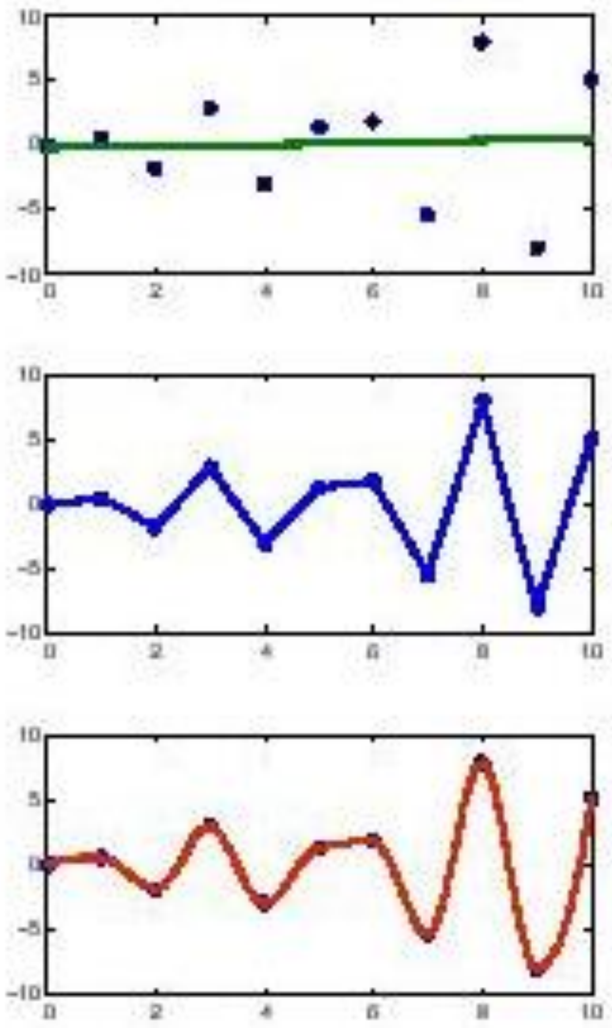
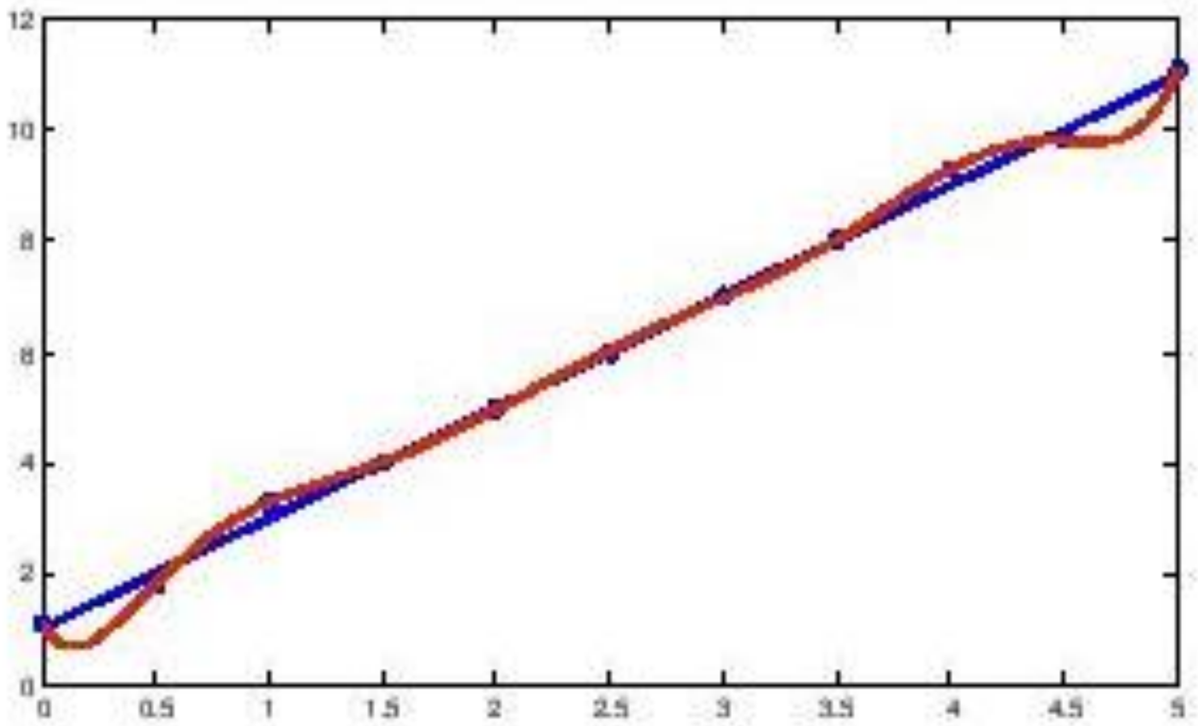
Regresija ir interpoliavimas



Aproksimuojančios kreivės

1. Regresija (mažiausių kvadratų metodas)
2. Tiesinis interpoliavimas
3. Interpoliavimas splainais

Interpoliavimas daugianariais netinka:



Tiesinis dėsnis: $y = 2x + 1$.

Duomenys su triukšmu.

Interpoliavimas daugianariais netinka!

Duomenų aproksimavimas

Tikslas: nustatyti funkciją y=f(x) (parinkti leidžiamų funkcijų klasę)

Interpoliavimas

Taškai $\{(x_k, y_k) \quad k = 1, \ldots, N\}$ žinomi tiksliai (5 reikšminiai skaitmenys ir daugiau)

Aproksimavimas mažiausių kvadratų metodu

Tikslumas 2-3 reikšminiai skaitmenys $\Rightarrow$ yra eksperimento paklaida ir realiai

$$f(x_k) = y_k + e_k, \quad e_k \text{ – matavimo paklaida.}$$

Netikslių duomenų aproksimavimas

Reikšmių lentelė
$(x_i, y_i),\ \ i = 0, 1, \cdots N$:

x_0	x_1	$\cdots$	x_N
y_0	y_1	$\cdots$	y_N

Tikslas: nustatyti funkciją $y = f(x)$ (parinkti leidžiamų funkcijų klasę).

- Egzistuoja eksperimento paklaida ir realiai

$$f(x_k) = y_k + e_k, \quad e_k - \text{matavimo paklaida.}$$

- Kaip rasti geriausią artinį (pvz., daugianarį)

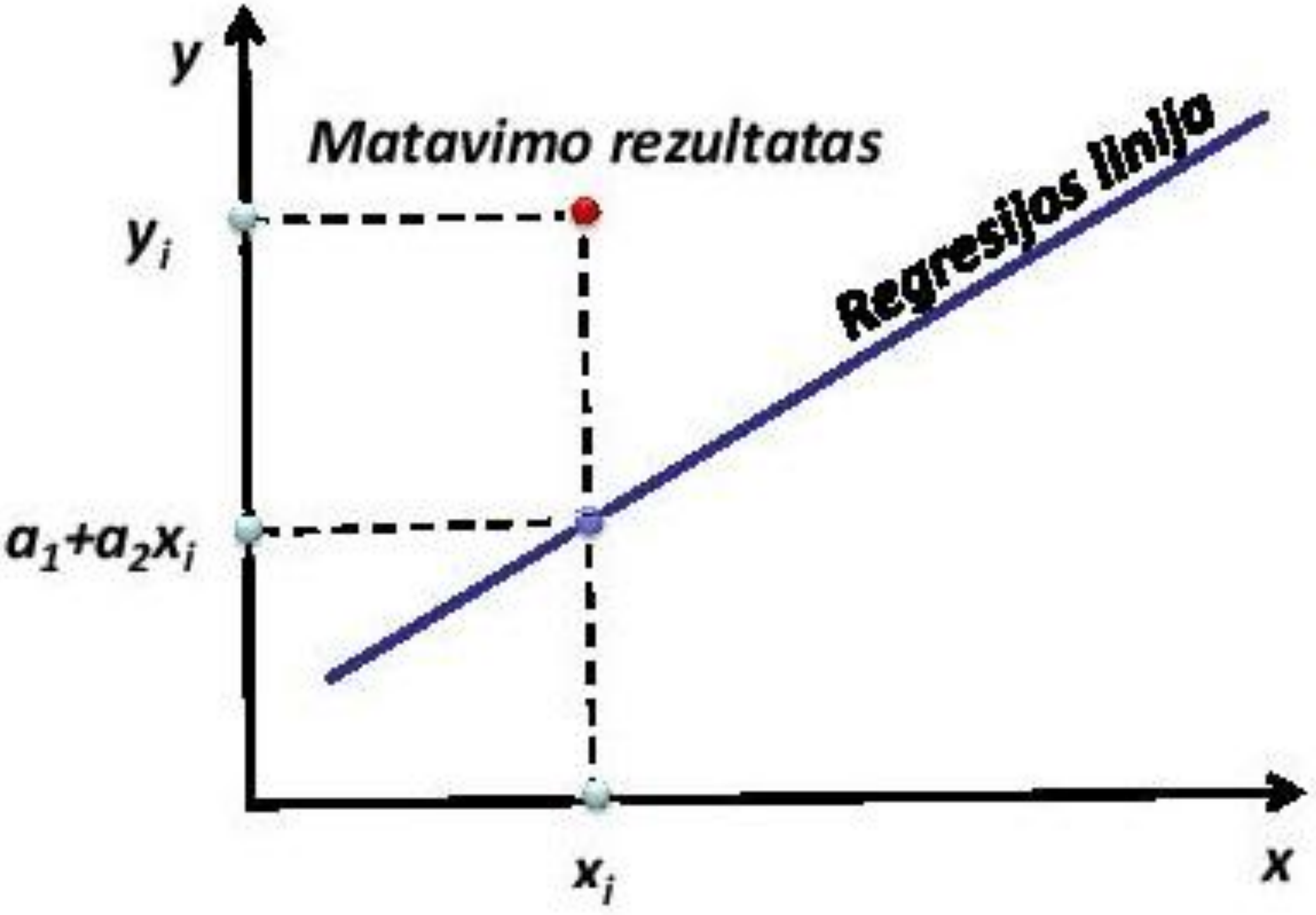
$$f(x) = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0,$$

einantį arti, bet ne visada per visus taškus?

- Turime analizuoti paklaidas (netiktį)

$$e_k = f(x_k) - y_k, \quad k = 1, \cdots, N.$$

Regresija ir matavimo paklaida



Paklaidos normos

1. **Maksimumo** (jei yra vienas blogas taškas, tai jis ir nustato paklaidos reikšmę):

$$E_\infty(f) = \max_{1 \leqslant k \leqslant N} |f(x_k) - y_k|.$$

2. **Vidurkinė** (suvidurkinta paklaida, dažnai naudojama dėl savo paprastumo):

$$E_1(f) = \frac{1}{N} \sum_{k=1}^{N} |f(x_k) - y_k|.$$

3. **Kvadratinė** (dažnai naudojama statistikoje):

$$E_2(f) = \left(\frac{1}{N} \sum_{k=1}^{N} |f(x_k) - y_k|^2 \right)^{\frac{1}{2}}.$$

Pavyzdys: paklaidų analizė

Palyginsime paklaidas tiesiniam artiniui $y = f(x) = 8,6 - 1,6x$, kai duoti taškai (x_i, y_i):

$$E_\infty(f) = \max_{1\leqslant k\leqslant N} |f(x_k) - y_k| = 0,8.$$

$$E_1(f) = \frac{1}{N}\sum_{k=1}^{N} |f(x_k) - y_k|$$

$$= \frac{1}{8}\cdot 2,6 = 0,325.$$

$$E_2(f) = \left(\frac{1}{N}\sum_{k=1}^{N} |f(x_k) - y_k|^2\right)^{\frac{1}{2}}$$

$$= \left(\frac{1,4}{8}\right)^{\frac{1}{2}} \approx 0,41833.$$

x_i	y_i	$f(x_i)$	$\lvert e_i\rvert$	$\lvert e_i\rvert^2$
-1	10	10,2	0,2	0,04
0	9	8,6	0,4	0,16
1	7	7,0	0,0	0,00
2	5	5,4	0,4	0,16
3	4	3,8	0,2	0,04
4	3	2,2	0,8	0,64
5	0	0,6	0,6	0,36
6	-1	-1	0,0	0,00
$\sum$			2,6	1,40

Paklaidų analizė

- Geriausia linija gaunama minimizuojant vieną iš paklaidų (1) - (3) $\Rightarrow$ yra trys geriausios linijos.
- Tradiciškai renkama $E_2(f)$, nes ją lengviau minimizuoti.

Tiesiniai modeliai

Tegul žinomi taškai $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$.

Tiesinis modelis

$$y_i = a_1\varphi_1(x_i) + \cdots + a_m\varphi_m(x_i) + e = \sum_{j=1}^{m} a_j\varphi_j(x_i).$$

- $\varphi_1(x), \cdots \varphi_m(x)$ - duotosios funkcijos.
- Koeficientai $a_1, \cdots, a_m$ - nežinomi parametrai.
- Tiesinė priklausomybė pagal a_j,
  bet $\varphi_j(x)$ dažniausiai netiesinės funkcijos.

Pavyzdžiai

$y \approx a_1x + a_2x^2 + a_3x^3, \quad y \approx a_0x^2 + a_1 \sin x$- tiesiniai modeliai;
$y \approx a_1e^{a_2x}$ - netiesinis modelis.

Tiesiniai modeliai

Bendroji lygtis matriciniu pavidalu

$$y = \Phi a + e, \text{ arba } y \approx \Phi a,$$

čia

$$\Phi = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_m(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_m(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_1(x_N) & \varphi_2(x_N) & \cdots & \varphi_m(x_N) \end{pmatrix}$$

Netiktis

$$e = y - \Phi a$$

$y = (y_1, \cdots, y_N)^T$ – stebėjimų vektorius;

$a = (a_1, \cdots, a_m)^T$ – regresijos koeficientai;

$e = (e_1, \cdots, e_N)^T$ – paklaidos.

Mažiausių kvadratų metodas

Parametrus $a_1, \cdots, a_m$ parinksime taip, kad netiktis

$$e = y - \Phi a$$

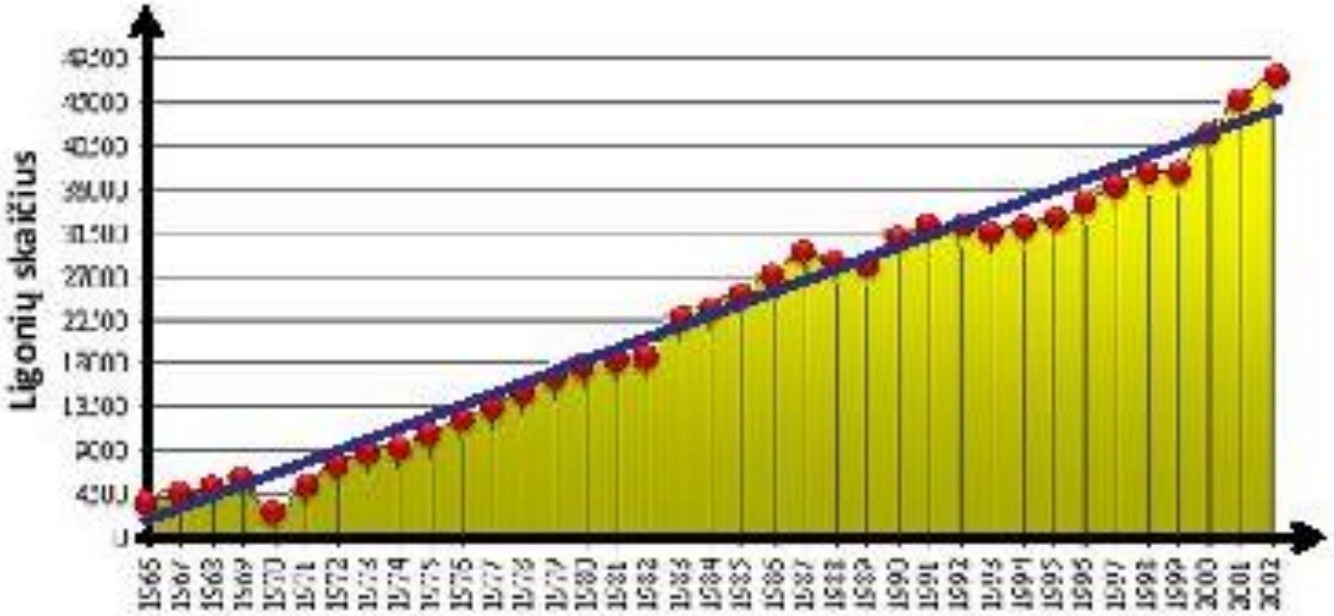
būtu mažiausia vienoje iš normų.

Mažiausių kvadratų metodas

$$\min_{a_j} \Big(\sum_{i=1}^{N} |y_i - \sum_{j=0}^{m} a_j \varphi_j(x_i)|^2 \Big).$$

Geometrinė interpretacija

- Minimizuojami vertikalieji atstumai nuo duomenų taškų iki regresijos kreivės.
- Visos klaidos yra tik matavimo paklaidos (x_i- be paklaidos).



Diagramoje - Lietuvos gydymo įstaigose užregistruoti diagnozuoti diabeto atvejai.

Normaliosios lygtys

- Uždavinys vektoriniu pavidalu:

$$\begin{aligned} S_r(a) =& \|e\|^2 = \|y - \Phi a\|^2 = (y - \Phi a)^T (y - \Phi a) \\ =& y^T y - y^T \Phi a - a^T \Phi^T y + a^T \Phi^T \Phi a \\ =& y^T y - 2a^T \Phi^T y + a^T \Phi^T \Phi a \end{aligned}$$

- Tikslas: minimizuoti pagal a

$$\Rightarrow \ (S_r(a))'_a = 0 \ \Rightarrow \ -2\Phi^T y + 2\Phi^T \Phi a = 0$$

- TLS

$$\boxed{\Phi^T \Phi a = \Phi^T y} \qquad \Rightarrow \ a.$$

- Randame vektorių a (pvz., Choleckio metodu).

Tiesinė regresija $f(x) = a_1 + a_2 x$

Tegul žinomi taškai $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$.

Aproksimavimas tiese:

$$\begin{aligned} f(x) &= a_1 + a_2 x \\ y_i &\approx a_1 + a_2 x_i. \end{aligned}$$

$$\varphi_1(x) = 1, \ \varphi_2(x) = x.$$

Duomenų matrica

$$\Phi = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) \\ \vdots & \vdots \\ \varphi_1(x_N) & \varphi_2(x_N) \end{pmatrix} \Rightarrow \Phi = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}.$$

1 pavyzdys: aproksimavimas tiese I

Taškai:$(1; 1,00), (2; 1,50), (3; 0,75), (4; 1,25).$

$$
\begin{aligned}
f(x) =& a_1 + a_2 x \\
y_i \approx& a_1 + a_2 x_i.
\end{aligned}
$$

$$\varphi_1(x) = 1,\ \varphi_2(x) = x.$$

Duomenų matrica

$$\Phi = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) \\ \vdots & \vdots \\ \varphi_1(x_N) & \varphi_2(x_N) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}.$$

1 pavyzdys: aproksimavimas tiese II

$$\Phi^T \Phi a = \Phi^T y$$

$$f(x) = a_1 + a_2 x$$
$$y_i \approx a_1 + a_2 x_i.$$

$$\Phi^T \Phi = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix};$$

$$\Phi^T y = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1,00 \\ 1,50 \\ 0,75 \\ 1,25 \end{pmatrix} = \begin{pmatrix} 4,5 \\ 11,25 \end{pmatrix};$$

$$\Phi^T \Phi a = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 4,5 \\ 11,25 \end{pmatrix} = \Phi^T y$$

$$\Rightarrow \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1,125 \\ 0 \end{pmatrix}$$

$\Rightarrow$ y=1,125 geriausia linija.

Kvadratinė regresija $f(x) = a_1 + a_2 x + a_3 x^2$

Tegul žinomi taškai $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$.

Aproksimavimas parabole:

$$\begin{aligned} f(x) =& a_1 + a_2 x + a_3 x^2 \\ y_i \approx& a_1 + a_2 x_i + a_3 x_i^2. \end{aligned}$$

$$\varphi_1(x) = 1,\ \varphi_2(x) = x,\ \varphi_3(x) = x^2.$$

Duomenų matrica

$$\Phi = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \varphi_3(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \varphi_3(x_2) \\ \vdots & \vdots & \vdots \\ \varphi_1(x_N) & \varphi_2(x_N) & \varphi_3(x_N) \end{pmatrix} \Rightarrow \Phi = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{pmatrix}.$$

2 pavyzdys: aproksimavimas parabole I

Taškai: $(1;2), (2;7), (3;9), (4;6)$.

$$
\begin{aligned}
f(x) &= a_1 + a_2 x + a_3 x^2 \\
y_i &\approx a_1 + a_2 x_i + a_3 x_i^2 .
\end{aligned}
$$

$$\varphi_1(x) = 1,\ \varphi_2(x) = x,\ \varphi_3(x) = x^2.$$

Duomenų matrica

$$\Phi = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \varphi_3(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \varphi_3(x_2) \\ \vdots & \vdots & \vdots \\ \varphi_1(x_N) & \varphi_2(x_N) & \varphi_3(x_N) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{pmatrix}.$$

2 pavyzdys: aproksimavimas parabole II

$$\Phi^T \Phi a = \Phi^T y$$

$$f(x) = a_1 + a_2 x + a_3 x^2$$
$$y_i \approx a_1 + a_2 x_i + a_3 x_i^2.$$

$$\Phi^T\Phi = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{pmatrix} = \begin{pmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{pmatrix};$$

$$\Phi^T y = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \end{pmatrix} \begin{pmatrix} 2 \\ 7 \\ 9 \\ 6 \end{pmatrix} = \begin{pmatrix} 24 \\ 67 \\ 207 \end{pmatrix};$$

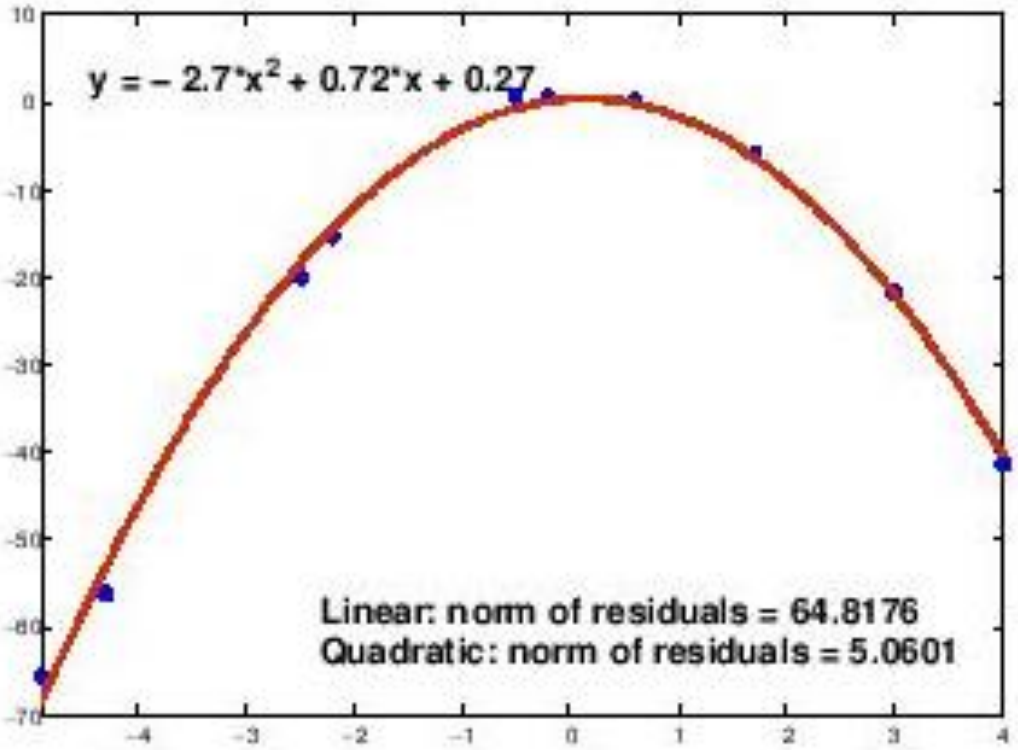
$$\Phi^T\Phi a = \begin{pmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 24 \\ 67 \\ 207 \end{pmatrix} = \Phi^T y$$

$$\Rightarrow \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} -7,5 \\ 11,4 \\ -2,0 \end{pmatrix} \Rightarrow \mathbf{y = -7,5 + 11,4x - 2x^2}$$

geriausia linija.

Pavyzdys

MATLAB sprendimas



x_i	y_i
$-4,9$	$-65,4$
$-2,5$	$-20,1$
$-2,2$	$-15,4$
$-0,5$	$0,6$
$-0,2$	$0,5$
$0,6$	$0,2$
$1,7$	$-6,0$
$3,0$	$-21,8$
$4,0$	$-41,3$
$4,3$	$-56,1$

Parabolė aproksimuoja geriau