STATISTIKA

REGRESIJA I KORELACIJA

REGRESIJA I KORELACIJA

- Jedan od zadataka statistike je da otkrije veze koje postoje među masovnim pojavama.
- U zavisnosti od broja posmatranih pojava **veza** među njima može biti **prosta i višestruka**.
- Veza je prosta ukoliko povezuje dve pojave, a višestruka ukoliko opisuje slaganje više od dve pojave, s tim što je uvek jedna pojava zavisno promenljiva veličina, a druga ili druge nezavisno promenljive ili veličine.

REGRESIJA I KORELACIJA

- **Prema intenzitetu**, međusobna zavisnost pojava može biti različita.
- Najjača je strogo određena ili **funkcionalna** veza pri kojoj svakoj vrednosti jedne pojave odgovara samo jedna vrednost druge pojave.
- Najslabija je potpuno **nedeterministička** veza kada između pojava ne postoji zavisnost.
- Sve veze između ova dva ekstrema nazivaju se **korelacionim,** stohastičkim ili statističkim.
- Kod stohastičkih veza jednoj vrednosti jedne pojave može odgovarati više vrednosti druge pojave.

REGRESIJA I KORELACIJA

- **Smer veza** među pojavama može biti **pozitivan ili negativan.**
- Kada se promene dve pojave kreću u istom smeru veza između njih je pozitivna, a kada porast jedne pojave prati pad druge, tj. kada se pojave kreću u suprotnom smeru njihova veza je negativna.
- Pored smera i jačine, veze među pojavama se mogu razlikovati i po obliku.
- ***Regresiona analiza otkriva oblik povezanosti posmatranih pojava. Metodom korelacije se meri jačina veze između pojava.***

REGRESIONA ANALIZA

- Dve kvantitativne promenljive mogu biti povezane na više načina. Svaka ta moguća veza predstavlja jedan regresioni model.
- *Pojava koja utiče i uslovljava veličinu druge pojave zove se nezavisno promenljiva, a pojava na koju ona utiče zove se zavisno promenljiva.*
- Cilj regresione analize je da se odredi onaj regresioni model koji najbolje opisuje vezu između pojava i da se na osnovu tog modela ocene i predvide vrednosti zavisne promenljive Y za odabrane vrednosti nezavisne promenljive X.

REGRESIONA ANALIZA

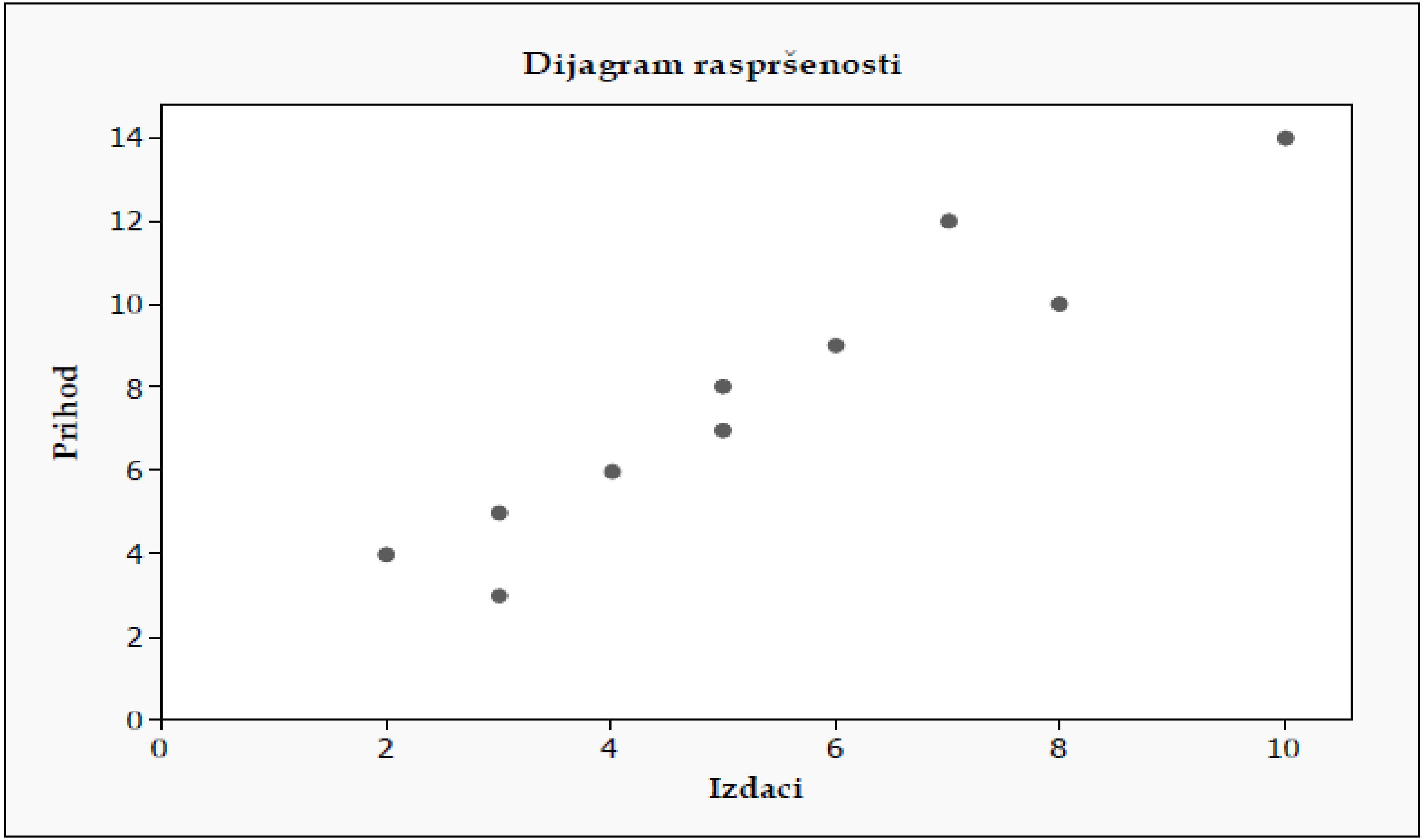
- Odnos promenljivih može biti različit i zato je prvi korak u određivanju veze između dve pojave prikupljanje parova njihovih vrednosti $(x_i;y_i)$, koji se zatim prikazuju u pravouglom koordinatnom sistemu tačkastim dijagramom pod nazivom **dijagram rasturanja ili dijagram raspršenosti.**
- Iz izgleda dijagrama rasturanja može se vizuelno odrediti oblik funkcije kojom se može aproksimirati veza analiziranih promenljivih.

REGRESIONA ANALIZA

Tabela 11.1 Izdaci za propagandu i prihod od prodaje 10 računarskih firmi, na osnovu slučajnog uzorka

Firma	A	B	C	D	E	F	G	H	I	J
Izdaci za propagandu	8	10	3	3	2	7	5	6	5	4
Prihod od prodaje	10	14	3	5	4	12	8	9	7	6

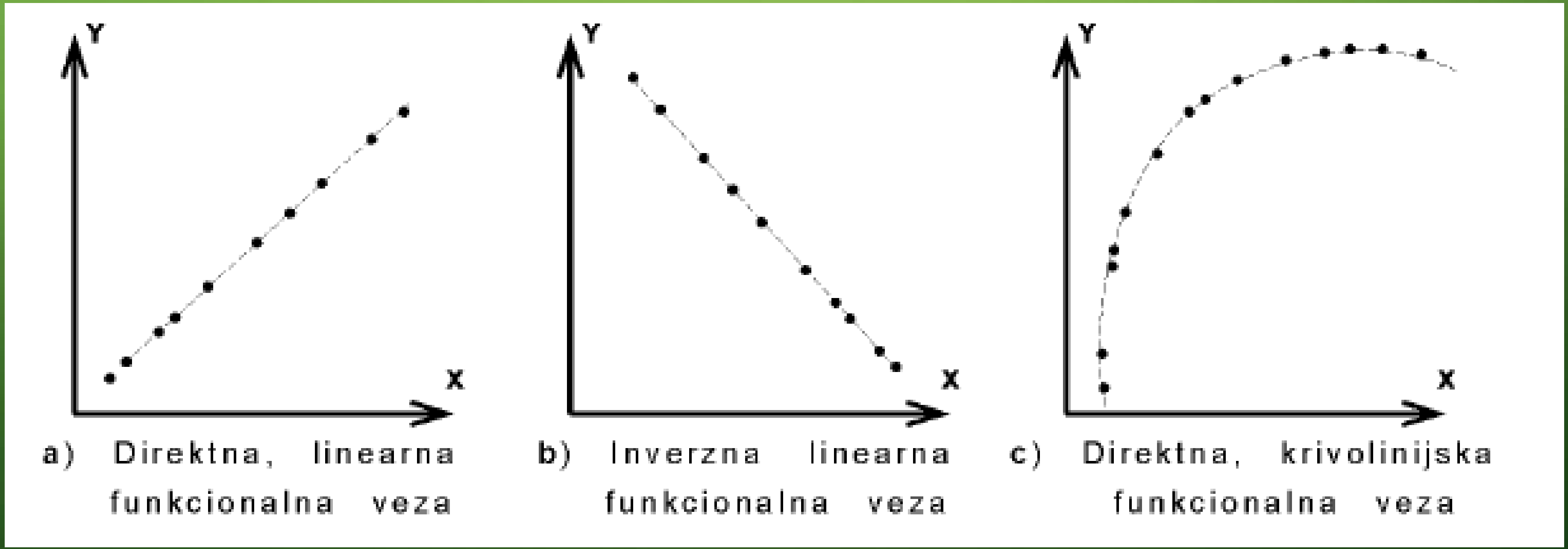
REGRESIONA ANALIZA



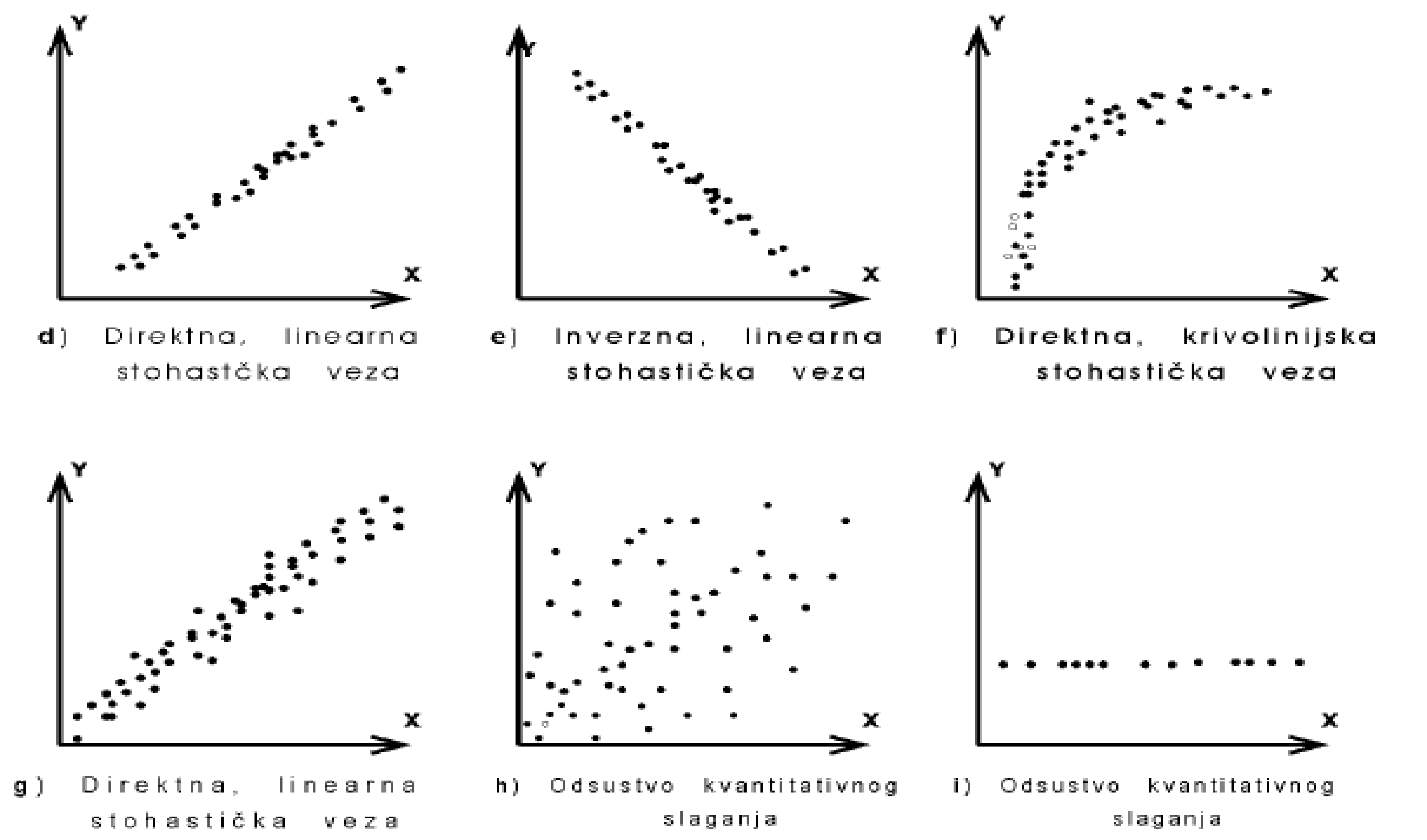
Slika 11.1 Dijagram raspršenosti za podatke Tabele 11.1

REGRESIONA ANALIZA

- Na narednim slikama su prikazani različiti oblici veze između dve pojave na dijagramima raspršenosti.



REGRESIONA ANALIZA



REGRESIONA ANALIZA

- *Jednačina funkcije koja odslikava vezu među pojavama zove se* ***jednačina regresije*****.**
- Prost linearni regresioni model je najjednostavniji i najčešće korišćeni regresioni model i ima oblik:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministički deo modela}} + \underbrace{\varepsilon_i}_{\text{stohastički deo modela}} \qquad i = 1,2,\ldots,N$$

REGRESIONA ANALIZA

Y_i i-ta zavisna promenljiva

x_i i-ta vrednost objašnjavajuće promenljive

β_0 i β_1 su regresioni parametri: β_0 je odsečak ili slobodni član, a β_1 nagib

ε_i stohastički član ili slučajna greška

- Regresioni parametar β_0 definiše se kao **prosečan početni nivo zavisne promenljive,** odnosno pokazuje prosečnu vrednost zavisne promenljive kada je nezavisna promenljiva jednaka nuli.

REGRESIONA ANALIZA

- Regresioni parametar β_1 (nagib) govori za koliko se jedinica u proseku promeni zavisno promenljiva Y kada se nezavisno promenljiva X promeni za jednu jedinicu.
- Ocenjivanje regresionog modela podrazumeva ocenjivanje nepoznatih parametara: slobodnog člana β_0 i koeficijenta nagiba β_1.
- Cilj je da se na osnovu uzorka dođe do najboljih mogućih ocena b_0 i b_1 i time postavi linija regresije u uzorku:

REGRESIONA ANALIZA

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

- Za ocenu parametara β_0 i β_1 najčešće se koristi metod najmanjih kvadrata koji se zasniva na minimiziranju kvadrata odstupanja stvarne vrednosti zavisne promenljive od njene ocenjene vrednosti iz uzorka:

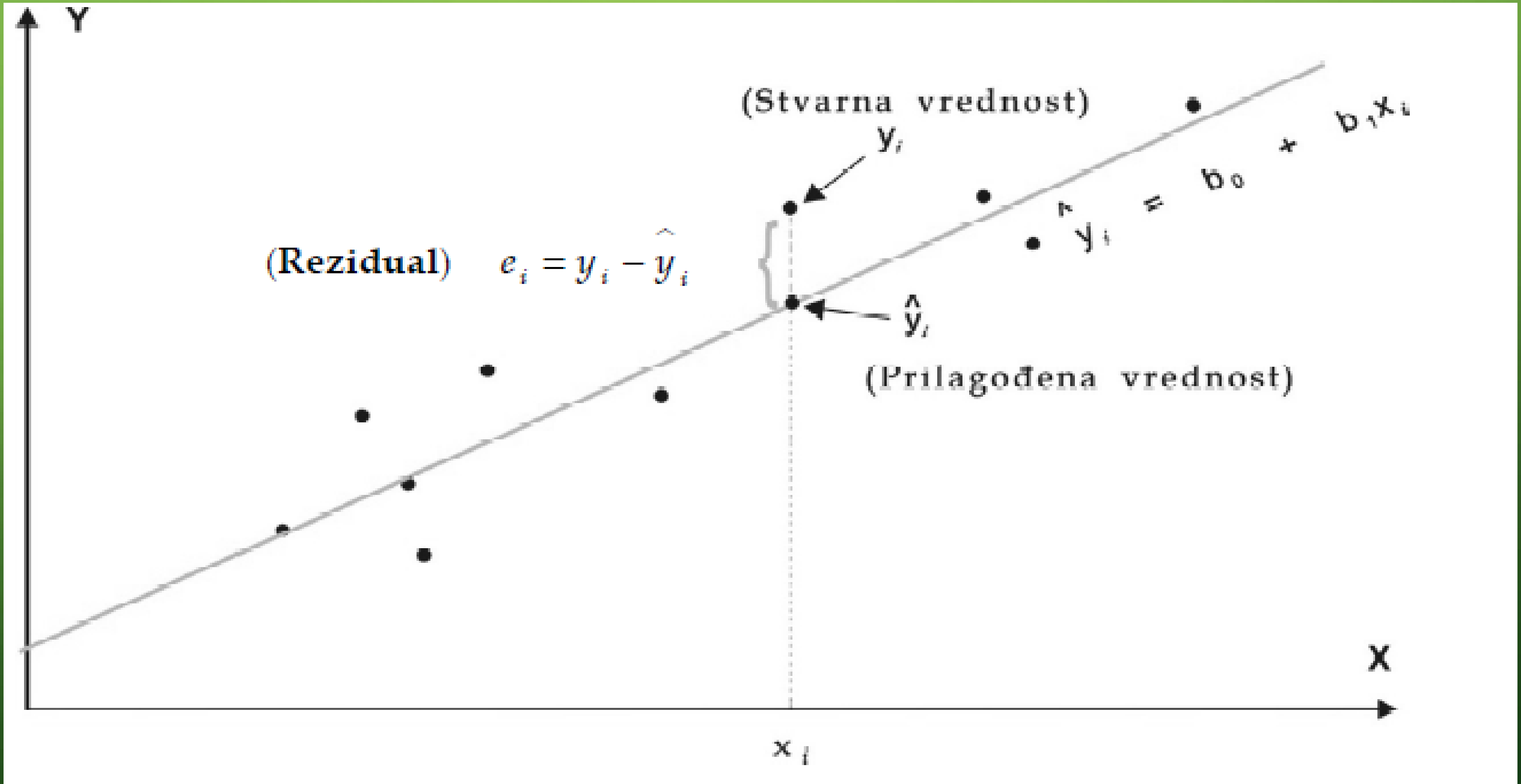
$$\sum (y_i - \hat{y}_i)^2 = \text{minimum}$$

REGRESIONA ANALIZA

- Razliku između stvarene vrednosti zavisne promenljive i njene ocenjene vrednosti nazivamo rezidualom i označavamo sa ei

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

REGRESIONA ANALIZA



REGRESIONA ANALIZA

- Primenom metoda najmanjih kvadrata dobija se sledeći sistem od dve jednačine sa dve nepoznate koje se nazivaju normalnim jednačinama

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

REGRESIONA ANALIZA

- Formule za ocenjivanje parametara regresionog modela metodom najmanjih kvadrata su:

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

$$b_0 = \overline{y} - b_1\overline{x}$$

REGRESIONA ANALIZA

Firma	Ulaganje u propagandu (x)	Prodaja (y)	xy	x^2	y^2
A	8	10	80	64	100
B	10	14	140	100	196
C	3	3	9	9	9
D	3	5	15	9	25
E	2	4	8	4	16
F	7	12	84	49	144
G	5	8	40	25	64
H	6	9	54	36	81
I	5	7	35	25	49
J	4	6	24	16	36
Σ	53	78	489	337	720

REGRESIONA ANALIZA

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 \cdot 489 - 53 \cdot 78}{10 \cdot 337 - 53^2} = 1,3476$$

$$b_0 = \overline{y} - b_1\overline{x} = \frac{78}{10} - 1,3476 \cdot \frac{53}{10} = 0,6577$$

- Ocenjeni regresioni model ima oblik

$$\hat{y}_i = 0,6577 + 1,3476x_i$$

REGRESIONA ANALIZA

- U prostoj linearnoj regresiji najvažnije je testirati hipotezu da li je parametar nagiba ß1 jednak nuli.
- Prema tome, postavljamo nultu hipotezu da između varijacija posmatranih pojava u osnovnom skupu ne postoji linearna veza, odnosno da *X* ne utiče na *Y*:

$$H_0 : \beta_1 = 0$$

REGRESIONA ANALIZA

- Protiv alternativne hipoteze $H_1 : \beta_1 \neq 0$
- Statistika testa ima oblik : $t = \dfrac{b_1}{s_{b_1}}$
- i sledi Studentov raspored sa (*n*-2) stepena slobode.
- Ocena standardne greške nagiba $s_{b_1} = \dfrac{s}{\sqrt{\sum x^2 - n\bar{x}^2}}$

REGRESIONA ANALIZA

- Pri čemu je s standardna greška regresije i dobija se po formuli

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

- **Standardna greška regresije** je apsolutna mera i pokazuje odstupanja empirijskih podataka u uzorku od regresione linije uzorka.

KORELACIONA ANALIZA

- Cilj korelacione analize je da se utvrdi da li između varijacija posmatranih pojava postoji kvantitativno slaganje (korelaciona veza) i, ako postoji, u kom stepenu.
- Ako se pri tome posmatraju dve pojave, govori se o **prostoj** korelaciji, a prilikom analize više pojava o **višestrukoj** korelaciji.
- Kao mera jačine proste linearne korelacione veze u uzorku koristi se P**irsonov koeficijent proste linearne korelacije**, ili samo koeficijent korelacije.

KORELACIONA ANALIZA

- Ovaj koeficijent pokazuje **stepen pravolinijskog kvantitativnog slaganja dve pojave.** Označava se sa *r* i izračunava po formuli:

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

- Koeficijent proste linearne korelacije uzima vrednosti **od -1 do +1**.

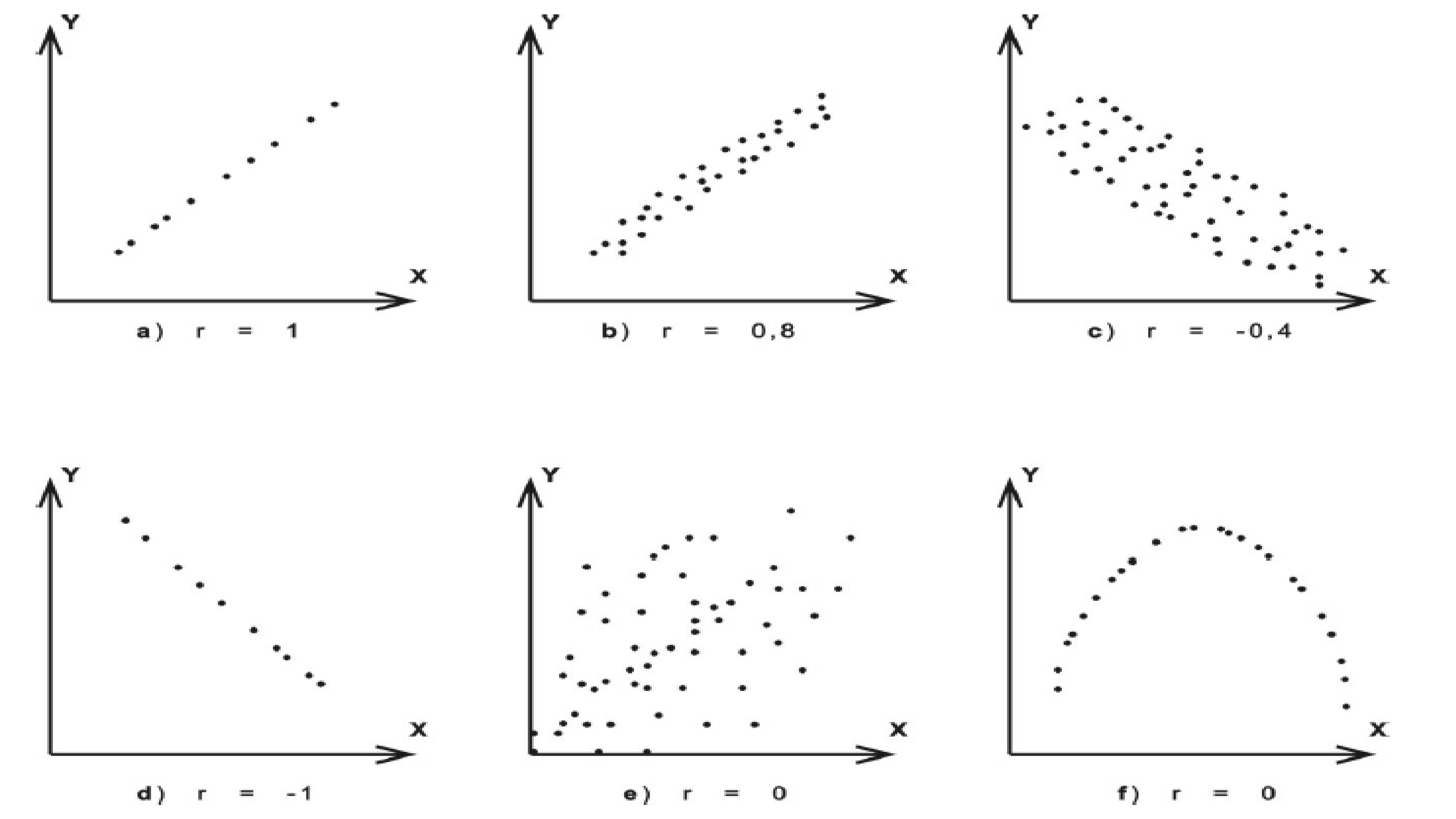
KORELACIONA ANALIZA

- Ukoliko uzima pozitivne vrednosti, korelacija između pojava je **direktna** ili **pozitivna** (obe pojave se kreću u istom smeru).
- U slučaju kada je $r < 0$, veza je **inverzna** ili **negativna** (kada jedna pojava raste druga opada, i obrnuto).
- Ako između posmatranih pojava postoji funkcionalna veza govorimo o **savršenoj** (perfektnoj) korelaciji. Tada koeficijent korelacije uzima vrednost -1 (ako je veza inverzna) ili +1 (ako je veza direktna).

KORELACIONA ANALIZA

- Što je koeficijent korelacije po apsolutnoj vrednosti bliži jedinici, sve je jača korelaciona veza između pojava. Nasuprot tome, što je bliži nuli **linearna** veza je slabija.
- Kada koeficijent korelacije uzme vrednost jednaku nuli, zaključuje se da **nema linearne veze** između pojava.

KORELACIONA ANALIZA



KORELACIONA ANALIZA

- Prilikom testiranja značajnosti ocenjenog koeficijenta proste linearne korelacije postavlja se nulta hipoteza u obliku

$$H_0 : \rho = 0$$

- odnosno, da u osnovnom skupu **ne postoji linearna** korelacija, ili, što je isto, da ocena, *r*, **nije** statistički značajna.
- Alternativna hipoteza ima oblik

KORELACIONA ANALIZA

$$H_1 : \rho \neq 0$$

- Za testiranje statističke značajnosti proste linearne korelacije koristi se *t* test sa *n* – 2 stepeni slobode, a test statistika glasi: $$t = \frac{r}{s_r}$$
- gde je *Sr* standardna greška ocene koeficijenta proste linearne korelacije i računa se prema formuli: $$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$