ALGORITMI ANALIZE MASIVNIH PODATKOV

DOMEN MONGUS

P04 – Analiza časovnih vrst

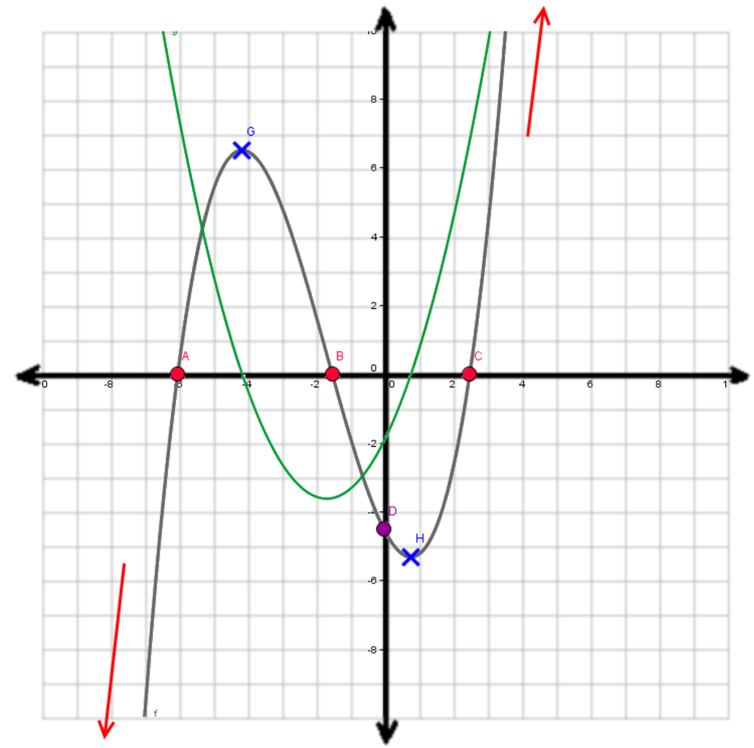
Motivacija - V 4 Velocity

- Časovna vrsta
  - Časovno urejena množica opazovanj

- Aplikacije:
  - Vremenske napovedi (temperatura, vlažnost, …)
  - Finančni trendi (vrednost valut, delnic …)
  - Povpraševanje po dobrinah (nakupi)
  - Medicina (srčni utrip, EEG,…)

Motivacija - V 4 Velocity

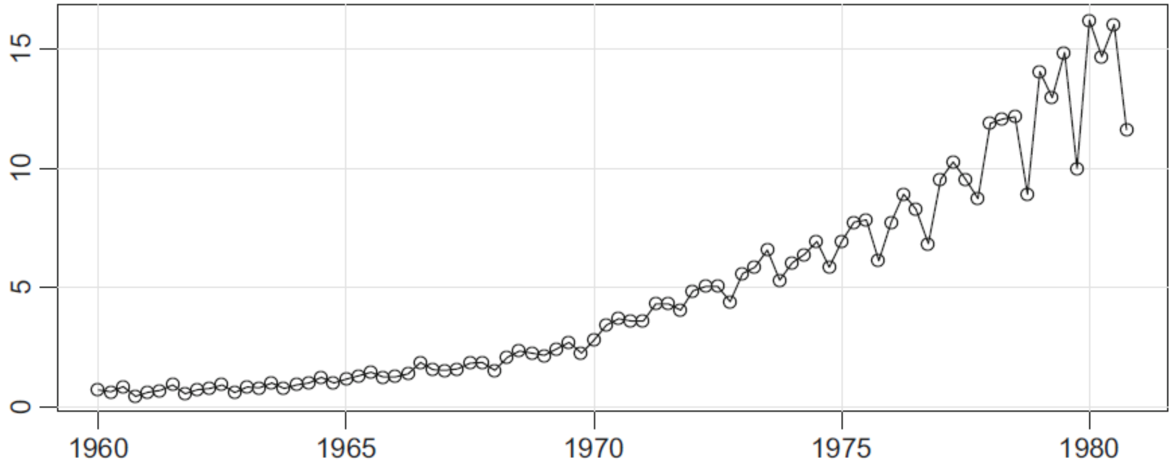
- Časovna vrsta
  - V čem je razlika?
- Regresija (tradicionalno)
  - Ciljna spremenljivka
  - Razlagalne spremenljivke



$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Motivacija - V 4 Velocity

- Časovna vrsta
  - Tradicionalna časovna vrsta (četrtletni zaslužki podjetja)



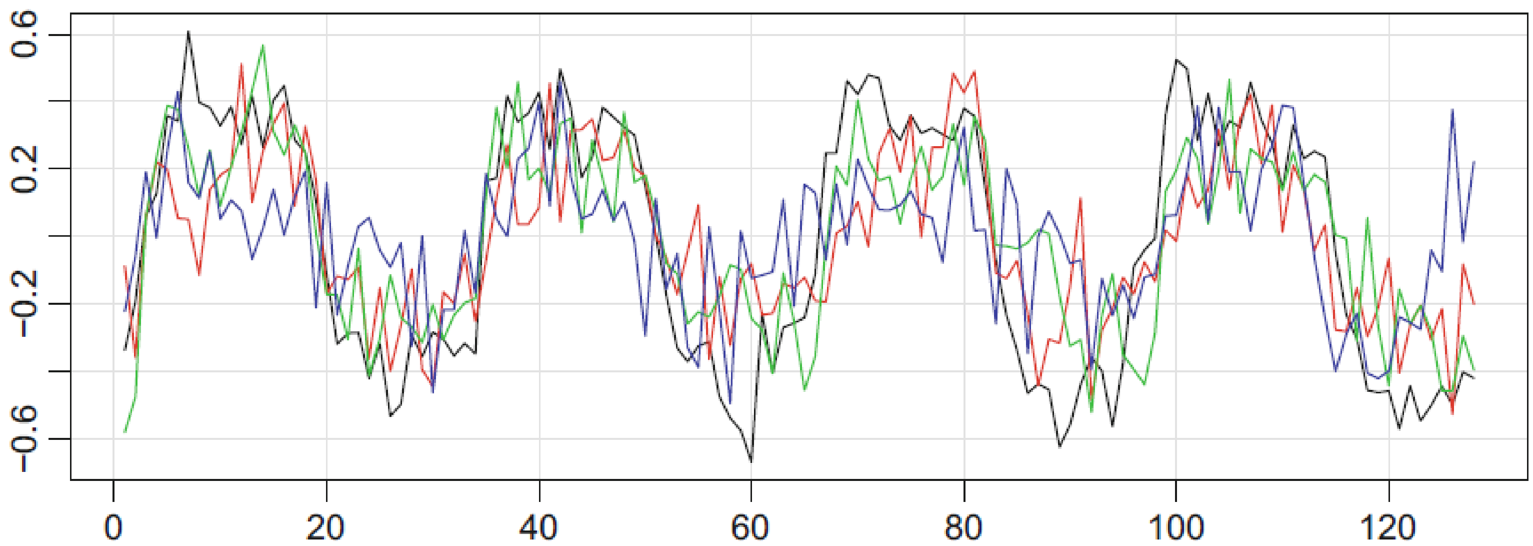
  - Analiza vsebovanih vzorcev za predvidevanje:
    - Trendi, cikli, šum, povezave z zunanjimi okoliščinami …

Vsebina

- Analiza in lastnosti časovnih vrst
- Napovedovanje vrednosti
  - Autokorelacija
  - ARIMA

Narava časovnih vrst

- Obravnavali bomo le **univariantne diskretne časovne vrste**
  - Univariantnost – spremljamo eno samo spremenljivko
  - Meritve izvajamo v enakomernih časovnih korakih
- Notacija
  - Naključna spremenljivka $X = \{x_t\}$, kjer t predstavlja čas
  - $t \in \{1, 2, \ldots, T\}$
- Variabilnost:

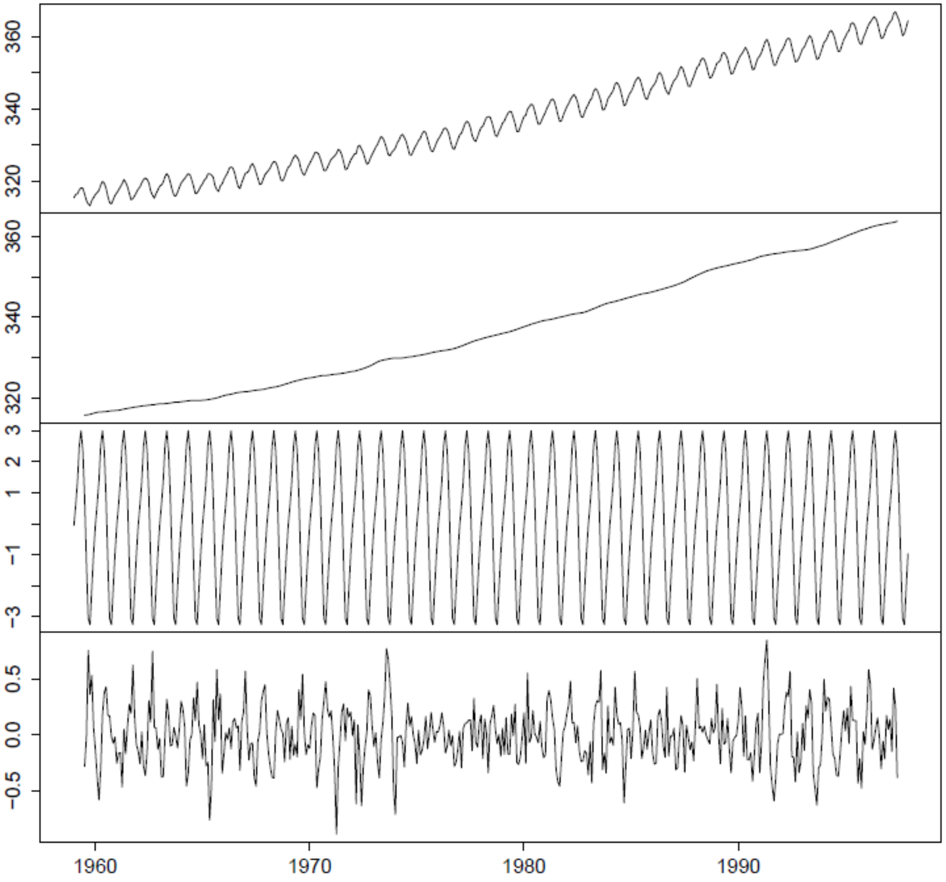


Osnovni matematični model

- Notacija
  - Naključna spremenljivka $X = \{x_t\}$, kjer t predstavlja čas
  - $t \in \{1, 2, \ldots, T\}$
- Naivna različica: $x_t = f(t)$
  - Zaradi visoke stopnje variabilnosti skoraj nikoli ni učinkovit
- Splošni model: $x_t = f(t) + \varepsilon$
  - $f(t)$ – deterministični del, ki sledi časovnim zakonitostim
  - ε – naključni del, ki sledi zakonom verjetnosti

Osnovni matematični model

- Razlogi za variabilnost vrednosti
  - Dekompozicija signala:
  - Trend:
  - Sezonski efekti:
  - Neregularne fluktuacije:



Stacionarna časovna vrsta

- Definicija:
  - Časovna vrsta je stacionarna, kadar je verjetnost pojavitve vsake vrednosti $X = \{x_t\}$ enaka verjetnosti pojavitve vsake vrednosti v drugem časovnem obdobju $X_h = \{x_{t+h}\}$,
  - Taka časovna vrsta je odvisna zgolj od časovne razlike in ne od dejanskega časa!
- Šibko stacionarna:
  - Povprečje je konstanta
- Zakaj je stacionarnost koristna?

Avtokorelacija

- Pearsonov korelacijski koeficient

$$r = r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\,\sqrt{(\sum y_i^2 - n\bar{y}^2)}}.$$

- kjer
  - *n* – število elementov
  - *x*,*y* – spremenljivki

Avtokorelacija

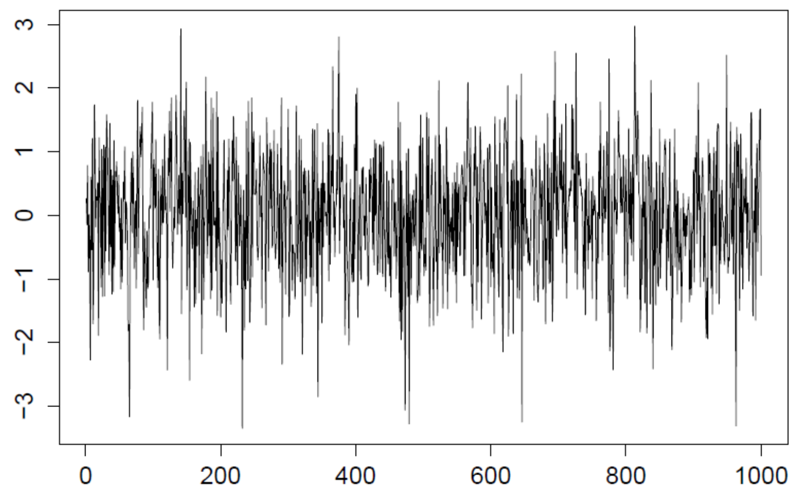
- Definicija
  - Korelacija med signalom $X = \{x_t\}$ in njegovo zakasnjeno kopijo $X_h = \{x_{t+h}\}$
- Naivni pristop k napovedovanju:
  - Poiskati najprimernejši h
  - $x_{t+1} = x_{t+h+1}$
- Določa sezonski efekt oz. periodičnost signala

Tradicionalne časovne vrste

Naključne vrednosti

- Nabor vrednosti iz območja [X,Y]
- Ima konstantno povprečje
- Konstantno varianco
- Je stacionaren

Beli šum (Gaussovo naključje)

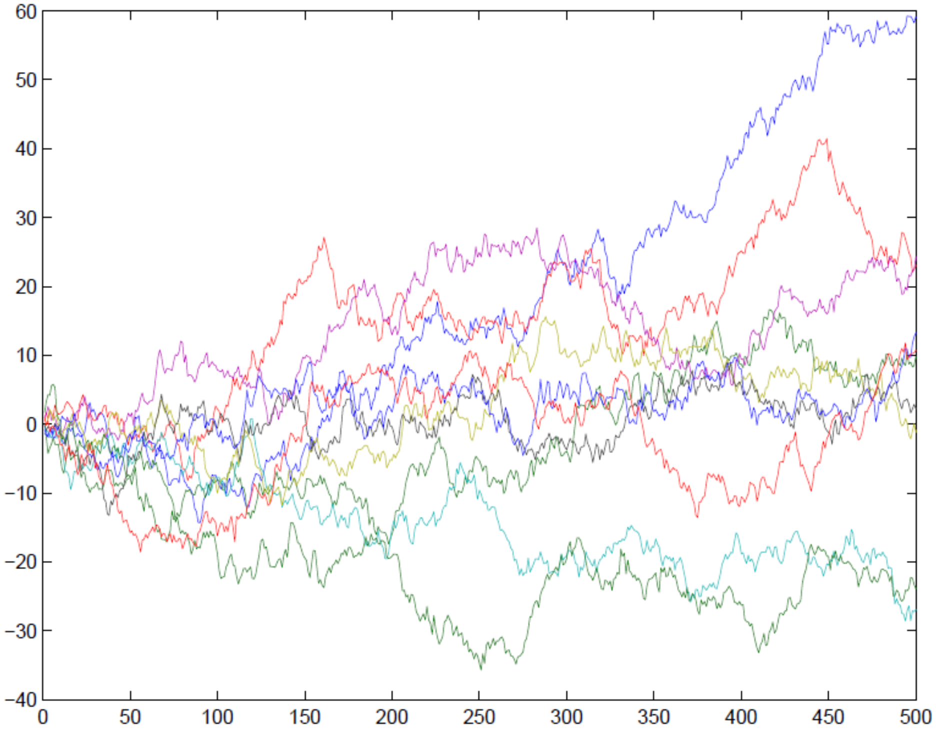


Tradicionalne časovne vrste

Naključni sprehod

- $x_{t+1} = x_t + w_t$, kjer
  - je w_t naključna vrednost
- Povprečje se spreminja
- Tudi varianca se spremenija
- Ni stacionaren

10 naključnih sprehodov:



Tradicionalne časovne vrste

Naključni sprehod

- $x_{t+1} = x_t + w_t$, kjer
  - je w_t naključna vrednost
- Povprečje se spreminja
- Tudi varianca se spremenija
- Ni stacionaren

Diferenciacija

- Odvod naključni sprehod
  - $\Delta x_{t+1} = x_{t+1} - x_t = w_t$
- Ker je w_t povsem naključna vrednost
  - je Δx_{t+1} stacionaren!

Ocena trenda – tradicionalna regresija

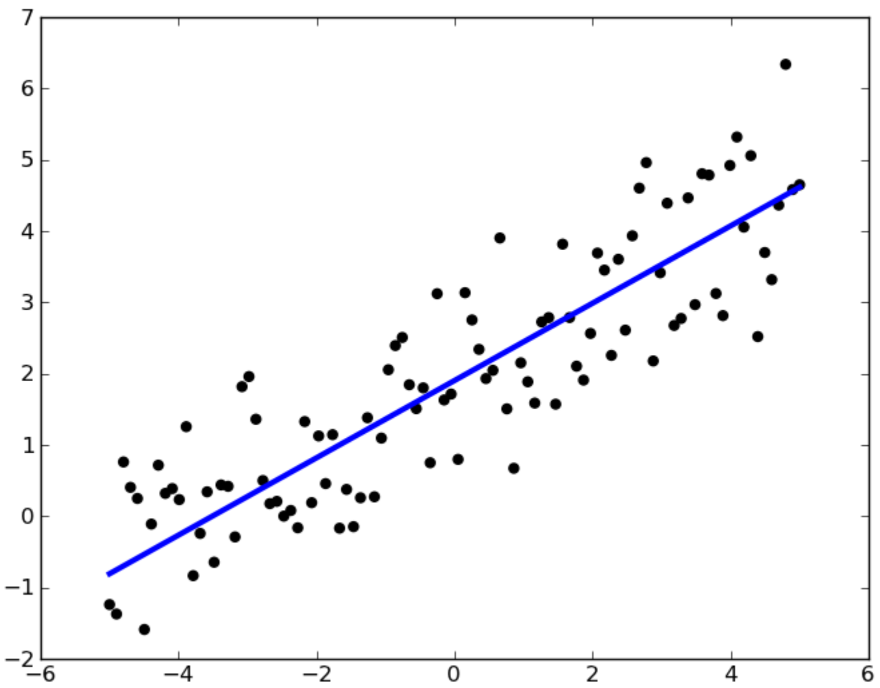
- Definicija
  - Korelacija med signalom $X = \{x_t\}$ in njegovo zakasnjeno kopijo $X_h = \{x_{t+h}\}$
- Naivni pristop k napovedovanju:
  - Poiskati najprimernejši h
  - $x_{t+1} = x_{t+h+1}$
- Takšen pristop lahko uporabimo zgolj nad stacionarno časovno vrsto.

Navadna linearna regresija

- Ocena dolgoročnega trenda
- Minimizacija napake
  - Differencialne enačbe
- Metoda najmanjših kvadratov

  (-) Poudari outlierje

  (+)Enostavna reševanje



Metoda najmanjših kvadratov

- Centriranje podatkov
  - Črta gre skozi koordinatno izhodišče

$$
\begin{aligned}
y_i &= b_0 + b_1 x_i \\
\overline{y} &= b_0 + b_1 \overline{x} \\
y_i - \overline{y} &= 0 + b_1 (x_i - \overline{x})
\end{aligned}
$$

- Splošni model

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k} + \varepsilon_i$$

  - k koeficientov za k parametrov
  - in napaka ε_i
  - V matrični obliki:

$$y_i = [x_{i,1},\ x_{i,2},\ \ldots,\ x_{i,k}] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon_i$$

$$y_i = \underbrace{x_i^T}_{(1\times k)} \underbrace{\beta}_{(k\times 1)} + \varepsilon_i$$

Metoda najmanjših kvadratov

- Če imamo več meritev, lahko izdelamo matriko
  - b predstavlja oceno dejanske vrednosti

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Generalizirano

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

- Velikosti matrik:

$\mathbf{y}$: $n \times 1$

$\mathbf{X}$: $n \times k$

$\mathbf{b}$: $k \times 1$

$\mathbf{e}$: $n \times 1$

Metoda najmanjših kvadratov

- Minimizacija kvadratov napak
- Rešitev: $$\frac{f(\mathbf{b})}{\partial \mathbf{b}} = 0$$

$$\begin{aligned} f(\mathbf{b}) &= \mathbf{e}^T\mathbf{e} \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{b} + \mathbf{b}\mathbf{X}^T\mathbf{X}\mathbf{b} \end{aligned}$$

- Po nekaj napora lahko z diferencialnimi enačbami ugotovimo
  - Ta formula je biblija!

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- Inverzna matrika:

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Metoda najmanjših kvadratov - Primer

- Ne pozabimo: $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$
- Vhod:

$$x_{1,\text{original}} = [1, \quad 3, \quad 4, \quad 7, \quad 9, \quad 9] \qquad x_1 = [-4.5,\ -2.5,\ -1.5,\ 1.5,\ 3.5,\ 3.5]$$

$$x_{2,\text{original}} = [9, \quad 9, \quad 6, \quad 3, \quad 1, \quad 2] \qquad x_2 = [4,\ 4,\ 1,\ -2,\ -4,\ -3]$$

$$y_{\text{original}} = [3, \quad 5, \quad 6, \quad 8, \quad 7, \quad 10] \qquad y = [-3.5,\ -1.5,\ -0.5,\ 1.5,\ 0.5,\ 3.5]$$

$$\mathbf{X} = \begin{bmatrix} -4.5 & 4 \\ -2.5 & 4 \\ -1.5 & 1 \\ 1.5 & -2 \\ 3.5 & -4 \\ 3.5 & -3 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} -3.5 \\ -1.5 \\ -0.5 \\ 1.5 \\ 0.5 \\ 3.5 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 55.5 & -57.0 \\ -57.0 & 62 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 36.5 \\ -36.0 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X}^{-1} = \frac{\begin{bmatrix} 62 & 57.0 \\ 57.0 & 55.5 \end{bmatrix}}{192} \qquad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 36.5 \\ -36.0 \end{bmatrix}$$

Inverzna matrika

- Ni enostavno!

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & I \end{bmatrix}^{\mathrm{T}} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} A & D & G \\ B & E & H \\ C & F & I \end{bmatrix}$$

- Bločna inverzija

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix}$$

Metoda najmanjših kvadratov - Primer

- Rezultat b1 = 1.01 in b2 = 0.43?