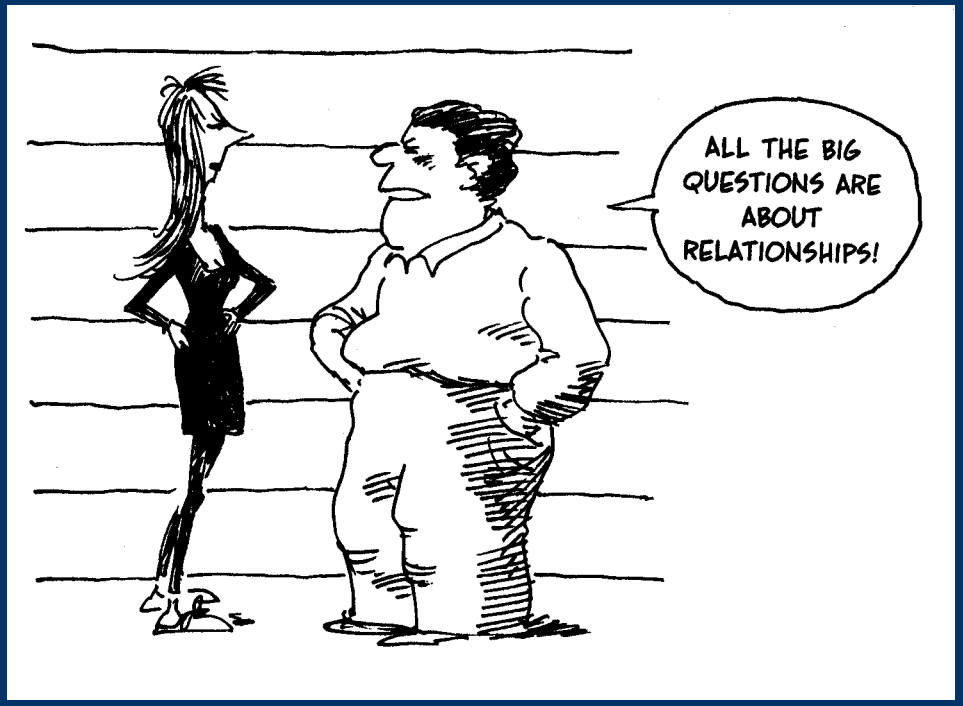
Univerza v Ljubljani
Fakulteta za farmacijo



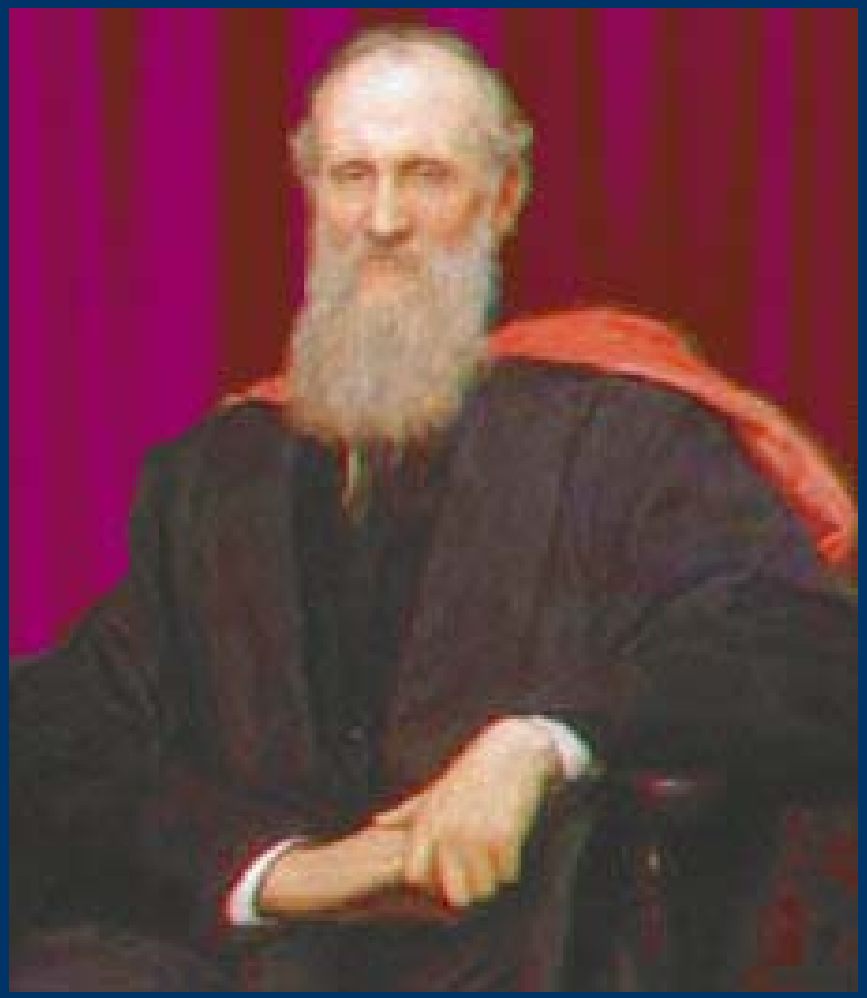
LINEARNA REGRESIJA in KORELACIJA

Iztok Grabnar
Univerza v Ljubljani, Fakulteta za farmacijo

Maj, 2009

Lord Kelvin

1824-1907



"In science there is only physics; all the rest is stamp collecting."

“When you can measure what you are speaking about and express it in numbers, you know something about it."

“They were so intent on making everything numerical, that they frequently missed seeing what was there to be seen."

Barbara McClintock

Matematični model

$$dx = k.x + m.y$$
$$dy = f.x$$

Regresijska analiza

- Univariatna, bivariatna, multivariatna analiza
- Statistična spremenljivka
  - Numerična zvezna
  - Atributivna
  - Dihotomna
- Linearna, logistična regresija
- Nelinearna regresija

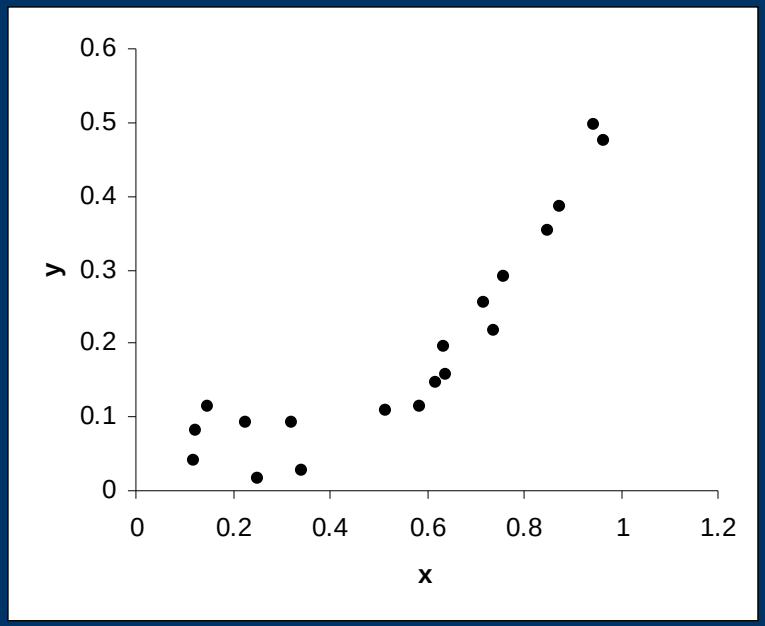
Statistično modeliranje

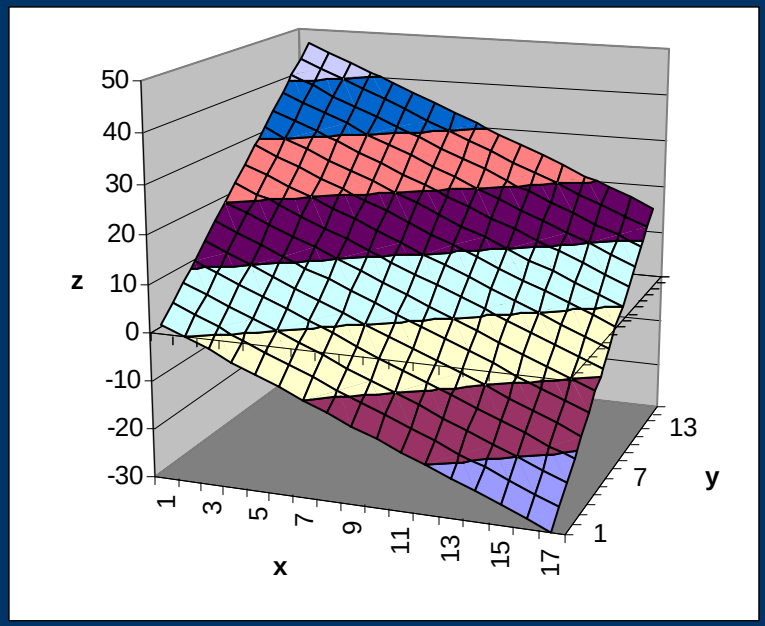
V procesu opazovanja povezanosti med statističnimi spremenljivkami se lahko osredotočimo na povezavo med opazovanim pojavom na eni strani (**odvisna spremenljivka**) in samo eno značilnostjo (**neodvisna spremenljivka**) na drugi - **preprosti modeli**, ali pa med opazovanim pojavom na eni strani in več značilnostmi na drugi **- multivariatni modeli**.

Oba postopka imata skupne značilnosti. Podobno kot pri opisovanju porazdelitve verjetnosti spremenljivk, ko dejansko porazdelitev opišemo z neko matematično funkcijo, tudi v analizi povezanosti opišemo odnose med spremenljivkami z nekim matematičnim modelom. Za analizo izberemo tistega, ki se nam zdi najprimernejši.

Grafični prikazi v analizi povezanosti

- Razsevni diagrami
- Odgovorne površine





Enostavna linearna regresija

- Matematični model
- Metoda najmanjših kvadratov
- Koeficienta korelacije in determinacije
- Predpostavke matematičnega modela
- Statistično sklepanje
- Uporaba modela
- Statistični programski paketi
- Analiza rezidualov: validacija predpostavk

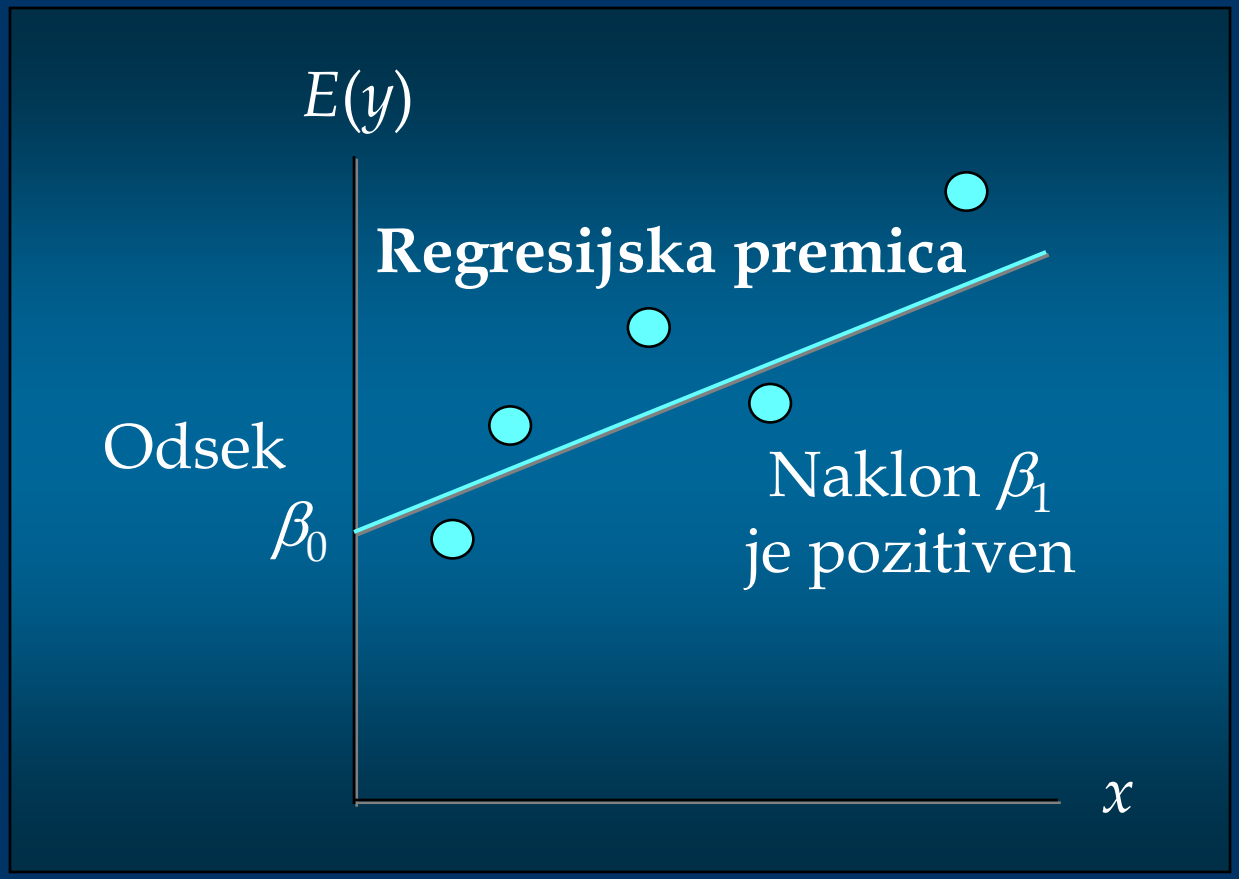
Enostavni linearni regresijski model

- Je enačba, ki opisuje odnos med odvisno (y) in neodvisno (x) spremenljivko in napako (ε).
- Enačba enostavnega linearnega regresijskega modela je:

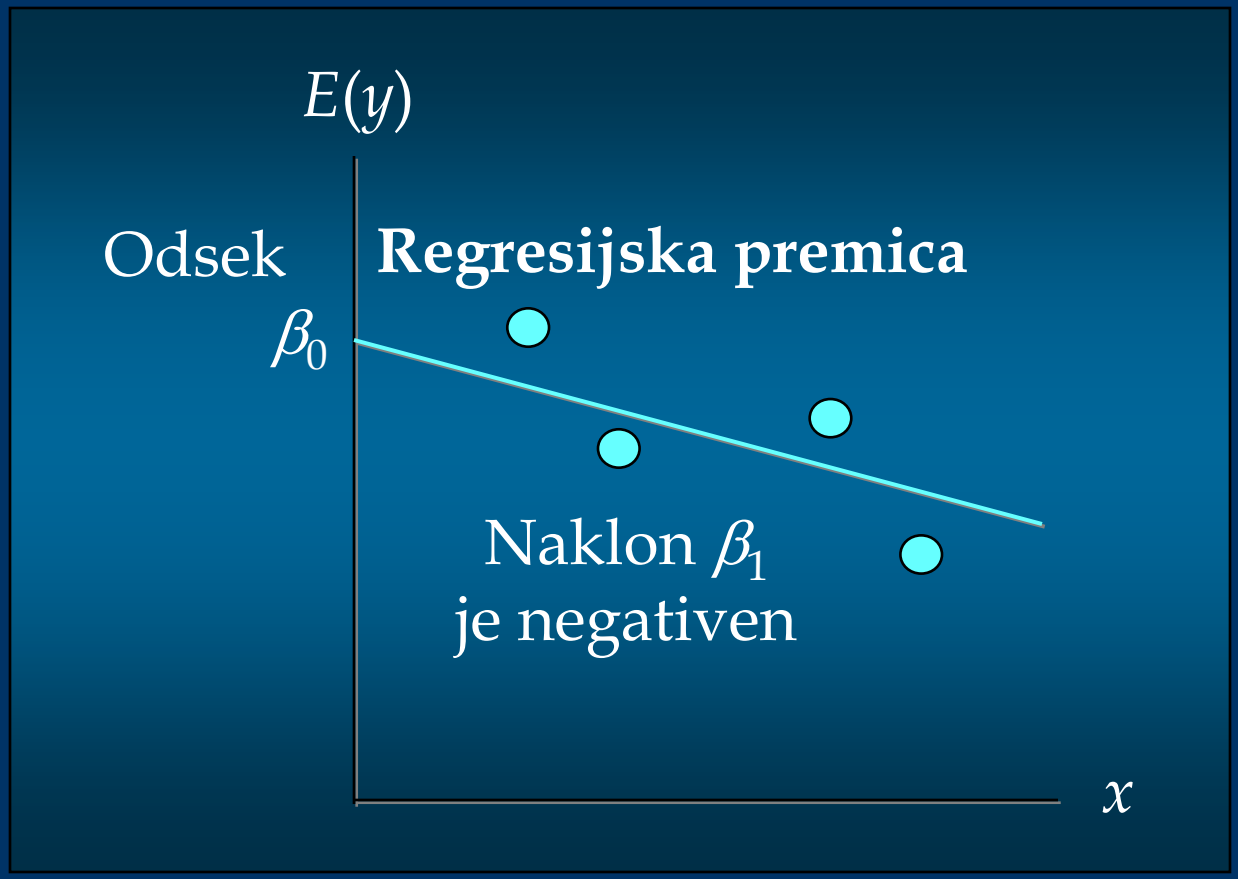
$$y = \beta_0 + \beta_1 x + \varepsilon$$

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

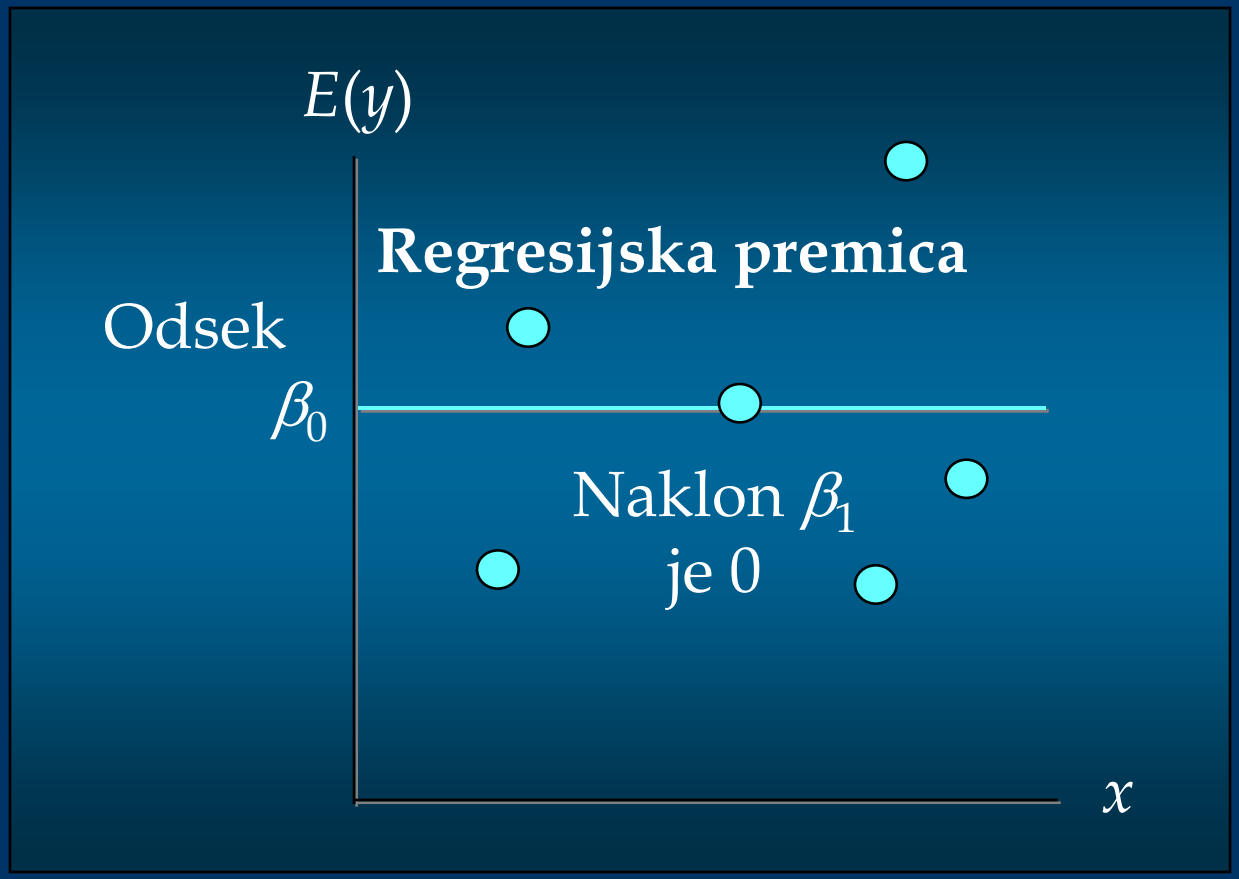
- Pozitiven linearni odnos



- Negativen linearni odnos



- Ni povezave

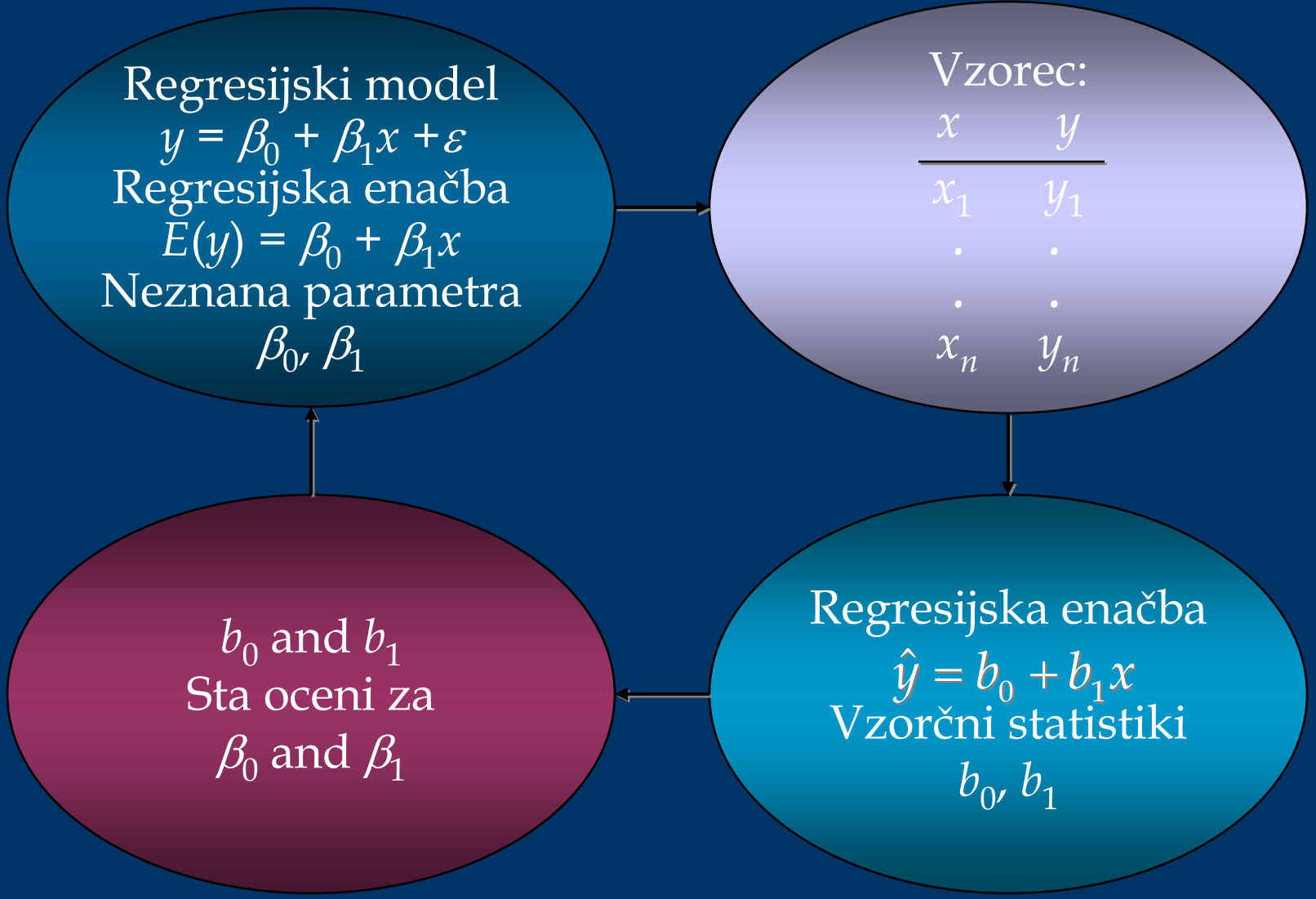


Vzorčni statistiki

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

b_0 in b_1 sta statistiki.

Določanje parametrov modela



Vsota najmanjših kvadratov

Gauss

...the most probable value of unknown quantities will be that in which the sum of squares of the differences between the actually observed and computed values multiplied by numbers that measure the degree of precision is a minimum...

$$\hat{\alpha}_{WLS} \rightarrow min \sum_{j=1}^{m} w_j \left(z(t_j) - y(\alpha, t_j) \right)^2$$

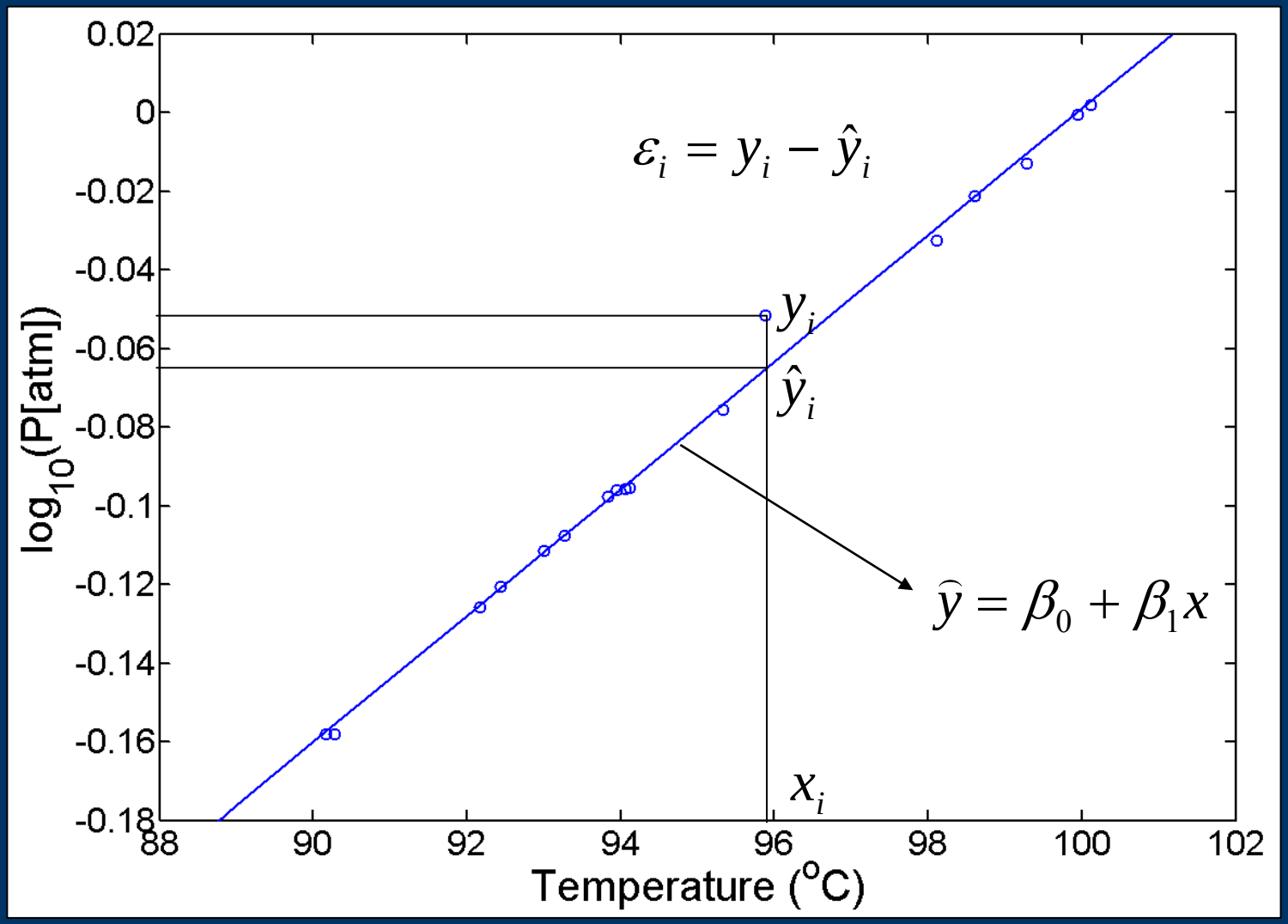
Johann Carl Friedrich Gauss (1777-1855)

Metoda najmanjših kvadratov

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

y_i = opažena i-ta vrednost odvisne spremenljivke

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


$\varepsilon_i = y_i - \hat{y}_i$
y_i
$\hat{y}_i$
$\hat{y} = \beta_0 + \beta_1 x$
x_i
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
-0.18
88
90
92
94
96
98
100
102
$\log_{10}$(P[atm])
Temperature (°C)

- Ocena naklona

$$b_1 = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}$$

- Ocena odseka

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

Koeficient determinacije

- Odnos med SST, SSR, SSE

$$SST = SSR + SSE$$

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

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

- Koeficient determinacije je:

$$r^2 = \text{SSR}/\text{SST}$$

kjer je:

SST = total sum of squares

SSR = sum of squares due to regression

Koeficient korelacije

- Vzorec

$$r_{xy} = (\text{predznak } b_1)\sqrt{\text{koeficient determinacije}}$$

$$r_{xy} = (\text{predznak } b_1)\sqrt{r^2}$$

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

Predpostavke modela

- Napaka ε
  1. ε je slučajna spremenljivka z aritmetično sredino 0.
  2. varianca ε, ki jo označimo $\sigma_e^{\,2}$, je enaka za vse vrednosti neodvisne spremenljivke.
  3. vrednosti ε so neodvisne.
  4. Porazdelitev ε je normalna

Statistično sklepanje

- Ničelna statistična hipoteza:
- $\beta_1 = 0$
- Dva pristopa
  - t Test
  - F Test
- Pri obeh je potrebna ocena $\sigma_e{}^2$.

- Ocena $\sigma_e^{\,2}$

  Varianca napake MSE ali s^2

$$s^2 = \text{MSE} = \text{SSE}/(n-2)$$

kjer:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

- Ocena σ_e
  - Kvadratni koren $\sigma_e{}^2$
  - s imenujemo standardna napaka ocene.

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

- Hipotezi

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

- Testna spremenljivka

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

$$s_{b_1} = \sqrt{\frac{\text{MSE}}{\sum (x_i - \bar{x})^2}}$$

t Test

- Zavrnitev ničelne hipoteze

Zavrni H_0 če $t < -t_{\alpha/2}$ ali $t > t_{\alpha/2}$

kjer je: $t_{\alpha/2}$ odklon v t porazdelitvi
z $(n - 2)$ stopinjami prostosti

Določanje intervala zaupanja za $\beta 1$

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

kjer je b_1 točkovna ocena naklona

$t_{\alpha/2} s_{b_1}$ kritična meja

$t_{\alpha/2}$ vrednost iz t-porazdelitve z $(n - 2)$ stopinjami prostosti

F Test

- Hipotezi

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

- Testna spremenljivka

$$F = \text{MSR}/\text{MSE}$$

- Zavrnitev ničelne hipoteze

Zavrni H_0 če $F > F_\alpha$

kjer je F_α odklon v F porazdelitvi
z 1 stopinjo prostosti v števcu ulomka in
$(n - 2)$ v imenovalcu ulomka

Previdnost pri interpretaciji statistične značilnosti

- Zavrnitev H_0: $\beta_1 = 0$ in sklep, da je povezava med spremenljivkama *x* in *y* značilna ne dopušča opredelitve povezave v smislu vzrok in posledica.
- Z zavrnitvijo H_0: $\beta_1 = 0$ nismo dokazali, da je odnos med spremenljivkama linearen.

Uporaba regresijske enačbe za napovedovanje

- Interval zaupanja $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$$

- Interval napovedovanja y_p

$$y_p \pm t_{\alpha/2} s_{\text{ind}}$$

Kjer je: $(1 - \alpha)$ koeficient zaupanja

Značinost koeficienta korelacije

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

$$t_{n-2,\alpha} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

H_0: $\rho = \rho_0$
H_1: $\rho \neq \rho_0$

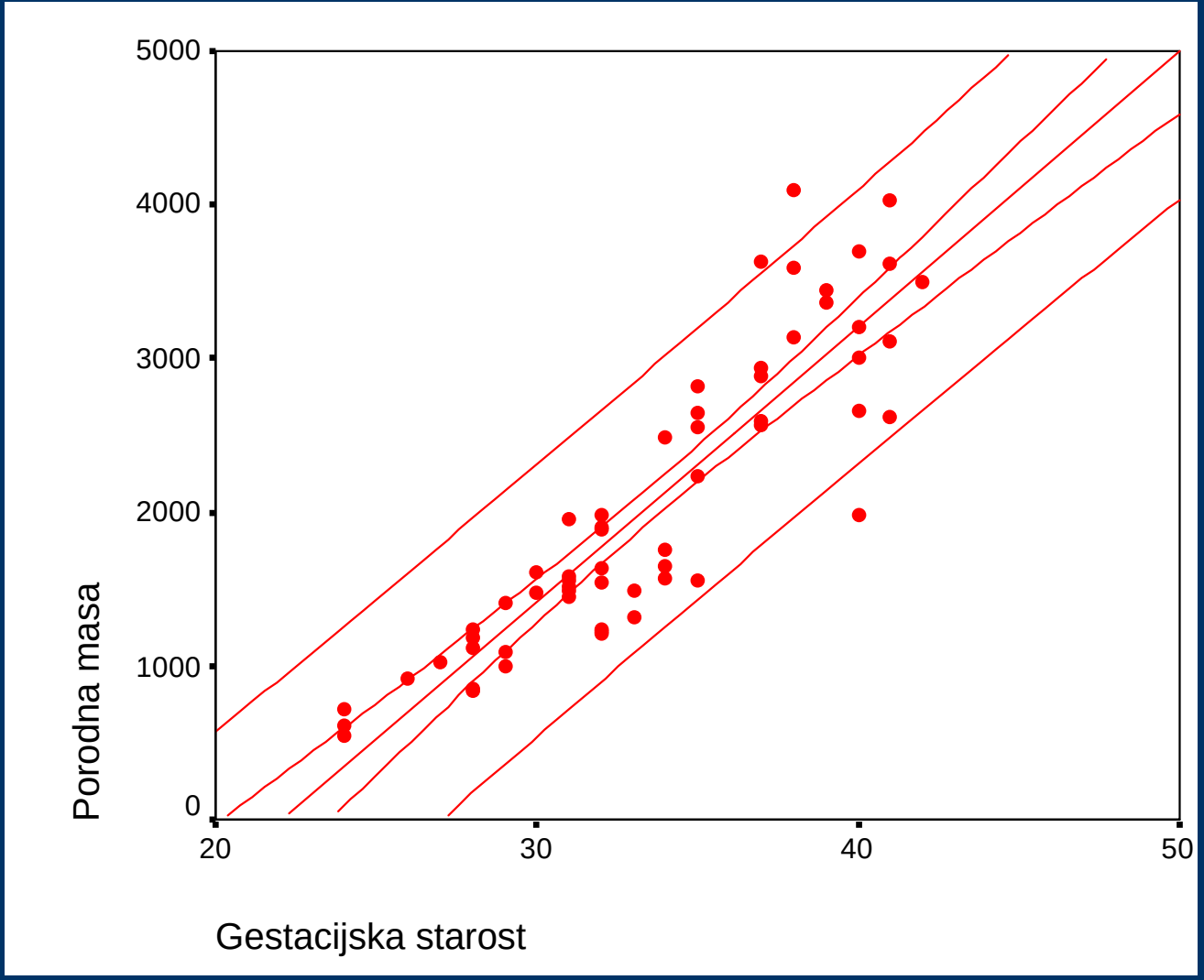
Fisherjeva transformacija

$$u = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$$s_u = \frac{1}{\sqrt{n-3}}$$

$$z_\alpha = (u - u_0)\sqrt{n-3}$$

Interval zaupanja in interval napovedovanja



- Interval zaupanja

$$\hat{y} \pm \mathrm{t}_{\mathrm{n}-2,\alpha} s \qquad s = \sqrt{\mathrm{MSE}\left(\frac{1}{\mathrm{n}} + \frac{(x-\bar{x})^2}{L_{xx}}\right)}$$

- Interval napovedovanja

$$\hat{y} \pm \mathrm{t}_{\mathrm{n}-2,\alpha} s \qquad s = \sqrt{\mathrm{MSE}\left(1 + \frac{1}{\mathrm{n}} + \frac{(x-\bar{x})^2}{L_{xx}}\right)}$$

$$L_{xx} = \sum\left[(x_i - \bar{x})^2\right]$$

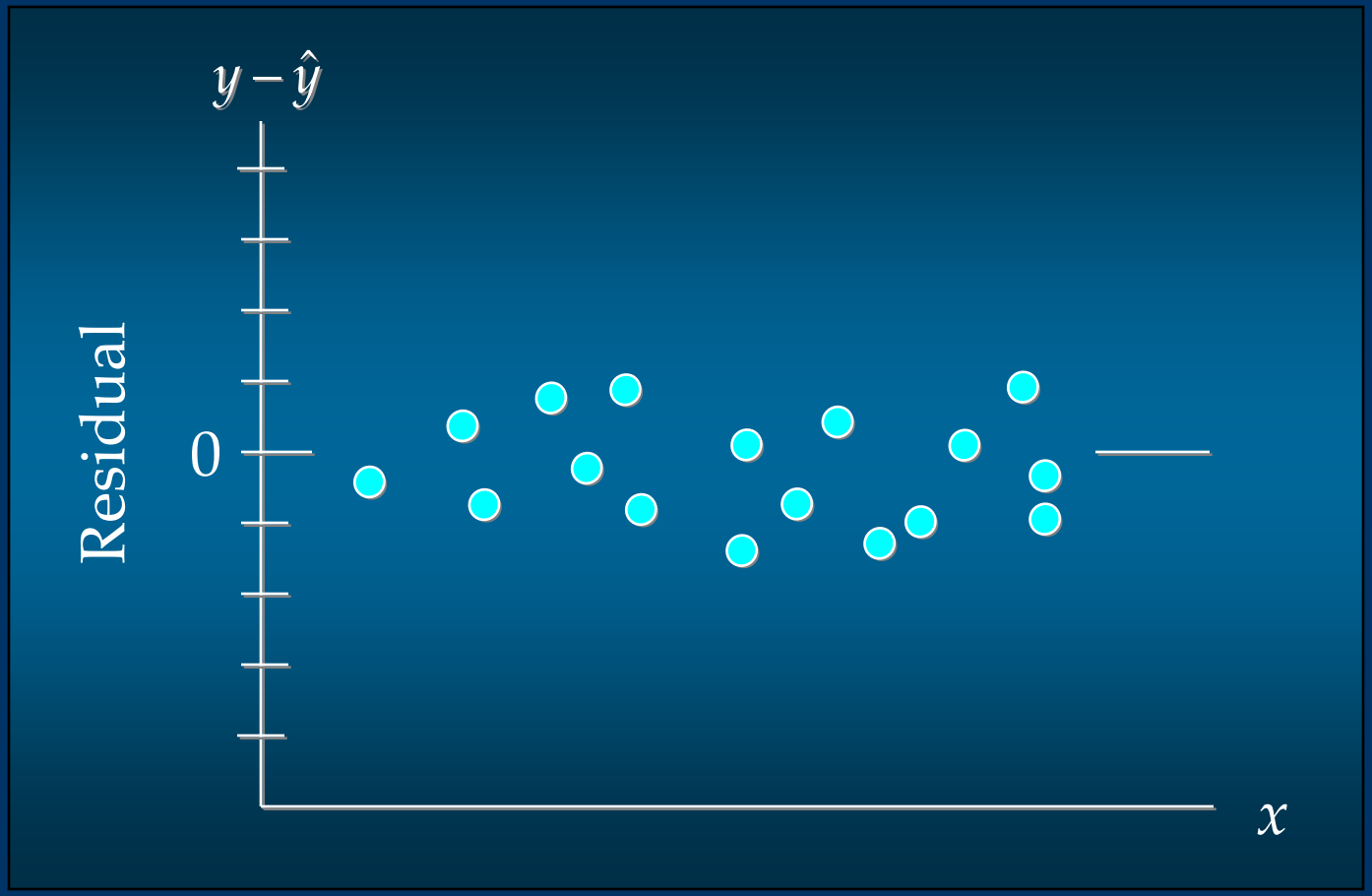
Analiza rezidualov

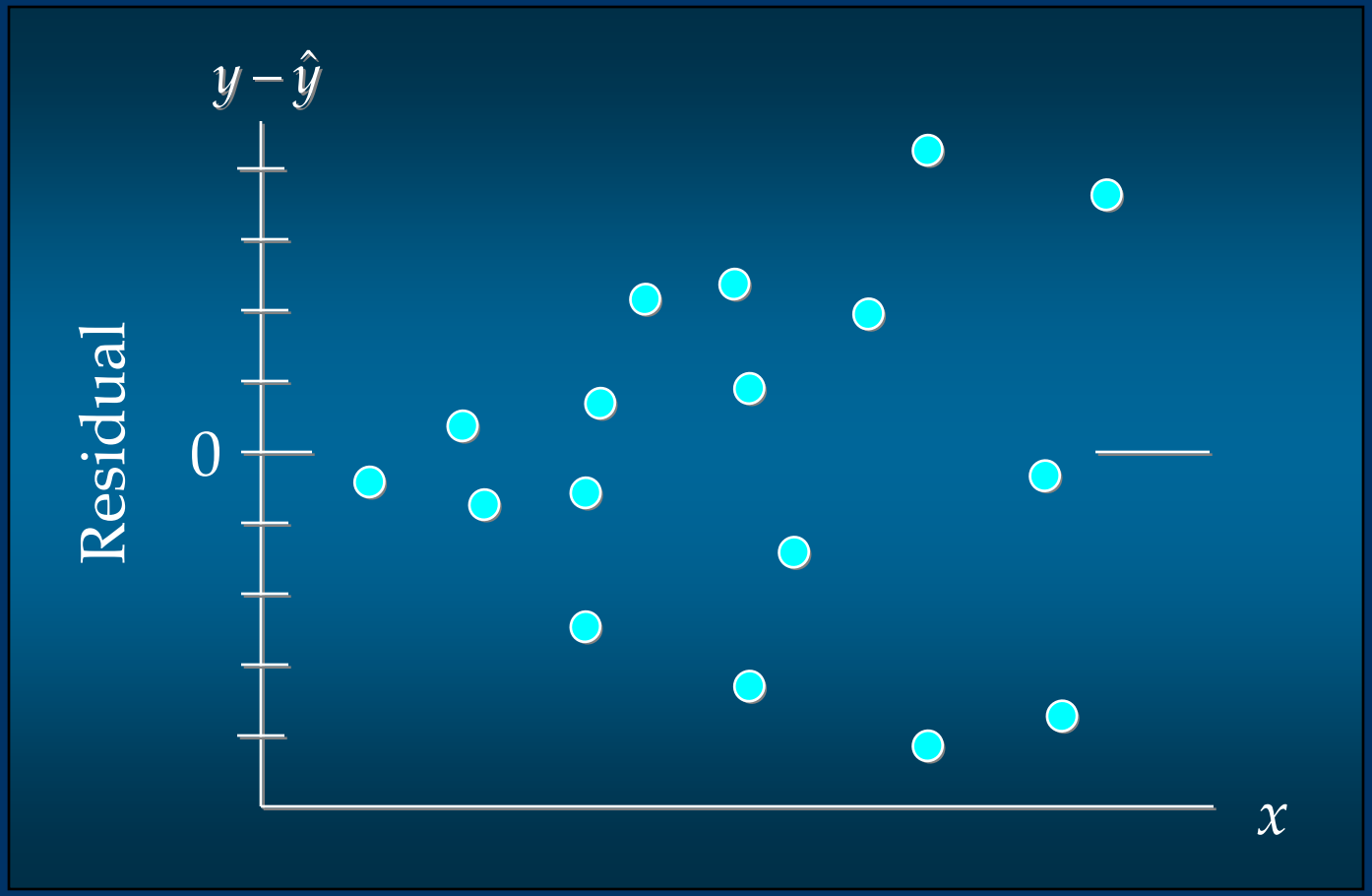
- Rezidual *i*-te vrednosti

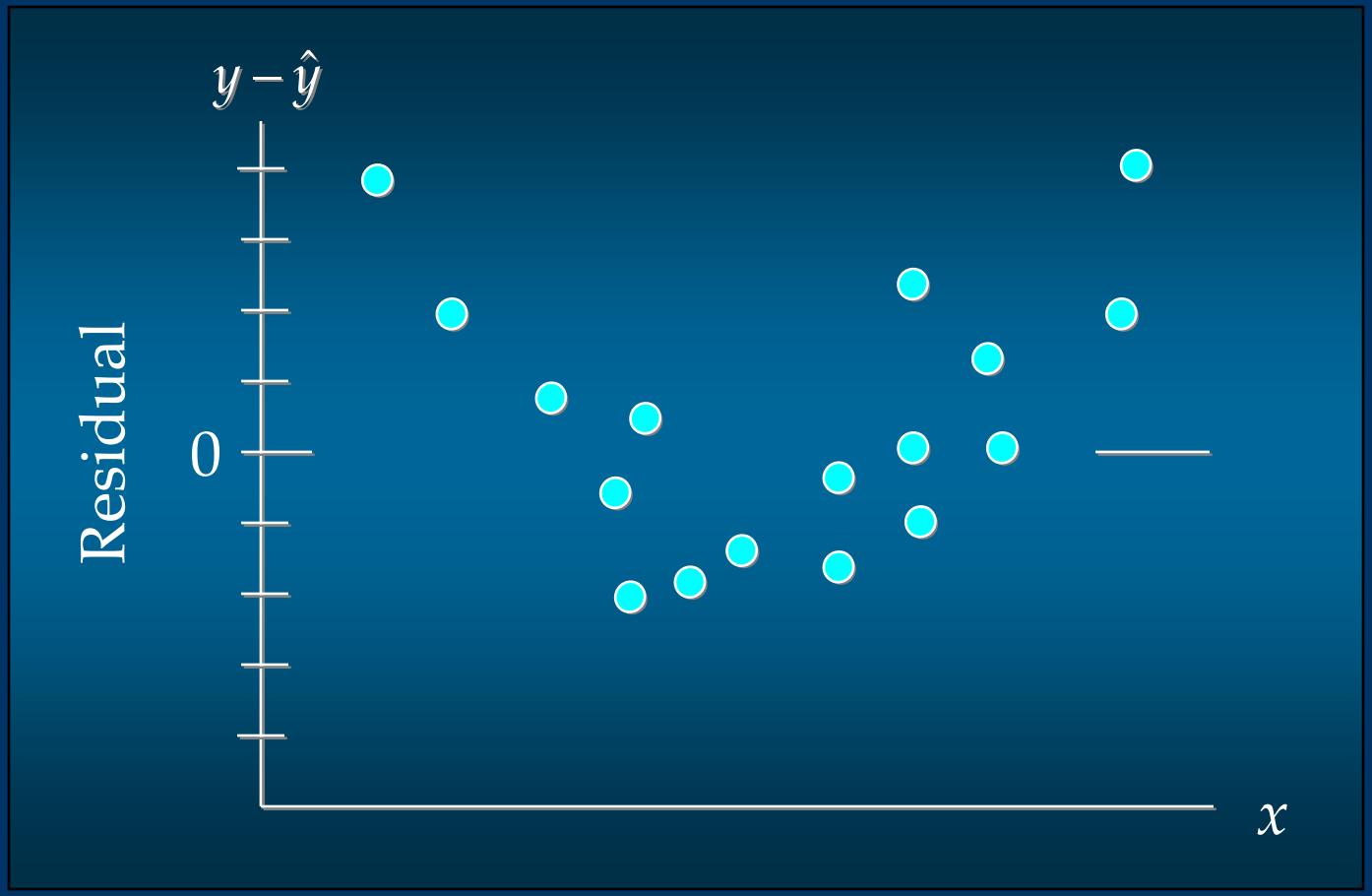
$$y_i - \hat{y}_i$$

- Standardizirani rezidual

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$


$y - \hat{y}$
Residual
0
x


$y - \hat{y}$
Residual
0
x


$y - \hat{y}$
Residual
0
x

Korelacija rangov

- Koeficient korelacije po Spearmanu

$$\rho = 1 - \frac{6\sum d_i^{\,2}}{n(n^2-1)}$$

d_i – razlika rangov x_i in y_i

- Koeficient korelacije po Kendallu