Predavanje 5

KORELACIJA I REGRESIJA

Pregled

- Regresijom i korelacijom analizira se povezanost (zavisnost, odnos) dve ili više varijabli .
- Dve varijable su povezane ako promenama jedne varijable odgovaraju promene kod druge varijable.
- Korelacija - analizira jačinu i smer povezanosti .
- Regresija - analizira oblik povezanosti.
- Regresioni model omogućava predikciju vrednosti zavisne varijable na osnovu poznavanja vrednosti nezavisnih varijabli.
- Posmatrane varijable treba da budu zavisne tj. da predstavljaju komponente višedimenzionalne slučajne promenljive.

Korelacija

Korelacija

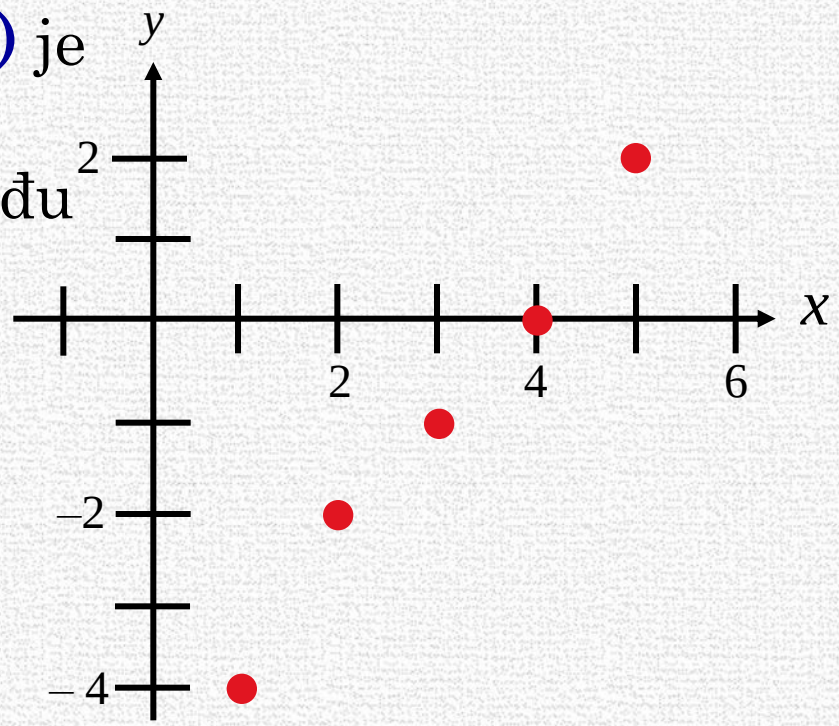
Korelacija je veza između dve varijable.

Dvodimenzionalni podaci su predstavljeni uređenim parovima (x, y).

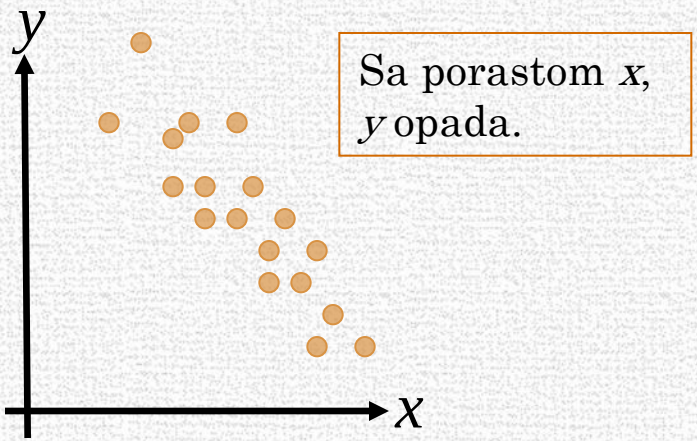
Dijagram rasipanja (scatter plot) je grafikon koji nam sugeriše postojanje i oblik zavisnosti između varijabli.

Primer:

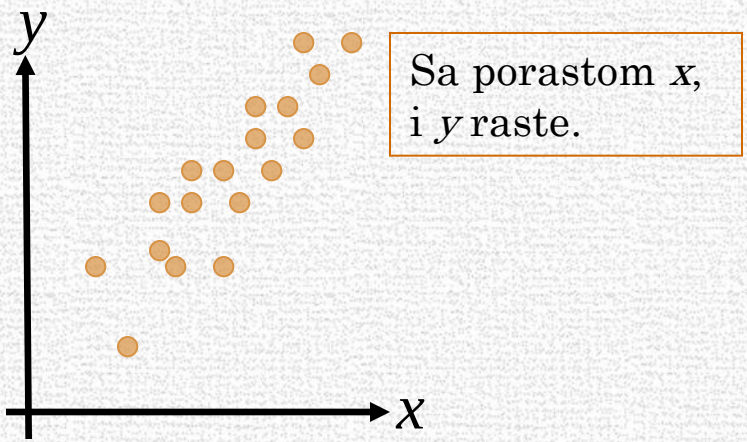
x	1	2	3	4	5
y	– 4	– 2	– 1	0	2



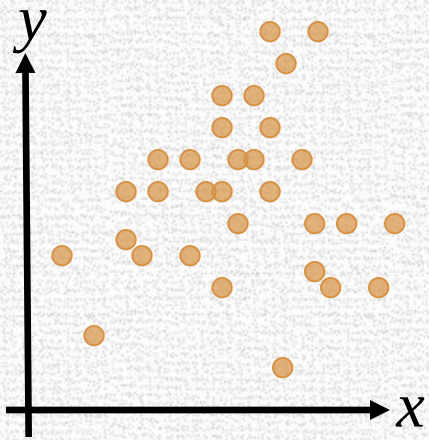
Linearna korelacija



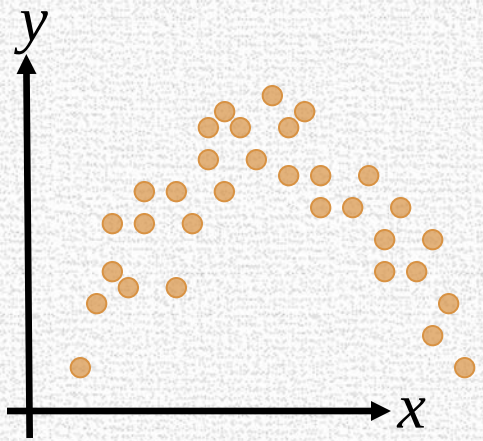
Negativna linearna korelacija



Pozitivna linearna korelacija



Nema korelacije



Nelinearna korelacija

Koeficijent korelacije

Koeficijent korelacije (Pirsonov, r) je mera jačine i smera linearne veze između dve varijable.

$$r = \frac{S_{XY}}{S_X S_Y}$$

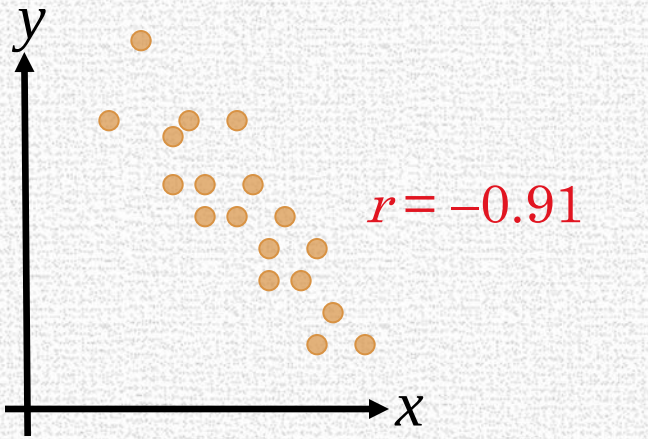
kovarijansa

Standardne devijacije

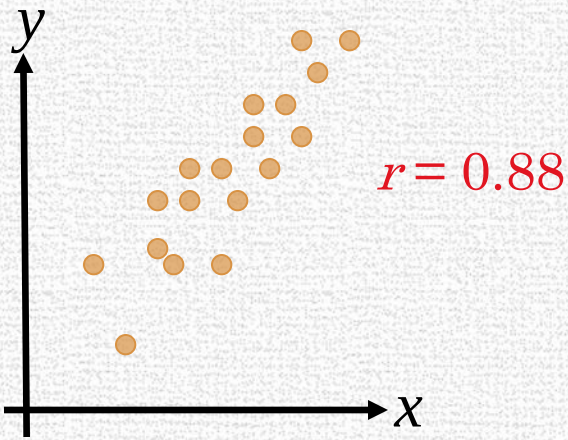
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

- Vrednost koeficijenta *r* je u intervalu od –1 do 1.
- Ako su *x* i *y* u jakoj pozitivnoj /negativnoj korelaciji, *r* je blizu 1/-1.
- Ako *x* i *y* nisu u linearnoj korelaciji, *r* je blizu 0.

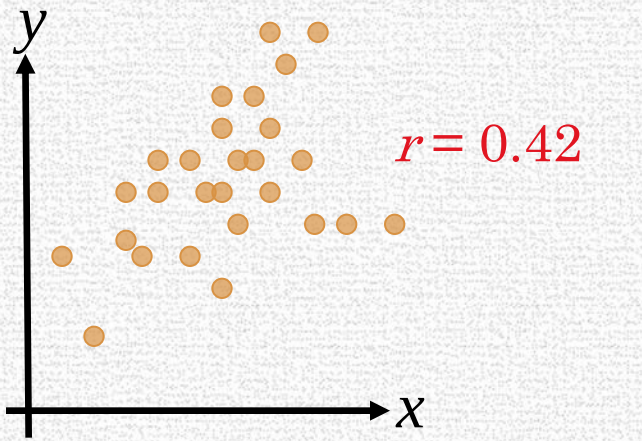
Primeri koeficijenta korelacije



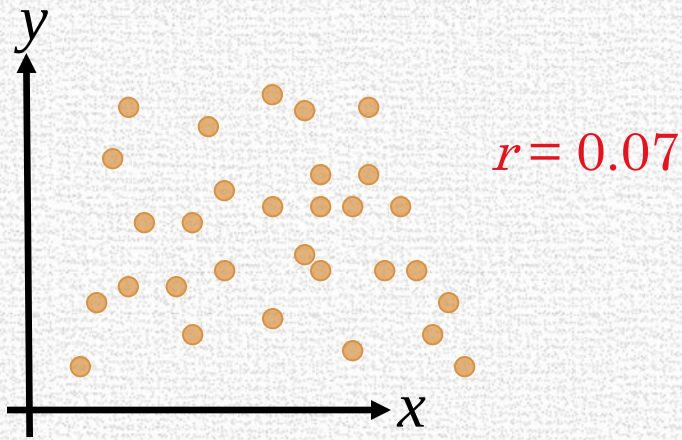
Jaka negativna korelacija



Jaka pozitivna korelacija



Slaba pozitivna korelacija



Nema linearne korelacije

Primer:

Izračunati i okarakterisati koeficijent korelacije:

x	y	xy	x^2	y^2
1	-3	-3	1	9
2	-1	-2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{5(9) - (15)(-1)}{\sqrt{5(55) - 15^2}\sqrt{5(15) - (-1)^2}} \approx 0.986$$

Postoji jaka pozitivna linearna korelacija između x i y.

Testiranje koeficijenta korelacije populacije

- Na osnovu izračunatog koeficijenta korelacije *r* možemo proceniti značajnost koeficijenta korelacije populacije ρ.
- Statistički test, sa nultom hipotezom $H_0(\rho=0)$.
- Ako je p-vrednost testa manja od zadatog praga značajnosti, odbacujemo nultu hipotezu i zaključujemo da postoji statistički značajna korelacija.
- Test je veoma senzitivan na obim uzorka (broj parova podataka), npr.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

Tabela prikazuje granične vrednosti značajnih koeficijenata korelacije $|r|$.

Test statistika

t-Test za koeficijent korelacije

Za određivanje značajnosti koeficijenta korelacije između dve promenljive koristimo sledeću test statistiku:

$$t = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Koja ima Studentovu raspodelu sa $n - 2$ stepena slobode.

Primer:
Podaci predstavljaju broj sati provedenih na FB tokom vikenda i broj osvojenih poena na testu u ponedeljak. Proveriti da li sa pragom značajnosti $\alpha = 0.01$ postoji značajna linearna korelacija?

Broj sati, x	0	1	2	3	3	5	5	5	6	7	7	10
Broj poena, y	96	85	82	74	95	68	76	84	58	65	75	50

- *Koeficijent korelacije* $r \approx -0.831$.

Nastavak primera:

H_0: $\rho = 0$ (nema korelacije) H_a: $\rho \neq 0$ (značajna korelacija)

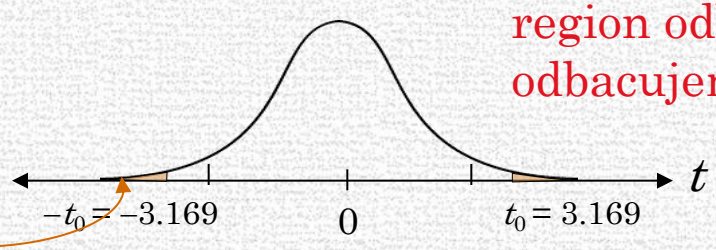
Prag značajnosti je $\alpha = 0.01$.

d.f. = 12 – 2 = 10.

Kritične vrednosti: $-t_0 = -3.169$ i $t_0 = 3.169$.

Test statistika:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.831}{\sqrt{\frac{1-(-0.831)^2}{12-2}}} \approx -4.72.$$

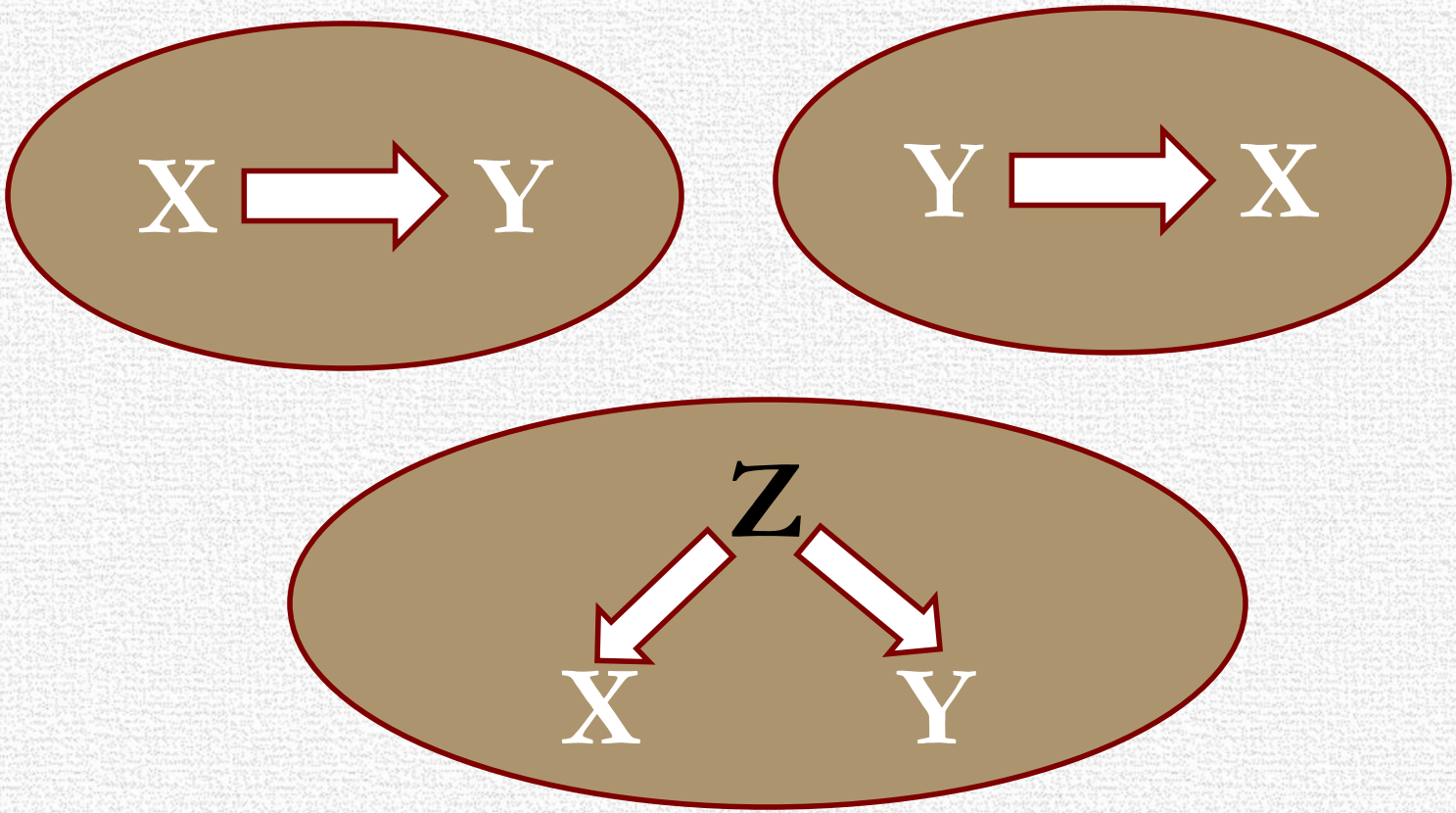


Test statistika upada u region odbacivanja, pa odbacujemo H_0.

Sa pragom značajnosti 1% zaključujemo da postoji značajna negativna korelacija između vremena provedenog na FB i ostvarenog uspeha na testu u ponedeljak.

Korelacija i kauzalnost

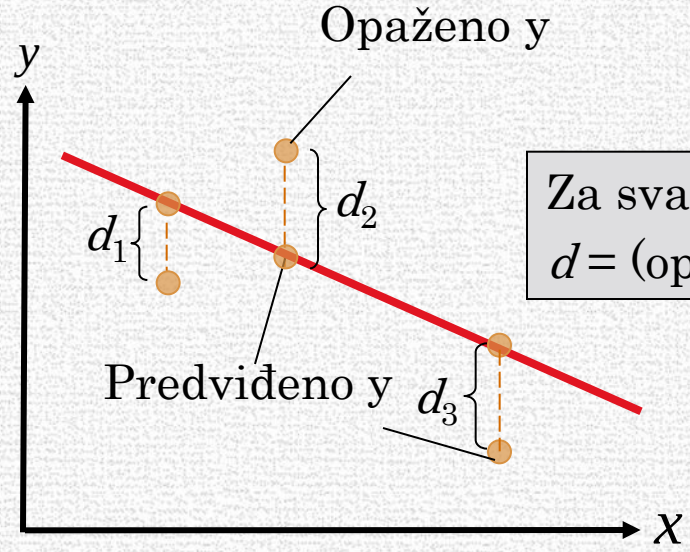
Činjenica da su dve varijable u jakoj korelaciji **ne** podrazumeva da između njih postoji uzročno-posledična tj. kauzalna veza.
Postoje 3 mogućnosti:



LINEARNA REGRESIJA

Reziduali

Ako između x i y postoji značajna linearna korelacija, možemo odrediti jednačinu prave koja tu zavisnost opisuje. Ta prava je prava linearne regresije.



Za svako dato x,
d = (opaženo y) – (predviđeno y)

Za svaki par (x_i , y_i) možemo izračunati razliku opažene i predviđene vrednosti d_i, koja se zove **rezidual**.

Regresiona prava

Regresiona prava je ona prava za koju je suma kvadrata reziduala minimalna.

Jednačina regresione prave

$$\hat{y} = mx + b$$

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \text{ and } b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

where $\bar{y}$ is the mean of the y-values and $\bar{x}$ is the mean of the x-values. The regression line always passes through $(\bar{x}, \bar{y})$.

Primer:

Odrediti jednačinu regresione prave.

x	y	xy	x^2	y^2
1	-3	-3	1	9
2	-1	-2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

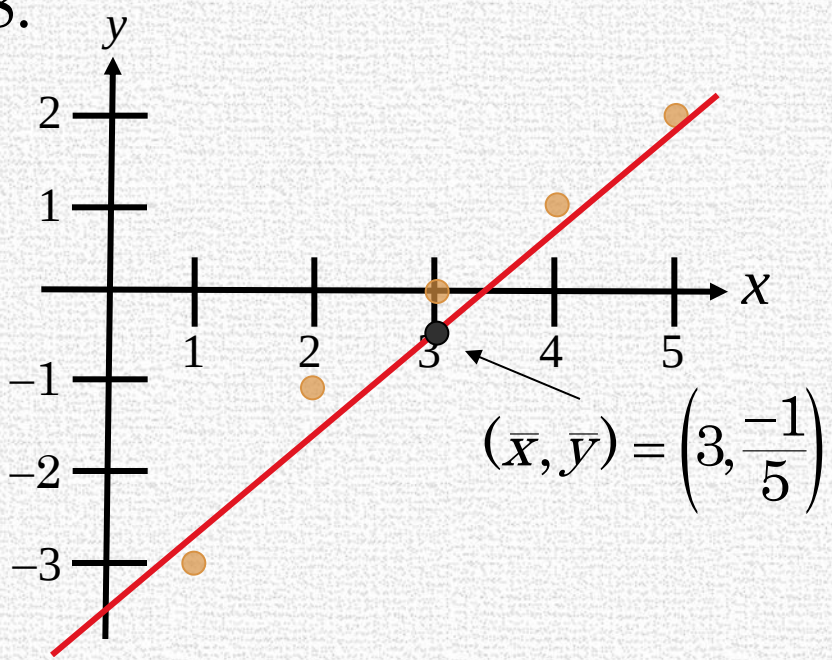
$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{5(9) - (15)(-1)}{5(55) - (15)^2} = \frac{60}{50} = 1.2$$

Nastavak primera:

$$b = \bar{y} - m\bar{x} = \frac{-1}{5} - (1.2)\frac{15}{5} = -3.8$$

Jednačina regresione prave glasi:

$$\hat{y} = 1.2x - 3.8.$$



Primer:

Podaci predstavljaju broj sati provedenih na FB tokom vikenda i broj osvojenih poena na testu u ponedeljak.

a) odrediti jednačinu regresione prave.

b) iskoristiti dobijenu jednačinu za predikciju broja bodova studenta koji je proveo 9h na FB.

Broj sati, x	0	1	2	3	3	5	5	5	6	7	7	10
Broj poena, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

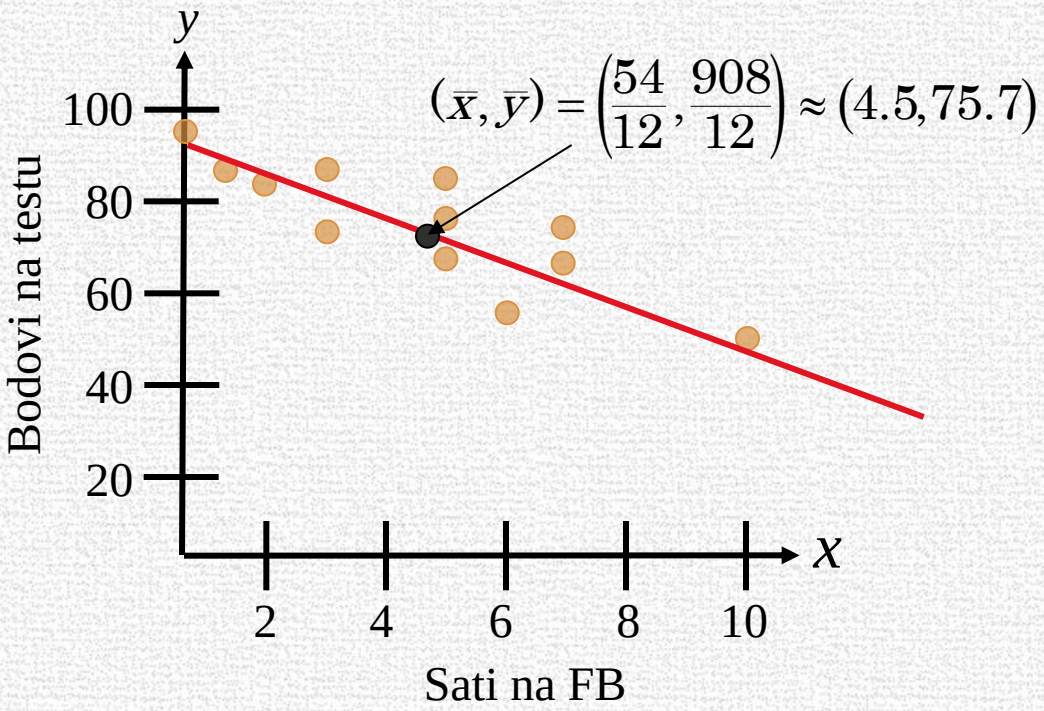
$\sum x = 54$ $\sum y = 908$ $\sum xy = 3724$ $\sum x^2 = 332$ $\sum y^2 = 70836$

Nastavak primera:

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{12(3724) - (54)(908)}{12(332) - (54)^2} \approx -4.067$$

$$b = \bar{y} - m\bar{x}$$
$$= \frac{908}{12} - (-4.067)\frac{54}{12}$$
$$\approx 93.97$$

$$\hat{y} = -4.07x + 93.97$$



Nastavak primera:

Koristeći jednačinu $\hat{y} = -4.07x + 93.97$, možemo predvideti bodove za provedenih 9 sati na FB.

$$\hat{y} = -4.07x + 93.97$$
$$= -4.07(9) + 93.97$$
$$= 57.34$$

Možemo očekivati da student koji provede 9 sati na FB tokom vikenda osvoji 57 bodova na testu u ponedeljak.

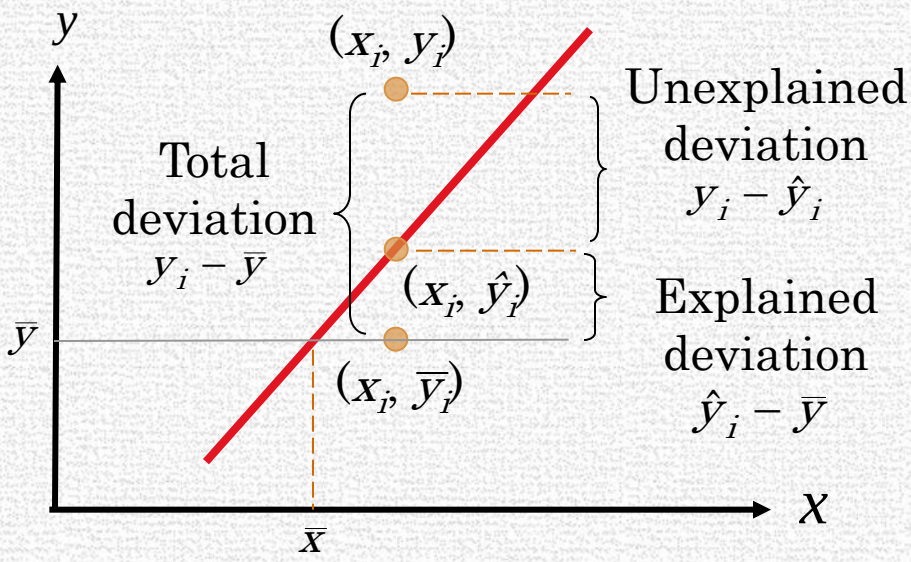
Predikcija može biti **interpolacija (unutar opsega X-a)** i **ekstrapolacija (van opsega X-a)**.

Varijacija regresione prave

Da bi izračunali totalnu varijaciju regresione prave, moramo izračunati: **totalnu devijaciju, objašnjenu devijaciju** i **neobjašnjenu devijaciju**.

Total deviation $= y_i - \bar{y}$ Explained deviation $= \hat{y}_i - \bar{y}$

Unexplained deviation $= y_i - \hat{y}_i$



Totalna varijacija regresione prave predstavlja sumu kvadrata totalnih devijacija (razlika opažene i izračunate srednje vrednosti y)

$$\text{Total variation} = \sum (y_i - \bar{y})^2$$

Objašnjena varijacija regresione prave predstavlja sumu kvadrata objašnjenih devijacija

$$\text{Explained variation} = \sum (\hat{y}_i - \bar{y})^2$$

Nebjašnjena varijacija regresione prave predstavlja sumu kvadrata neobjašnjenih devijacija (razlika opažene i predviđene vrednosti y)

$$\text{Unexplained variation} = \sum (y_i - \hat{y}_i)^2$$

Totalna var = Objašnjena var + neobjašnjena var

Koeficijent determinacije

Koeficijent determinacije r^2 predstavlja odnos objašnjene i ukupne varijacije:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Primer:
Ako je koeficijent korelacije između vremena na FB i uspeha na testu $r \approx -0.831$, izračunaj i prokomentariši koeficijent determinacije.

$$r^2 \approx (-0.831)^2$$
$$\approx 0.691$$

Oko 69.1% varijacije kod uspeha na testu je objašnjeno vremenom na FB. Ostalih 30.9% varijacije je neobjašnjeno.

Standardna greška procene

Kada se procenjena $\hat{y}$-vrednost računa na osnovu prave za datu vrednost x, u pitanju je tačkasta ocena.

Moguće je konstruisati i interval poverenja za y.

Standardna greška procene s_e je srazmerna neobjašnjenoj varijaciji, i izračunava se po sledećoj formuli (gde je n broj parova podataka):

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$$

Što su bliže predviđene i opažene vrednosti, to je manja standardna greška procene.

Primer:
Za date parove podataka je određena regresiona prava

$$\hat{y} = 1.2x - 3.8.$$

Izračunati standardnu grešku procene.

x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	– 3	– 2.6	0.16
2	– 1	– 1.4	0.16
3	0	– 0.2	0.04
4	1	1	0
5	2	2.2	0.04
			$\sum = 0.4$

Neobjašnjena varijacija

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{0.4}{5-2}} \approx 0.365$$

Standardna greška procene iznosi približno 0.365.

Interval poverenja za y

Za datu jednačinu regresione prave $\hat{y} = mx + b$ i za datu vrednost x_0 (konkretnu vrednost promenljive x), ***c*-interval poverenja** za y je

$$\hat{y} - E < y < \hat{y} + E$$

Gde je

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}.$$

E je margina greške, $\hat{y}$ je tačkasta predviđena vrednost. Verovatnoća da interval sadrži pravu vrednost y iznosi *c*.

Interval poverenja

Primer:
Za podatke koji predstavljaju vreme provedeno na FB tokom vikenda i broj osvojenih bodova na testu u ponedeljak, konstruisati 95% interval poverenja za predikciju broja bodova studenta koji je 4 sata proveo na FB.

sati, x	0	1	2	3	3	5	5	5	6	7	7	10
poeni, y	96	85	82	74	95	68	76	84	58	65	75	50

$$\hat{y} = -4.07x + 93.97$$

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{658.25}{12-2}} \approx 8.11$$

Nastavak primera:

Broj stepeni slobode je $n - 2 = 12 - 2 = 10$.

Tačkasta ocena broja bodova je:

$$\hat{y} = -4.07x + 93.97 = -4.07(4) + 93.97 = 77.69.$$

Kritična vrednost je $t_c = 2.228$, i margina $E = 18.83$.

$$\hat{y} - E < y < \hat{y} + E$$

$77.69 - 18.83 = 58.86$ $77.69 + 18.83 = 96.52$

Sa sigurnošću 95% možemo tvrditi da će broj bodova kod studenta koji provede 4h na FB biti između 58.86 i 96.52.

Regresiona analiza - rezime

- Regresija koju smo videli je najjednostavniji vid regresione analize – **jednostruka linearna regresija**.
- Postoje mnoga uopštenja:
  - **Višestruka linearna regresija** – više nezavisnih varijabli, jedna zavisna varijabla, linearna veza.
  - **Polinomijalna regresija** – jedna ili više nezavisnih varijabli, polinomna veza.
  - **Logistička regresija** – koristi se za kombinaciju neprekidnih i kategoričkih nezavisnih varijabli i jednu, obično dihotomnu zavisnu varijablu (npr. predikcija pojave bolesti u zavisnosti od pojave više simptoma).
  - ...