

# POGLAVLJE 13



## **PROSTA LINEARNA REGRESIJA**

# 13.1 PROST LINEARNI REGRESIONI MODEL

---

- Prosta regresija
- Linearna regresija

# Prosta regresija

---

## Definicija

Regresioni model je matematički model koji opisuje vezu između dvije ili više promjenljivih.

**Prost regresioni** model obuhvata samo dvije promjenljive: jednu objašnjavajuću i jednu zavisnu. Zavisna promjenljiva je promjenljiva čije varijacije treba da objasnimo na osnovu kretanja objašnjavajuće promjenljive.

# Linearna regresija

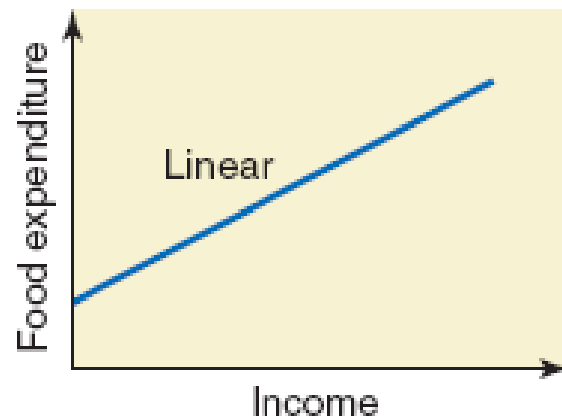
---

## Definicija

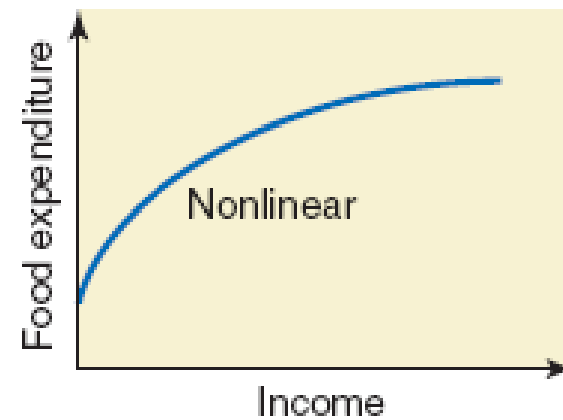
Regresioni model kojim se opisuje linearna međuzavsinost između dvije promjenljive naziva se **prost linearni regresioni** model.

Slika 13.1 Veza između izdataka za hranu i dohotka. (a) Linearna veza. (b) Nelinearna veza.

---



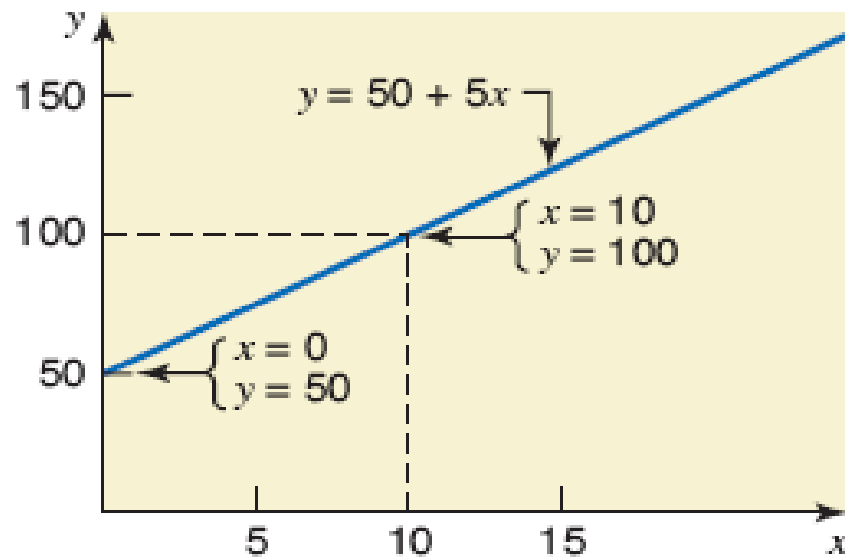
(a)



(b)

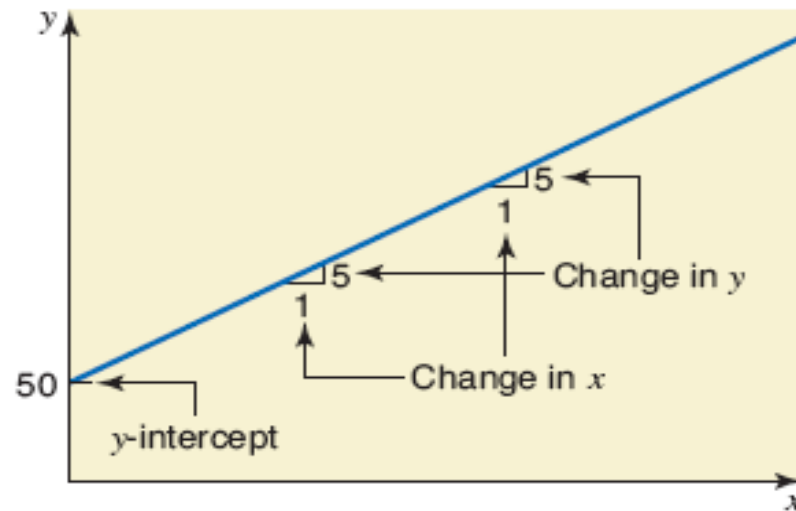
## Slika 13.2 Grafički prikaz jednačine prave.

---



## Slika 13.3 Odsječak i nagib prave linije.

---



## 13.2 PROSTA LINEARNA REGRESIONA ANALIZA

---

- Dijagram raspršenosti (rasturanja)
- Metod najmanjih kvadrata
- Interpretacija ocijenjenih vrijednosti  $a$  i  $b$
- Pretpostavke prostog linearnog regresionog modela



# PROSTA LINEARNA REGRESIONA ANALIZA

---

Constant term or y-intercept

Slope

$$y = A + Bx$$

Dependent variable

Independent variable

The diagram illustrates the components of the linear regression equation  $y = A + Bx$ . The equation is centered on the page. Above the equation, the text 'Constant term or y-intercept' is connected to the letter 'A' by a horizontal line that then turns down into an arrow pointing to 'A'. Similarly, the text 'Slope' is connected to the 'Bx' term by a horizontal line that then turns down into an arrow pointing to 'B'. Below the equation, the text 'Dependent variable' is connected to the letter 'y' by a vertical arrow pointing up to 'y'. The text 'Independent variable' is connected to the letter 'x' by a vertical arrow pointing up to 'x'.

# PROSTA LINEARNA REGRESIONA ANALIZA

---

## Definicija

U **regresionom modelu**  $y = A + Bx + \varepsilon$ ,  $A$  je odsječak ili konstanta,  $B$  koeficijent nagiba, a  $\varepsilon$  slučajna greška. Zavisna i objašnjavajuća promjenljiva su  $y$  i  $x$ , respektivno.

# PROSTA LINEARNA REGRESIONA ANALIZA

---

## Definicija

U regresionom modelu uzorka  $\hat{y} = a + bx$ , koeficijenti  $a$  i  $b$  nazivaju se ocjene parametara A i B, respektivno.

## Tabela 13.1 Dohodak (u stotinama dolara) i izdaci za hranu sedam domaćinstava

---

<b>Income</b>	<b>Food Expenditure</b>
55	14
83	24
38	13
61	16
33	9
49	15
67	17

# Dijagram raspršenosti (rasturanja)

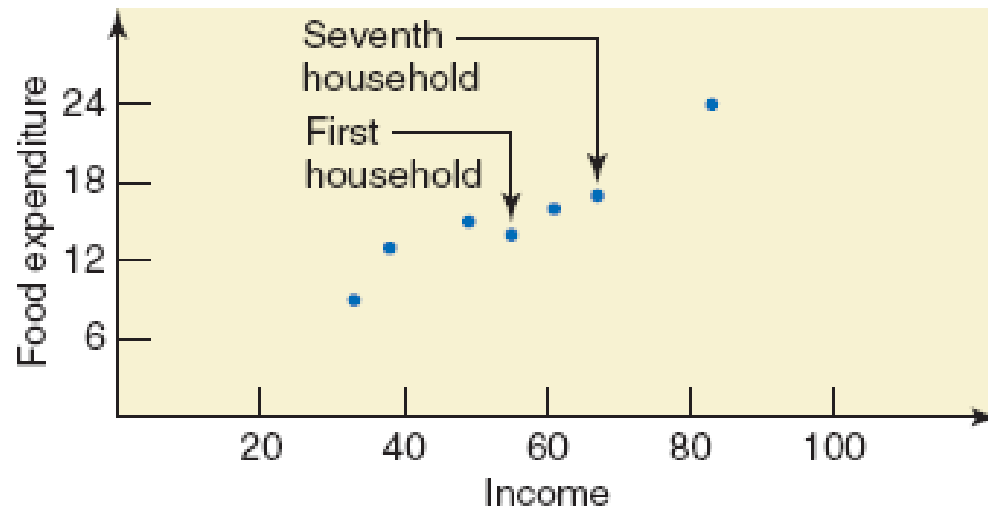
---

## Definicija

Grafički prikaz parova podataka  $X$  i  $Y$  u osnovnom skupu (ili uzorku) naziva se **dijagram raspršenosti (rasturanja)**.

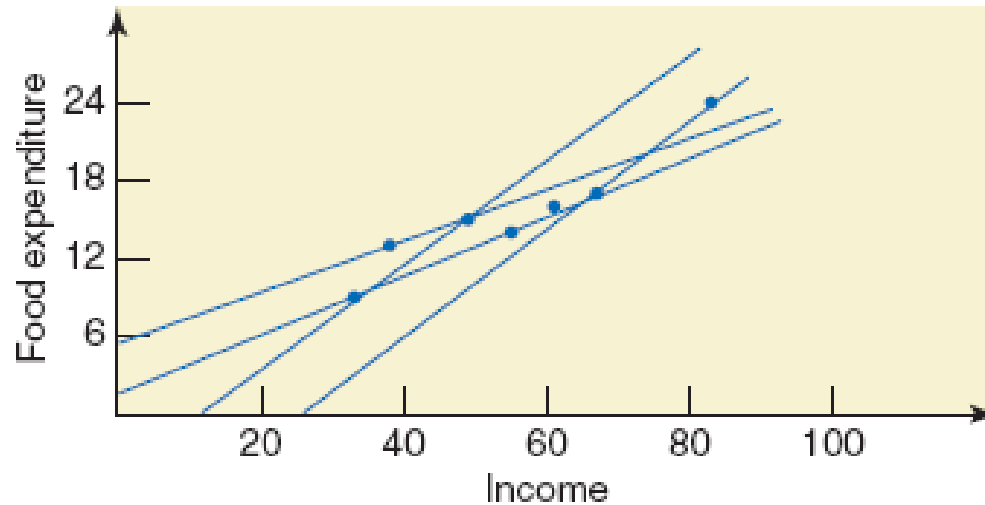
## Slika 13.4 Dijagram raspršenosti.

---



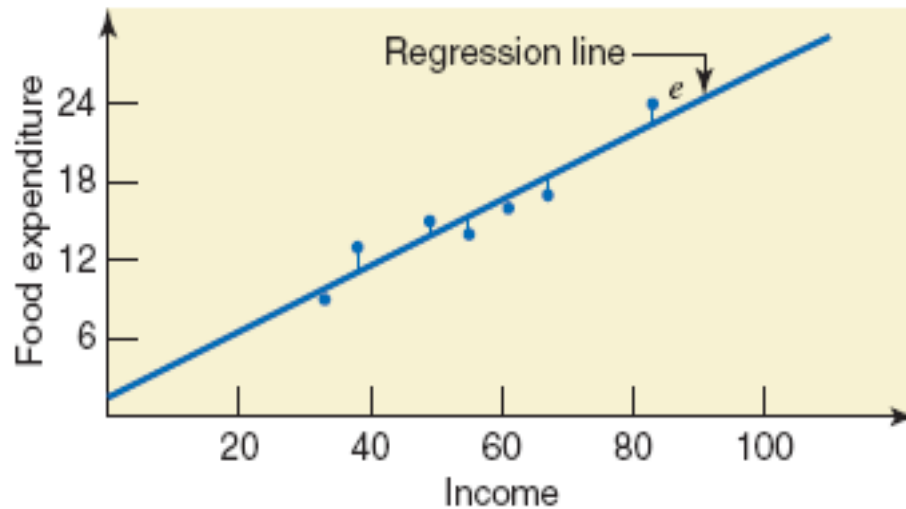
# Slika 13.5 Dijagram raspršenosti i regresione prave.

---



## Slika 13.6 Regresiona prava i reziduali.

---





## Suma kvadrata reziduala (SSE)

---

**Suma kvadrata reziduala**, u oznaci SSE, je

$$\mathbf{SSE} = \sum \mathbf{e}^2 = \sum (\mathbf{y} - \hat{\mathbf{y}})^2$$

Minimiziranjem sume kvadrata reziduala dobijaju se  $a$  i  $b$  kao **ocjene regresionih parametara**  $A$  i  $B$ , a regresiona prava koja se na osnovu tih ocjena dobija naziva se **regresiona prava uzorka**.

## Linija regresije uzorka

---

Koeficijenti regresione prave uzorka  $\hat{y} = a + bx$ , odnosno ocjene po metodu najmanjih kvadrata glase:

$$b = \frac{SS_{xy}}{SS_{xx}} \quad \text{i} \quad a = \bar{y} - b\bar{x}$$

## Linija regresije uzorka

---

gdje je

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} \quad \text{i} \quad SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

i gdje SS označava odgovarajuću "sumu kvadrata". Linija regresije uzorka  $\hat{y} = a + bx$  se takođe naziva regresija od  $y$  na  $x$ .

## Primjer 13-1

---

Na osnovu podataka slučajnog uzorka od sedam domaćinstava, prikazanih u Tabeli 13.1, ocijenite regresioni model primjenom metoda najmanjih kvadrata.

Objašnjavajuća promjenljiva je dohodak, a zavisna promjenljiva izdaci za hranu.

## Tabela 13.2

<b>Income</b>	<b>Food Expenditure</b>		
$x$	$y$	$xy$	$x^2$
55	14	770	3025
83	24	1992	6889
38	13	494	1444
61	16	976	3721
33	9	297	1089
49	15	735	2401
67	17	1139	4489
$\Sigma x = 386$	$\Sigma y = 108$	$\Sigma xy = 6403$	$\Sigma x^2 = 23,058$

## Primjer 13-1: Rješenje

---

$$\sum \mathbf{x} = 386 \qquad \sum \mathbf{y} = 108$$

$$\bar{\mathbf{x}} = \sum \mathbf{x} / n = 386 / 7 = 55.1429$$

$$\bar{\mathbf{y}} = \sum \mathbf{y} / n = 108 / 7 = 15.4286$$

## Primjer 13-1: Rješenje

---

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 6403 - \frac{(386)(108)}{7} = 447.5714$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 23,058 - \frac{(386)^2}{7} = 1772.8571$$

## Primjer 13-1: Rješenje

---

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{447.5714}{1772.8571} = .2525$$

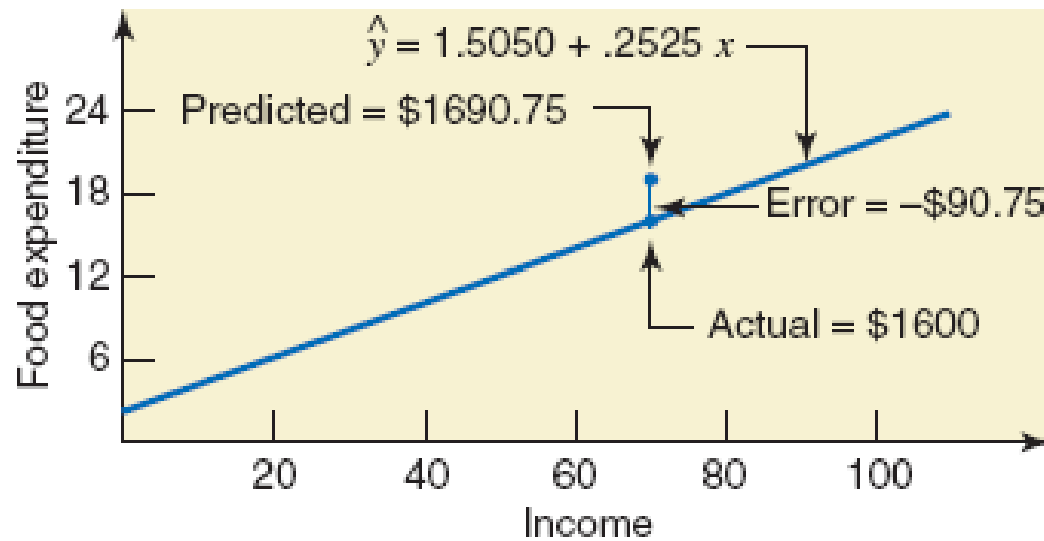
$$a = \bar{y} - b\bar{x} = 15.4286 - (.2525)(55.1429) = 1.5050$$

Dakle, ocijenjeni regresioni model glasi

$$\hat{y} = 1.5050 + 0.2525 x$$



## Slika 13.7 Rezidual.



## Interpretacija ocijenjenih vrijednosti $a$ i $b$

---

### Interpretacija ocijenjene vrijednosti $a$

- Ukoliko razmatramo domaćinstvo sa nulnim nivoom dohotka, na osnovu ocijenjene regresione prave iz primjera 13-1, dobijamo ocijenjenu vrijednost  $y$  za  $x=0$ :
  - $\hat{y} = 1.5050 + 0.2525(0) = \$1.5050$  stotina
- Dakle, možemo zaključiti da domaćinstvo koje ne ostvaruje nikakav dohodak troši \$150.50 mjesečno na hranu
- Regresiona prava važi samo za vrijednosti  $x$  između 33 i 83

## Interpretacija ocijenjenih vrijednosti $a$ i $b$

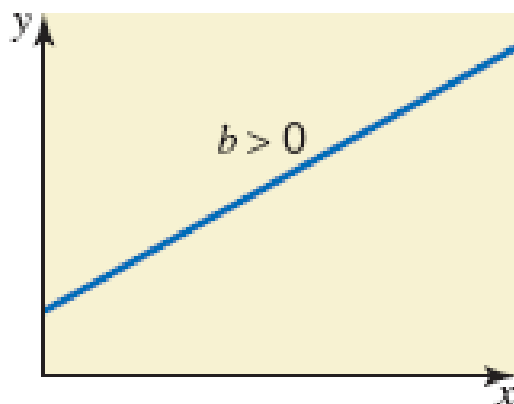
---

### Interpretacija ocijenjene vrijednosti $b$

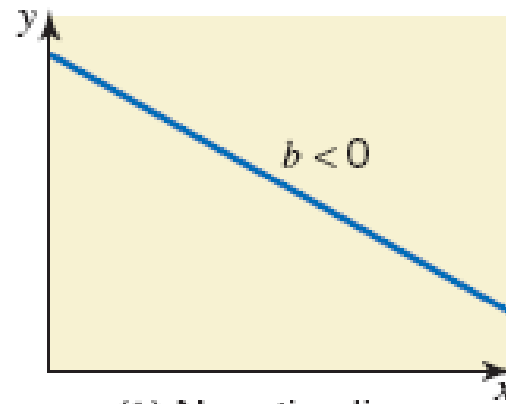
- Vrijednost  $b$  u regresionom modelu pokazuje koliko se u prosjeku promijeni  $y$  (zavisna promjenljiva) ako se  $x$  (objašnjujuća promjenljiva) poveća za jednu svoju jedinicu.
- Možemo zaključiti, u prosjeku, da će rast dohotka od \$100 (ili \$1) imati za rezultat porast izdataka za hranu za \$25.25 (ili \$0.2525).

# Slika 13.8 Pozitivna i negativna linearna veza između $x$ i $y$ .

---



(a) Positive linear relationship



(b) Negative linear relationship

# Pretpostavke prostog linearnog regresionog modela

---

## Pretpostavka 1:

Očekivana vrijednost slučajne greške  $\epsilon$  jednaka je nuli za svaku vrijednost promjenljive  $x$

## Pretpostavka 2:

Slučajne greške različitih opservacija su međusobno nezavisne

# Pretpostavke prostog linearnog regresionog modela

---

## Pretpostavka 3:

Za datu vrijednost promjenljive  $x$ ,  
raspodjela slučajnih grešaka je normalna

## Pretpostavka 4:

Slučajne greške za svaku vrijednost  
promjenljive  $x$  imaju konstantnu  
standardnu devijaciju  $\sigma_\epsilon$

## 13.3 STANDARDNA DEVIJACIJA SLUČAJNE GREŠKE

---

Broj stepeni slobode u prostom linearnom regresionom modelu

**Broj stepeni slobode** u prostom linearnom regresionom modelu je

$$df = n - 2$$

# STANDARDNA GREŠKA REGRESIJE

---

**Standardna greška regresije** izračunava se kao

$$s_e = \sqrt{\frac{SS_{yy} - bSS_{xy}}{n - 2}}$$

gdje

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$



## Primjer 13-2

---

Na osnovu podataka datih u Tabeli 13.1, izračunajte standardnu grešku regresije.

# Tabela 13.3

---

<b>Income</b>	<b>Food Expenditure</b>	
<i>x</i>	<i>y</i>	<i>y</i> <sup>2</sup>
55	14	196
83	24	576
38	13	169
61	16	256
33	9	81
49	15	225
67	17	289
$\Sigma x = 386$	$\Sigma y = 108$	$\Sigma y^2 = 1792$

## Primjer 13-2: Rješenje

---

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 1792 - \frac{(108)^2}{7} = 125.7143$$

$$s_e = \sqrt{\frac{SS_{yy} - bSS_{xy}}{n-2}} = \sqrt{\frac{125.7143 - .2525(447.5714)}{7-2}} = 1.5939$$

# KOEFICIJENT DETERMINACIJE

---

Koeficijent determinacije

**Koeficijent determinacije**, označen sa  $r^2$ , predstavlja učešće objašnjenog u ukupnom varijabilitetu, odnosno  $r^2$  je

$$r^2 = \frac{b \text{ SS}_{xy}}{\text{SS}_{yy}}$$

i

$$0 \leq r^2 \leq 1$$

## Primjer 13-3

---

Na osnovu podataka u Tabeli 13.1 o mjesečnom dohotku i izdacima za hranu, izračunati koeficijent determinacije.

## Primjer 13-3: Rješenje

---

- Na osnovu rezultata iz primjera 13-1 i 13-2,
- $b = 0.2525$ ,  $SS_{xx} = 447.5714$ ,  $SS_{yy} = 125.7143$

$$r^2 = \frac{b SS_{xy}}{SS_{yy}} = \frac{(.2525)(447.5714)}{125.7143} = .90$$

## 13.5 STATISTIČKO ZAKLJUČIVANJE O PARAMETRU $B$

---

Interval povjerenja za parametar  $B$

$(1 - \alpha)100\%$  interval povjerenja za parametar  $B$  glasi

gdje je

$$s_b = \frac{b \pm t s_b}{\sqrt{SS_{xx}}}$$

Vrijednost  $t$  statistike dobijamo iz Tablice Studentove  $t$  raspodjele za površinu od  $\alpha/2$  na oba kraja raspodjele i  $n-2$  stepena slobode.

## Primjer 13-4

---

Formirati 95% interval povjerenja za parametar  $B$  za podatke o dohotku i izdacima za hranu sedam domaćinstava datim u Tabeli 13.1.



## Primjer 13-4: Rješenje

---

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{1.5939}{\sqrt{1772.8571}} = .0379$$

$$df = n - 2 = 7 - 2 = 5$$

$$\alpha / 2 = (1 - .95) / 2 = .025$$

$$t = 2.571$$

$$b \pm ts_b = .2525 \pm 2.571(.0379)$$

$$= .2525 \pm .0974 = .155 \quad \text{do} \quad .350$$

## Testiranje hipoteze o regresionom parametru $B$

---

### Statistika testa za $b$

vrijednost statistike testa  $t$  za  $b$  računa se kao

$$t = \frac{b - B}{S_b}$$

Vrijednost regresionog parametra  $B$  u prethodnom izrazu zamjenjuje se hipotetičkom vrijednošću koja je definisana nultom hipotezom.

## Primjer 13-5

---

Testirati pri nivou značajnosti 1% da li je koeficijent nagiba regresione prave osnovnog skupa za primjer o dohotku i izdacima za hranu sedam domaćinstava statistički značajan.

## Primjer 13-5: Rješenje

---

- Etapa 1:
  - $H_0: B = 0$  (regresioni parametar je nula)
  - $H_1: B \neq 0$  (regresioni parametar je statistički značajan)
  
- Etapa 2:
  - $\sigma_\epsilon$  nije poznato
  - Dakle, koristimo  $t$  raspodjelu za test o  $B$

## Primjer 13-5: Rješenje

---

- Etapa 3:
- $\alpha = 0.01$
- Površina na oba kraja raspodjele =  $\alpha/2 = 0.01/2 = 0.005$
- $df = n - 2 = 7 - 2 = 5$
- Kritična vrijednost  $t$  je 4.032

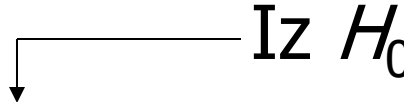
## Primjer 13-5: Rješenje

---

Etapa 4:

$$t = \frac{b - B}{s_b} = \frac{.2525 - 0}{.0379} = 6.662$$

Iz  $H_0$



## Primjer 13-5: Rješenje

---

- Etapa 5:
- Vrijednost statistike testa  $t = 6.662$ 
  - Veća je od kritične vrijednosti  $t = 4.032$
  - Nalazi se u oblasti odbacivanja nulte hipoteze
- Dakle, odbacujemo nultu hipotezu
- Zaključujemo da promjenljiva  $x$  (dohodak) statistički značajno utiče na promjenljivu  $y$  (izdaci za hranu).

## 13.6 LINEARNA KORELACIJA

---

- Koeficijent linearne korelacije
- Testiranje hipoteze o koeficijentu proste linearne korelacije



# Koeficijent linearne korelacije

---

Vrijednost koeficijenta proste linearne korelacije

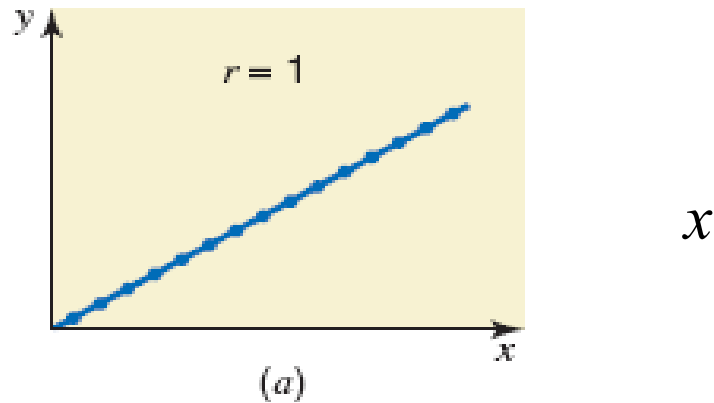
**Vrijednost koeficijenta proste linearne korelacije** je uvijek u intervalu od  $-1$  do  $1$ ; odnosno,

$$-1 \leq \rho \leq 1 \quad \text{i} \quad -1 \leq r \leq 1$$

# Slika 13.18 Linearna korelacija između dvije promjenljive.

---

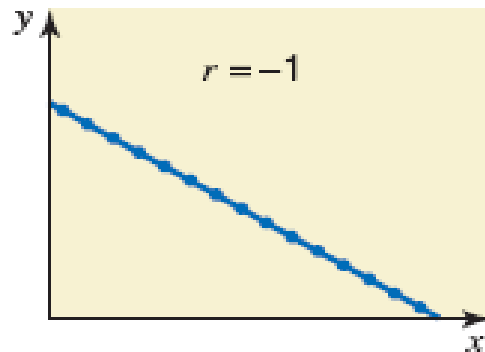
(a) Perfektna pozitivna linearna korelacija,  $r = 1$



# Slika 13.18 Linearna korelacija između dvije promjenljive.

---

(b) Perfektna negativna linearna korelacija,  $r = -1$



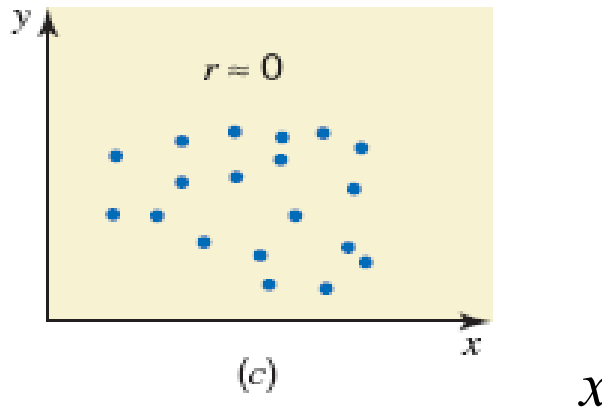
(b)

$x$

# Slika 13.18 Linearna korelacija između dvije promjenljive.

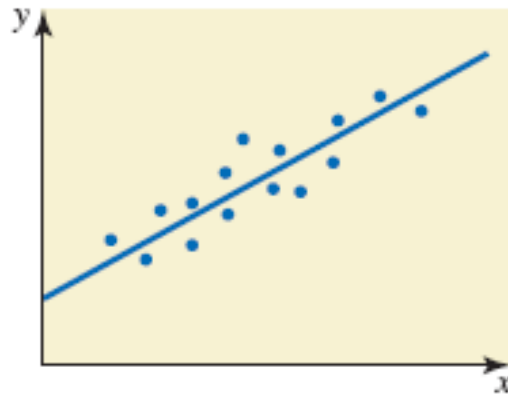
---

(c) Nepostojanje linearne korelacije,  $r \approx 0$



# Slika 13.19 Linearna korelacija između dvije promjenljive.

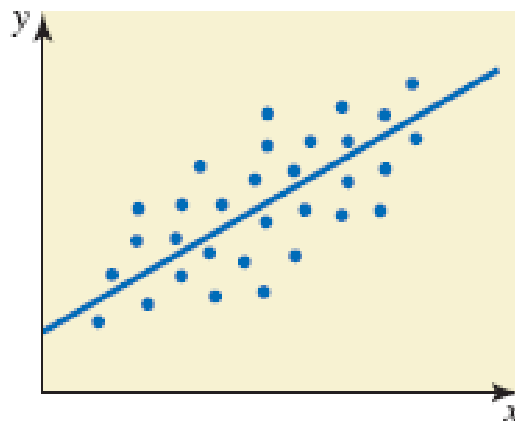
---



(a) Strong positive linear correlation  
( $r$  is close to 1)

# Slika 13.19 Linearna korelacija između dvije promjenljive.

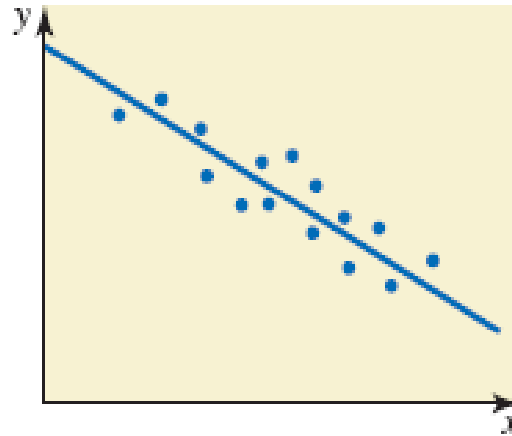
---



(b) Weak positive linear correlation  
( $r$  is positive but close to zero)

# Slika 13.19 Linearna korelacija između dvije promjenljive.

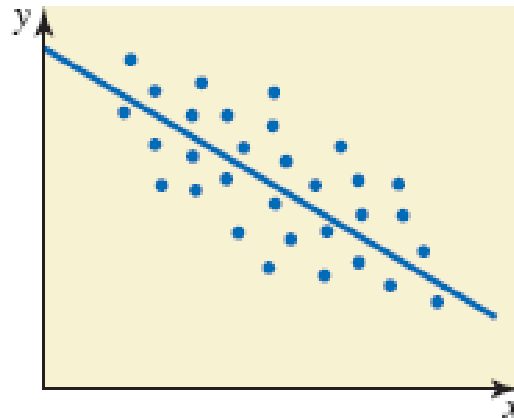
---



(c) Strong negative linear correlation  
( $r$  is close to  $-1$ )

# Slika 13.19 Linearna korelacija između dvije promjenljive.

---



(d) Weak negative linear correlation  
( $r$  is negative and close to zero)



## Koeficijent proste linearne korelacije

---

Koeficijent proste linearne korelacije

**Koeficijent proste linearne korelacije**,

označen sa  $r$ , mjeri jačinu linearne veze između dvije promjenljive u uzorku i glasi:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

## Primjer 13-6

---

Izračunati koeficijent korelacije za primjer o dohotku i izdacima za hranu sedam domaćinstava.

## Primjer 13-6: Rješenje

---

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$
$$= \frac{447.5714}{\sqrt{(1772.8571)(125.7143)}} = .95$$

# Testiranje hipoteze o koeficijentu proste linearne korelacije

---

## Statistika testa za $r$

Pod pretpostavkom da je zajednički raspored dvije promjenljive normalan i za  $H_0: \rho = 0$ , statistika  $t$  testa glasi

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

Ovdje oznaka  $n - 2$  predstavlja stepene slobode.

## Primjer 13-7

---

Koristeći nivo značajnosti 1% i podatke iz primjera 13-1, testirati da li je koeficijent linearne korelacije između dohotka i izdataka za hranu statistički značajan. Pretpostavka je da je zajednička raspodjela dvije promjenljive normalna.

## Primjer 13-7: Rješenje

---

- Etapa 1:
- $H_0: \rho = 0$  (koeficijent linearne korelacije je jednak nuli)
- $H_1: \rho \neq 0$  (koeficijent linearne korelacije je statistički značajan)
  
- Etapa 2: Pod pretpostavkom da je zajednička raspodjela dvije promjenljive normalna, za testiranje hipoteze o koeficijentu proste linearne korelacije koristimo  $t$  raspodjelu.

## Primjer 13-7: Rješenje

---

- Etapa 3:
- Površina na oba kraja raspodjele =  $0.01/2 = 0.005$
- $df = n - 2 = 7 - 2 = 5$
- Kritična vrijednost  $t = 4.032$

## Primjer 13-7: Rješenje

---

Etapa 4:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$
$$= .95 \sqrt{\frac{7-2}{1-(.95)^2}} = 6.803$$



## Primjer 13-7: Rješenje

---

- Etapa 5:
- Vrijednost statistike testa  $t = 6.803$ 
  - Veća je od kritične vrijednosti  $t=4.032$
  - Nalazi se u oblasti odbacivanja
- Dakle, odbacujemo nultu hipotezu
- Zaključujemo da postoji statistički značajna pozitivna veza između dohotka i izdataka za hranu.