoji, obično se zaključuje da je provedeni eksperiment vrlo nestabilan pošto ponovljeni eksperiment ne može ponoviti prethodne rezultate

REGRESIJA I REGRESIONE KRIVE

- **regresija** – opis zavisnost jedne slučajne promjenljive od druge:
  - jednostavna (prosta) regresija – model koji sadrži jednu zavisnu i jednu nezavisnu promjenljivu
  - višestruka regresija – model sa dvije ili više nezavisnih promjenljivih
- Opšti model zavisnosti slučajnih promjenljivih X i Y je:

$$Y = f(X) + \varepsilon = E(Y|X = x) + \varepsilon$$

gdje je:

- X i Y slučajne promjenljive
- ε- slučajna greška (odstupanje), ovo je stohastički dio modela regresije
- *f(X)* –regresiona kriva kojom se objašnjava veza između X i Y, tako da možemo za svaki x_i iz populacije aproksimirati vrijednost slučajne promjenljive Y (ovo je deterministički dio modela regresije)
- *E(Y|X=x)* – matematičko očekivanje slučajne promjenljive Y za vrijednost x slučajne promjenljive X. Ovo se takođe označava i sa $\mu_{Y|X=Xi}$ (prosjek Y pod uslovom da je X=xi)

- Regresiona kriva f(x) može bitii prava ili kriva linija (kvadratna parabola, hiperbola, eksponencijalna funkcija, lgaritamska itd.). Ako je prava, onda se radi o linearnoj regresiji, a ako je definisana samo jedna nezavisna promjenljiva onda se radi o prostoj linearnoj regresiji.

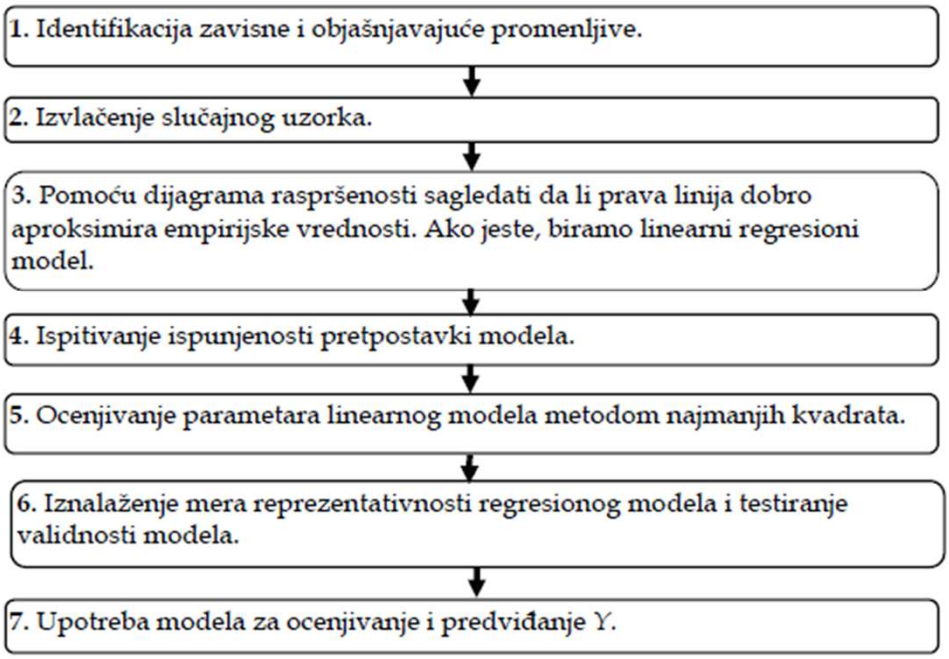
https://dlscrib.com/download/statistika_5899a18e6454a74548b1e91c_pdf

1) http://www.ekfak.kg.ac.rs/sites/default/files/nastava/Novi Studijski Programi/I godina/Osnovi statistike/Materijali/udzbenik/11_OS_Regresija_2009.pdf

„Kod regresione analize nužno je unapred identifikovati koja pojava će imati ulogu zavisne promenljive, a koja nezavisne promenljive. U statistici se kod regresije najčešće ne koristi termin "nezavisna promenljiva", već objašnjavajuća promenljiva ili regresor. Naziva se objašnjavajuća jer pomoću nje pokušavamo da objasnimo varijacije zavisne promenljive. Koja promenljiva će biti izabrana za objašnjavajuću utvrđuje se na osnovu prethodnih teorijskih ili empirijskih saznanja, ili pretpostavki o prirodi analiziranih pojava.

Kod regresije se izbegava izraz “nezavisna promenljiva“ jer to implicira da je X uzrok, a Y posledica. Međutim, regresionom analizom je nemoguće dokazati uzročnu vezu između pojava.“

ETAPE U PROSTOJ REGRESIONOJ ANALIZI



PROSTA KORELACIONA I REGRESIONA ANALIZA
http://www.ekfak.kg.ac.rs/sites/default/files/nastava/Novi%20Studijski%20Programi/I%20godina/Osnovi%20statistike/Materijali/udzbenik/11_OS_Regresija_2009.pdf

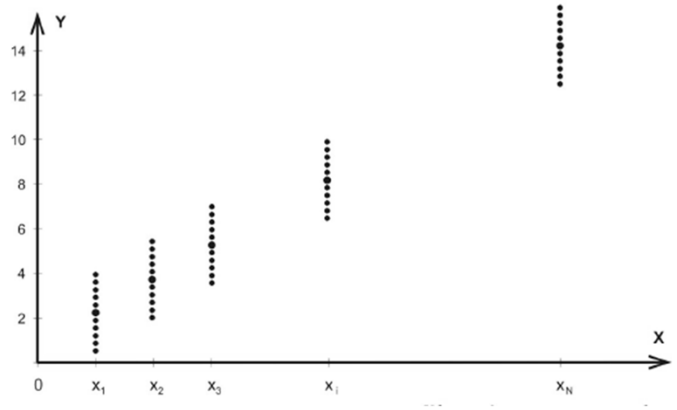
PROST LINEARNI REGRESIONI MODEL

- **Prost linearni regresioni model** (u osnovnom skupu):

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i \qquad , i=1,2,....,N$$

gdje je:

- Y_i - i-ta zavisna promjenljiva (kod stohastićkih veza jednoj vrijednosti nezavisne promjenljive X odgovara čitav niz mogućih vrijednosti zavisne promjenljive)
- x_i - i-ta vrjednost obašnjavajuće promjenljive
- β_0 i β_1 – regresioni parametri, koji određuju koeficijente prave linije (slobodni član i koeficijent pravca)
- ε_i- slučajna greška (odstupanje) modela regresije uz pretpostavke da za ε_i važi:
  - $\varepsilon_i \sim N(0, \sigma)$ – ima normalnu raspodjelu
  - međusobno nezavisni
  - nezavisni od vrijednosti promenljive x_i
- N - veličina osnovnog skupa
- i –i-ta vrijednost u osnovnom skupu



REGRESIONA ANALIZA

- **Cilj regresione analize** je da se odredi regresioni model koji najbolje opisuje vezu između pojava i da se na osnovu toga modela ocijene i predvide vrijednosti zavisne promjenljive Y za odabrane vrijednosti objašnjavajuće (nezavisne)[1] promjenljive X. Sprovodi se na osnovu stvarnih vrijednosti pojava (promjenljivih) u uzorku. Nije svejedno koja se promjenljiva klasifikuje kao nezavisna, a koja kao zavisna
- **Regresiona linija osnovnog skupa –** linija koja prolazi kroz sve prosječne vrijednosti Y je najviše prilagođena podacima iz osnovnog skupa. Njena jednačina je:

$$\mu_{YIX=x_i} = E(Y_i) = \beta_0 + \beta_1 \cdot x_i \qquad , i=1,2,....,N$$

- Ako bi bili poznati β_0 i β_1, onda bismo mogli za svako pojedinacno x_i predvidjeti prosječne vrijednosti za Y_i
- β_0 i β_1 –nepoznati parametri koje treba ocijeniti na osnovu podataka iz uzorka, kao ocjene b_0 i b_1
- Kod linearne zavisnosti promjenljivih X i Y treba ocijeniti nepoznate parametre β_0 i β_1 u regresionoj liniji osnovnog skupa:

$$E(Y_i) = \beta_0 + \beta_1 \cdot x_i$$

- **Linija regresije u uzorku -** definiše se kao prava sa ocjenom parametara β_0 i β_1

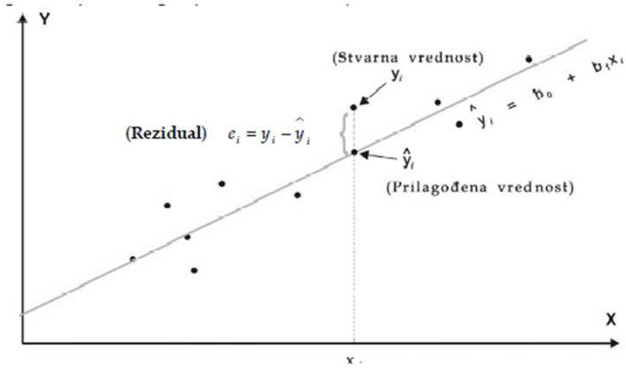
$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

gdje je:

$\hat{y}_i$ – prilagođena vrijednost Y= vrijednost Y koja se nalazi na najbolje prilagođenoj liniji regresije uzorka

b_0 i b_1 - ocjene slobodnog člana i koeficijenta nagiba, dobijaju se metodom najmanjih kvadrata

REGRESIONA ANALIZA



- **Rezidual** – vertikalno odstupanje između stvarne vrijednosti y_i i prilagođene vijednosti /predstavlja ocjenu stohastičkog člana ε_i

$$\boldsymbol{e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 \cdot x_i)}$$

- Od svih mogućih pravih linija treba odabrati onu kod koje je najmanja suma kvadrata vertikalnih odstupanja, odnosno naći minimum izraza:

$$\boldsymbol{\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (b_0 + b_1 \cdot x_i)]^2}$$

- Postupak podrazumijeva nalaženje parcijalnih izvoda po b_0 i b_1 i izjednačavanje tih izvoda sa nulom. Odatle se dobija:

$$b_1 = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \cdot \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2} = r \cdot \sqrt{\frac{s_y}{s_x}} = \frac{s_{xy}}{s_x^2} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

http://www.ekfak.kg.ac.rs/sites/default/files/nastava/Novi%20Studijski%20Programi/I%20godina/Osnovi%20statistike/Materijali/udzbenik/11_OS_Regresija_2009.pdf

MJERE REPREZENTATIVNOSTI REGRESIONOG MODELA

- Pokazatelji uspješnosti modela, odnosno pokazatelji koliko model kvalitetno (uspješno) opisuje zavisnost između dvije pojave:
  - **standardna greška regresije-** apsolutna mjera odstupanja empirijskih tačaka (stvarnih vrijednostiniz uzorka) – što je raspršenost tačaka oko prave linije manja, model je bolji

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

  - **koeficijent determinacije (određenosti)** mjeri koliko je regresiona linija koju smo dobili na osnovu uzorka reprezentativna (odnosno koliko odgovara) regresionoj liniji koju bi konstruisali na populaciji.
    - relativna mjera koja pokazuje koliko je jaka zavisnost između promjenljivih X i Y (jačina modela)
    - vrijednost varira od $0 \leq r^2 \leq 1$
    - ako je $r^2=1$ - sve uzoračke (empirijske) vrijednosti y_i se nalaze na liniji regresije: postoji čvrsta funkcionalna veza između X i Y
    - ako je $r^2=0$ - ne postoji linearna zavisnost između X i Y

$$r^2 = r_{xy}{}^2 = \frac{s_{xy}{}^2}{s_x{}^2 \cdot s_y{}^2}$$

    gdje je:

    r_{xy} - koeficijent korelacije

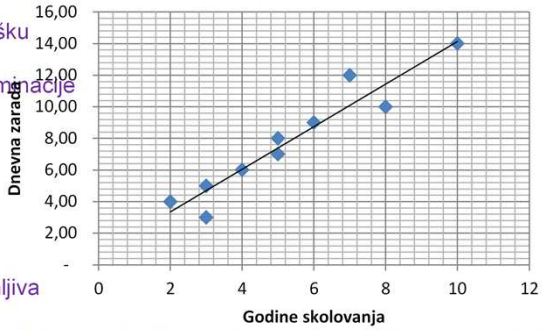
    s_{xy} – uzoračka kovarijansa

    $s_x{}^2$ – srednje kvadratno odstupanje vrijednosti promjenljive X od njene aritmetičke sredine

    $s_y{}^2$ – srednje kvadratno odstupanje vrijednosti promjenljive Y od njene aritmetičke sredine
  - testiranje parametarske hipoteze da je nagib regresione prave jednak nuli. Ako je tako, onda postoji linearna zavisnost (ovime se nećemo baviti)

Primjer 2. Za promjenljive čije su vrijednosti date u tabeli (iz primjera 1)
a) definisati regresionu krivu koja će pokazati u kakvoj su vezi dnevna zarada i godine školovanja
b) sračunati standardnu grešku regresije
c)sracunati koeficijent determinacije

Redni br.	Godine školovanja X	Dnevna zarada Y
1	8	10,00
2	10	14,00
3	3	3,00
4	3	5,00
5	2	4,00
6	7	12,00
7	5	8,00
8	6	9,00
9	5	7,00
10	4	6,00



1. X- objašnjavajuća promjenljiva
2. Y - zavisna promjenljiva
3. Iz dijagrama rasipanja se vidi da se podaci iz uzorka grupišu približno u pravoj liniji (proračunom Pirsonovog koeficijenta *r* smo dokazali da postoji jaka linearna korelacija između ovih velićina (X i Y))
4. Opšti oblik linije regresije u uzorku $\hat{y}_i = b_0 + b_1 \cdot x_i$
5. proračun koeficijenata

$$b_1 = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \cdot \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2} = 1{,}3476$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 0{,}6577$$

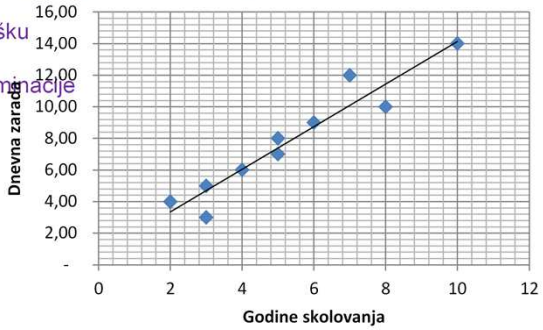
Redni br.	X	Y	XY	X^2	Y^2
1	8,00	10,00	80,00	64,00	100,00
2	10,00	14,00	140,00	100,00	196,00
3	3,00	3,00	9,00	9,00	9,00
4	3,00	5,00	15,00	9,00	25,00
5	2,00	4,00	8,00	4,00	16,00
6	7,00	12,00	84,00	49,00	144,00
7	5,00	8,00	40,00	25,00	64,00
8	6,00	9,00	54,00	36,00	81,00
9	5,00	7,00	35,00	25,00	49,00
10	4,00	6,00	24,00	16,00	36,00
SUMA	53,00	78,00	489,00	337,00	720,00

6. regresiona prava
   $\hat{y}_i = b_0 + b_1 \cdot x_i$=0,6577+1,3476$x_i$
7. Nacrtati je na dijagramu rasipanja
   (naci dvije tacke po prethodnoj formuli i kroz njih provuci pravu)
8. standardna greska regresije:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}} = 1{,}102$$

http://people.dmi.uns.ac.rs/~zlc/fajlovi/Regresija.pdf

Primjer 2. Za promjenljive čije su vrijednosti date u tabeli (iz primjera 1)
a) definisati regresionu krivu koja će pokazati u kakvoj su vezi dnevna zarada i godine školovanja
b) sračunati standardnu grešku regresije
c)sracunati koeficijent determinacije



Redni br.	Godine školovanja X	Dnevna zarada Y
1	8	10,00
2	10	14,00
3	3	3,00
4	3	5,00
5	2	4,00
6	7	12,00
7	5	8,00
8	6	9,00
9	5	7,00
10	4	6,00

9. proračun koeficijenta determinacije r^2

U prvom primjeru smo sračunali Pirsonov koeficijent. Ovaj koeficijent determinacije je kvadrat Pirsonovog koeficijenta

r^2=0,955*0,955=0,912=91,2% - 91,2% - ovoliko procenata ukupnog varijabiliteta se može objasniti godinama rad

Redni br.	X	Y	XY	X^2	Y^2
1	8,00	10,00	80,00	64,00	100,00
2	10,00	14,00	140,00	100,00	196,00
3	3,00	3,00	9,00	9,00	9,00
4	3,00	5,00	15,00	9,00	25,00
5	2,00	4,00	8,00	4,00	16,00
6	7,00	12,00	84,00	49,00	144,00
7	5,00	8,00	40,00	25,00	64,00
8	6,00	9,00	54,00	36,00	81,00
9	5,00	7,00	35,00	25,00	49,00
10	4,00	6,00	24,00	16,00	36,00
SUMA	53,00	78,00	489,00	337,00	720,00

http://people.dmi.uns.ac.rs/~zlc/fajlovi/Regresija.pdf

Literatura

- PROSTA KORELACIONA I REGRESIONA ANALIZA
  http://www.ekfak.kg.ac.rs/sites/default/files/nastava/Novi%20Studijski%20Programi/I%20godina/Osnovi%20statistike/Materijali/udzbenik/11_OS_Regresija_2009.pdf

- Osnovi statistike
  http://www.ekfak.kg.ac.rs/OASOE_predmet_osnovi_statistike_materijali