Regresiona i korelaciona analiza

Vježbe
Darko Milunović, MA

Šta ćemo znati nakon ovog poglavlja?

- Shvatiti razliku između funkcionalne i stohastičke veze
- Objasniti ulogu regresionog modela
- Standardna greška regresije (uloga i značaj)
- Koristiti koeficijent korelacije

Regresiona i korelaciona analiza

PROSTA

Prilikom istraživanja međusobnih veza dvije promjenljive

VIŠESTRUKA

Prilikom istraživanja međusobnih veza više promjenljivih

PROSTA LINEARNA REGRESIJA

Jednačina koja pokazuje odnos između dvije varijable data je na sljedeći način:

$$Y = \beta_0 + \beta_1 X$$

gdje Y predstavlja zavisnu varijablu,

X predstavlja nezavisnu varijablu

β_0 odsječak na Y-osi

β_1 nagib

PROSTA LINEARNA REGRESIJA

Ocjene za koeficijente β_0 i β_1 se određuju pomoću metode najmanjih kvadrata

Linija regresije u uzorku:

$$\hat{y} = b_0 + b_1 x$$

Gdje se ocjene dobiju pomoću formula:

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

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

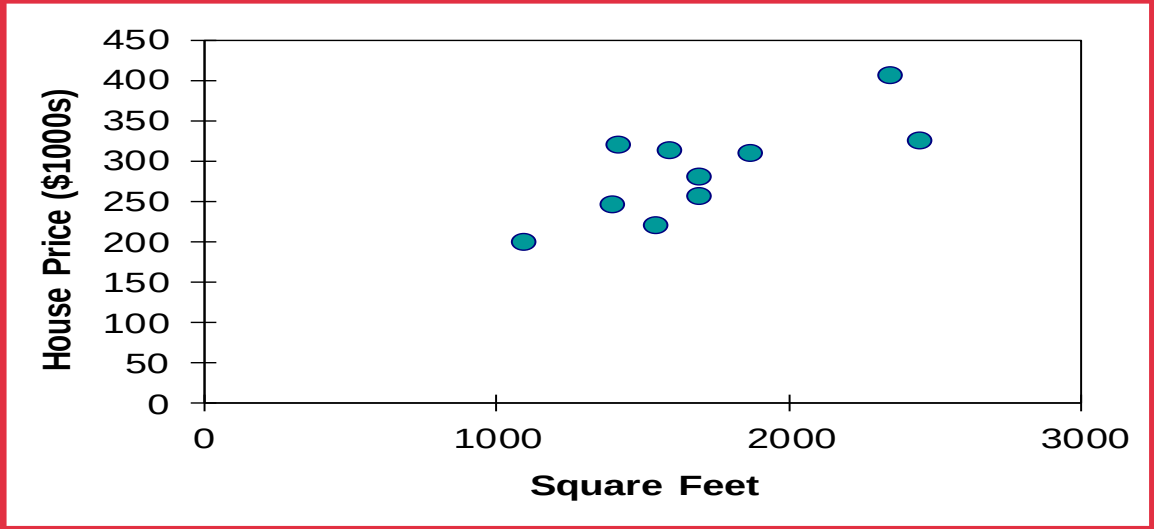
Zadatak 1.

Prodajni agent želi ispitati odnos između prodajne cijene kuće i veličine (mjerene u kvadratnim metarima). Slučajno je izabrano 10 kuća.

Cijena u 1000 KM (Y)	Površina (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Zadatak 1.

Nakon definisanja varijabli, poželjno je prikazati odnos tih veličina dijagramu rasipanja (scatterplot)



Cijena = 98.24833 + 0.10977 (površina)

Zadatak 2.

Dati su podaci o ostvarenom prometu i troškovima reklame jednog trgovinskog preduzeća:

Ostvareni promet (000 KM.)	Troškovi reklame (000 KM.)
60	5
120	10
140	16
180	21
200	25
250	30
300	38
1250	145

1. Izračunati prosječan zakonomjeran kvantitativni odnos troškova reklame i ostvarenog prometa pomoću jednačine regresije. Testirati značajnost nepoznatih parametara;

2. Ocjeniti uz 95% pouzdanosti prosječni obim ostvarenog prometa za troškove reklame od 35.000 KM;

3. Odrediti jednačinu inverznog regresionog modela;

4. Odrediti stepen i smjer međusobne povezanosti varijacija troškova reklame i ostvarenog prometa kao i mjeru u kojoj su varijacije u obimu ostvarenog određene varijacijama u troškovima reklame posmatranog trgovinskog preduzeća. Testirati dobijene parametre uz 95% pouzdanosti.

n – veličina uzorka, broj parova podataka

Ostvareni promet (000 KM.)- y	Troškovi reklame (000 KM.) - x		
60	5	300	25
120	10	1200	100
140	16	2240	256
180	21	3780	441
200	25	5000	625
250	30	7500	900
300	38	11400	1444
1250	145	31420	3791

$$\hat{y}_i = 33{,}173 + 7{,}019 \cdot x_i$$

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

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

$$b_1 = \frac{7 \cdot 31420 - 145 \cdot 1250}{7 \cdot 3791 - 145^2} = 7{,}019$$

$$b_0 = 178{,}57 - 7{,}019 \cdot 20{,}71 = 33{,}173$$

$$\bar{y} = 178{,}57$$

$$\bar{x} = 20{,}71$$

Testiranje značajnosti dobijenih parametara b_0 i b_1 uz 95% pouzdanosti predstavlja testiranje značajnosti postojanja ustanovljene linearne regresione veze:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

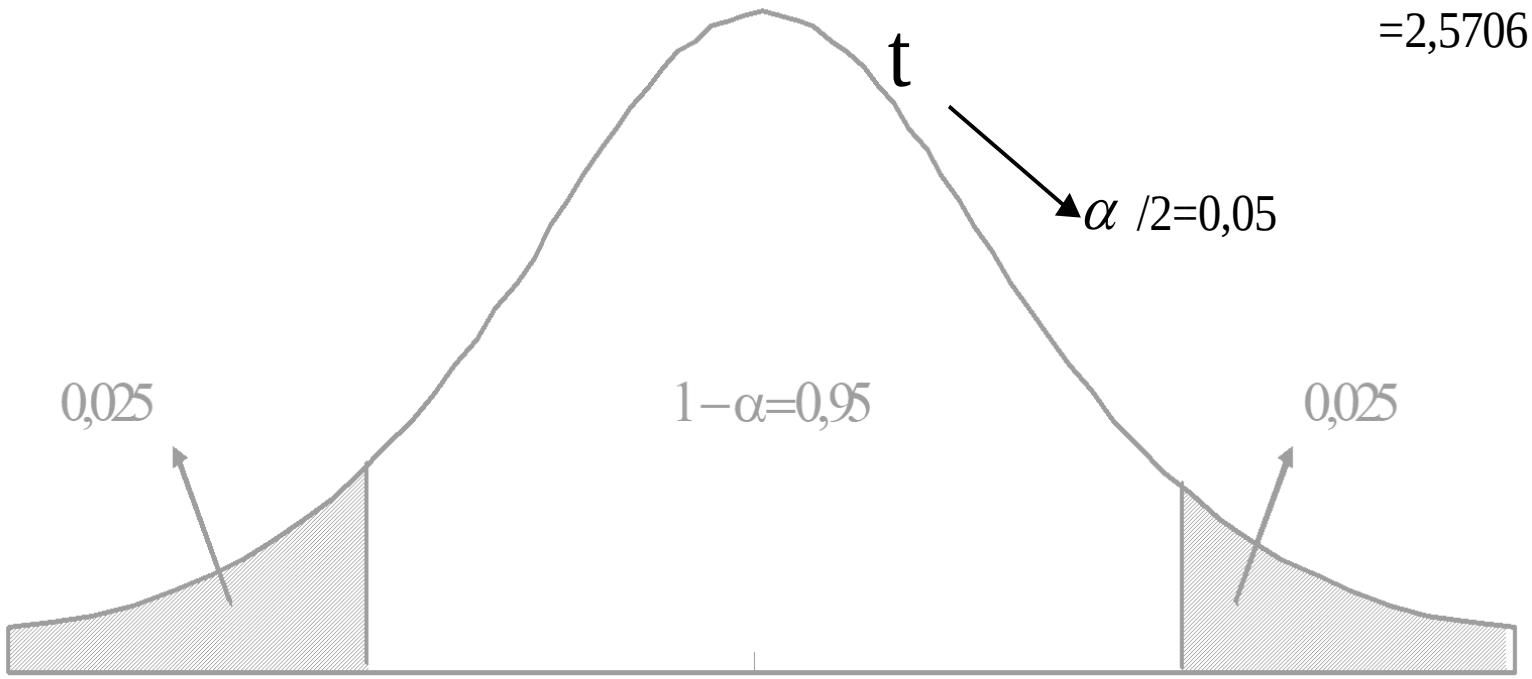
$i = 0,1$

n-2=7-2=5

=2,5706

t

α /2=0,05



$$t_0 = \frac{b_0}{S_{b_0}} = \frac{b_0}{s \cdot \sqrt{\frac{\sum x^2}{n \cdot \left(\sum x^2 - n\bar{x}^2\right)}}} = \frac{33{,}173}{9{,}894 \cdot \sqrt{\frac{3791}{7 \cdot \left(3791 - 7 \cdot 20{,}71^2\right)}}} = 4{,}046$$

$$t_1 = \frac{b_1}{S_{b_1}} = \frac{b_1}{\frac{s}{\sqrt{\sum x^2 - n\bar{x}^2}}} = \frac{7{,}019}{\frac{9{,}894}{\sqrt{3791 - 7 \cdot 20{,}71^2}}} = 19{,}92$$

S_{b_0} - standardna greška ocjene odsječka b_0

S_{b_1} - standardna greška ocjene odsječka b_1

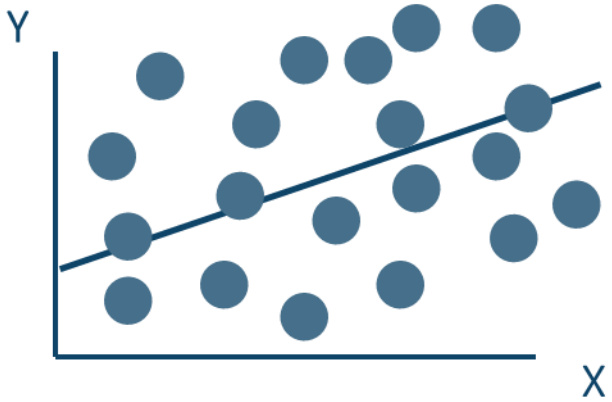
$$t_0, t_1 > t_{n-2,\alpha/2}(2{,}5706) \Rightarrow$$ **Odbacujemo H_0**

Standardna greška regresije

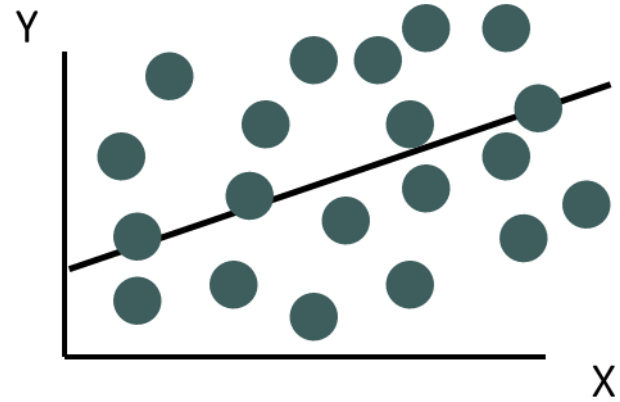
Apsolutna mjera varijacija empirijskih podataka od regresione linije uzorka predstavlja ocjenu varijanse greške (korijen te ocjene) i dobije se kao odnos sume kvadrata odstupanja i broja stepeni slobode:

$$s = \hat{\sigma} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{489{,}423}{7-2}} = 9{,}894$$

$$s = \hat{\sigma} = \sqrt{\frac{\sum y^2 - b_o \sum y - b_1 \sum xy}{n-2}} = 9{,}894$$



mala vrijednost SE



velika vrijednost SE

Ocjena prosječne vrijednosti obima ostvarenog prometa za troškove rekalame od 35.000 KM (odnosi se na ocjenu u osnovnom skupu):

$$\hat{y}_p - t_{n-2,\alpha/2} \cdot s_{\hat{y}_p} \leq E(Y_p) \leq \hat{y}_p + t_{n-2,\alpha/2} \cdot s_{\hat{y}_p}$$

$$\hat{y}_p = 33{,}173 + 7{,}019 \cdot 35 = 278{,}838$$

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

$$9{,}894 \cdot \sqrt{\frac{1}{7} + \frac{(35 - 20{,}71)^2}{3791 - 7 \cdot 20{,}71^2}} = 6{,}271$$

$$278{,}838 - 2{,}5706 \cdot 6{,}271 \leq E(Y_p) \leq 278{,}838 + 2{,}5706 \cdot 6{,}271$$

$$262{,}72 \leq E(Y_p) \leq 294{,}96$$

Inverzni model je...

$$\hat{y}_i = \boxed{\bar{y} - \bar{x} \cdot \frac{C_{xy}}{\sigma_x^2}} \quad \boxed{\frac{C_{xy}}{\sigma_x^2}} \cdot x_i$$

b_0 b_1

$$\sigma_x^2 = 112,48982$$

$$\boxed{\hat{y}_i = 33,178 + 7,019 \cdot x_i}$$

$$\hat{x}_i = \boxed{\bar{x} - \bar{y} \cdot \frac{C_{xy}}{\sigma_y^2}} \quad \boxed{\frac{C_{xy}}{\sigma_y^2}} \cdot y_i$$

b_0' b_1'

$$\sigma_y^2 = 5612,248$$

$$\boxed{\hat{x}_i = -4,409 + 0,141 \cdot y_i}$$

$$C_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum xy}{n} - \bar{x}\,\bar{y} = 789,5922$$

Koeficijent korelacije predstavlja stepen i smjer međusobne povezanosti posmatranih pojava ostvarenog prometa i troškova reklame (Pearson-ov koeficijent proste linearne korelacije):

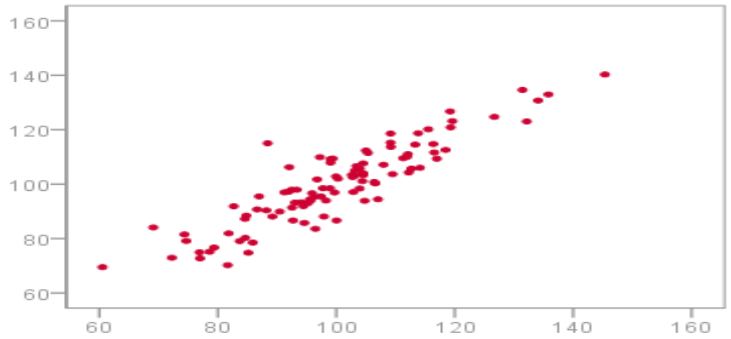
$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n\sum y^2 - (\sum y)^2}} =$$

$$\frac{7 \cdot 31420 - 145 \cdot 1250}{\sqrt{7 \cdot 3791 - 145^2} \cdot \sqrt{7 \cdot 262500 - 1250^2}} \Rightarrow$$

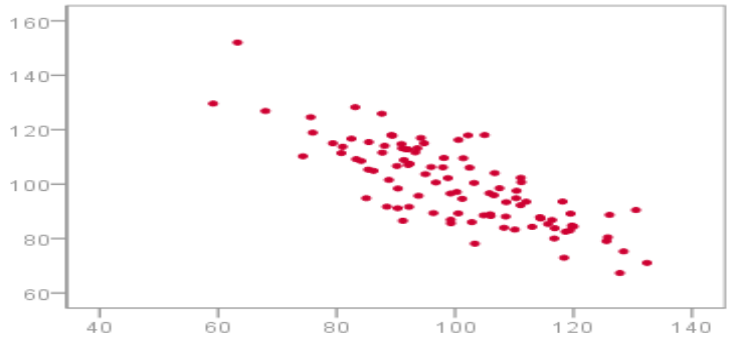
$$r = \frac{38690}{\sqrt{5512} \cdot \sqrt{275000}} = 0{,}99375$$

$$r = \pm\sqrt{b_1 \cdot b_1'} = \sqrt{7{,}019 \cdot 0{,}141} = 0{,}994$$

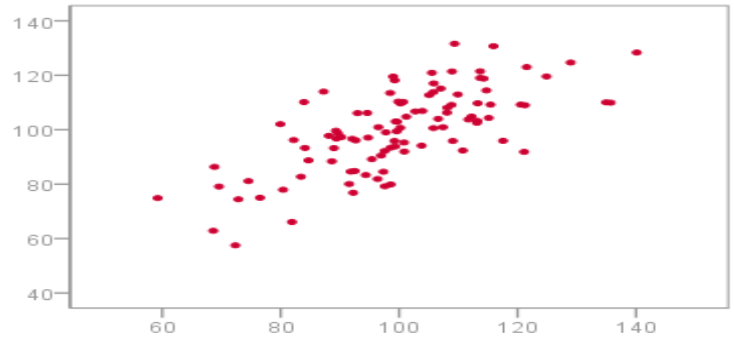
PEARSON CORRELATION (r) VISUALIZED AS SCATTERPLOT



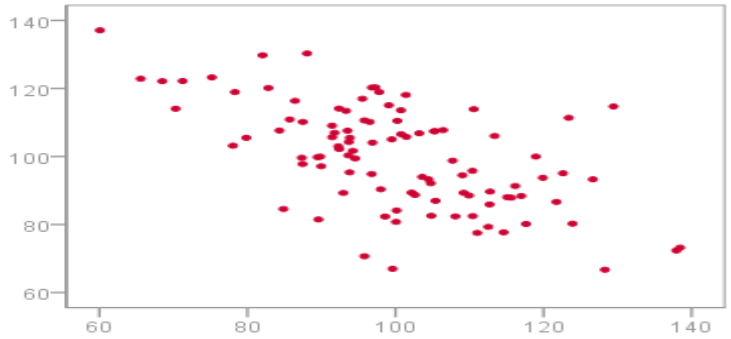
r = 0.9



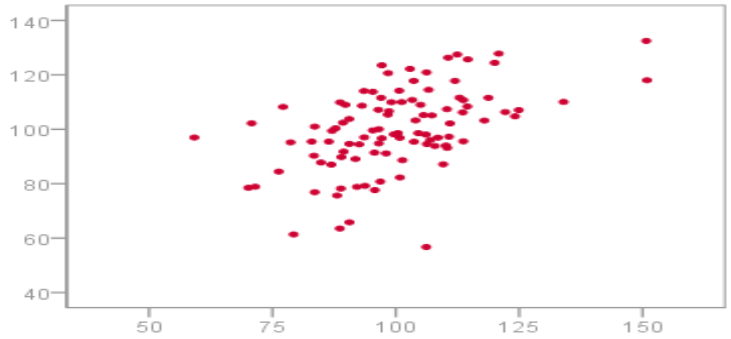
r = -0.8



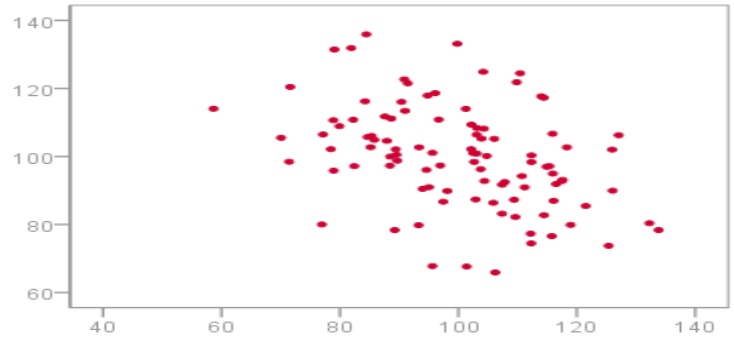
r = 0.7



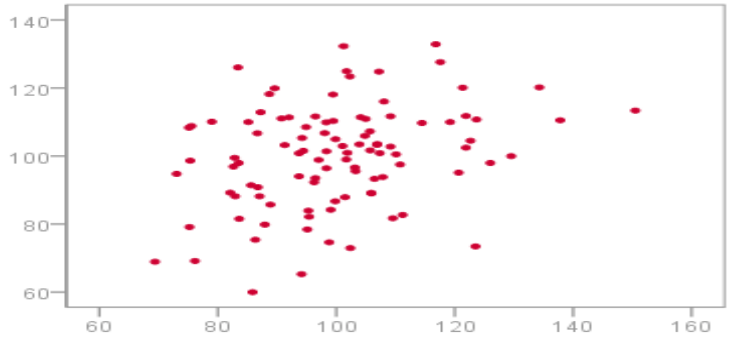
r = -0.6



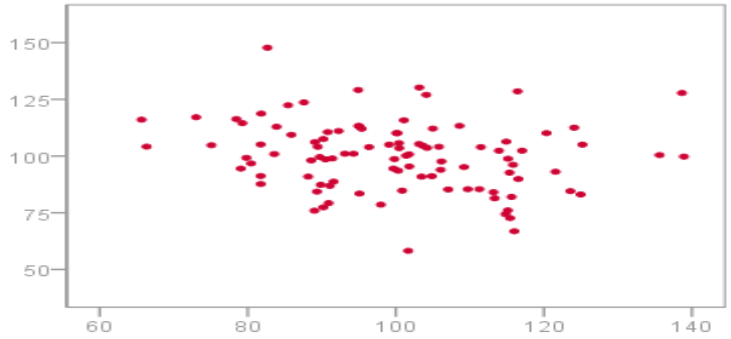
r = 0.5



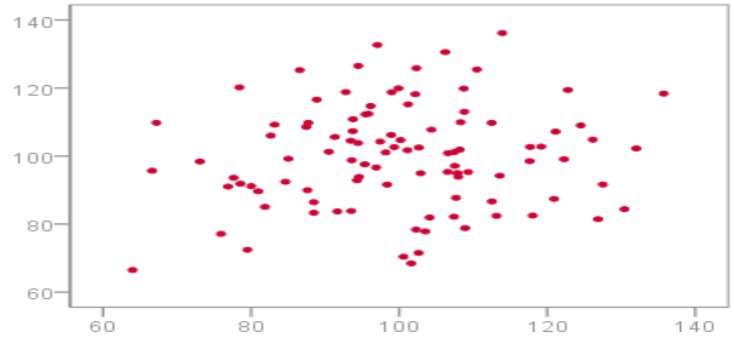
r = -0.4



r = 0.3



r = -0.2



r = 0.1

Koeficijent determinacije pokazuje udio varijabiliteta koji je objašnjen modelom...

$$0 < r^2 < 1$$

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

$$7{,}019 \cdot \frac{3791 - 7 \cdot 20{,}714^2}{262500 - 7 \cdot 178{,}571^2} = 0{,}9875$$

ZAKLJUČAK 98,75% varijabiliteta obima ostvarenog prometa je objašnjeno troškovima reklame dok je ostatak varijabiliteta nastao pod uticajem nekontrolisanih faktora.

Testiranje koeficijent korelacije – r

predstavlja testiranje značajnosti ocjene koeficijenta proste linearne korelacije između posmatrane 2 promjenjive:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

$$t_{0,05;5} = 2,5706$$

$$t = \frac{r}{s_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,99375}{\sqrt{\frac{1-0,9875}{7-2}}} = 19,875$$

Odbacujemo H_0 i zaključujemo uz 5% rizika da u osnovnom skupu postoji značajna linearna regresiona veza, odnosno dobijeni koeficijent korelacije koji pokazuje direktnu međuzavisnost je statistički značajan.


regression
variables
analysis
models
normal
equations
Statistics
Galton
set
varepsilon
linear
Jump
Regression
Statistical
function
used
dependent
unknown
different
sometimes
example
values
assumptions
beta
data
parameters
method
population
Gauss
may
variable
statistical
squares
independent
extrapolation
model
least
also
edit
Linear
methods
hat
number
packages
HVALA NA PAŽNJI!