PROSTA LINEARNA REGRESIJA

PREDAVANJE BR.13

Regresija – terminološko pojašnjenje

- Termin “regresija” ima različita značenja u psihologiji.
- Regressio (lat.) – kretanje unazad
- U psihologiji označava vraćanje na raniji ili manje zreo način ponašanja
- Ovdje ćemo termin (jednostruka linearna) regresija koristiti u statističkom smislu
- Statistički smisao se sastoji u pronalaženju najadekvatnijeg linearnog modela za opisivanje veze između zavisne i nezavisne promjenljive.

Osnovni cilj regresione analize

Definisanje regresionog modela koji može, na osnovu poznavanja rezultata nezavisne varijable (prediktorske), manje ili više precizno da predvidi (ocijeni) rezultat zavisne (kriterijumske) varijable.

Dakle, ciljevi su:

- Predviđanje jedne numeričke karakteristike preko druge numeričke karakteristike
- Ispitivanje zavisnosti jednog parametra od drugog

Napomena: Ovdje pretpostavljamo da su obije varijable kvantitativne.

Linearni model za predviđanje

- Linearni regresioni model za populaciju:

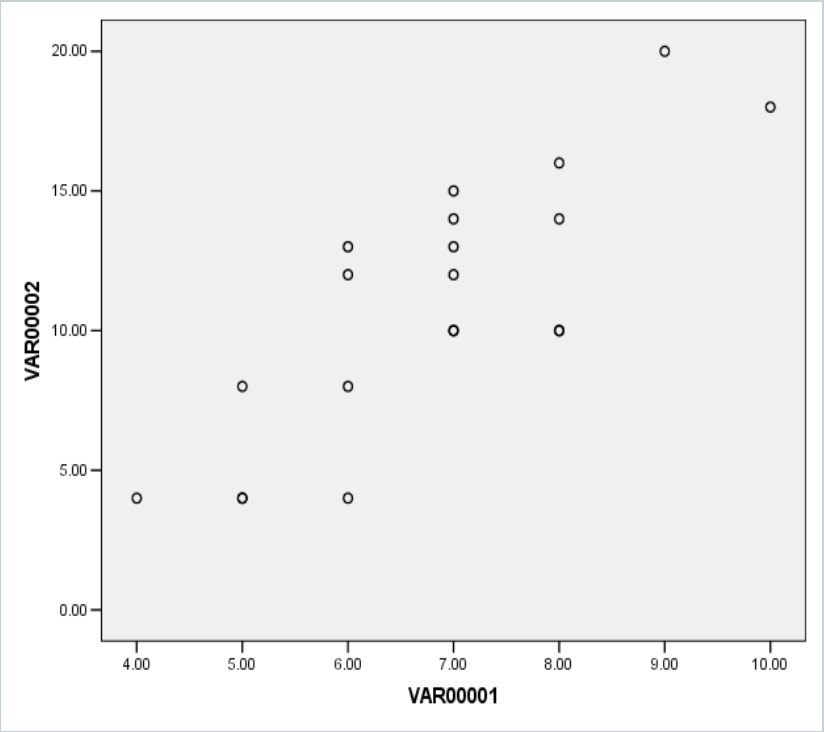
$$y_i = \beta_0 + \beta_1 x$$

- Ocjena linearnog regresionog modela na uzorku:

$$y_i^* = b_0 + b_1 x$$

Ocjenjivanje regresionih parametara

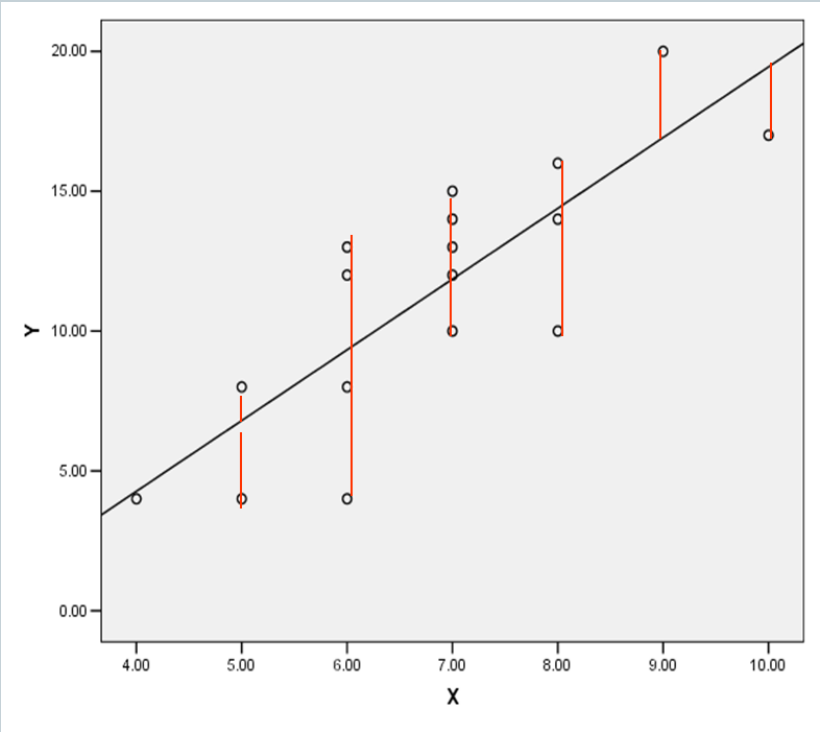
- Zadatak je ucrtati regresionu liniju, kroz raspored tačaka na dijagramu raspršenosti, tako da empirijske tačke budu “što bliže” regresionoj pravoj.



- Tačke su jedinice posmatranja. Koordinate tačaka su određene rezultatima jedinica posmatranja na X i Y varijabli.

Ocjenjivanje regresionih parametara: metoda najmanjih kvadrata

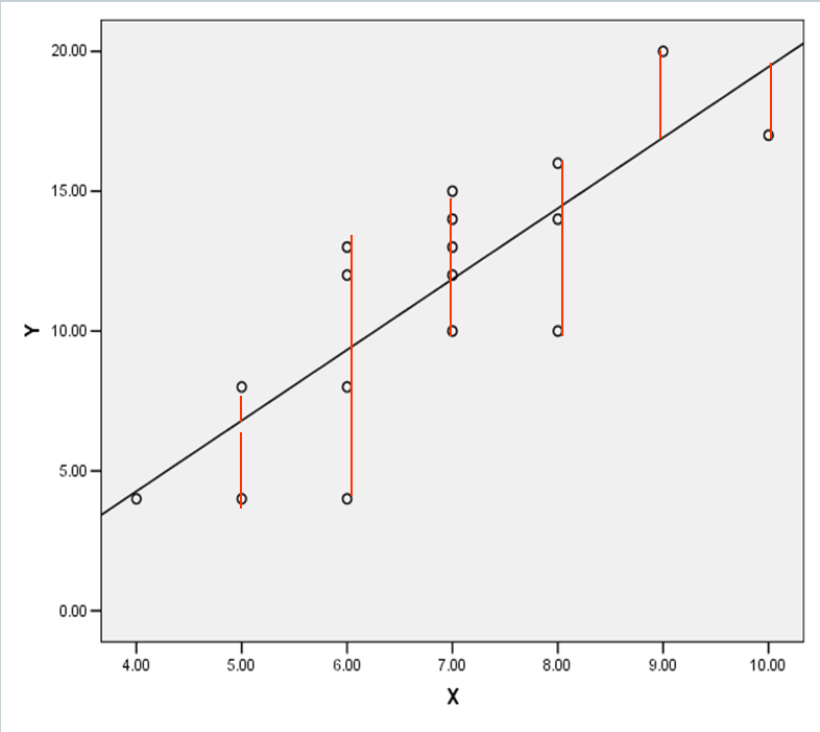
- MNK "što bliže" – naći minimum kvadrata odstupanja svih empirijskih tačaka od regresione linije.



- Crvene linije predstavljaju odstupanja empirijskih tačaka od regresione prave. Na reg.pravoj nalaze se predviđene vrijednosti y.

Ocjenjivanje regresionih parametara: metoda najmanjih kvadrata

- Vertikalna rastojanja između empirijskih tačkaka od regresione linije predstavljaju greške predviđanja.



Ocjenjivanje regresionih parametara: metoda najmanjih kvadrata

- Predviđena (ocijenjena) vrijednost zavisne promjenljive za jedinicu posmatranja

$$y_i^* = b_0 + b_1 x_i$$

- Empirijski dobijena vrijednost zavisne promjenljive za jedinicu posmatranja

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

- Greška predviđanja (rezidual)

$$\varepsilon_i = y_i - y_i^*$$

Ocjenjivanje regresionih parametara: metoda najmanjih kvadrata

- Matematički, problem se svodi na određivanje minimuma funkcije:

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - y_i^*)^2 = \sum_{i=1}^{N} (y_i - b_0 - b_1 x_{1i})^2$$

- Dakle, treba odrediti b0 i b1 kojima se postiže minimum funkcije. Do rješenja se dolazi nalaženjem parcijalnih izvoda funkcije po b0 i b1 i njihovim izjednačavanjem sa nulom.

Ocjenjivanje regresionih parametara: metoda najmanjih kvadrata

Intercept (slobodan član) regresione prave:

$$b_0 = M_y - b_1 M_x$$

Nagib regresione prave (koeficijent pravca)

$$b_1 = r \frac{SD_y}{SD_x}$$

$$b_1 = \frac{\sum x \sum y - n \sum xy}{(\sum x)^2 - n \sum x}$$

Koeficijent determinacije

- Koeficijent determinacije u linearnoj regresiji je kvadrat koeficijenta linearne korelacije
- Koeficijent determinacije daje informaciju o stepenu varijabiliteta u zavisnoj promjenljivoj koji se može objasniti varijacija u nezavisnoj promjenljivoj.
- Ukupan varijabilitet je jednak zbiru objašnjenog i neobjašnjenog varijabiliteta

$$r_{yx}^2 = \frac{SS_r}{SS_t} = \frac{\sum (y^* - M_y)^2}{\sum (y - M_y)^2}$$

$$r_{yx}^2 = b_1^2 \frac{\sum x^2 - NM_x^2}{\sum y^2 - NM_y^2}$$

Standardna greška regresije

- Standardna greška ocjene (regresije) je standardna devijacija distribucije reziduala.
- Govori o prosječnoj grešci koju pravimo kada predviđamo zavisnu promjenljivu na osnovu nezavisne.

$$SD_{y(x)} = SD_y \sqrt{1 - r_{xy}^2}$$

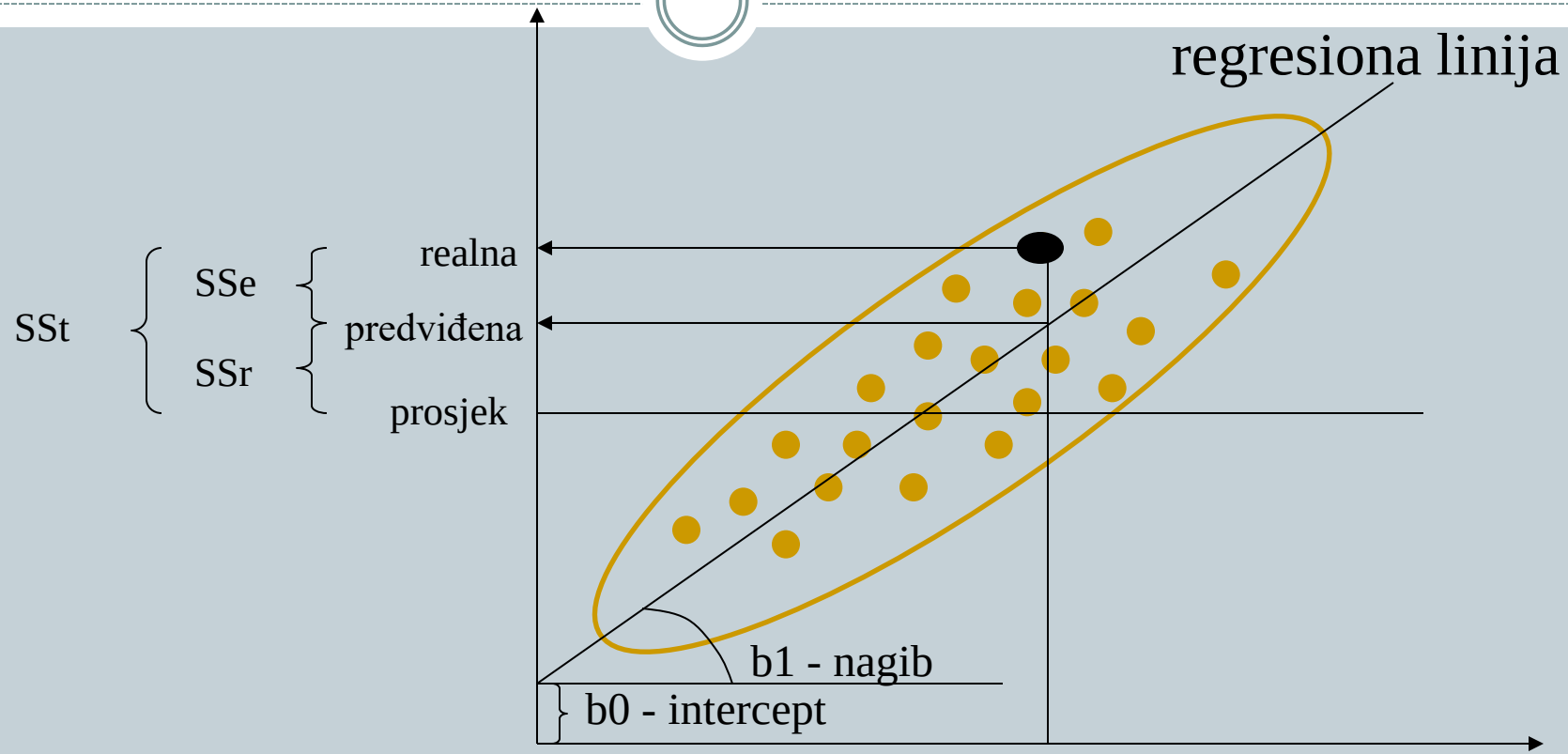
Testiranje nulte hipoteze o koeficijentu determinacije: F test

- H_0: $\rho^2 = 0$ (ρ^2 je koeficijent linearne determinacije u populaciji)
- F statistik:

$$F = r^2 \frac{N-2}{1-r^2}$$

- Ako je nulta hipoteza tačna, F statistik ima Snidikorovu F distribuciju uzorkovanja čiji stepeni slobode su 1 i n-2.

Šta je predikcija?



$Y=b1*X+b0$

SSt = SSr (predviđeno) + SSe (greška)

Testiranje značajnosti koeficijenta b_0

- Ho: $\beta_0 = 0$ (β_0 je intercept u populaciji)
- T statistik je:

$$t = \frac{b_0}{SE_{b_0}}$$

- Ako je Ho tačna, onda t-stat ima Studentovu distribuciju sa n-2 broja stepeni slobode
- Nulta hipoteza o interceptu se rijetko testira u psihologiji

Testiranje značajnosti koeficijenta b_1

- Ho: $\beta_1 = 0$ (β_1 je nagib u populaciji)
- T statistik je:

$$t = \frac{b_1}{SE_{b_1}}$$

- Ako je Ho tačna, onda t-stat ima Studentovu distribuciju sa n-2 broja stepeni slobode
- Testiranje nulte hipoteze o koeficijentu nabiga u jednostrukoj linearnoj regresiji isto je što i test nulte hipoteze o koeficijentu linearne korelacije.

O pogodnosti modela za predviđanje

- Što je koeficijent determinacije bliži 1, to je model "bolji"
- Što je veličina standardne greške regresije bliža nuli ili što je manja u odnosu na standardnu devijaciju zavisne varijable, to je model bolji.

Predviđanje pomoću regresionog modela

- Interpolacija – predviđanje vrijednosti zavisne varijable za jedinice posmatranja nezavisne varijable koje su korišćene u kontrukciji regresionog modela

- Ekstrapolacija - predviđanje vrijednosti zavisne varijable za jedinice posmatranja nezavisne varijable koje nijesu korišćene u kontrukciji regresionog modela

Uslovi za primjenu linearne regresije

- Bivarijanta normalna raspodjela varijabli u populaciji
- Kvantitativne varijable
- Linearan odnos među varijablama
- Nezavisna varijabla mjerena bez greške

Primjer

- 15 studenata je bilo upitano koliko su se sati pripremali za jedan kolokvijum iz statistike. Njihovi odgovori na to pitanje upoređeni su sa bodovima koje su dobili na kolokvijumu (max broj bodova je 100). U narednoj tabeli su prikazani ti rezultati.

Sati učenja (X)	0,5	0,8	1	1,3	1,5	1,8	2	2,3	2,5	2,8	3	3,3	3,5	3,8	4	2,3
Bodovi (Y)	57	64	59	68	74	76	79	83	85	86	88	89	90	94	96	79

Mx	2,25
My	79,2
SDx	1,12
SDy	12,48
r	0,973

a. Ocijeniti koeficijente linearne regresije.
b. Testirati statističku značajnost regresionog koeficijenta b1 (ako je SEb1=0,719 i tablična vrijednost za t=2,16).
c. Izračunati koeficijent determinacije i dati tumačenje.
d. Izračunati standardnu grešku regresije.
e. Ako je neki student pripremao ispit 0,25 sati, koji je njegov najvjerovatniji broj bodova na kolokvijumu?

Primjer

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.973[a]	.946	.942	3.00695

a. Predictors: (Constant), sati

$$b_0 = M_y - b_1 M_x$$

$$SD_{y(x)} = SD_y \sqrt{1 - r_{xy}^2}$$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2062.857	1	2062.857	228.148	.000[b]
	Residual	117.543	13	9.042		
	Total	2180.400	14			

a. Dependent Variable: bodovi

b. Predictors: (Constant), sati

$$b_1 = r \frac{SD_y}{SD_x}$$

$$y_i^* = b_0 + b_1 x_i = 54,7 + 10,8 x_i$$

$$y(0,25) = 57,4$$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	54.771	1.794		30.530	.000
	sati	10.857	.719	.973	15.105	.000

a. Dependent Variable: bodovi

Za vježbu

- Pretpostavimo da smo u jednom istraživanju varijabli X i Y dobili ove rezultate:

  Mx=600 My=4,8 r=0,58 SDx=100
  SDy=0,4 N=200

a. Koliko iznosi prognozirana vrijednost Y ako je X=70?
b. Koliko iznosi stepen objašnjenosti varijable Y u modelu?
c. Za koliko će se promijeniti Y ako se X promijeni za jednu jedinicu?