Co-funded by the Erasmus+ Programme of the European Union

TOPIC PLAN		
Partner organization	Univerzitet Metropolitan	
Topic	Linearna regresiona i korelaciona analiza	
Lesson title	Metod najmanjih kvadrata	
Learning objectives	Student razume pojam linearne korelacije Student je u stanju da za dva procesa ili pojave proceni da li su korelirani i u kojoj meri Student razume pojam linearna regresija Student razume metod najmanjih kvadrata. Student ume da primenom metode najmanjih kvadrata odredi regresionu pravu	**Strategies/Activities** ☐Graphic Organizer ☐Think/Pair/Share ☑Modeling ☑Collaborative learning ☑Discussion questions ☐Project based learning ☑Problem based learning
Aim of the lecture / Description of the practical problem	Svrha ove lekcije je da student usvajajući metod najmanjih kvadrata bude u stanju da odredi regresionu pravu koja najbolje odražava linearnu korelaciju između dve pojave ili procesa. Praktični problem: Promena jednog obeležja statističkog skupa često utiče na promenu drugih obeležja zbog međusobne povezanosti. Povezanost između obeležja može se razlikovati i po smeru i po jačini povezanosti. Statističko istraživanje veza između pojava vrši se multivarijacionom analizom. Ona se deli na na regresionu analizu i korelacionu analizu. Predmet korelacione analize jeste otkrivanje karaktera i stepena (čvrstine) kvantitativnog slaganja varijacija pojava. Predmet regresione analize jeste otkrivanje forme korelacione veze, odnosno forme slaganja varijacija pojave. Jedna i druga analiza se međusobno dopunjuju.	**Assessment for learning** ☑Observations ☑Conversations ☑Work sample ☐Conference ☐Check list ☐Diagnostics
Previous knowledge assumed:	Funkcije dve promenljive, parcijalni priraštaji i izvodi funkcije dve promenljive, Silvesterov kriterijum (postupak za određivanje lokalnih ekstrema funkcije dve promenjlive)	**Assessment as learning** ☑Self-assessment ☐Peer-assessment ☐Presentation ☐Graphic Organizer ☐Homework

Introduction / Theoretical basics	Funkcija dve promenljive	Assessment of learning
	Definicija 1. Realna funkcija dve realne promenljive je bilo koje pravilo ili zakon po kome se svakom uređenom paru (x,y) iz nekog skupa $A \subseteq \mathbb{R}^2$ pridružuje tačno jedan broj $z \in B \subseteq \mathbb{R}$. **Definicija 2.** Totalni priraštaj funkcije *f* u tački A (sa priraštajima argumenata Δx i Δy), u oznaci Δf, je veličina $\Delta f = f(a_1+\Delta x, a_2+\Delta y) - f(a_1,a_2)$, gde su Δx i Δy realne veličine različite od nule. U slučaju kada u prethodnoj relaciji važi da je: 1. $\Delta y = 0$ - tada $\Delta f = f(a_1+\Delta x, a_2) - f(a_1, a_2)$ nazivamo parcijalnim priraštajem funkcije *f* u tački A po prvoj promenljivoj (po *x*) i označavamo sa $\Delta_x f$. 2. $\Delta x=0$ - tada $\Delta f = f(a_1, a_2+\Delta y) - f(a_1, a_2)$ nazivamo parcijalnim priraštajem funkcije *f* u tački A po drugoj promenljivoj (po y) i označavamo sa $\Delta_y f$. Koristeći pojmove uvedene u prethodnoj definiciji, može uočiti sledeće granične vrednosti: $$\lim_{\Delta x \to 0} \frac{\Delta_x f}{\Delta x} \quad (1)$$ $$\lim_{\Delta x \to 0} \frac{\Delta_y f}{\Delta x} \quad (2)$$ Granične vrednosti (1) i (2) mogu postojati ili ne u $\mathbb{R} \cup \{-\infty, +\infty\}$. Ako su konačne, nazivamo ih parcijalni izvodi prvog reda funkcije *f* u tački A po prvoj, odnosno po drugoj promenljivoj, respektivno. U tom slučaju koristimo jednu od sledećih oznaka $$f'_x(x,y)|_A, \frac{\partial f}{\partial x} \; ili \; f'_x(A)$$ $$f'_y(x,y)|_A, \frac{\partial f}{\partial y} \; ili \; f'_y(A)$$ Za ova dva parcijalna izvoda prvog reda možemo kreirati funkcije tih parcijalnih izvoda $f'_x(x,y)$ i $f'_y(x,y)$ za (x,y)	☑Test ☑Quiz ☐Presentation ☐Project ☐Published work

iz domena funkcije.
Prethodno date funkcije su opet funkcije po promenljivim x i y.
Za funkciju *f* u tački A iz domena funkcije D možemo formirati parcijalne izvode drugog reda po jednoj, odnosno po drugoj promenljivoj na sledeći način: neka su date izvodne funkcije parcijalnih izvoda prvog reda funkcije *z*=*f*(x, y), za (x, y) ∈ D sa izvoda $f'_x(x,y)$ i $f'_y(x,y)$ za (x,y) iz domena funkcije D.

Za njih je moguće (ponaosob) potražiti parcijalne izvode prvog reda u tački A ∈ D (ako su za to obezbeđeni uslovi) i time dobijamo

$$(f'_x(x,y))'_x|_A,\ (f'_x(x,y))'_y\big|_A,\ (f'_y(x,y))'_x\big|_A,\ (f'_y(x,y))'_y\big|_A.$$

Standardne oznake za prethodno date parcijalne izvode su

$$f''_{xx}(x,y)|_A, f''_{xy}(x,y)|_A,\ f''_{yx}(x,y)|_A,\ f''_{yy}(x,y)|_A$$

ili

$$\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial x \partial y}, \frac{\partial^2 f}{\partial y \partial x}, \frac{\partial^2 f}{\partial y^2}.$$

Često se umesto $f''_{xx}(x,y)|_A$ koristi oznaka $f''_{x^2}(x,y)|_A$. Slično, umesto $f''_{yy}(x,y)|_A$ koristi oznaka $f''_{y^2}(x,y)|_A$. Takođe, umesto $f''_{xx}(x,y)|_A$ često se koristi oznaka $f''_{xx}(A)$. Analogno se označavaju i ostali parcijalni drugog reda u tački A.

Lokalne ekstremne vrednosti

Neka je data funkcija z = *f*(x, y) na oblasti $D \subseteq \mathbb{R}^2$. Tada za tačku $M_1(x_1, y_1) \in D$ kažemo da je lokalni minimum}} ako postoji barem jedna njena ε-okolina, u oznaci O_1, gde je $O_1 \subseteq D$, takva da je $f(x_1, y_1) \leq f(x, y)$, za svako (x, y) ∈ O_1. Takođe, za tačku $ $M_2 (x_2, y_2)$) ∈ D kažemo da je lokalni maksimum ako postoji barem jedna njena ε-okolina, u oznaci O_2, gde je $O_2 \subseteq D$, takva da je $f(x_2, y_2)$

≥ f(x, y), za svako (x, y) ∈ O_2. Tačke lokalnih minimuma i maksimuma se nazivaju i lokalni ekstremi funkcije f na D.

U narednom razmatranju ćemo dati jedan postupak za određivanje tačaka lokalnih ekstrema posmatrane funkcije (ako ih ona uopšte ima).

Neka je funkcija z = f(x, y) definisana na oblasti $D \subseteq \mathbb{R}^2$. Za tačku $ $M_0(x_0, y_0)$ ∈ D kažemo da je stacionarna tačka funkcije f ako je rešenje sistema

$$f'_x(x, y) = 0$$

$$f'_y(x, y) = 0$$

na oblasti D. Skup rešenja prethodnog sistema označimo sa S. On može biti prazan ili neprazan.

Stav. Neka je funkcija z = f(x, y) definisana na oblasti $D \subseteq \mathbb{R}^2$ i neka je S skup stacionarnih tačaka za tu funkciju na D. Tada svaka tačka lokalnog ekstrema funkcije na oblasti D pripada skupu S.

Napomena. Iz prethodnog stava možemo zaključiti da ako je skup S prazan, tada funkcija f na D nema lokalne ekstreme. Međutim, on ne važi u suprotnom smeru. To znači da sve tačke koje pripadaju skupu S ne moraju biti lokalni ekstremi. Stoga se nameće pitanje, ako je skup S neprazan, kako ćemo od svih tačaka koje mu pripadaju, izdvojiti one koje su lokalni ekstremi i kako ćemo znati da li su one lokalni maksimumi ili minimumi. O tome govorimo u nastavku.

Stav (Silvesterovo pravilo). Neka je funkcija z = f(x, y) definisan na $D \subseteq \mathbb{R}^2$ i neka je S skup njenih stacionarnih tačaka na D takav da je skup S neprazan. Dalje, neka je $M_0(x_0, y_0)$ ∈ S.Takođe, neka je

$$f''_{x^2}(M_0) = A,\ f''_{xy}(M_0) = f''_{yx}(M_0) = B, f''_{y^2}(M_0) = C.$$

	Označimo sa $\gamma = A \cdot C - B^2$. Tada a) ako je γ > 0 i A > 0, tačka M_0 je lokalni minimum funkcije *f* na D; b) ako je γ > 0 i A < 0, tačka M_0 je lokalni maksimum funkcije *f* na D; c) ako je γ < 0, funkcija *f* u tački M_0 nema lokalnih ekstrema; d) ako je γ = 0, za tačku M_0 nemamo nikakav odgovor po pitanju lokalnih ekstrema.	
Action	Pojam stohastika je starogrčka reč i predstavlja sinonim s pojmom slučajnost. On se koristi u situacijama kada se želi istaći da se neka pojava ili više njih posmatraju sa stanovišta slučajnosti. Ovde ćemo posmatrati povezanost između dve pojave. Obično se jedna od njih posmatra kao nezavisna promenljiva i označava se sa X, dok se druga smatra zavisno promenljivom od X i označava sa Y. Međusobne veze između ovakvih pojava mogu biti determinističke ili stohastičke. Diskusija sa studentima o ovim pojmovima. Korelacija pretpostavlja stohastičku (verovatnosnu) vezu među pojavama, zbog čega pojave moraju imati osobinu slučajnosti. Najčešće je pretpostavka da slučajne promenljive (pojave) *X* i *Y* imaju normalne raspodele i tada govorimo o teoriji Normalne korelacije. Teorija Normalne korelacije pretpostavlja, dakle, postojanje normalnih raspodela korelirajućih slučajnih promenljivih. Normalne raspodele se često sreću u praksi, a takođe, kada se povećava broj posmatranih jedinica statističkog skupa, veliki broj raspodela teži normalnoj. Korelacione veze, u osnovi prirode pojava kod kojih se proučavaju, moraju imati uzročne veze. Uzročne veze među pojavama pretpostavka su postojanja korelacionih veza. Korelacija koja se ispituje između dve pojave se naziva prosta korelacija. Ona može biti linearna i nelinearna. Mi ćemo se ovde baviti utvrđivanjem linearne veze. U statističkim ispitivanjima najčešće se koristi tzv. koeficijent proste linearne korelacije ili Pirsonov koeficijent koji služi kao mera linearne zavisnosti	

promenljivih X i Y. Ovaj koeficijent se definiše preko varijansi i kovarijansi slučajnih promenljivih X i Y. Kovarijansa slučajnih promenljivih X i Y, u oznaci Cov(X, Y) se definiše na sledeći način

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y) .$$

Dakle, za slučajne promenljive X i Y koeficijent korelacije, u oznaci, $\rho_{X,Y}$, definiše se sa

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{\sigma^2(X) \cdot \sigma^2(Y)}}.$$

Koeficijent proste linearne korelacije u osnovnom skupu obeležavamo ρ, a u uzorku obeležavamo sa $r_{X,Y}$. Koeficijent proste linearne korelacije može uzeti vrednost samo iz intervala [−1, 1] .
Za ovaj koeficijent važi da je $|\rho_{X,Y}| \leq 1$ pri čemu je $|\rho_{X,Y}| = 1$ ako i samo ako je Y = α X + β, gde je α > 0 za ρ = 1, a α < 0 za ρ = −1, ako su slučajne promenljive X i Y nezavisne slučajne promenljive, tada je $\rho_{X,Y} = 0$ (obrnuto ne mora da važi).

Dakle, kada na istoj populaciji posmatramo dve numeričke karakteristike prirodno se postavlja pitanje da li se među ovim vrednostima može ustanoviti postojanje neke veze (ovde govorimo, kao što je već rečeno o linearnoj zavisnosti dve slučajne promenljive). Ako (x_1, y_1), (x_2, y_2),…, (x_n, y_n) predstavlja jedan dvodimenzionalni uzorak, tada se koeficijent linearne korelacije traži po sledećem obrascu

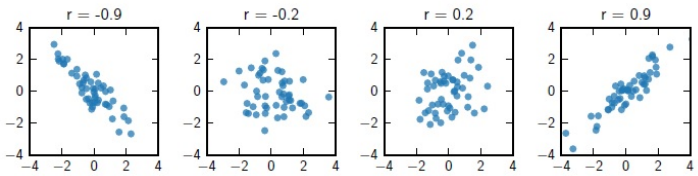
$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

gde su $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ i $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ izračunate srednje vrednosti prvih i drugih elemenata uređenih parova (x_i, y_i), i = 1, 2, 3,…, n.

Pritom se smatra da je za:

- 0 < $|\rho_{X,Y}|$ ≤ 0,3 veza nije linearna, ali se ne isključuju nelinearni oblici povezanosti,
- 0,3 < $|\rho_{X,Y}|$ ≤ 0,5 linearna veza je slaba,
- 0,5 < $|\rho_{X,Y}|$ ≤ 0,7 linearna veza je značajna,

- 0,7 < $|\rho_{X,Y}|$ ≤ 0,9 linearna je jaka,
- 0,9 < $|\rho_{X,Y}|$ ≤ 1 linearna veza je vrlo jaka.



Slika 1. Dijagram rasturanja za dve pojave u zavisnosti od vrednosti njihovog koeficijenta linearne korelacije.

Na slici je predstavljeno nekoliko primera dijagrama rasturanja u zavisnosti od veličine za r. Vidimo da što je vrednost za r po apsolutnoj vrednosti bliža broju 1, time je povezanost između posmatranih obeležja jača. Negativna vrednost za linearnu korelacija ukazuje na da su posmatrana obeležja obrnuto proporcionalna, dok pozitivna vrednost ukazuje na to da su direktno proporcionalna.

Regresioni model je matematički model koji opisuje vezu između dve ili više promenljivih. Prost regresioni model obuhvata samo dve promenljive: jednu objašnjavajuću i jednu zavisnu. Zavisna promenljiva je promenljiva čije varijacije treba da objasnimo na osnovu kretanja objašnjavajuće promenljive. U regresionom modelu, objašnjavajuća promenljiva se obično obeležava sa X, a zavisna promenljiva sa Y. Međuzavisnost između dve promenljive u regresionoj analizi izražava se matematičkom jednačinom koja se naziva regresiona jednačina ili regresioni model. Veza između promenljivih može biti različitog oblika, uključujući i linearnu vezu. Regresioni model kojim izražavamo linearnu vezu između dve promenljive se naziva linearni regresioni model. Regresioni model kojim se opisuje linearna međuzavisnost između dve promenljive naziva se prost linearni regresioni model. Ovde ćemo se njime baviti. U ostalim slučajevima se radi o nelinearnom regresionom modelu.

Prost linearni regresioni model je oblika

$$Y = a + b \cdot X,$$

gde koeficijent *b* predstavlja koeficijent pravca prave, a *a* odsečak na Oy osi. Ovaj model se naziva deterministički model i pokazuje determinističku (egaktnu ili funkcionalnu) vezu između promenljivih X i Y. Preciznije, ovaj model podrazumeva da je promenljiva Y deterministički (egzaktno) određena promenljivom X, odnosno da za datu vrednostl postoji jedna i samo jedna vrednost Y. Kao što smo već rekli, u praksi veza između ovih promenljivih je stohastička, jer postoji uticaj i mnogih drugih faktora, koji dovode do varijacija u vrednostima koje uzima promenljiva Y, za iste vrednosti promenljive X. Uticaj tih faktora u prostom linearnom regresionom modelu obuhvatamo dodatnim članom, koji nazivamo slučajna greška koji ćemo označiti sa δ i on je, zapravo, oblika

$$Y = a + b \cdot X + \delta.$$

Sada ćemo ukazati na to kako se dolazi do prostog linearnog regresionog modela i koje su to etape u njegovom dobijanju.

U regresionoj analizi, najpre, mora da se izvrši identifikacija promenljivih. To znači da je potrebno odrediti koja je pojava zavisna promenljiva, a koja pojava je objašnjavajuća promenljiva. Nakon toga potrebno je izabrati slučajni uzorak (x_1, y_1), (x_2, y_2),…, (x_n, y_n) i na osnovu dijagrama rasturanja proveriti da li za ove vrednosti prava predstavlja najbolji oblik povezanosti između posmatrane dve pojave. U tom slučaju, ako sa y_i označimo zavisno promenljivu veličinu dobijenu iz uzorka, a sa x_i objašnjavajuće veličine dobijene iz uzorka, za i = 1,2,…, n, tada imamo da je odgovarajući prost linearni regresioni model, definisan na sledeći način

$$y_i = a + b \cdot x_i + \delta_i, \; i = 1,2,\ldots, n, \qquad (*)$$

gde su *a* i *b* regresioni parametre koje treba odrediti, a δ_i, i = 1,2,…, n, je stohastički član ili slučajna greška, pri i-tom članu uzorka. Naime, u slučajnom uzorku, za iste vrednosti po veličini x^*, dobijaćemo različite vrednosti za y. Kada se bude definisala regresiona prava, ona će prolaziti tačku $(x^*, \bar{y})$, gde $\bar{y}$ predstavlja prosečnu vrednost dobijenu po veličini y. Na ovoj logici se zasniva

određivanje parametara *a* i *b*. Kada odredimo *a* i *b* tada tako dobijenu pravu koristimo kao model za ponašanje čitave populacije. Ovo nam, dalje, omogućava da za poznate vrednosti po veličini x, predviđamo prosečnu vrednost za y.

Sledeći korak jeste provera da li su ispunjene pretpostavke za primenu ovog modela. On podrazumeva da u formuli (*) važi $\delta_i \in N(0,\sigma^2)$, i = 1,2,…, n, tj. da se slučajna greška jeste slučajna promenljiva koje ima normalnu raspodelu, za koju je $E(\delta_i)=0$, za svako i = 1,2,…, n. Takođe, bilo koje dve slučajne greške δ_i i δ_j $(i ≠ j, i, j = 1, 2,…, n ne smeju biti korelirane. Na kraju, veličine za x su veličine koje istraživač sam zadaje, dakle one nemaju karakter slučajnost, a veličine za Y se određuju na osnovu uzorka.

Na kraju, treba izvršiti ocenjivanje parametara *a* i *b* u datom modelu. Postoji egzaktan matematički način za to koji se bazira na uslovu da je zbir kvadrata vertikalnih udaljenosti tačaka od pravca najmanji. Taj metod se naziva metod najmanjih kvadrata. Osnove ove metode predložio je Gaus. U slučaju regresione prave, ideja je da za posmatrani uzorak (x_1, y_1), (x_2, y_2),…, (x_n,y_n) uočimo i-to odstupanje, u oznaci D_i, (i=1,2,…,n) za koje važi

$$D_i = y_i - (a + b\cdot x_i),$$

i da odredimo koeficijente *a* i *b* tako da zbir kvadrata svih ovih odstupanja bude minimalan. Stoga, uočimo funkciju

$$F(a, b) = [y_1 - (a + b\cdot x_1)]^2 + [y_2 - (a + b\cdot x_2)]^2 + \ldots + [y_n - (a + b\cdot x_n)]^2$$

i odredimo njen minimum. Očigledno je posmatrana funkcija, realna funkcija dve realne promenljive za koju možemo primeniti Silvesterov kriterijum prilikom određivanja lokalnog minimum. Njegovom primenom dobijamo da se za

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

	$$b = \frac{n \cdot \sum_{i=1}^{n}(x_i \cdot y_i) - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$ postiže minimimum funkcije F(*a*, *b*). Kao što smo rekli, regresiona linija izražava se jednačinom y = *a* + *b*·x, gde je y – zavisno promenljiva, x – nezavisno promenljiva, *a* – regresiona konstanta, *b* – koeficijent regresije. U linearnom regresionom modelu se za koeficijent regresije *b* koristi formula određena pomoću metode najmanjih kvadrata dok se koeficijent *a* određuje na sledeći način $$a = \bar{y} - b \cdot \bar{x},$$ gde su $\bar{x}$ i $\bar{y}$ aritmetičke sredine odgovarajućih veličina iz uzorka. Kada se odrede koeficijenti za prostu linearnu regresiju metodom najmanjih kvadrata, može se postaviti pitanje koliko je dobro ocenjena regresiona linija ovim parametrima, u odnosu na regresionu liniju polaznog skupa? O tome govori sledeći stav. **Stav.** Ako su ispunjene sve pretpostavke prostog linearnog regresionog modela, ocene dobijene metodom najmanjih kvadrata su najbolje, nepristrasne linearne ocene.	
Materials / equipment / digital tools / software	*Materials*: Dodatni materijali za učenje su dati u spisku referenci na kraju ovog plana. *Equipment*: učionica, tabla, kreda; *Digital tools*: laptop, projektor;	
Consolidation	• Nastavnik pokreće diskusiju sa studentima postavljajući im odgovorajuća pitanja; • Studenti rešavaju jednostavnije zadatke pod nadzorom nastavnika; • Nastavnik navodi prigodne primere prilikom uvođenja odgovorajućih pojmova u saradnji sa studentima; • Zadavanje domaćeg zadatka od strane nastavnika, pri čemu studenti treba da ga urade do narednog predavanja.	
Reflections and next steps		

Activities that worked	Parts to be revisited
Nakon časa nastavnik popunjava ovaj deo prema ličnoj proceni uspeha časa.	Na osnovu uspeha studenata u izradi domaćih zadataka, kao i na osnovu pitanja i diskusije na početku narednog časa, nastavnik dolazi do zaključka koje delova gradiva treba da obnovi/ponovi sa studentima.

References

1. Dr Rale Nikolić, Elektronski materijali predavanja za učenje, Metropolitan Univerzitet, 2021. godina, Beograd
2. https://web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/MethodLeastSquares.pdf
3. https://personal.utdallas.edu/~Herve/Abdi-LeastSquares06-pretty.pdf