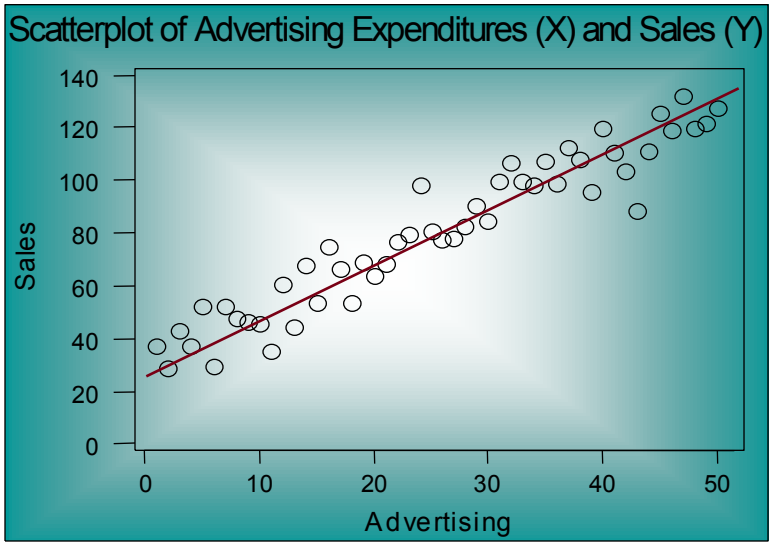
12 Prosta linearna regresija i korelacija

- Model proste linearne regresije
- Ocjena: Metod najmanjih kvadrata
- Standardna greška regresije
- Testovi hipoteza o regresionoj vezi
- Koliko je dobar regresioni model?
- Korelacija
- Korišćenje regresionog modela za prognozu

12-1 Statistike

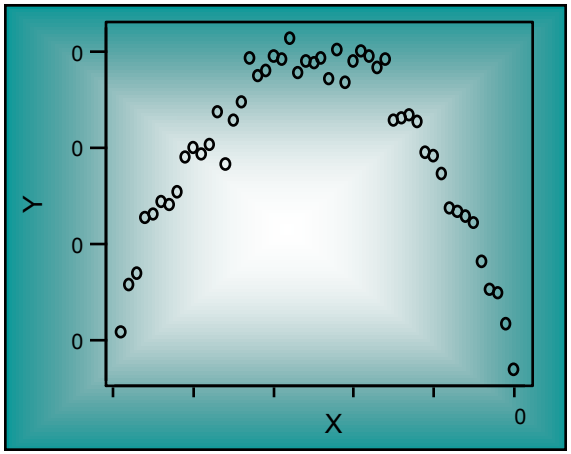
Dijagram rasturanja locira parove podataka troškova reklame na x-osi i prodaja na y-osi.

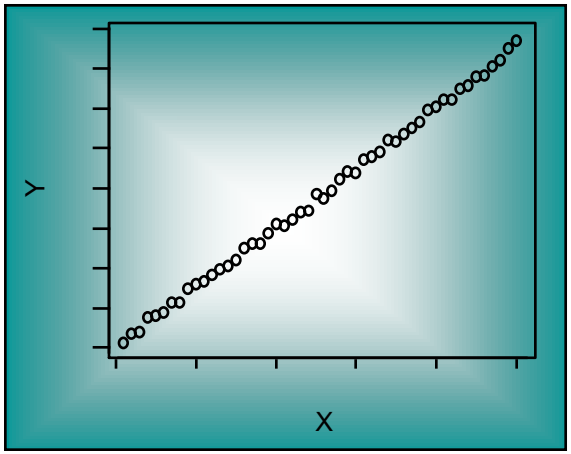


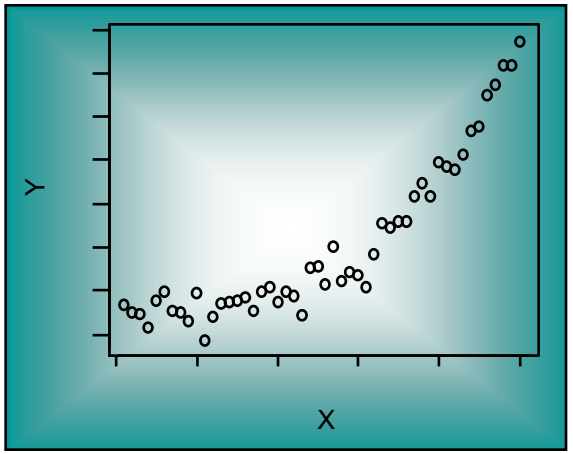
Veće (manje) vrijednosti prodaja *se pridružuju* većim (manjim) vrijednostima reklamiranja.

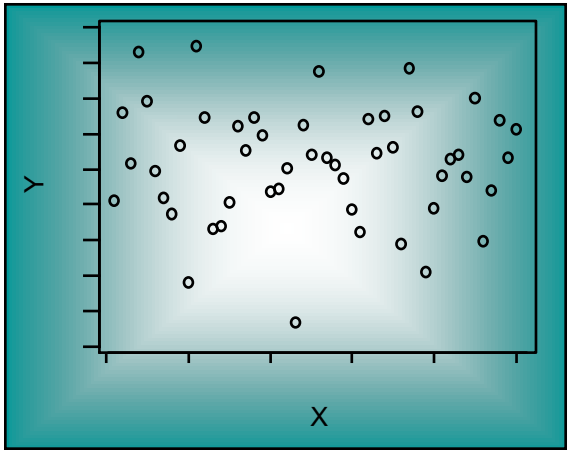
Tendencija – ka pravoj liniji pozitivnog nagiba – *linearna veza*.

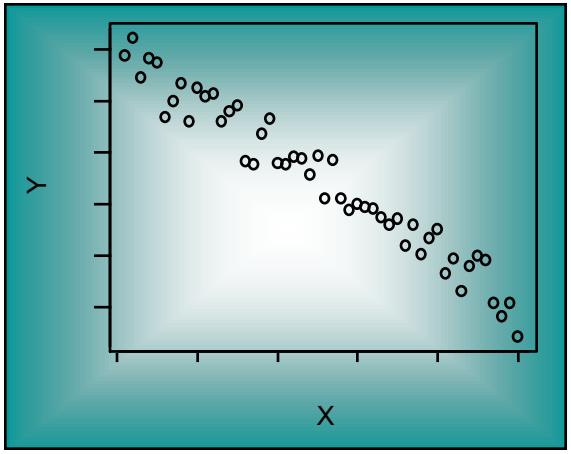
Primjeri dijagrama

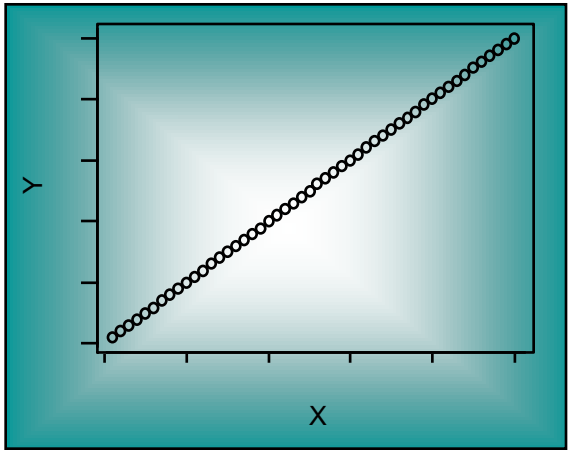












12-3 Ocjena: Metod najmanjih kvadrata

Ocijenjena linija regresije u uzorku:

$$Y = b_0 + b_1 X + e$$

Y - **zavisna promjenljiva**, koja se objašnjava ili predviđa; *X* - **nezavisna promjenljiva**

b_0 – ocjena parametra β_0 ;

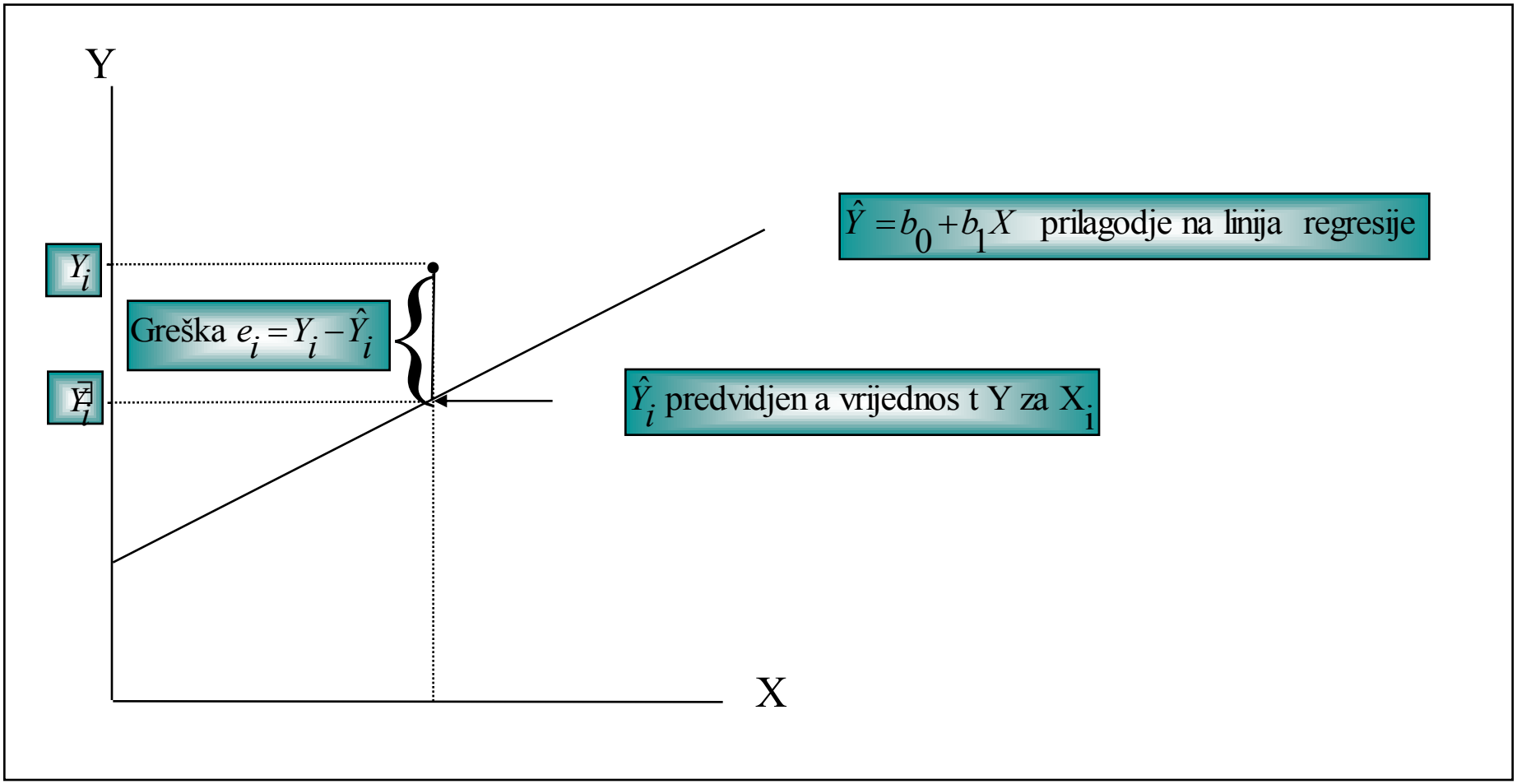
b_1 – ocjena parametra β_1;

i ***e*** je opažena greška - reziduali prilagođavanja ocijenjene linije regresije $b_0 + b_1 X$ nizu *n* podataka.

Ocijenjena linija regresije:

$$\hat{Y} = b_0 + b_1 X$$

Greške u regresiji



Metod najmanjih kvadrata

- Minimizirati sumu kvadrata odstupanja:

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Normalne jednacine:

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

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

Metod najmanjih kvadrata

Koeficijenti:

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

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

$$\hat{y} = b_o + b_1 x$$

Primjer 12-1

Posmatra se zavisnost iznosa troškova od pređenih milja. Ocijeniti liniju regresije, ako su, na osnovu 25 podataka date sledeće sume:

$$\sum x = 79448$$

$$\sum y = 106605$$

$$\sum x^2 = 293426944$$

$$\sum xy = 390185024$$

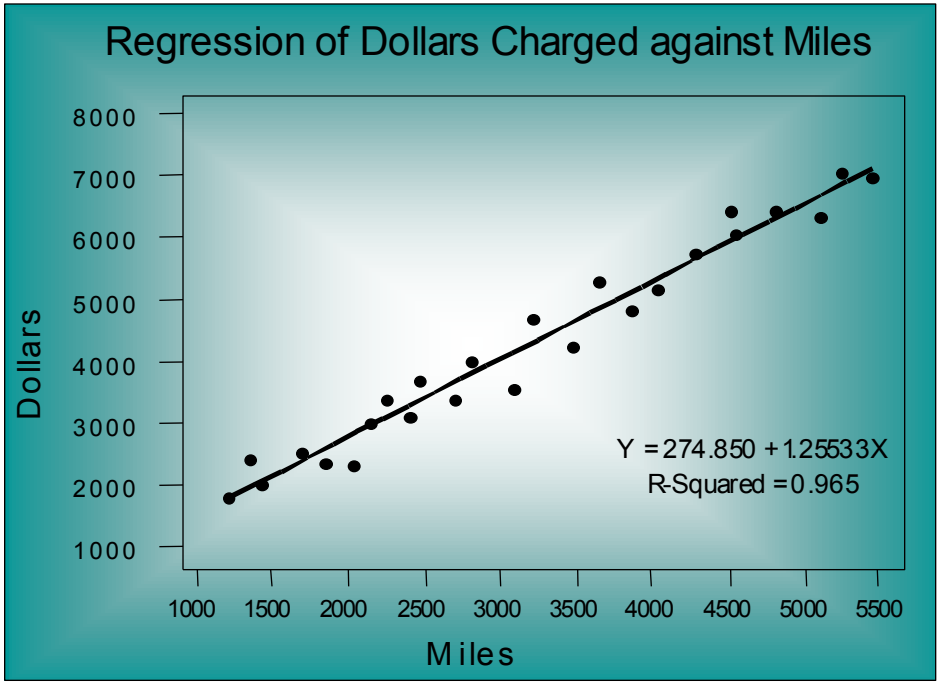
$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} =$$

$$= \frac{25 * 390185024 - 79448 * 106605}{25 * 293426944 - 79448 * 79448} = 1.255333776$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{106605}{25} - (1.255333776)\left(\frac{79448}{25}\right)$$

$$= 274.85$$

Primjer 12-1


Regression of Dollars Charged against Miles
Dollars
8000
7000
6000
5000
4000
3000
2000
1000
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
Miles
Y = 274.850 + 1.25533X
R-Squared = 0.965

Primjer 12-2

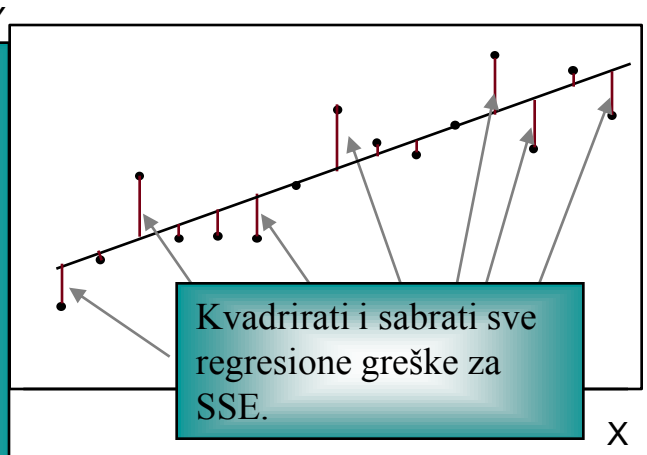
- Na osnovu podataka za 15 godina o per capita raspoloživom dohotku (x) i per capita ličnoj potrošnji (y) u SAD-u ($\sum x = 64022, \sum y = 57980,$, $\sum x^2 = 275132696$, $\sum y^2 = 225955018$, $\sum xy = 249318631$) ocijenjena je linija regresije $\hat{y} = -343{,}71033 + 0{,}986156x$. Zaokružiti tačnu konstataciju:
  - Ako se per capita lična potrošnja poveća za 1$, per capita dohodak će porasti za 0.986156$
  - Ako se per capita dohodak poveća za 1$, per capita lična potrošnja će porasti za 0.986156$
  - Ako se per capita dohodak poveća za 1$, per capita lična potrošnja će ostati nepromijenjena
  - Prosječna lična potrošnja u SAD-u je 343,71033$

12-4 Standardna greška regresije

Stepeni slobode u regresiji :

df = (n - 2) (n uk. podataka manje 2 ocijenjena parametra (b_0 i b_1))

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



Primjer 12-2

- Izračunati st. grešku regresije!

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

$$= \sqrt{\frac{225995018 + 343{,}71033 \cdot 57980 - 0{,}986156 \cdot 249318631}{15-2}} = 65{,}8$$

Standardna greška koef. pravca regresije

$$s_{b_1} = \frac{s}{\sqrt{\sum x^2 - n\bar{x}^2}}$$

Primjer 12-1:

$$s(b_1) = \frac{s}{\sqrt{SS_X}} = \frac{318.158}{\sqrt{40947557.84}} = 0.04972$$

Intervali povjerenja za regresione parametre

$(1-\alpha)\,100\%$ interval povjerenja za b_0:

$$b_0 \pm t_{\left(\frac{\alpha}{2},(n-2)\right)} s(b_0)$$

$(1-\alpha)\,100\%$ interval povjerenja za b_1:

$$b_1 \pm t_{\left(\frac{\alpha}{2},(n-2)\right)} s(b_1)$$

Primjer 12-1

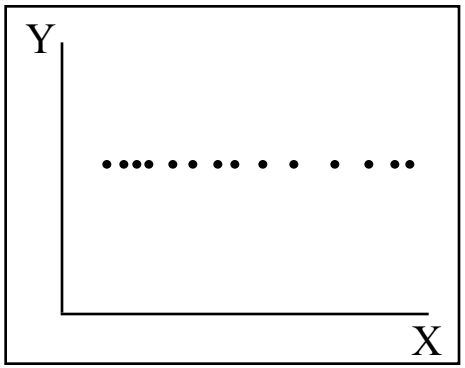
95% Intervali povjerenja:

$$b_0 \pm t_{(0.025,(25-2))} s(b_0)$$
$$= 274.85 \pm (2.069)\,(170.338)$$
$$= 274.85 \pm 352.43$$
$$= [-77.58, 627.28]$$

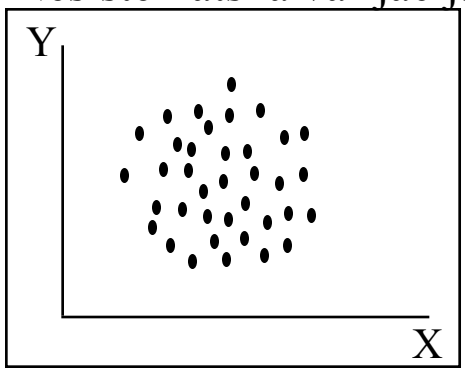
$$b_1 \pm t_{(0.025,(25-2))} s(b_1)$$
$$= 1.25533 \pm (2.069)\,(0.04972)$$
$$= 1.25533 \pm 0.10287$$
$$= [1.15246, 1.35820]$$

Test hipoteza regresione veze

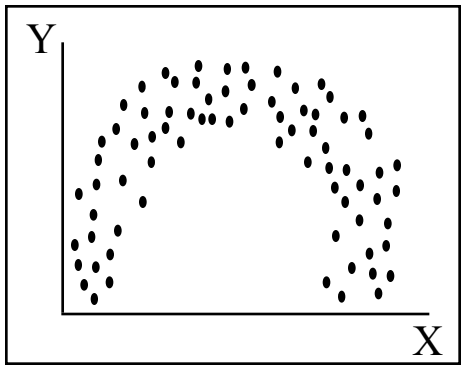
Konstantno Y



Nesistematska varijacija



Nelinearna veza



Test hipoteza: Za postojanje linearne veze između X i Y:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Test statistika:

$$t_{(n-2)} = \frac{b_1}{s(b_1)}$$

Test hipoteza za regresioni koeficijent

Primjer 12-1:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$t_{(n-2)} = \frac{b_1}{s(b_1)}$$

$$= \frac{1.25533}{0.04972} = 25.25$$

$$t_{(0.005,23)} = 2.807 < 25.25$$

H_0 se odbacuje pri 1% nivou

zakljucak - postoji veza izmedju troskova i predjenih milja.

$$s_{b_1} = \frac{s}{\sqrt{\sum x^2 - n\bar{x}^2}} = 0{,}04972$$

Test hipoteza za regresioni koeficijent

Primjer 12-2: *Testiranje_proporcionalne_regresione_veze*

$n = 60,\ b_1 = 1.24, s(b_1) = 0.21, \alpha = 0.1$

$H_0 : \beta_1 = 1$

$H_1 : \beta_1 \neq 1$

$$t_{(n-2)} = \frac{b_1 - 1}{s(b_1)}$$

$$= \frac{1.24 - 1}{0.21} = 1.14$$

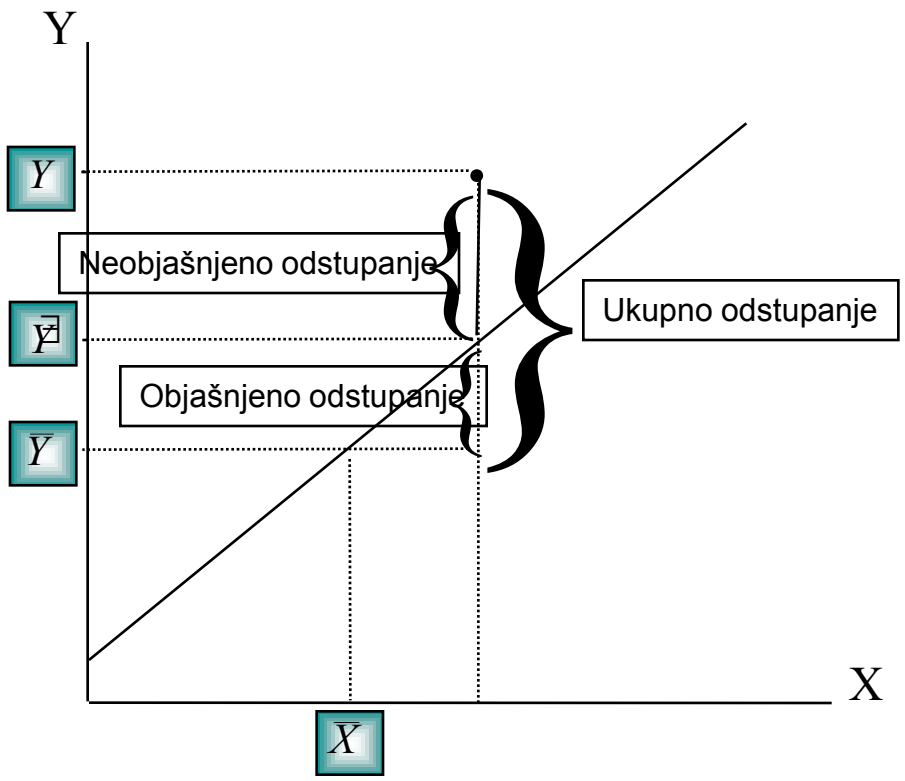
$$t_{(0.05,58)} = 1.671 > 1.14$$

H_0 se ne odbacuje pri 10% nivou.

Mozemo zakljuciti da je beta koef. jednak 1.

12-6 Koliko je dobar regresioni model?

Koeficijent determinacije, r^2, je deskriptivna mjera jačine regresione veze, koja mjeri koliko se dobro regresiona linija prilagođava podacima.



$$(y-\bar{y}) = (y-\hat{y}) + (\hat{y}-\bar{y})$$

Ukupno odstupanje = Neobjasnjeno odstupanje (greska) + Objasnjeno odstupanje (Regresija)

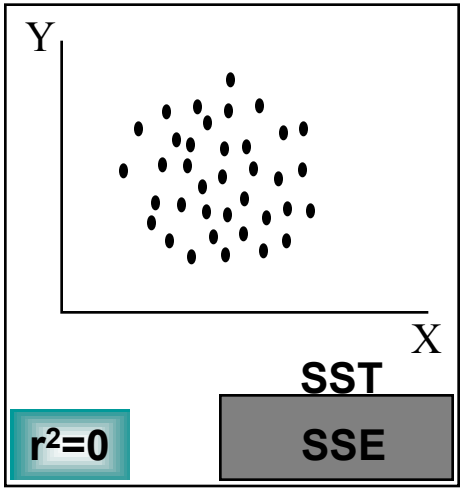
$$\sum(y-\bar{y})^2 = \sum(y-\hat{y})^2 + \sum(\hat{y}-\bar{y})^2$$

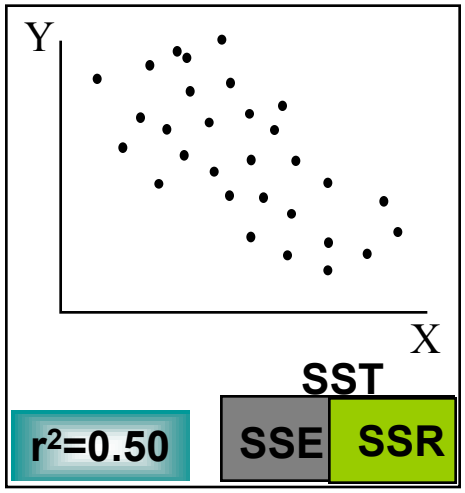
SST = SSE + SSR

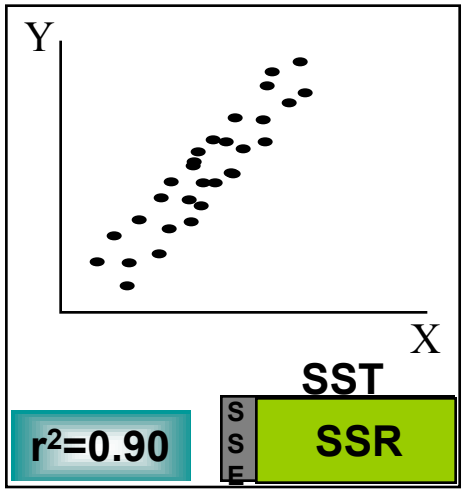
$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Procenat ukupne varijacije koja je objašnjena regresijom.

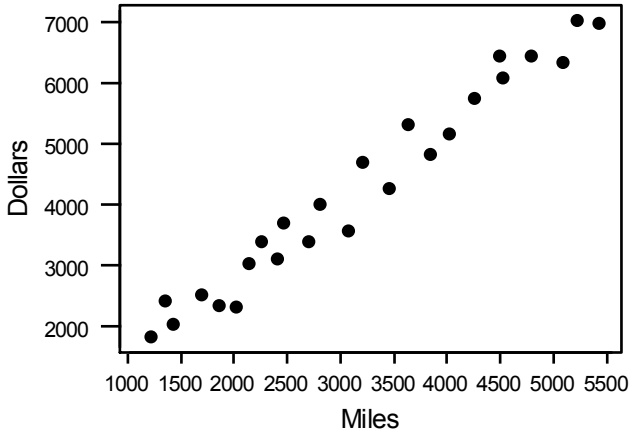
Koeficijent determinacije







$$r^2 = b_1{}^2 \frac{\sum x^2 - n\bar{x}^2}{\sum y^2 - n\bar{y}^2}$$



Primjer 12-2

- Izračunati procenat odstupanja koji je objašnjen modelom.
- r^2!

$$r^2 = b_1{}^2 \frac{\sum x^2 - n\bar{x}^2}{\sum y^2 - n\bar{y}^2} = 0{,}986156^2 \frac{275132696 - 15 \cdot (\frac{64022}{15})^2}{225995018 - 15(\frac{57980}{15})^2} = 0{,}9725 \frac{1878263{,}733}{1882991{,}33} =$$

$$= 0{,}97$$

12-5 Korelacija

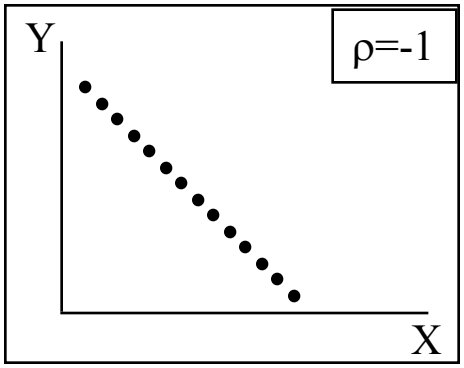
Korelacija između dvije sl. promjenljive, X i Y, je mjera *stepena linearne veze* između njih.

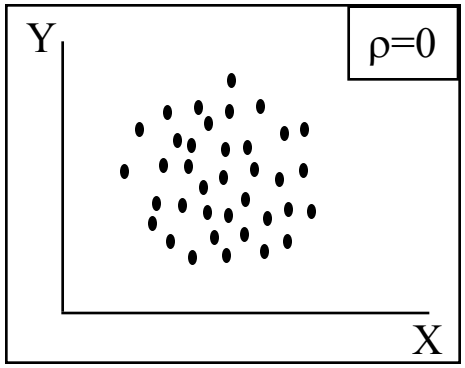
Populaciona korelacija, u oznaci ρ, može uzeti vrijednost između -1 i 1.

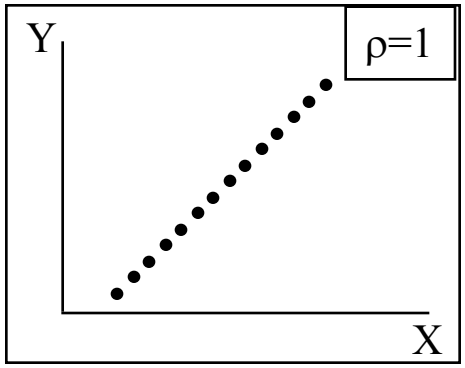
$\rho=-1$	označava perfektnu negativnu linearnu vezu
$-1<\rho<0$	označava negativnu linearnu vezu
$\rho=0$	označava odsustvo linearne veze
$0<\rho<1$	označava pozitivnu linearnu vezu
$\rho=1$	označava perfektnu pozitivnu linearnu vezu

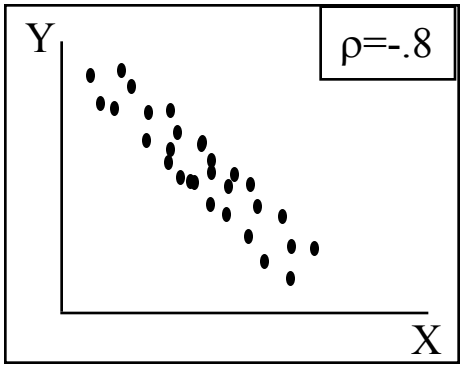
Apsolutna vrijednost ρ pokazuje jačinu ili stepen slaganja veze.

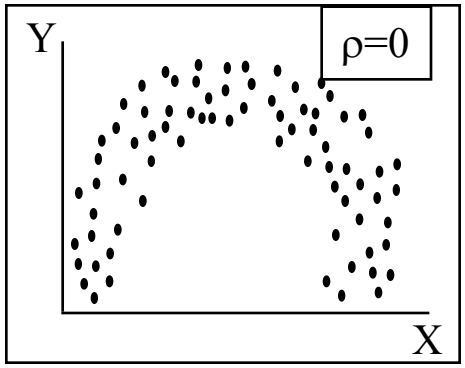
Ilustracija korelacije

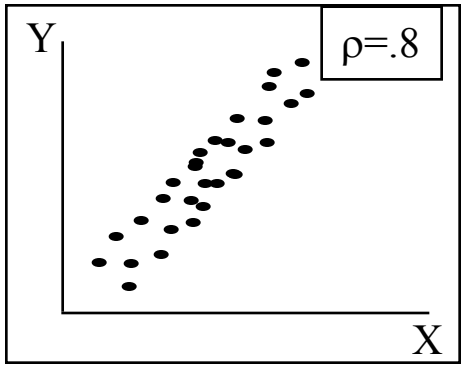












Kovarijansa i korelacija

Kovarijansa za X i Y:

$$Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]$$

gdje su μ_X i μ_Y populacioni prosjeci X i Y.

Populacioni koeficijent korelacije:

$$\rho=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

Koeficijent korelacije uzorka*:

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

*Napomena: Ako je $r<0$, $b_1<0$ Ako $r=0$, $b_1=0$ Ako je $r>0$, $b_1>0$

Test hipoteza za koeficijent korelacije

- Testirati koeficijent korelacije od 0,9824, za seriju od 25 podataka, uz nivo značajnosti 99%. (t=2,807)

H_0: $\rho=0$ (Nema linearne veze)
H_1: $\rho\neq 0$ (Postoji linearna veza)

Test statistika:

$$t_{(n-2)} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

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

Zadatak:

$$t_{(n-2)} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$= \frac{0.9824}{\sqrt{\frac{1-0.9651}{25-2}}}$$

$$= \frac{0.9824}{0.0389} = 25.25$$

$t_{0.005} = 2.807 < 25.25$

H_0 se odbacuje pri 1% nivou

1.-4. zadatak

1. Za 9 radnika jedne fabrike posmatra se zavisnost procenta škarta u njihovoj proizvodnji od dužine radnog staža (u mjesecima), i dobijeni su sledeći podaci: $\sum x = 57$, $\sum y = 50$, $\sum x^2 = 409$, $\sum y^2 = 304$ i $\sum xy = 284$. Na osnovu linije regresije, koeficijent pravca iznosi:

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = -0.68$$

2. Na osnovu prethodnog zadatka, može se zaključiti:

- Sa većim radnim stažom, veći je procenat škarta
- **Sa većim radnim stažom, manji je procenat škarta**
- Sa manjim radnim stažom, manji je procenat škarta
- Ništa od navedenog

3. Na osnovu podataka iz 1. zadatka, ako je slobodan član regresione jednačine 9,866, procenat odstupanja koji je objašnjen modelom je:

$$r^2 = b_1{}^2 \frac{\sum x^2 - n\bar{x}^{-2}}{\sum y^2 - n\bar{y}^{-2}} = 0{,}8465 = 84{,}65\%$$

4. Na osnovu podataka iz 1. zadatka, koeficijent proste linearne korelacije, uz rizik greške od 5%:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \pm\sqrt{r^2} = -\sqrt{0.8465} = -0.92$$

$$t_{0.025;7} = 2{,}365 \qquad t = \frac{r}{s_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = -6.216$$

Je statistički značajan!

Primjer 12-2

- Na osnovu podataka za 15 godina o per capita raspoloživom dohotku (x) i per capita ličnoj potrošnji (y) u SAD-u ($\sum x = 64022, \sum y = 57980,$, $\sum x^2 = 275132696$, $\sum y^2 = 225955018$, $\sum xy = 249318631$) ocijenjena je linija regresije $\hat{y} = -343{,}71033 + 0{,}986156x$. Zaokružiti tačnu konstataciju:
  - Ako se per capita lična potrošnja poveća za 1$, per capita dohodak će porasti za 0.986156$
  - Ako se per capita dohodak poveća za 1$, per capita lična potrošnja će porasti za 0.986156$
  - Ako se per capita dohodak poveća za 1$, per capita lična potrošnja će ostati nepromijenjena
  - Prosječna lična potrošnja u SAD-u je 343,71033$

Primjer 12-2

Odgovor: Ako se per capita dohodak poveća za 1$, per capita lična potrošnja će porasti za 0.986156$

- Izračunati st. grešku regresije!

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

$$= \sqrt{\frac{225995018 + 343{,}71033 \cdot 57980 - 0{,}986156 \cdot 249318631}{15-2}} = 65{,}8$$

Primjer 12-2

- Izračunati procenat odstupanja koji nije objašnjen modelom.
- 1-r^2!

$$r^2 = b_1{}^2 \frac{\sum x^2 - n\bar{x}^2}{\sum y^2 - n\bar{y}^2} = 0{,}986156^2 \frac{275132696 - 15 \cdot (\frac{64022}{15})^2}{225995018 - 15(\frac{57980}{15})^2} = 0{,}9725 \frac{1878263{,}733}{1882991{,}33} =$$

$$= 0{,}97$$

- 100-97=3%

Primjer 12-2

- Uz 95% nivo pouzdanosti, testirati da li je koeficijent proste linearne korelacije značajan.

$$r = +\sqrt{r^2} = 0{,}9849$$

H_0: $\rho=0$ (Nema linearne veze)
H_1: $\rho\neq 0$ (Postoji linearna veza)

Test statistika: $$t_{(n-2)} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Primjer 12-1:

$$t_{(n-2)} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.9849}{\sqrt{\frac{1-0.97}{15-2}}} = \frac{0.9849}{0.048} = 20.52$$

$t_{0.025} = 2.16 < 20.52$

H_0 se odbacuje pri 5% nivou

12-3. zadatak

- Za 9 parova vrijednosti broja stanovnika u hiljadama (promjenljiva X) i broja ekspozitura poslovnih banaka (promjenljiva Y) izračunate su sledeće vrijednosti: $\sum x = 1380$, $\sum y = 405$, $\sum x^2 = 225250$, $\sum y^2 = 19331$, $\sum xy = 65960$. Pri povećanju broja stanovnika za hiljadu broj ekspozitura se linearno povećava u prosjeku za:

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = 0{,}283$$

12-3. zadatak

2. Parametar b_0 za regresiju iz prethodnog zadatka iznosi:

$$b_0 = \bar{y} - b_1 \bar{x} = 1{,}61$$

12-3. zadatak

3. Na osnovu podataka iz zadatka 12-3. standardna greška regresije iznosi:

$$s = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}} = 1{,}32$$

12-3. zadatak

4. Na osnovu date regresije, uz vjerovatnoću od 0,95, prosječan broj mogućih ekspozitura poslovnih banaka u gradu koji ima 300 hiljada stanovnika je:

$$t_{0.025;7} = 2{,}365 \quad \hat{y} = b_o + b_1 x = 1{,}61 + 0{,}283 * 300 = 86{,}51$$

$$s_{\hat{y}_p} = \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x^2 - n\bar{x}^2}}$$

$$\hat{y}_p - t_{\frac{\alpha}{2},n-2} s_{\hat{y}_p} \le E(Y_p) \le \hat{y}_p + t_{\frac{\alpha}{2},n-2} s_{\hat{y}_p} = (83{,}46;89{,}56)$$

12-4. zadatak

1. Na bazi istraživanja o godišnjem prihodu u hiljadama eura (x) i izdacima za otplatu stambenog kredita u hiljadama eura (y) 8 klijenata jedne banke dobijeni su sledeći rezultati:

$$\sum y = 28{,}5 \quad \sum x = 220 \quad \sum x^2 = 7100 \quad \sum y^2 = 114{,}75 \quad \sum xy = 897{,}5$$

Ako nema prihoda, izdaci za otplatu stambenog kredita iznose:

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = 0{,}11$$

$$b_0 = \bar{y} - b_1\bar{x} = 0{,}5375 * 1000 = 537{,}5 eura$$

12-4. zadatak

3. Na bazi podataka iz zadatka 12-4., standardna greška regresije je oko:

$$s = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}} = 0{,}34$$

12-4. zadatak

3. Na bazi podataka iz zadatka 12-4., uz rizik greške od 5%, možemo zaključiti da je parametar b_1 :

$$s_{b_1} = \frac{s}{\sqrt{\sum x^2 - n\bar{x}^2}} = 0{,}01 \qquad t_{0.025;6} = 2{,}447$$

$$t = \frac{b_1}{s_{b_1}} = 11$$

Statistički značajan!