Primer 1.

Koristeći podatke iz primera koji smo obradili u prethodnim predavanjima/vežbama oceniti prosečnu vrednost obeležja *Y* i predvideti individualnu vrednost obeležja *Y* za promet x_0=115 mil.din. uz rizik od 5%.

Promet x_i	Dobit y_i	$x_i y_i$	x_i^2	Regresijske vred. $\hat{y}_i$	y_i^2
20	1	20	400	1,05	1
30	3	90	900	2,35	9
40	3,5	140	1600	3,65	12,25
50	5	250	2500	4,95	25
70	7	490	4900	7,55	49
80	8,5	680	6400	8,85	72,25
90	9	810	8100	10,15	81
100	13	1300	10000	11,45	169
480	50	3780	34800	50.00	418,5

$\bar{x} = \frac{\sum x_i}{n} = \frac{480}{8} = 60$; $\bar{y} = \frac{\sum y_i}{n} = \frac{50}{8} = 6{,}25$.

Parametre linearne regresije *a* i *b*:

$$b = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{3780 - 8 \cdot 60 \cdot 6{,}25}{34800 - 8 \cdot 60^2} = 0{,}13$$

$$a = \bar{y} - b\,\bar{x} = 6{,}25 - 0{,}13 \cdot 60 = -1{,}55$$

Koristićemo dobijenu jednačinu linearne regresije

$\hat{y}_i = a + bx_i = -1{,}55 + 0{,}13x_i$.

Srednja mera odstupanja dobiti od linije regresije je:

$$S_e = \sqrt{\frac{\sum y_i^2 - a\sum y_i - b\sum x_i y_i}{n-2}} = 0{,}8756$$

Sledi da je:

$\hat{y}_{p(x=115)} = -1{,}55 + 0{,}13 \cdot 115 = 13{,}4$

$$S_{\hat{y}_p} = S_e \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2}} = 0{,}875 \cdot \sqrt{\frac{1}{8} + \frac{(60-115)^2}{34800 - 8 \cdot 60^2}} = -0{,}6945.$$

- Interval ocene prosečne vrednosti zavisne promenljive uz rizik od 0,05 i već ranije dobijenu tabličnu vrednost iz *t* rasporeda za 6 stepeni slobode ($t_{6;0.025}$=2,4469≈2,45) je:

13,4−2,45·0,6945≤$E(y_p)$≤13,4+2,45·0,6945, odnosno 11,70≤ E(y_p)≤15,10.

Zaključujemo da uz rizik od 0,05 dobijeni interval obuhvata prosečno kretanje dobiti posmatranog preduzeća za promet od 115 mil.din.

- Kako je sada vrednost standardna greška predviđanja S*p*, individualne vrednosti zavisno promenljive jednaka:

$$S_p = S_e \cdot \sqrt{1+\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{\sum x_i^2 - n\bar{x}^2}} = 0{,}875 \cdot \sqrt{1+\frac{1}{8}+\frac{(60-115)^2}{34800-8\cdot 60^2}} = 1{,}1176.$$

Sledi, interval predviđanja individualne vrednosti je:

13,4−2,45·1,1176≤ y_p ≤13,4+2,45·1,1176,

10,66 ≤ y_p ≤ 16,14.

Predviđamo, uz rizik od 0,05, da će se pri godišnjem prometu od 115 mil.din. posmatranog preduzeća, godišnja dobit kretati u intervalu od 10,66 do 16,14 mil. din.

Kao posledica veće standardne greške dobija se i širi interval, što je i očekivano jer su individualne vrednosti podložnije većim varijacijama nego prosečne.

Primer 2.

Neka je za 12 parova podataka vrednosti promenljivih *X* i *Y* izračunata vrednost koeficijenta korelacije $\boldsymbol{r_{xy}}$=0,54. Da li to znači da postoji jaka pozitivna (direktna) korelacija?

Rešenje:

Definišu se hipoteze

H_o**:** $r_{xy} = 0$, H_1: $r_{xy} \neq 0$

Za stepen slobode *n*-2=10 i α=0,05 granična vrednost Studentove raspodele je $t_{10;\,0{,}025}$=2,23.

Iz jednakosti

$$r \cdot \frac{\sqrt{n-2}}{\sqrt{1-r^2}} = 2{,}23,$$

sledi: $r \cdot \sqrt{10} = 2{,}23 \cdot \sqrt{1 - r^2}$, tj. r =0,576.

Kako je dobijena vrednost $\boldsymbol{r}_{xy}$=0,54 manja od granične tablične vrednosti, $r_{10;\,0{,}025}$ =0,576, za 10 stepeni slobode i prag značajnosti α=0,05, možemo zaključiti da se ne odbacuje nulta odnosno ne prihvata alternativna hipoteza sa greškom α<0,05 tj.sa sigurnošću 95% tvrdimo da ne postoji jaka pozitivna korelaciona veza između datih parova.

Primer 3. (*Spearmanov koeficijent korelacije* ranga)

Dati su bodovi studenata koji su položili dva kolokvijuma is statistike. Naći usklađenost znanja statistike studenata na oba kolokvijuma pomoću Spearmanovog koeficijenta.

Redni br. studenta	Bodovi sa prvog kolokvijuma	Bodovi sa drugog kolokvijuma
1	53	48
2	15	32
3	30	62
4	47	64
5	60	70
6	75	65
7	14	17
8	25	28
9	25	30
10	19	16

Rešenje:

Rangovi za promenljivu X su prikazani u radnoj tabeli u četvrtoj koloni.

- najmanjoj vrednosti promenljive X =14, pridružen je rang 1.
- sledeći su po veličini bodova 15 i 19, pa su njima pridruženi rangovi 2 i 3.
- slede dva po veličini jednaka broja bodova, 25, a kako su na redu rangovi 4 i 5, to je svakoj vrednosti pridružena aritmetička sredina tih dvaju rangova, tj. 4.5.
- sledi po veličini 30 bodova, kojima je pridružen rang 6 itd.

Slično za promenljivu *Y* nađeni su rangovi u petoj koloni.

Redni br. studenta	Bodovi na kolokvijumu		Rang vrednosti promenljive *X*	Rang vrednosti promenljive *Y*	Razlika rangova	Kvadrati razlike rangova
	Prvi	Drugi				
	x_i	y_i	$r(x_i)$	$r(y_i)$	d_i	d_{i2}
1	53	48	8	6	2	4
2	15	32	2	5	−3	9
3	30	62	6	7	−1	1
4	47	64	7	8	−1	1
5	60	70	9	10	−1	1
6	75	65	10	9	1	1
7	14	17	1	2	−1	1
8	25	28	4,5	3	1.5	2.25
9	25	30	4,5	4	0.5	0.25
10	19	16	3	1	2	4
Ukupno	**-**	**-**	**55**	**55**	**0**	**24.5**

Spearmanov koeficijent korelacije ranga je:

$$r_s = 1 - \frac{6\Sigma d_i^2}{n^3 - n} = 1 - \frac{6 \cdot 24,5}{10^3 - 10} = 0,8515$$

Koeficijent korelacije ranga je dosta blizu jedinice, što znači da je veza među rangovima dveju promenljivih pozitivna i dosta jaka. Student koji je dobro uradio prvi kolokvijum, uradio je dobro i drugi i obrnuto. To upućuje na dosta dobru usklađenost znanja statistike studenata iz oblasti oba kolokvijuma

Primer 4.

Rangirano je 7 studenata po broju položenih ispita (*X*) i stepenu prisutnosti na časovima (*Y*). Rang 1 je najpovoljnija, a rang 7 najnepovoljnija ocenu za oba modaliteta.

Student	A	B	C	D	E	F	G
Rang za *X*	4	2	6	1	3	7	5
Rang za *Y*	3	1	6	2	4	7	5

Da li je broj položenih ispita u međuzavisnosti sa stepenom prisutnosti na časovima ?

Rešenje:
Definišu se hipoteze:

H_0: Između broja položenih ispita i stepena prisutnosti na časovima ne postoji međuzavisnost.
H_1: Između broja položenih ispita i stepena prisutnosti na časovima postoji međuzavisnost.

Konstruišemo radnu tabelu i sređujemo rangove po veličini:

Student	Rang za X	Rang za Y	d	$d2$
A	4	3	1	1
B	2	1	1	1
C	6	6	0	0
D	1	2	−1	1
E	3	4	−1	1
F	7	7	0	0
G	5	5	0	0
∑	-	-	-	**4**

Spirmanov koeficijent ranga korelacije je:

$$r = 1 - \frac{6\Sigma d_i^2}{n^3 - n} = 1 - \frac{6 \cdot 4}{7^3 - 7} = 0,93$$

Statistika testa je:

$$t_\gamma = \frac{r_{xy}}{\sqrt{\frac{1-r_{xy}^2}{n-2}}} = 0,93 \frac{\sqrt{7-2}}{\sqrt{1-0,93^2}} = 5,6577.$$

Kako je za n=7 , n −2=5, i α= 0,01, iz tablica t- rasporeda vrednost: $t_{n-2;\frac{\alpha}{2}} = t_{5;0,005} = 4,032$.

Dobija se da je vrednost statistike testa 5,6577 > 4,032. To znači da ćemo odbaciti nultu hipotezu. Dakle, između broja položenih ispita i stepena prisutnosti na časovima postoji jaka međuzavisnost, što tvrdimo sa verovatnoćom 99%.

Pitanja i zadaci za vežbu:

Primer 1. Na osnovu podataka tržišne statistike o kretanju cena i tražnje jedne vrste proizvoda datih u tabeli:

Cena (din.)	10	12	15	18	20	23	25
Tražnja (000kg)	80	76	71	65	60	55	45

a) Odrediti f-ju linearne regresije između cene i tražnje posmatranog proizvoda.
b) Izračunati standardnu grešku regresije.
c) Uz verovatnoću od 95% predvideti obim tražnje pri nivou cene od 40 dinara.

d) Izračunati relativnu meru reprezentativnosti regresionog modela i objasniti rezultat.

Primer 2. Troškovi i profit preduzeća prikazani su sledećom tabelom:

Troškovi (mil. din.)	3	4	6	6	7	8
Profit (mil. din.)	8	10	12	11	13	14

e) Na osnovu datih podataka nacrtaj dijagram raspršenosti.
f) Oceniti linearnu regresiju i grafički je predstaviti.
g) Oceniti profit pri troškovima od 11 mil. din. sa 95% pouzdanosti.

Primer 3. Na bazi podataka datih u tabeli ispitati , pomoću korelacije ranga, da li između dužine radnog staža radnika i kvaliteta proizvedenih proizvoda postoji kvantitativna zavisnost i ako postoji u kojoj meri?

Radnik	M.K.	N.T.	A.S.	A.D.	E.T.	G.Z.	K.L.	R.D.
Dužina radnog staža (god.)	2	3	5	7	1	4	6	8
Kvalitet proizvoda (ocean kvaliteta od 1 do 10)	2	1	5	8	3	4	6	7

Pitanja: (*odnose se na lekcije koje smo obradili na predavanju/vežbama prethodne i ove nedelje*)

1. Koja je razlika između determinističke i stohastičke zavisnosti?
2. Objasni razliku između regresione i korelacione analize.
3. Navesti cilj regresione i cilj korelacione analize.
4. Šta se može zaključiti na bazi dijagrama raspršenosti?
5. Kakva je razlika između proste i višestruke regresije?
6. Navesti etape u prostoj linearnoj regresiji.
7. Kako se odabira tip regresionog modela?
8. Šta pokazuje parameter $\boldsymbol{a}$ *i* $\boldsymbol{b}$ kod regresione prave $\hat{y}_i = a + bx$?
9. Kakva je razlika između ocenjivanja i predviđanja u regresionoj analizi?
10. Koji factor utiče na veličinu standardne greške regresije?
11. Kako se tumači koeficijent proste linearne korelacije?
12. Šta je koeficijent determinacije?
13. Ako je $r = -0.84$, objasniti datu vrednost.
14. Ako je $r^2 = 0.92$, objasniti datu vrednost.
15. Kada je koeficijent korelacije pozitivan, a kada negativan?

dr Slavica Dabetić