Moderne biostatistične metode- multivariabilne metode

Univerza v Ljubljani
Fakulteta za farmacijo

doc. dr. Mitja Kos, *mag. farm.*

Katedra za socialno farmacijo
Univerza v Ljubljani- Fakulteta za farmacijo

Analiza povezanosti

- Opazovani pojav= odvisna spremenljivka
- Napovedni dejavnik= neodvisna spremenljivka
- Statistični modeli:
  - Univariabilni:
    - en napovedni dejavnik
    - Povezava kot pomembna pokaže:
      - zaradi dejanske povezanosti napovednega dejavnika s pojavom
      - lahko tudi zaradi povezanosti z nekim drugim napovednim dejavnikom.
  - Multivariabilni:
    - več napovednih dejavnikov

Linearna regresija

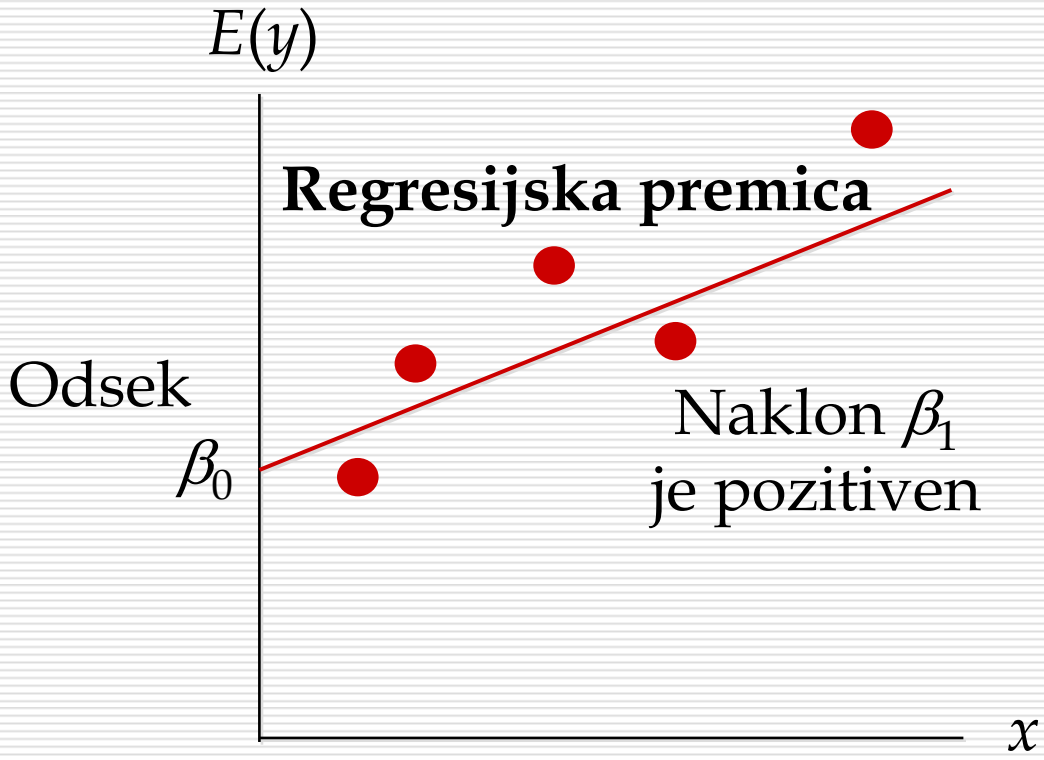
- Preprosta LR: matematični model= premica
  - za vsak posamezni y

$$y = \beta 0 + \beta 1 x + \varepsilon$$

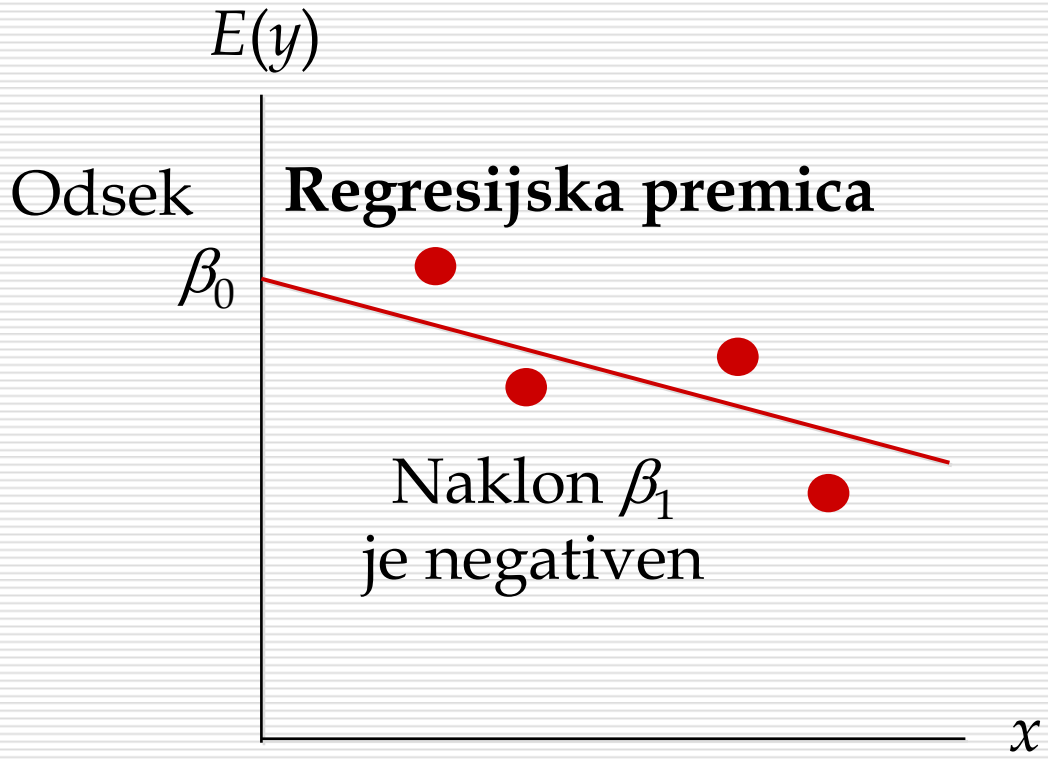
  - β_0 in β_1 sta parametera modela.
  - ε je napaka $N(0, \sigma_e^2)$
- Pričakovana vrednost (povprečen y)

$$E(y) = \beta 0 + b1x \qquad \beta 1 = \frac{\Delta y}{\Delta x}$$

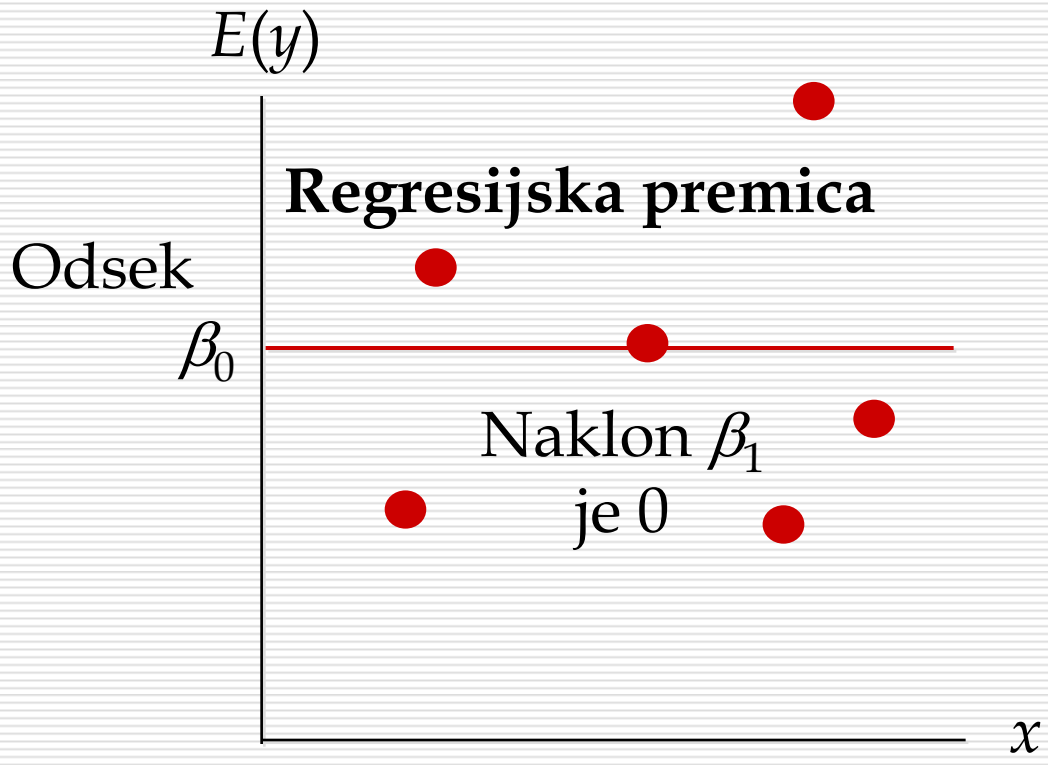
Pozitiven linearni odnos



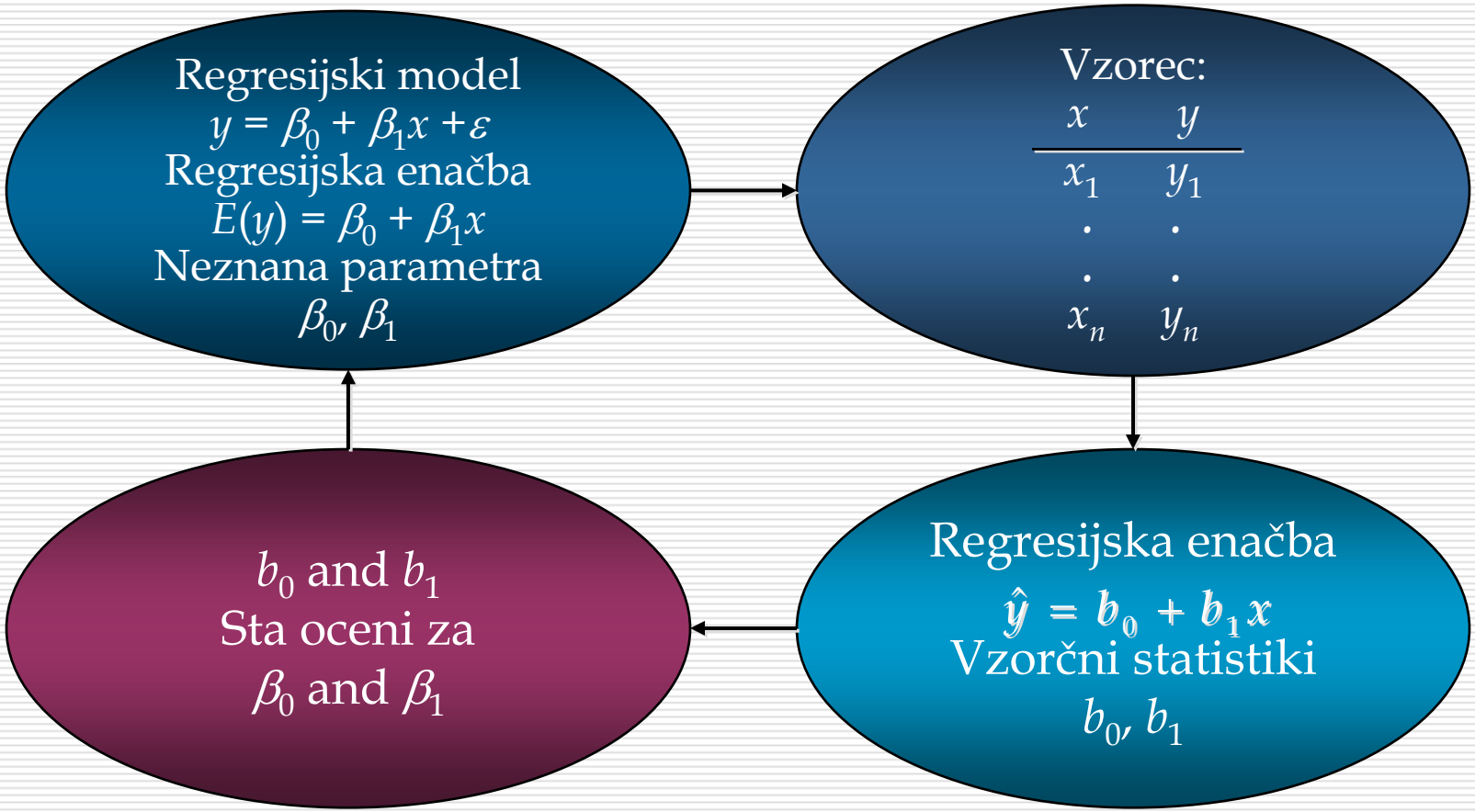
Negativen linearni odnos



Ni povezave



Določanje parametrov modela

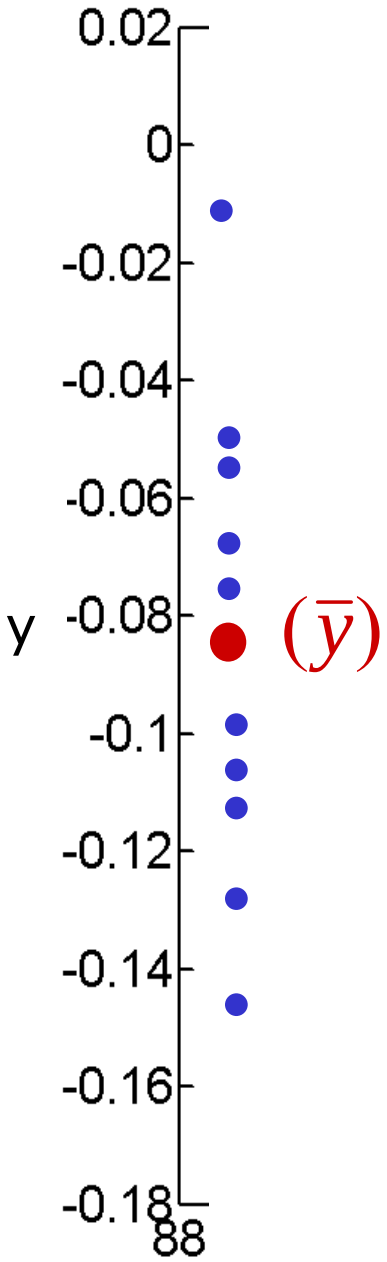


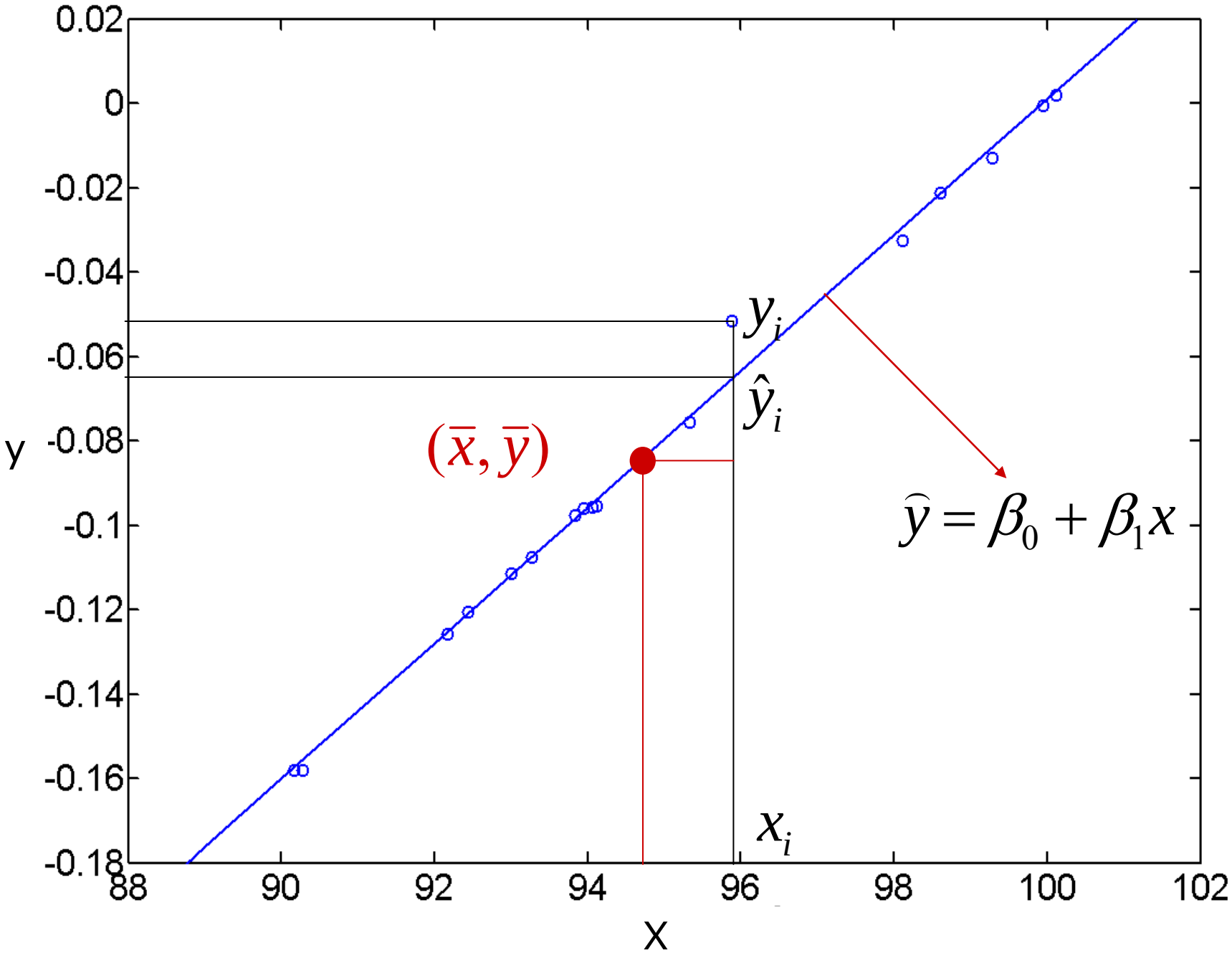
Metoda najmanjših kvadratov

$$\min \sum \left(y_i - \hat{y}_i \right)^2$$

$\hat{y}_i$ = ocena i-te vrednosti odvisne spremenljivke

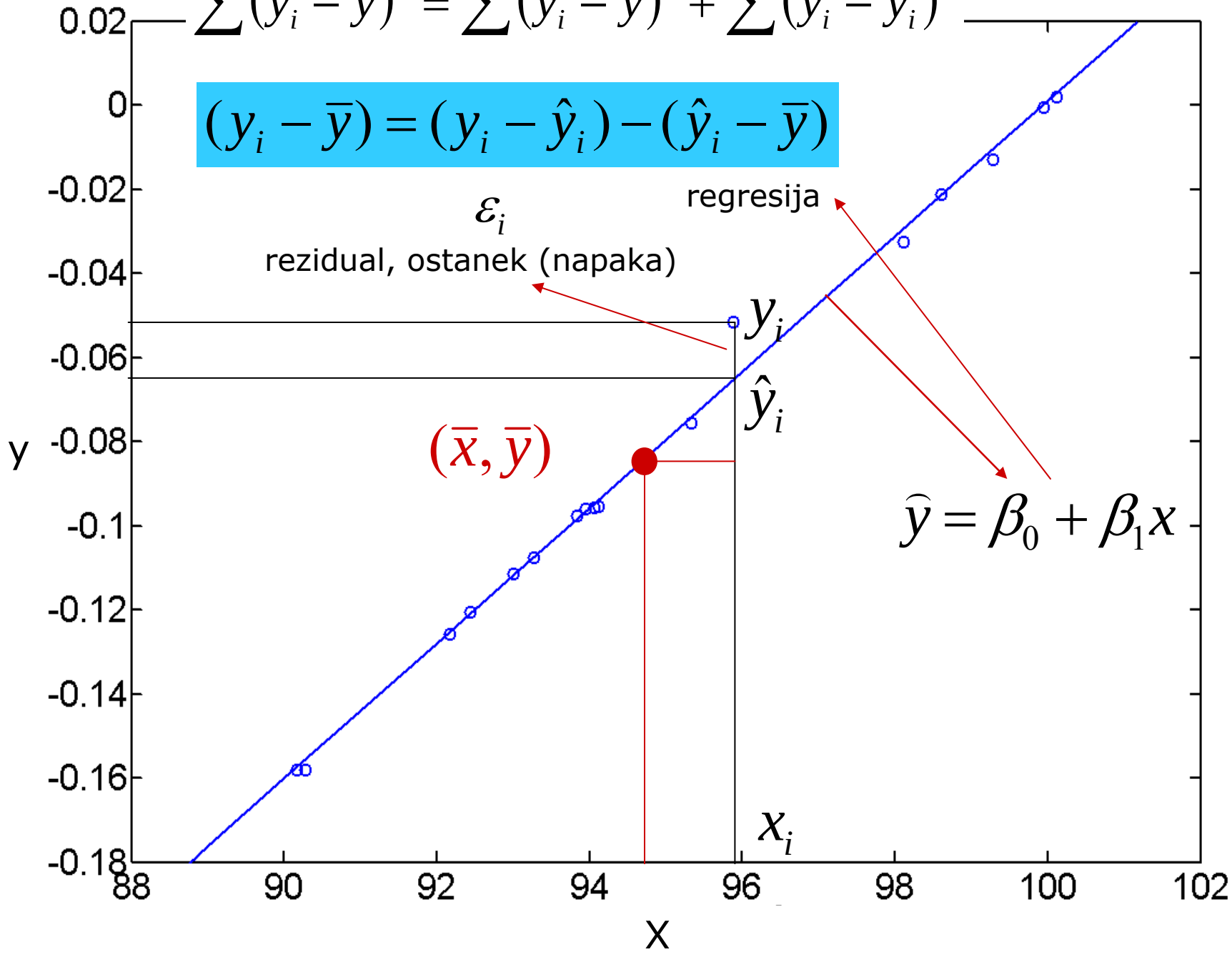
y_i = opažena i-ta vrednosti odvisne spremenljivke


0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
-0.18
y
$(\bar{y})$
88


0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
-0.18
y
88
90
92
94
96
98
100
102
X
y_i
$\hat{y}_i$
$(\bar{x}, \bar{y})$
x_i
$\hat{y} = \beta_0 + \beta_1 x$

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

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) - (\hat{y}_i - \bar{y})$$



Multipli regresijski model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ so parametri.
- Standardizirani parametri ($x_i \rightarrow z_i$)
- ε je napaka, ki je slučajna spremenljivka.

Metoda najmanjših kvadratov

- Kriterij $\min \sum (y_i - \hat{y}_i)^2$
- Določitev koeficientov

  Kompleksna algebra. Uporaba statističnih programskih paketov.

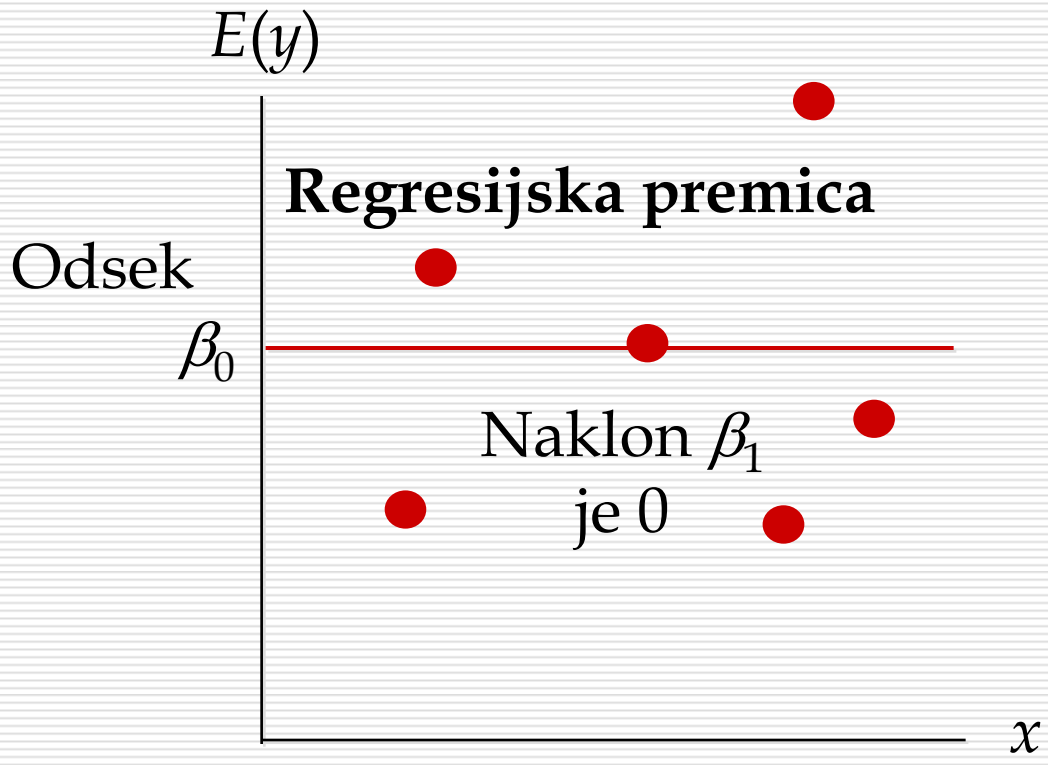
- Interpretacija koeficientov

$$E(y) = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \qquad b_k = \frac{\Delta y}{\Delta x_k}$$

Statistično sklepanje

- Pri enostavni linearni regresiji nas *F* and *t* test vodita k istim sklepom.
- V multipli regresiji uporabimo *F* in *t* test za različna namena.

Ni povezave



F test

- Pomembnost modela kot celote.
- Test for overall significance.
- Hipotezi:

  H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

  H_a: Najmanj eden izmed parametrov ni enak 0

t test

- Če je izid *F* testa za model kot celota značilen, uporabimo *t* teste za ugotavljanje značilnosti vplivov posameznih neodvisnih spremenljivk.
- Za vsako neodvisno spremenljivko izvedemo en *t* test.
- Vsak t test imenujemo tudi test posamične značilnosti (test for individual significance).
- Hipotezi:

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

Koeficient determinacije

- Multipli koeficient determinacije R2
- Kako dobro model opisuje podatke?

Primer:
Vpliv kajenja na pljučno funkcijo

Z raziskavo želimo opredeliti vpliv kajenja na pljučno funkcijo. Naključno smo izbrali 1000 ljudi obeh spolov in spremljali pljučno funkcijo (parameter FEV), poleg tega smo beležili še njihovo starost in telesno višino. Za analizo rezultatov raziskave smo uporabili metodo multiple regresije. Neodvisne spremenljivke v analizi so bile:

starost [leta]
telesna višina [cm]
spol (0 = ženski, 1 = moški)
kajenje (0 = nekadilec, 1 = kadilec)
Odvisna spremenljivka pa je bila FEV [liter].

Primer:
Vpliv kajenja na pljučno funkcijo

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.497[a]	.247	.243	.25269

a. Predictors: (Constant), kajenje, t. višina (cm), starost (leta), spol

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20.786	4	5.196	81.383	.000[a]
	Residual	63.533	995	.064		
	Total	84.319	999			

a. Predictors: (Constant), kajenje, t. višina (cm), starost (leta), spol

b. Dependent Variable: FEV (l)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.666	.118		22.568	.000
	starost (leta)	-1.9E-03	.001	-.063	-2.296	.022
	t. višina (cm)	-5.6E-04	.001	-.024	-.844	.399
	spol	.219	.017	.376	13.208	.000
	kajenje	-.275	.023	-.327	-11.857	.000

a. Dependent Variable: FEV (l)


www.rtvslo.si

Tveganje, obeti ter razmerja

Relativno tveganje

ang. Relative Risk

$$RR = \frac{p_1}{p_2}$$

Razmerje obetov

ang. Odds Ratio

$$\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

Povezava med razmerjem obetov ter relativnim tveganjem:

$$\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1}{p_2} \times \frac{1-p_2}{1-p_1} = RR \times \frac{1-p_2}{1-p_1}$$

Preprosta logistična regresija

- Opazovani pojav= binarna spremenljivka 0/1
- Kako uporabiti linearni model za opisovanje odnosa?

Preprosta logistična regresija

Vrednost binarne spremenljivke zapisati kot:

- Verjetnost, da zavzame vrednost 1 pri danem X:
  - π (x)- populacija
  - p(x)- vzorec
  - Zavzame le vrednosti med 0 in 1- preozek razpon
- Razmerje verjetnosti, da dogodek zgodi in da se ne zgodi:

$$\frac{p(x)}{1-p(x)}$$

  - Zavzame vrednosti med 0 in + neskončnostjo

Preprosta logistična regresija

- Logaritmiranje=> med – in + neskončnostjo

$$\ln\left[\frac{p(x)}{1-p(x)}\right]=\log it$$

Populacija:

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]=a+bx=\log it$$

Ocena parametrov iz vzorca:

$$\ln\left[\frac{p(x)}{1-p(x)}\right]=a+bx=\log it$$

$$p(x)=\frac{e^{a+bx}}{1+e^{a+bx}}$$

Verjetnost, da dogodek zgodil pri danem X

Pomen regresijskega koef. b

- Lin. regresija: $b = y_{(x+1)} - y_{(x)}$

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = a + bx = \log it$$

- Log. regresija: $b = \log it_{(x+1)} - \log it_{(x)}$
- Binarni X (0/1) => x=0, x+1=1.
- Obet da bo y=1:

$$O_{x=1} = \frac{p(1)}{1-p(1)} \qquad O_{x=0} = \frac{p(0)}{1-p(0)}$$

- Logaritmiranje:

$$\log it_{x=1} = \ln\left[\frac{p(1)}{1-p(1)}\right] \qquad \log it_{x=0} = \ln\left[\frac{p(0)}{1-p(0)}\right]$$

Pomen regresijskega koef. b

- Log. razmerja obetov:

$$RO = \frac{O_{x=1}}{O_{x=0}} = \frac{\left[\frac{p(1)}{1-p(1)}\right]}{\left[\frac{p(0)}{1-p(0)}\right]}$$

$$O_{x=1} = \frac{p(1)}{1-p(1)}$$

$$O_{x=0} = \frac{p(0)}{1-p(0)}$$

$$\ln RO = \ln \frac{\left[\frac{p(1)}{1-p(1)}\right]}{\left[\frac{p(0)}{1-p(0)}\right]} = \log it(1) - \log it(0)$$

$$\ln RO = b = \log it_{(1)} - \log it_{(0)}$$

$$RO = e^{b}$$

Ocenjevanje b v vzorcu

- p(x)? =>b in a ?

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = a + bx = \log it$$

- Lin. regresija: metoda najmanjših kvadratov ostankov.
- Log. regresija: metoda največjega verjetja (maximum likelihood method).
  - Funkcija največjega verjetja
  - Oz. logaritem funkcije verjetja ("log likelihood"):
    - Nelinearna funkcija parametrov modela a in b.
    - Iteracijska metoda, več ocen parametrov. Nove ocene, dokler še zveča funkcijo največjega verjetja. Lokalna/globalna točka največjega verjetja. Start?

Ocenjevanje b v populaciji

- Vzorčna porazdelitev b- ja: normalna porazdelitev = > interval zaupanja:

$$\beta = b \pm z * SE_{(b)}$$

- Vzorčna ocena razmerja obetov: nenormalna, nesimetrična porazdelitev
- Interval zaupanja za RO iz b (spodnjo/zgornjo mejo):

$$e^{b \pm z * SE_b}$$

Multipla logistična regresija

- Matematični model:

$$p(x) = \frac{e^{a+b_1x_1+b_2x_2+\ldots+b_kx_k}}{1+e^{a+b_1x_1+b_2x_2+\ldots+b_kx_k}}$$

- b(k): sprememba v logitu, ki spremlja spremembo X(k) za 1 enoto, medtem ko se ostale sprem. ne spreminjajo.
- Regresijski koef. b = logaritem razmerja obetov
- Razmerje obetov= antilogaritem b: e^b
- Metoda največjega verjetja

Testiranje pomembnosti modela kot celote

- Ali se log verjetja modela s spremenljivkami X statistično značilno poveča v primerjavi z log verjetja modela brez njih.
- G- statistika- test razmerja dveh verjetij:

$$G = 2\ln\left[\frac{verjetje_{SPREM.X}}{verjetje_0}\right]$$

- Porazdeljuje po hi- kvadrat, df: k-1
- k: št. vseh spremenljivk v modelu
- V SPSS-u: “Omnibus test”

Vrednotenje prileganja modela

- Nagelkerkejev R^2

- Test hi- kvadrat
- Hosmer- Lemeshov test:
  - Oba primerjata opazovano število enot, pri katerem se je opazovani dogodek zgodil, s pričakovanim številom, ki temelji na enačbi logistične regresije.

Testiranje pomembnosti b

- Ho: b=0, Ha: b≠ 0
- Testi:
  - Z: $$z = \frac{b - 0}{SE_b}$$
  - Waldova statistika: $$\chi^2 = \left(\frac{b}{SE_b} \right)^2$$
  - G- statistika: $$G = 2 \ln \left[\frac{verjetje_x}{verjetje_0} \right]$$

G- statistika: Model, ki vključuje X, pove več o Y.