Vedecko-technické výpočty

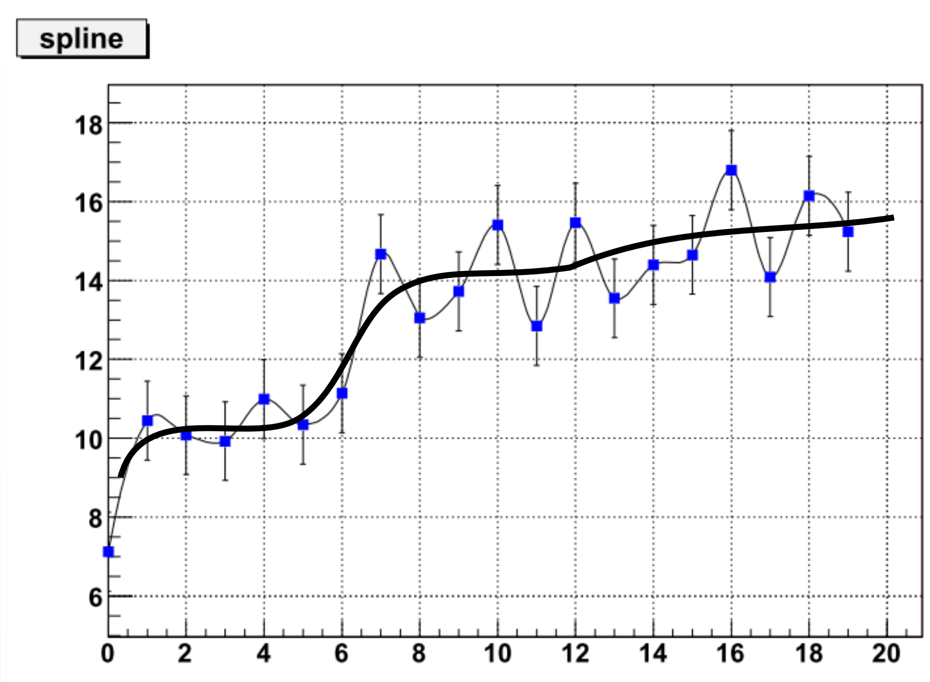
Regresia

Prehľad

- Motivácia a čo to je regresia?
- Lineárna regresia
  - Odvodenie lineárnej regresie
  - Lineárna regresia cez nulu
- Nelineárna regresia
  - Mocninová funkcia
  - Exponenciálna funkcia
  - Funkcia saturovaného rastu
  - Polynomiálna funkcia

Motivácia

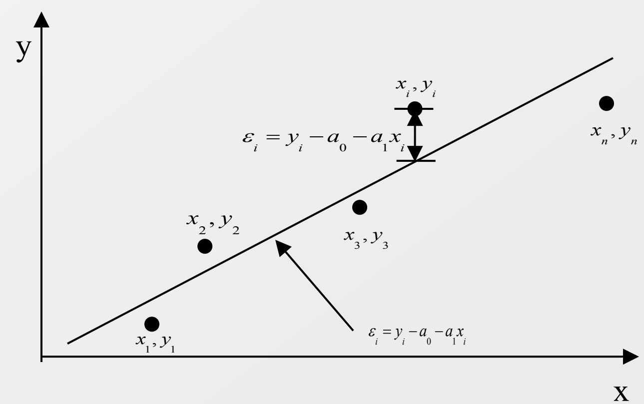
Pri veľkom počte meraní je nezriedka chybným krokom, použitie interpolácie polynómom vysokého stupňa.



História

- Za zakladateľa štatistickej matematiky sa pokladá Karl (Carl) Pearson (1857–1936).
- V roku 1896 navrhnutý spôsob lineárnej extrapolácie v Philosophical Transactions of the Royal Society of London

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

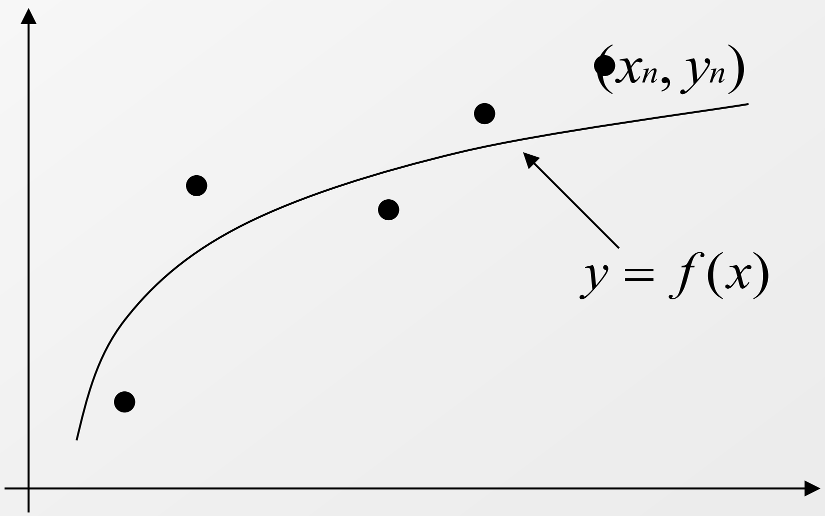


Regresia (fit, extrapolácia)

Majme zadaných n bodov $(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)$ a hľadáme funkciu f(x), ktorá najlepšie vystihuje dané dáta. Vo všeobecnosti sa hľadá funkcia ktorá má minimum druhých mocnín odchýlok dát, reziduí, od danej funkcie.

Odchýlky bodov od funkcie

$$\varepsilon_i = y_i - f(x_i)$$



Najjednoduchší možný prípad.

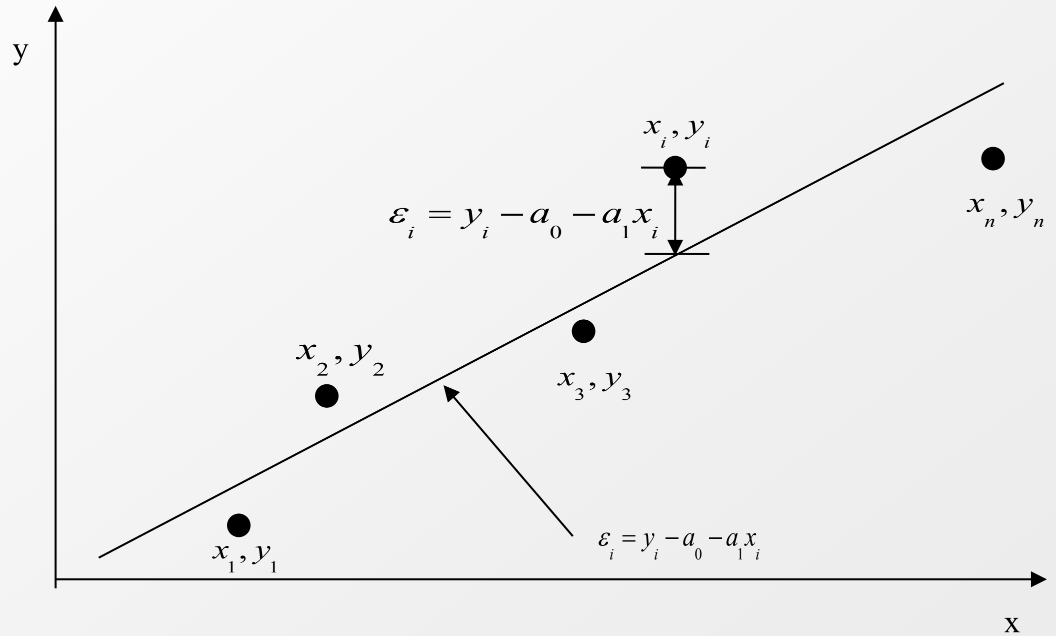
LINEÁRNA REGRESIA

Kritérium A: odchýlky hodnôt of funkcie

Majme zadaných n bodov (x_1,y_1), (x_2,y_2),…,(x_n,y_n) a hľadáme funkciu $f(x)=a + bx$, ktorá najlepšie vystihuje dané dáta.

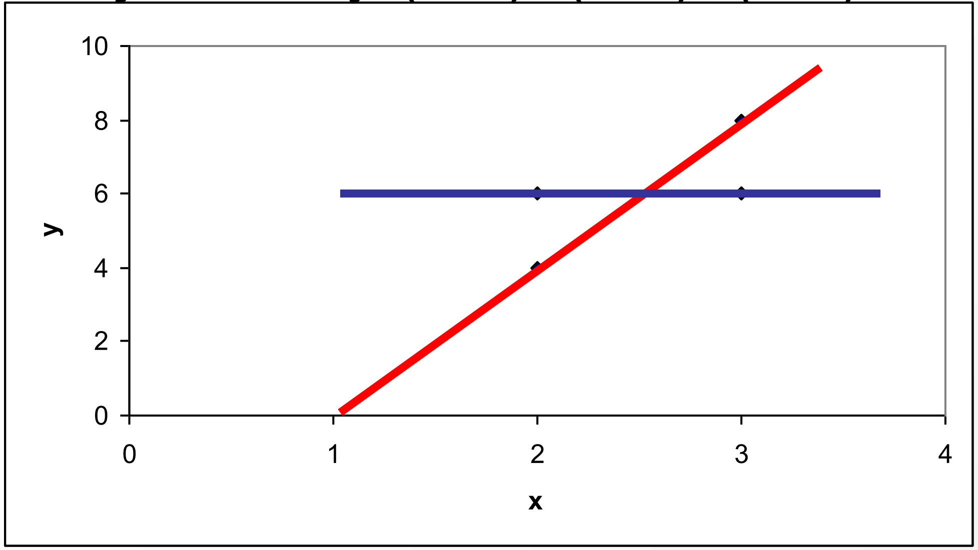
Vezmime si ako kritériu kvality fitu $\sum_{i=1}^{n} \varepsilon_i$

Pričom $\varepsilon_i = y_i - (a_0 + a_1 x_i)$



Kritérium A Test

Majme body (2,4), (3,6), (2,6) and (3,8) otestujme kritérium A.



x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

Otestujme dve funkcie
$f(x)=4x - 4$ a $f(x) = 6$

x	y	y_{cal}	$\varepsilon = y - y_{cal}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
y = 4x – 4			$\sum_{i=1}^{4} \varepsilon_i = 0$

Napriek prakticky nulovej sume odchýlok, máme nejednoznačný fit

Failed

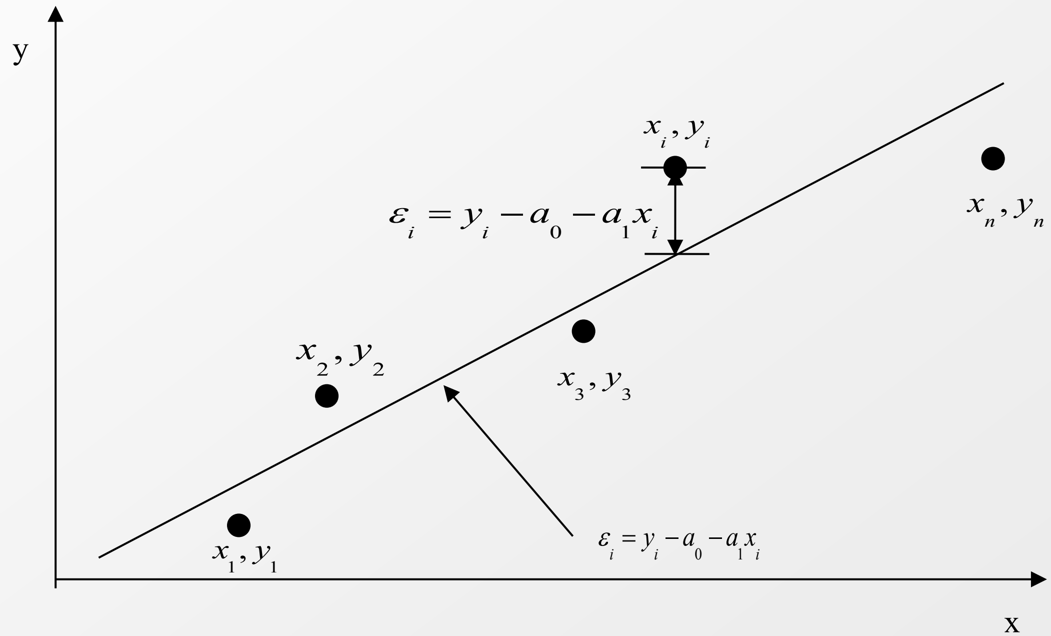
x	y	y_{cal}	$\varepsilon = y - y_{cal}$
2.0	4.0	6.0	-2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
y=6			$\sum_{i=1}^{4} \varepsilon_i = 0$

Kritérium B: absolútna hodnota odchýlky

Majme zadaných n bodov $(x_1,y_1), (x_2,y_2),\dots,(x_n,y_n)$ a hľadáme funkciu $f(x)=a + bx$, ktorá najlepšie vystihuje dané dáta.

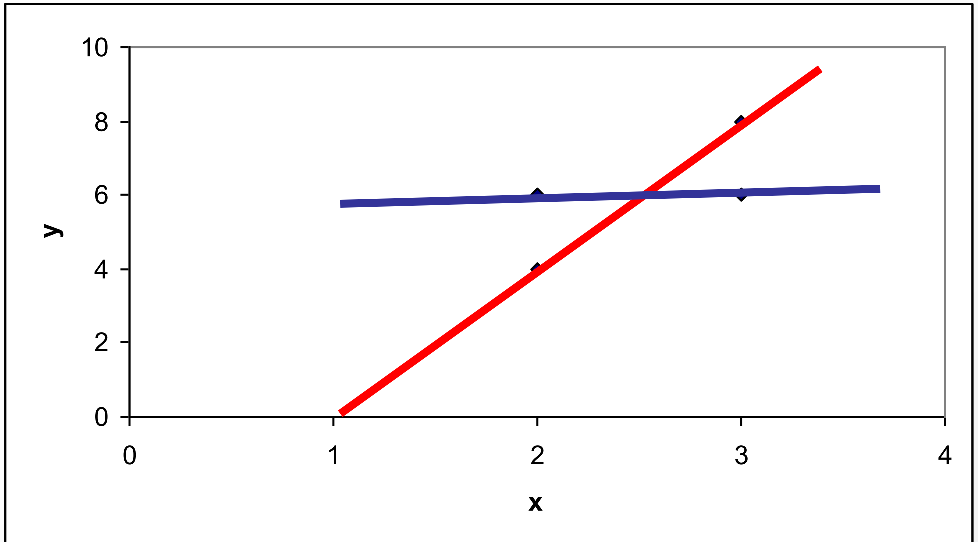
Vezmime si ako kritériu kvality fitu $\sum_{i=1}^{n} |\varepsilon_i|$

Pričom $\varepsilon_i = y_i - (a_0 + a_1 x_i)$



Kritérium B Test

Majme body (2,4), (3,6), (2,6) and (3,8) otestujme kritérium A.



x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

Otestujme dve funkcie

f(x)=4x – 4 a *f(x)* = *6*

x	y	y_{cal}	$\|\varepsilon\|=\|y-y_{cal}\|$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
y = 4x – 4			$\sum_{i=1}^{4}\left\|\varepsilon_i\right\| = 4$

Máme nejednoznačný fit

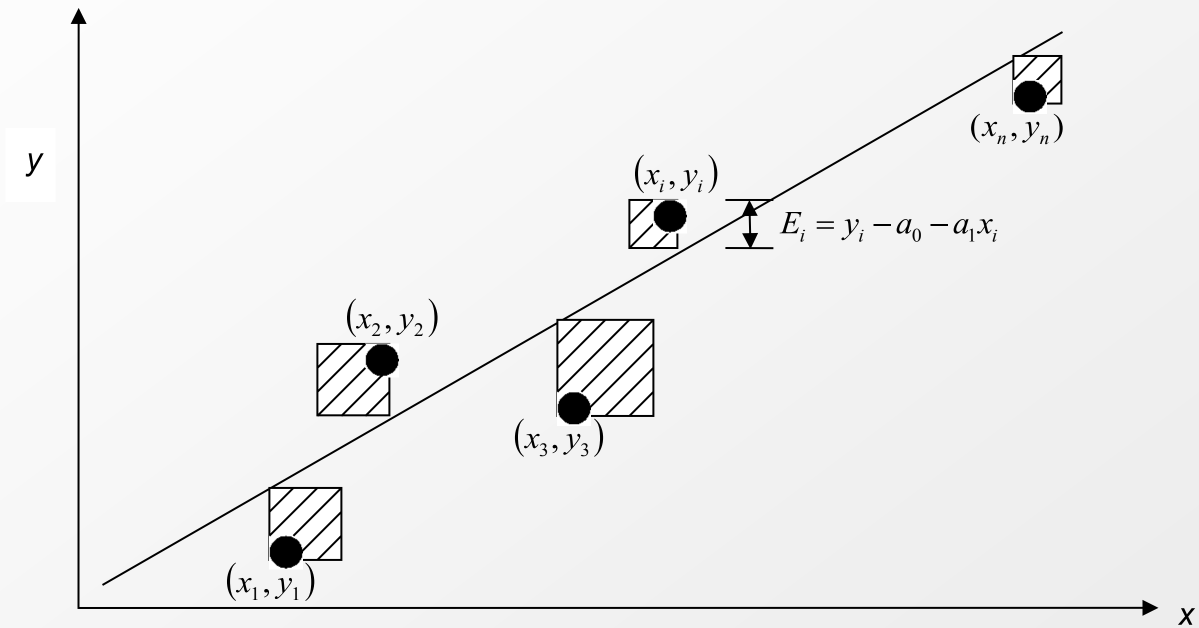
x	y	y_{cal}	$\|\varepsilon\|=\|y-y_{cal}\|$
2.0	4.0	6.0	2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
y=6			$\sum_{i=1}^{4}\left\|\varepsilon_i\right\| = 4$

Failed

Kritérium C: kvadrát hodnoty odchýlky

Majme zadaných n bodov $(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)$ a hľadáme funkciu $f(x)=a_0 + a_1x$, ktorá najlepšie vystihuje dané dáta.

Vezmime si ako kritériu kvality fit $S_r = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$



Hľadanie koeficientov

Hľadanie koeficientov pre funkciu $f(x)=a_0 + a_1x$ sa nám zmení na problém minimalizácie funkci

$$S_r = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$

Vieme, že pri hľadaní lokálnych miním musí byť prvá derivácia nulová

$$\frac{\partial S_r}{\partial a_0} = 2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)\left(-1\right) = 0$$

$$\frac{\partial S_r}{\partial a_1} = 2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)\left(-x_i\right) = 0$$

Čím dostávame

$$\sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = \sum_{i=1}^{n} y_i \qquad \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^{\,2} = \sum_{i=1}^{n} y_i x_i$$

Hľadanie koeficientu a$_1$

$$\frac{\partial S_r}{\partial a_0} = 2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i)(-1) = 0$$

$$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = 0$$

$$n a_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a_0 = \frac{\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_i}{n}$$

$$a_0 = \frac{\sum_{i=1}^{n} y_i}{n} - a_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$(a_0 = \bar{y} - a_1 \bar{x})$$

$$\sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$\sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\frac{\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_i}{n} \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - a_1 \left(\sum_{i=1}^{n} x_i\right)^2 + n a_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i y_i$$

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Hľadanie koeficientu a$_0$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

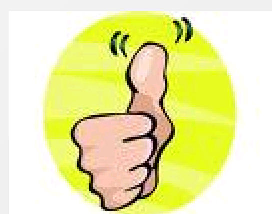
$$a_1 = \frac{\sum_{i=1}^{n} x_i y_i - a_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

$$a_0 = \frac{\sum_{i=1}^{n} y_i}{n} - a_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$a_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} x_i y_i - a_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \frac{\sum_{i=1}^{n} x_i}{n}$$

$$a_0 n \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \left(\sum_{i=1}^{n} x_i y_i - a_0 \sum_{i=1}^{n} x_i \right)$$

$$a_0 n \sum_{i=1}^{n} x_i^2 - a_0 \left(\sum_{i=1}^{n} x_i \right)^2 = \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i$$



$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

Metóda najmenších štvorcov (lineárny prípad)

- Pre koeficienty funkcie $f(x)=a_0 + a_1 x$ dostávame

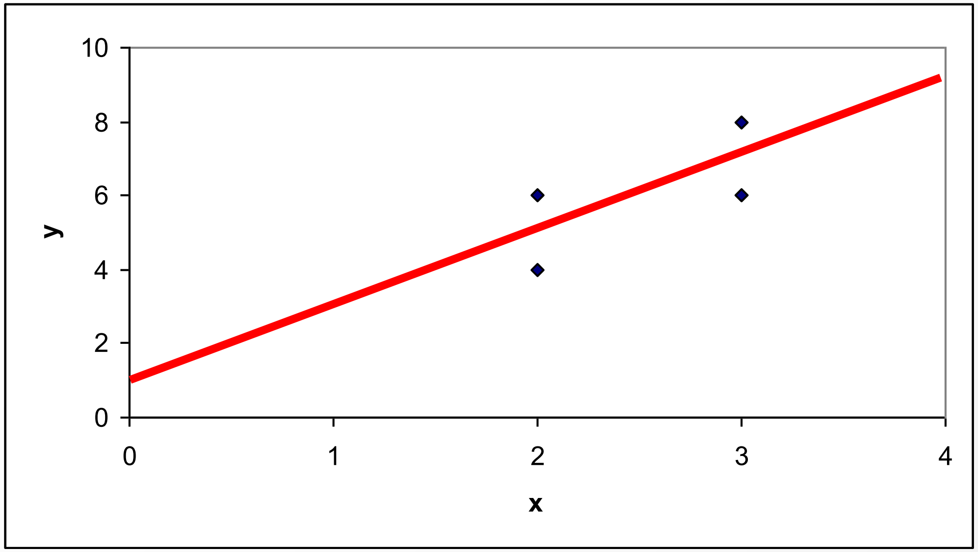
$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Oba vzťahy si netreba samozrejme pamätať, je však fajn ovládať princíp odvodenia.

Aplikácia

Majme body (2,4), (3,6), (2,6) and (3,8) otestujme kritérium A.



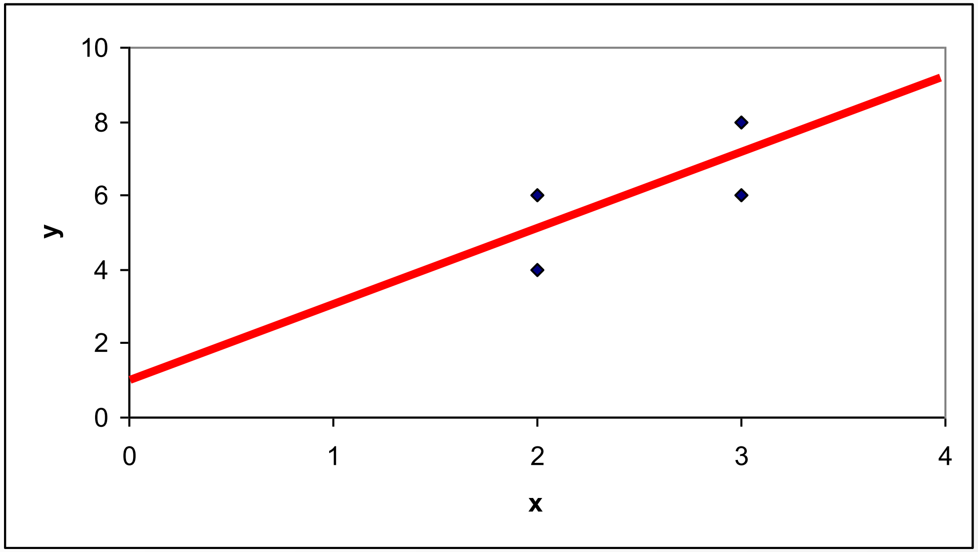
x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Koeficienty a_0 = 2 a a_1 = 1 $\Rightarrow$ Hľadaná funkcia je f(x)=2x+1

Aplikácia

Majme body (2,4), (3,6), (2,6) and (3,8) otestujme kritérium A.



x	y	y_{cal}	$\|\varepsilon\|=\|y-y_{cal}\|$
2.0	4.0	3.0	1.0
3.0	6.0	7.0	1.0
2.0	6.0	3.0	1.0
3.0	8.0	7.0	1.0
y = 2x + 1			$\sum_{i=1}^{4}\|\varepsilon_i\| = 4 \quad \sum_{i=1}^{4}\|\varepsilon_i\|^2 = 4$

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Koeficienty a_0 = 2 a a_1 = 1 $\Rightarrow$ Hľadaná funkcia je f(x)=2x+1

ŠPECIÁLNY PRÍPAD $Y=A_1X$

Ak očakávame $a_0=0$

Zoberme si napríklad roztažnosť telies so zmenou teploty. Pri nulovej zmene, je rozťažnosť nulová, takže môžeme použiť lineárnu regresiu fixovanú bodom (0,0).

Odvodili sme si:

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

A súčastne sme získali aj:

$$a_0 = \frac{\sum_{i=1}^{n} y_i}{n} - a_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - a_1 \bar{x}$$

Hypotéza: a_1 môžem získať ako:

$$\cancel{a_1 = \frac{\bar{y}}{\bar{x}}}$$

Minimalizovali sme funkciu $f(x)=a_0 + a_1 x$ a nie $f(x)=a_1 x$!!!

Odvodenie pre prípad $a_0=0$

Minimalizujeme

$$S_r = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_1 x_i\right)^2$$

Prvá derivácia nulová

$$\frac{\partial S_r}{\partial a_1} = 2\sum_{i=1}^{n} \left(y_i - a_1 x_i\right)\left(-x_i\right) = 0$$

$$\sum_{i=1}^{n} a_1 x_i^{\,2} = \sum_{i=1}^{n} y_i x_i$$

$$a_1 \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} y_i x_i$$

$$a_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^{\,2}}$$

Upozornenie

- Prvou deriváciou sme iba získali informáciu o tom že nájdené parametre a0 a a1 sú pre funkciu reprezentujúcu lokálny extrém, ale neoverili sme že ide o minimum
- Na overenie minima by sme museli urobiť druhú deriváciu.

NELINEÁRNA REGRESIA

Najrozšírenejšie modely

1. Exponenciálny model: $(y = ae^{bx})$

2. Mocninový model: $(y = ax^{b})$

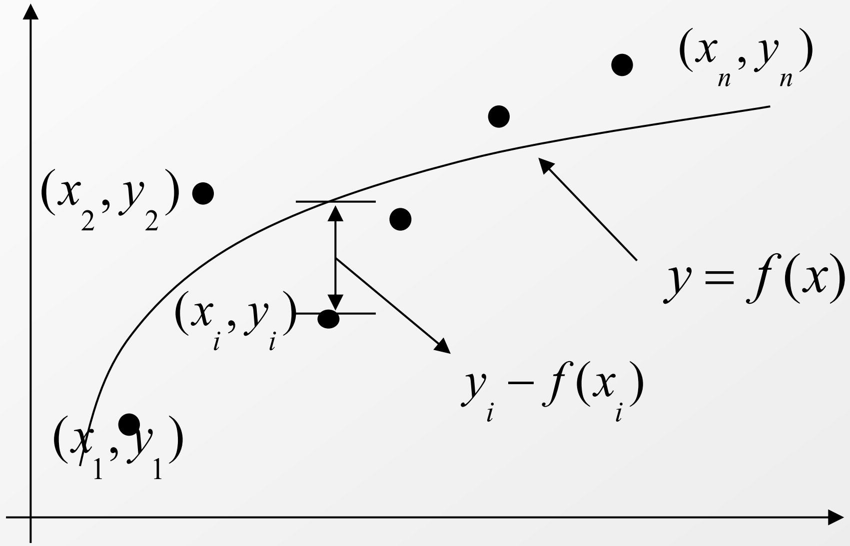
3. Model saturovaného rastu: $\left(y = \dfrac{ax}{b+x}\right)$

4. Polynómny model: $(y = a_0 + a_1 x + \ldots + a_m x^m)$

5. Regresia gaussom.

Definícia

Majme zadaných n bodov $(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)$ a hľadáme funkciu $f(x)$, ktorá najlepšie vystihuje dané dáta.



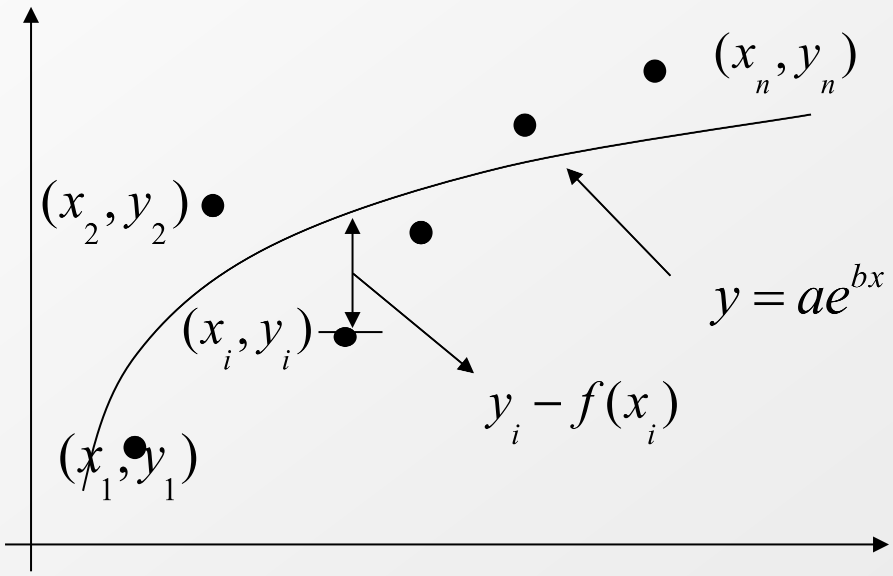
NELINEÁRNA REGRESIA EXPONENCÁLNOU FUNKCIOU

Exponenciálna regresia

Majme zadaných n bodov (x_1,y_1), (x_2,y_2),...,(x_n,y_n) a hľadáme funkciu $f(x)=ae^{bx}$, ktorá najlepšie vystihuje dané dáta.

Minimalizujme nasledujúce kritérium kvality fitu

$$S_r = \sum_{i=1}^{n} \left(y_i - ae^{bx_i} \right)^2$$



Minimalizácia

$$S_r = \sum_{i=1}^{n} \left(y_i - a e^{b x_i} \right)^2$$

Opäť využijeme fakt, že pri hľadaní lokálnych miním musí byť prvá derivácia nulová

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^{n} 2\left(y_i - a e^{b x_i} \right)\left(- e^{b x_i} \right) = 0$$

$$\frac{\partial S_r}{\partial b} = \sum_{i=1}^{n} 2\left(y_i - a e^{b x_i} \right)\left(- a x_i e^{b x_i} \right) = 0$$

Hľadáme konštantu a

Riešením prvej rovnice dostávame

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^{n} 2\left(y_i - ae^{bx_i}\right)\left(-e^{bx_i}\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i\right)\left(-e^{bx_i}\right) + \sum_{i=1}^{n} ae^{bx_i}e^{bx_i} = 0$$

$$\sum_{i=1}^{n} ae^{2bx_i} = \sum_{i=1}^{n}\left(y_i\right)\left(e^{bx_i}\right)$$

$$a = \frac{\sum_{i=1}^{n} y_i e^{bx_i}}{\sum_{i=1}^{n} e^{2bx_i}}$$

Našli sme vyjadrenie a v závislosti od b

Hľadáme konštantu b

Dosadením do druhej rovnice dostávame

$$\frac{\partial S_r}{\partial b} = \sum_{i=1}^{n} 2\left(y_i - ae^{bx_i}\right)\left(-ax_i e^{bx_i}\right) = 0$$

$$\sum_{i=1}^{n} \left(y_i - ae^{bx_i}\right)\left(-ax_i e^{bx_i}\right) = 0$$

$$\sum_{i=1}^{n} y_i x_i e^{bx_i} - \frac{\sum_{i=1}^{n} y_i e^{bx_i}}{\sum_{i=1}^{n} e^{2bx_i}} \sum_{i=1}^{n} x_i e^{2bx_i} = 0$$

Nelineárna rovnica pre parameter b – je potrebné ju riešiť numericky. Môžeme použiť niektorú numerickú metódu.

Príklad na exponenciálnu regresiu

Pri počítačovej tomografii sa sa často využíva izotop ^{99m}Tc s polčasom rozpadu cca 5 hodín. Počas jedného zo zákrokov boli namerané nasledujúce aktivity.

t(hrs)	0	1	3	5	7	9
γ	1.000	0.891	0.708	0.562	0.447	0.355

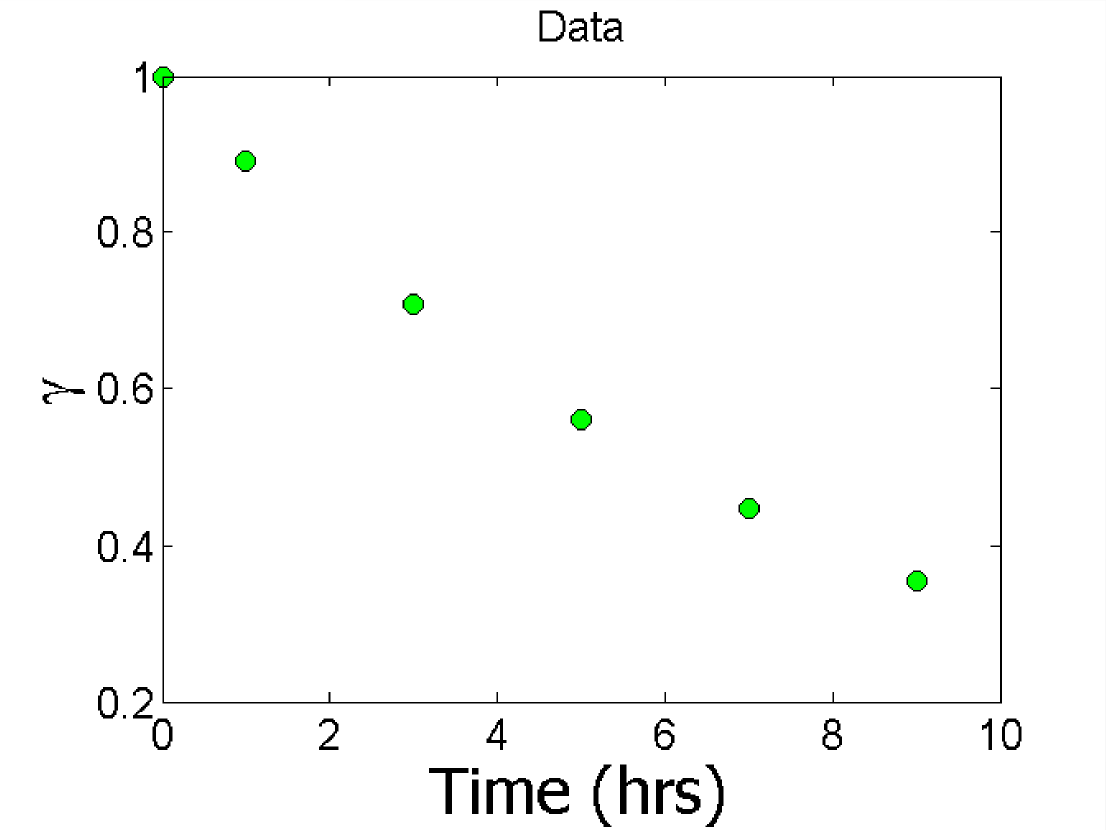
Úlohou je:

a) Získať funkciu opisujúcu danú závislosť

b) Odhadnúť dobu polpremeny pre tento izotop

c) Odhadnúť aktivitu po 24 hodinách

Voľba vhodnej funkcie



Z jadrovej fyziky vieme, že vhodná funkcia je $\gamma = Ae^{\lambda t}$

Hľadá me vhodné parametre

Máme funkciu $f(x)=ae^{bx}$ a pre jej parametre platí a a b

$$a = \frac{\sum_{i=1}^{n} y_i e^{bx_i}}{\sum_{i=1}^{n} e^{2bx_i}}$$

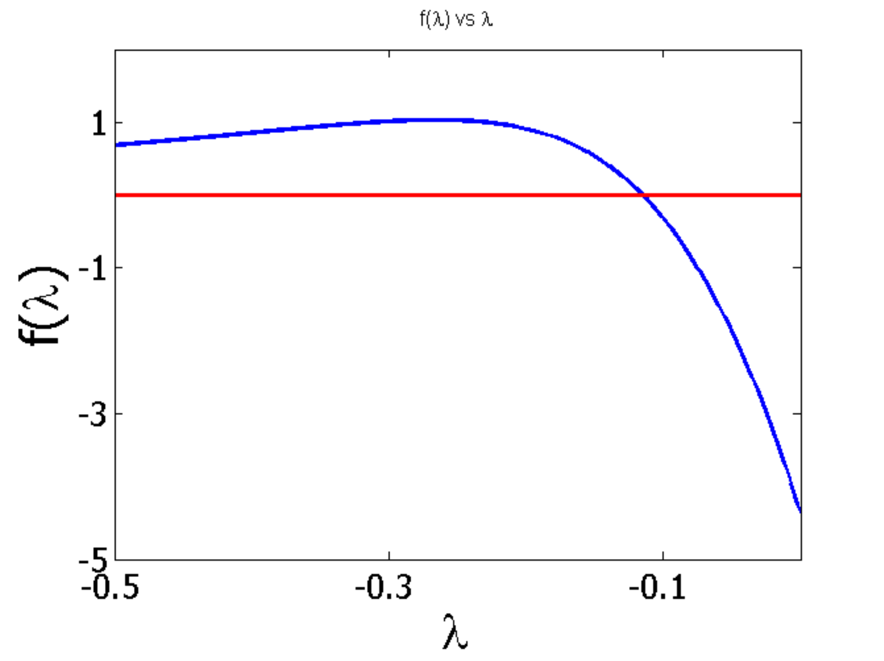
$$\sum_{i=1}^{n} y_i x_i e^{bx_i} - \frac{\sum_{i=1}^{n} y_i e^{bx_i}}{\sum_{i=1}^{n} e^{2bx_i}} \sum_{i=1}^{n} x_i e^{2bx_i} = 0$$

$$\gamma = Ae^{\lambda t}$$

$$A = \frac{\sum_{i=1}^{6} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{6} e^{2\lambda t_i}}$$

$$f(\lambda) = \sum_{i=1}^{n} \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^{n} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}} \sum_{i=1}^{n} t_i e^{2\lambda t_i} = 0$$

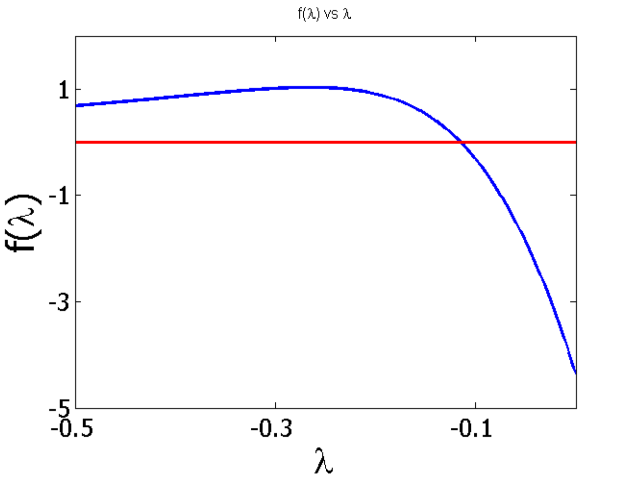
Hľadá sa parameter λ



$$f(\lambda)=\sum_{i=1}^{n}\gamma_i t_i e^{\lambda t_i}-\frac{\sum_{i=1}^{n}\gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n}e^{2\lambda t_i}}\sum_{i=1}^{n}t_i e^{2\lambda t_i}=0$$

Riešenie rovnice hľadáme numerickými metódami.
Viac o numerických metódach riešenia nelineárnych rovníc v niektorej ďalšej prednáške

Hľadá sa parameter λ

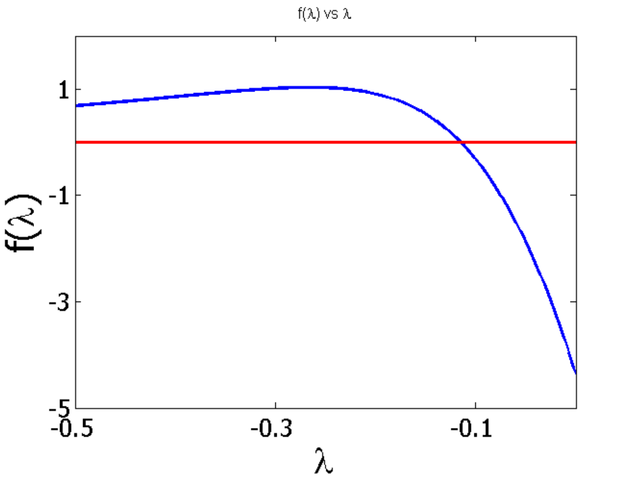


$$f(\lambda) = \sum_{i=1}^{n} \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^{n} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}} \sum_{i=1}^{n} t_i e^{2\lambda t_i} = 0$$

Odhadnime si λ = -0.120

i	t_i	γ_i	$\gamma_i t_i e^{\lambda t_i}$	$\gamma_i e^{\lambda t_i}$	$e^{2\lambda t_i}$	$t_i e^{2\lambda t_i}$
1	0	1	0.00000	1.00000	1.00000	0.00000
2	1	0.891	0.79205	0.79205	0.78663	0.78663
3	3	0.708	1.4819	0.49395	0.48675	1.4603
4	5	0.562	1.5422	0.30843	0.30119	1.5060
5	7	0.447	1.3508	0.19297	0.18637	1.3046
6	9	0.355	1.0850	0.12056	0.11533	1.0379
$\sum_{i=1}^{6}$			6.2501	2.9062	2.8763	6.0954

Hľadá sa parameter λ



$$f(\lambda)=\sum_{i=1}^{n}\gamma_i t_i e^{\lambda t_i}-\frac{\sum_{i=1}^{n}\gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n}e^{2\lambda t_i}}\sum_{i=1}^{n}t_i e^{2\lambda t_i}=0$$

Získali sme pre λ = -0.120

$$\sum_{i=1}^{6}\gamma_i t_i e^{-0.120 t_i}=6.2501$$

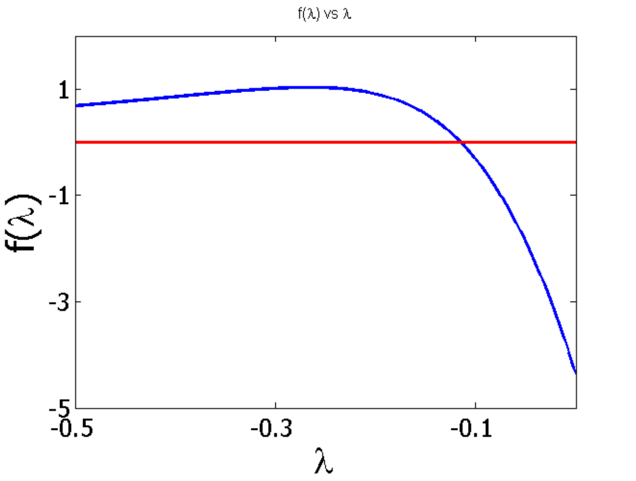
$$\sum_{i=1}^{6}\gamma_i e^{-0.120 t_i}=2.9062$$

$$\sum_{i=1}^{6}e^{2(-0.120)t_i}=2.8763$$

$$\sum_{i=1}^{6}t_i e^{2(-0.120)t_i}=6.0954$$

$$f(-0.120)=(6.2501)-\frac{2.9062}{2.8763}(6.0954)=0.091357$$

Hľadá sa parameter λ



$$f(\lambda)=\sum_{i=1}^{n}\gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^{n}\gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}}\sum_{i=1}^{n} t_i e^{2\lambda t_i} = 0$$

Rovnakým spôsobom pre λ = -0.110

$$f(-0.110) = -0.10099$$

$f(-0.120)\times f(-0.110)<0$ koreň leží v intervale (-0.120;-0.110)

ďalší odhad je hodnota pre -0.115 (v strede intervalu)

Metóda bisekcie

Po 20 interáciách získame s presnosťou 0.000008%

$$\lambda = -0.11508$$

Hľadá sa parameter A

$$A=\frac{\sum_{i=1}^{6}\gamma_i e^{\lambda t_i}}{\sum_{i=1}^{6} e^{2\lambda t_i}}=\frac{\begin{array}{c}1\times e^{-0.11508(0)}+0.891\times e^{-0.11508(1)}+0.708\times e^{-0.11508(3)}+\\ 0.562\times e^{-0.11508(5)}+0.447\times e^{-0.11508(7)}+0.355\times e^{-0.11508(9)}\end{array}}{\begin{array}{c}e^{2(-0.11508)(0)}+e^{2(-0.11508)(1)}+e^{2(-0.11508)(3)}+\\ e^{2(-0.11508)(5)}+e^{2(-0.11508)(7)}+e^{2(-0.11508)(9)}\end{array}}$$

$$=\frac{2.9373}{2.9378}=0.99983$$

Oproti predchádzajúcemu postupu, bola metóda najmenších štvorcov pre lineárny prípad triviálny prípad. Ako zjednodušiť komplikovaný problém...

LINEARIZÁCIA PROBLÉMU

Stanislav.Antalic@fmph.uniba.sk

Odvodenie linearizácie

Nezriedka je možné nelineárny problém zapísať pomocou lineárnej funkcie a aplikovať lineárnu metódu najmenších štvorcov.

$$y = ae^{bx} \quad \Rightarrow \quad \ln y = \ln a + bx$$

Zadefinujme si $z = \ln y \quad a_0 = \ln a \quad a_1 = b$

$$z = a_0 + a_1 x$$

Pozor: transformujeme zadané hodnoty a nie model!!!

Takže získané hodnoty nemusia byť úplne optimálne

$$a_1 = \frac{n\sum_{i=1}^{n} x_i z_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \Rightarrow \quad a_0 = \bar{z} - a_1 \bar{x} \quad \Rightarrow \quad \begin{aligned} b &= a_1 \\ a &= e^{a_0} \end{aligned}$$

POWER MODEL

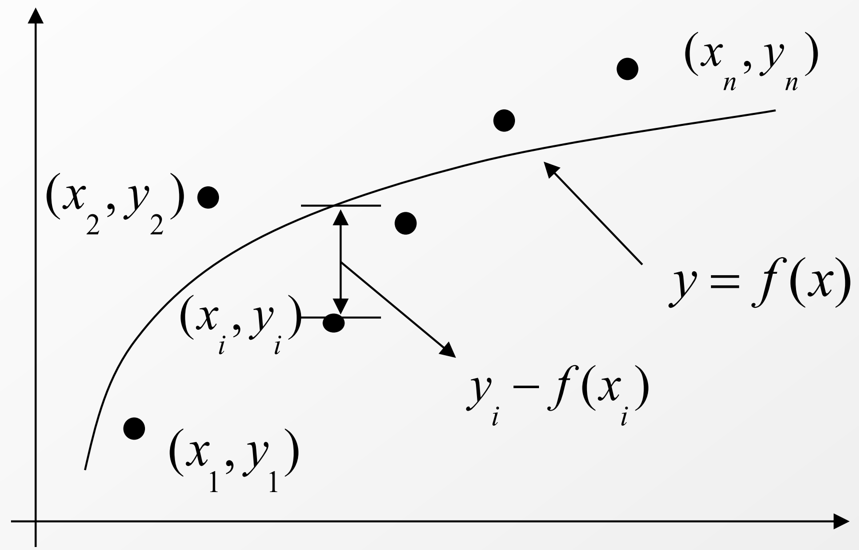
Odvodenie minimalizácie

Majme zadaných n bodov $(x_1,y_1), (x_2,y_2),\dots,(x_n,y_n)$ a hľadáme funkciu $f(x)=ax^b$, ktorá najlepšie vystihuje dané dáta.

Príkladom môže byť napr. keď z hadice strieka pod tlakom voda

$$prietok = a \cdot tlak^b$$

$$F = a \cdot p^b$$



Minimalizujeme

$$S_r = \sum_{i=1}^{n} \left(y_i - ax_i^b \right)^2$$

Získané hodnoty

Predpovedané hodnoty

Minimalizácia

$$S_r = \sum_{i=1}^{n} \left(y_i - a x_i^{\,b}\right)^2$$

Využívame

$$\frac{d}{dx}\left(u^2\right) = 2u \frac{du}{dx} \qquad \frac{d}{dx}\left(a^x\right) = a^x \ln(a)$$

Opäť využijeme fakt, že pri hľadaní lokálnych miním musí byť prvá derivácia nulová

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^{n} 2\left(y_i - a x_i^{\,b}\right)\left(-x_i^{\,b}\right) = 0$$

$$\frac{\partial S_r}{\partial b} = \sum_{i=1}^{n} 2\left(y_i - a e^{b x_i}\right)\left(-a x_i^{\,b} \ln\left(x_i\right)\right) = 0$$

Minimalizácia (pokr.)

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^{n} 2\left(y_i - a x_i^{\ b}\right)\left(-x_i^{\ b}\right) = 0 \qquad \frac{\partial S_r}{\partial b} = \sum_{i=1}^{n} 2\left(y_i - a x^b\right)\left(-a x_i^{\ b} \ln(x_i)\right) = 0$$

$$\frac{\partial S_r}{\partial a} = -2\sum_{i=1}^{n} y_i x_i^{\ b} + 2a\sum_{i=1}^{n} x_i^{2b} = 0$$

$$\frac{\partial S_r}{\partial b} = -2a\sum_{i=1}^{n} y_i x_i^{\ b} \ln(x_i) + 2a^2\sum_{i=1}^{n} x_i^{2b} \ln(x_i) = 0$$

$$\frac{\partial S_r}{\partial a} = -\sum_{i=1}^{n} y_i x_i^{\ b} + a\sum_{i=1}^{n} x_i^{2b} = 0$$

$$\frac{\partial S_r}{\partial b} = -\sum_{i=1}^{n} y_i x_i^{\ b} \ln(x_i) + a\sum_{i=1}^{n} x_i^{2b} \ln(x_i) = 0$$

$\uparrow$

Funkcia $a=f(b)$!

Minimalizácia (výsledok)

$$\frac{\partial S_r}{\partial a} = -\sum_{i=1}^{n} y_i x_i^{\,b} + a\sum_{i=1}^{n} x_i^{2b} = 0 \quad\Rightarrow\quad a = \frac{\sum_{i=1}^{n} y_i x_i^{\,b}}{\sum_{i=1}^{n} x_i^{2b}}$$

$$\frac{\partial S_r}{\partial b} = -\sum_{i=1}^{n} y_i x_i^{\,b}\ln(x_i) + a\sum_{i=1}^{n} x_i^{2b}\ln(x_i) = 0$$

$$\Downarrow$$

$$-\sum_{i=1}^{n} y_i x_i^{\,b}\ln(x_i) + \frac{\sum_{i=1}^{n} y_i x_i^{\,b}}{\sum_{i=1}^{n} x_i^{2b}}\sum_{i=1}^{n} x_i^{2b}\ln(x_i) = 0$$

Nelineárna rovnica riešiteľná numerickými metódami

Opäť jeden pokus o zjednodušenie problému

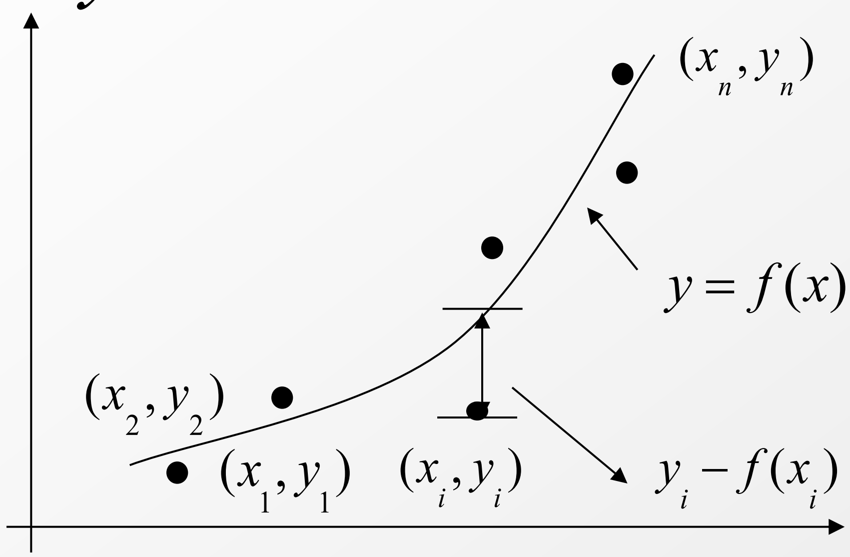
POWER MODEL - LINEARIZÁCIA

Odvodenie transformácie

Majme zadaných n bodov $(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)$ a hľadáme funkciu $f(x)=ax^b$, ktorá najlepšie vystihuje dané dáta.

hľadáme funkciu $\ln(f(x))=\ln(ax^b)$, s ktorou sa pracuje pohodlnšie

$$y = a \cdot x^b \Rightarrow \ln y = \ln a + b \ln x \Rightarrow z = a_0 + a_1 w$$



$$z_i = \ln(y_i)$$

$$w_i = \ln(x_i)$$

$$b = a_1$$

$$a_0 = \ln a \Rightarrow a = e^{a_0}$$

Odvodenie transformácie (pokr.)

$$z = a_0 + a_1 w$$

$$a_1 = \frac{n\sum_{i=1}^{n} w_i z_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} w_i}{n\sum_{i=1}^{n} w_i^2 - \left(\sum_{i=1}^{n} w_i\right)^2} \quad \Rightarrow \quad a_0 = \bar{z} - a_1 \bar{w}$$

Pozor – neminimalizujeme pôvodné kritérium

$$\cancel{S_r = \sum_{i=1}^{n}\left(y_i - a x_i^{\,b}\right)^2}$$

Minimalizujeme kritérium

$$S_r = \sum_{i=1}^{n}\left(z_i - a_o - a_1 w_i\right)^2 = \sum_{i=1}^{n}\left(\ln(y_i) - a_o - a_1 \ln(x_i)\right)^2$$

Konkrétny príklad

Z hadice strieka pod tlakom voda s prietokom

$$F = a \cdot p^b \quad \Rightarrow \quad \ln F = \ln a + b \ln p = a_0 + a_1 w$$

p(bar)	10	16	25	40	60
F(l/min)	94	118	147	180	230

Príklad (pokr.)

p	F	w=ln(p)	z=ln(F)	wi * zi	wi^2
10	94	2,303	4,543	10,461	5,302
16	118	2,773	4,771	13,227	7,687
25	147	3,219	4,990	16,064	10,361
40	180	3,689	5,193	19,156	13,608
60	230	4,094	5,438	22,265	16,764
Sum		16,077	24,935	81,174	53,722

$$a_1 = \frac{n\sum_{i=1}^{n} w_i z_i - \sum_{i=1}^{n} z_i \sum_{i=1}^{n} w_i}{n\sum_{i=1}^{n} w_i^2 - \left(\sum_{i=1}^{n} w_i\right)^2} = \frac{5\cdot 81.174 - 24.935\cdot 16.077}{5\cdot 53.722 - (16.077)^2} = 0.4910$$

$$a_0 = \bar{z} - a_1 \bar{w} = 24.935/5 - 0.490\cdot 16.077/5 = 3.4083$$

Našli sme koeficienty pre $z = a_0 + a_1 w$

Príklad (pokr.)

Pokračujeme v transformácii naspäť

$$z = a_0 + a_1 w \qquad a_0 = 3.4083 \qquad a_1 = 0.4910$$

$$F = a \cdot p^b \qquad \begin{aligned} a &= e^{a_0} = e^{3.4083} = 30.213 \\ b &= a_1 = 0.4910 \end{aligned}$$

$$F = ap^b = 30.213 p^{0.4910}$$

Ďalší z modelov využívaných pri regresii

POLYNOMIAL MODEL

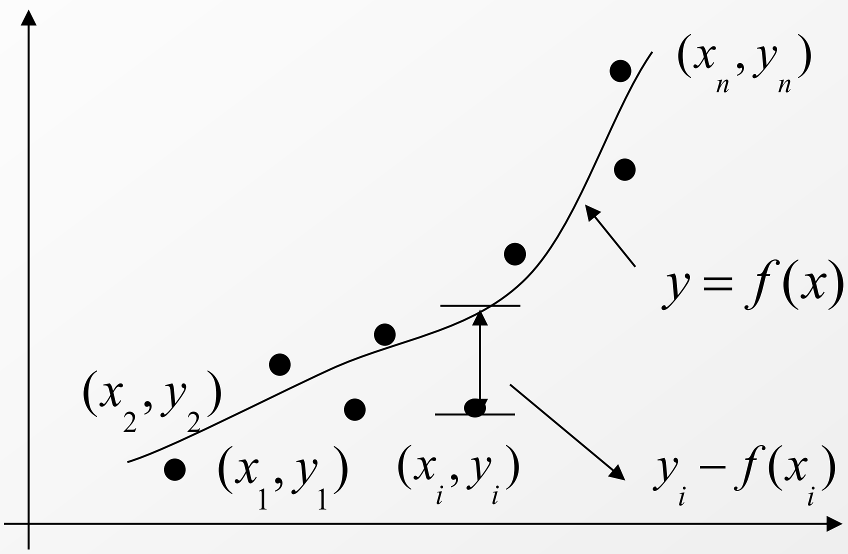
Zadanie regresie

Majme zadaných n bodov $(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)$ a hľadáme funkciu $f(x)=a_0+a_1x+\ldots a_m x^m$, ktorá najlepšie vystihuje dané dáta pre $m \leq n-2$.

Pre $m=n-1$ by sme hľadali interpoláciu

Minimalizujeme

$$S_r = \sum_{i=1}^{n}\left(y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m\right)^2$$



Vyjadrenie regresie

$$S_r = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - . \ . \ . - a_m x_i^m \right)^2$$

Hľadaná funkcia musí zodpovedať podmienke

$$\frac{\partial S_r}{\partial a_0} = \sum_{i=1}^{n} 2.\left(y_i - a_0 - a_1 x_i - . \ . \ . - a_m x_i^m \right)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = \sum_{i=1}^{n} 2.\left(y_i - a_0 - a_1 x_i - . \ . \ . - a_m x_i^m \right)(-x_i) = 0$$

$$\vdots \qquad \vdots \qquad \qquad \vdots \qquad \qquad \qquad \vdots$$

$$\frac{\partial S_r}{\partial a_m} = \sum_{i=1}^{n} 2.\left(y_i - a_0 - a_1 x_i - . \ . \ . - a_m x_i^m \right)(-x_i^m) = 0$$

Týmto získavame m rovníc s m neznámymi

Vyjadrenie regresie

$$\frac{\partial S_r}{\partial a_0} = \sum_{i=1}^{n} 2.\left(y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m\right)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = \sum_{i=1}^{n} 2.\left(y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m\right)(-x_i) = 0$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\frac{\partial S_r}{\partial a_m} = \sum_{i=1}^{n} 2.\left(y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m\right)(-x_i^m) = 0$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i + \ldots \sum_{i=1}^{n} a_m x_i^m$$

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 + \ldots \sum_{i=1}^{n} a_m x_i^{m+1}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\sum_{i=1}^{n} y_i x_i^m = \sum_{i=1}^{n} a_0 x_i^m + \sum_{i=1}^{n} a_1 x_i^{m+1} + \ldots \sum_{i=1}^{n} a_m x_i^{2m}$$

Vyjadrenie regresie

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i + \ldots \sum_{i=1}^{n} a_m x_i^m$$

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 + \ldots \sum_{i=1}^{n} a_m x_i^{m+1}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\sum_{i=1}^{n} y_i x_i^m = \sum_{i=1}^{n} a_0 x_i^m + \sum_{i=1}^{n} a_1 x_i^{m+1} + \ldots \sum_{i=1}^{n} a_m x_i^{2m}$$

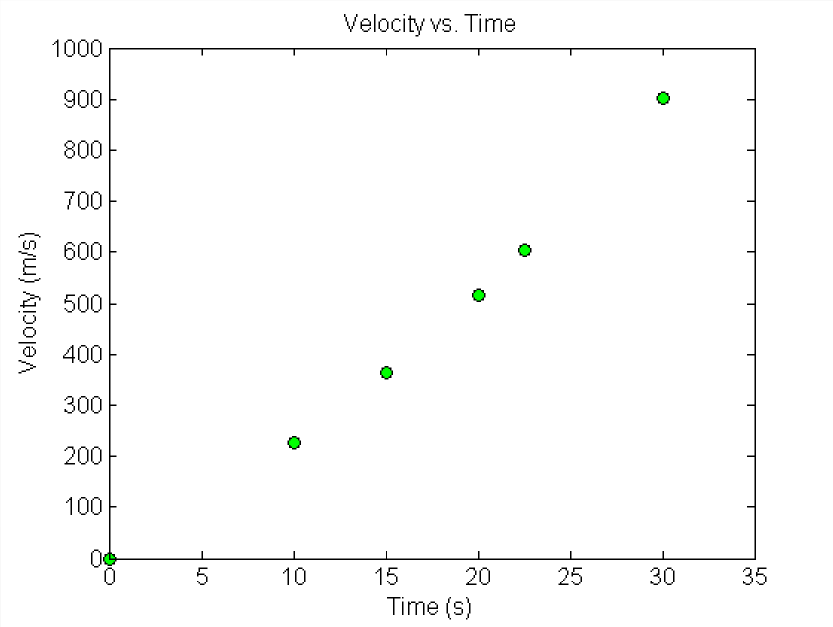
Riešim to ako systém rovníc

$$\begin{bmatrix} n & \left(\sum_{i=1}^{n} x_i\right) & \cdots & \left(\sum_{i=1}^{n} x_i^m\right) \\ \left(\sum_{i=1}^{n} x_i\right) & \left(\sum_{i=1}^{n} x_i^2\right) & \cdots & \left(\sum_{i=1}^{n} x_i^{m+1}\right) \\ \cdots & \cdots & \cdots & \cdots \\ \left(\sum_{i=1}^{n} x_i^m\right) & \left(\sum_{i=1}^{n} x_i^{m+1}\right) & \cdots & \left(\sum_{i=1}^{n} x_i^{2m}\right) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \cdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \cdots \\ \sum_{i=1}^{n} x_i^m y_i \end{bmatrix}$$

Príklad

Vrátime sa k starému príkladu s raketou

$t, (\mathrm{s})$	$v(t), (\mathrm{m/s})$
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67



Nájdite kvadratickú funkciu repreznetujúcu dané dáta
Nájdite rýchlosť v čase t = 16

Riešenie

t, (s)	$v(t)$, (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i + \ldots \sum_{i=1}^{n} a_m x_i^m$$

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 + \ldots \sum_{i=1}^{n} a_m x_i^{m+1}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\sum_{i=1}^{n} y_i x_i^m = \sum_{i=1}^{n} a_0 x_i^m + \sum_{i=1}^{n} a_1 x_i^{m+1} + \ldots \sum_{i=1}^{n} a_m x_i^{2m}$$

$$\sum_{i=1}^{6} y_i = \sum_{i=1}^{6} a_0 + \sum_{i=1}^{6} a_1 x_i + \sum_{i=1}^{6} a_2 x_i^2$$

$$\sum_{i=1}^{6} y_i x_i = \sum_{i=1}^{6} a_0 x_i + \sum_{i=1}^{6} a_1 x_i^2 + \sum_{i=1}^{6} a_2 x_i^3$$

$$\sum_{i=1}^{6} y_i x_i^2 = \sum_{i=1}^{n} a_0 x_i^2 + \sum_{i=1}^{n} a_1 x_i^3 + \sum_{i=1}^{n} a_2 x_i^4$$

Riešenie (pokr.)

t, (s)	$v(t)$, (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

$$\sum_{i=1}^{6} v_i = \sum_{i=1}^{6} a_0 + \sum_{i=1}^{6} a_1 t_i + \sum_{i=1}^{6} a_2 t_i^2$$

$$\sum_{i=1}^{6} v_i t_i = \sum_{i=1}^{6} a_0 t_i + \sum_{i=1}^{6} a_1 t_i^2 + \sum_{i=1}^{6} a_2 t_i^3$$

$$\sum_{i=1}^{6} v_i t_i^2 = \sum_{i=1}^{n} a_0 t_i^2 + \sum_{i=1}^{n} a_1 t_i^3 + \sum_{i=1}^{n} a_2 t_i^4$$

$$\sum_{i=1}^{6} v_i = n a_0 + a_1 \sum_{i=1}^{6} t_i + a_2 \sum_{i=1}^{6} t_i^2$$

$$\sum_{i=1}^{6} v_i t_i = a_0 \sum_{i=1}^{6} t_i + a_1 \sum_{i=1}^{6} t_i^2 + a_2 \sum_{i=1}^{6} t_i^3$$

$$\sum_{i=1}^{6} v_i t_i^2 = a_0 \sum_{i=1}^{n} t_i^2 + a_1 \sum_{i=1}^{n} t_i^3 + a_2 \sum_{i=1}^{n} t_i^4$$

$$\sum_{i=1}^{6} v_i = 2611.81 \quad \sum_{i=1}^{6} t_i = 97.5 \quad \sum_{i=1}^{6} t_i^{\,2} = 2131.25$$

$$\sum_{i=1}^{6} v_i t_i = 5.87\times10^4 \quad \sum_{i=1}^{6} t_i^{\,3} = 5.08\times10^4$$

$$\sum_{i=1}^{6} v_i t_i^2 = 1.43\times10^6 \quad \sum_{i=1}^{6} t_i^{\,4} = 1.29\times10^6$$

Riešenie (pokr.)

$$2611.81 = 6a_0 + 97.5a_1 + 2131.25a_2$$

$$5.87\times10^4 = 97.5a_0 + 2131.25a_1 + 5.08\times10^4 a_2$$

$$1.43\times10^6 = 2131.25a_0 + 5.08\times10^4 a_1 + 1.29\times10^6 a_2$$

$$a_0 = 435.301 - 16.25a_1 - 355.208a_2$$

$$601.8 = a_0 + 21.86a_1 + 520.67a_2$$

$$670.04 = a_0 + 23.82a_1 + 603.82a_2$$

$$601.8 = 435.301 - 16.25a_1 - 355.21a_2 + 21.86a_1 + 520.67a_2$$

$$670.04 = 435.301 - 16.25a_1 - 355.21a_2 + 23.82a_1 + 603.82a_2$$

$$166.49 = 5.61a_1 + 165.46a_2$$

$$234.74 = 7.57a_1 + 248.61a_2$$

Riešenie (pokr.)

$$166.49 = 5.61a_1 + 165.46a_2$$
$$234.74 = 7.57a_1 + 248.61a_2$$

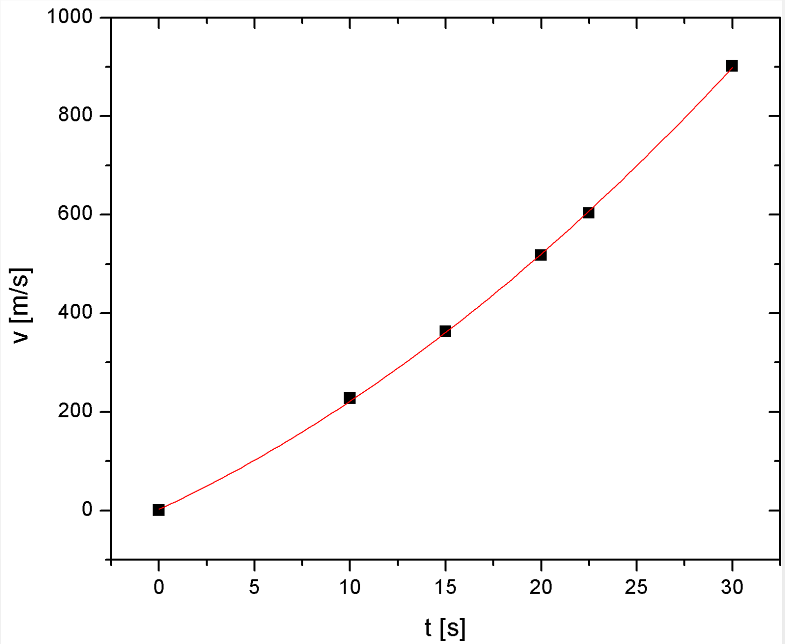
$$29.6774 = a_1 + 29.4938a_2$$
$$31.0092 = a_1 + 32.8415a_2$$

$$1.3318 = 3.3485a_2$$
$$a_2 = 0.3977$$

$$a_1 = 29.6774 - 29.4938a_2$$
$$a_1 = 17.9477$$

$$a_0 = 435.301 - 16.25a_1 - 355.208a_2$$
$$a_0 = 2.1226$$

$$v(t) = 2.1226 + 17.9477t + 0.3977t^2$$



$$v(16) = 2.1226 + 17.9477 \cdot 16 + 0.3977 \cdot 16^2 = 391.097$$

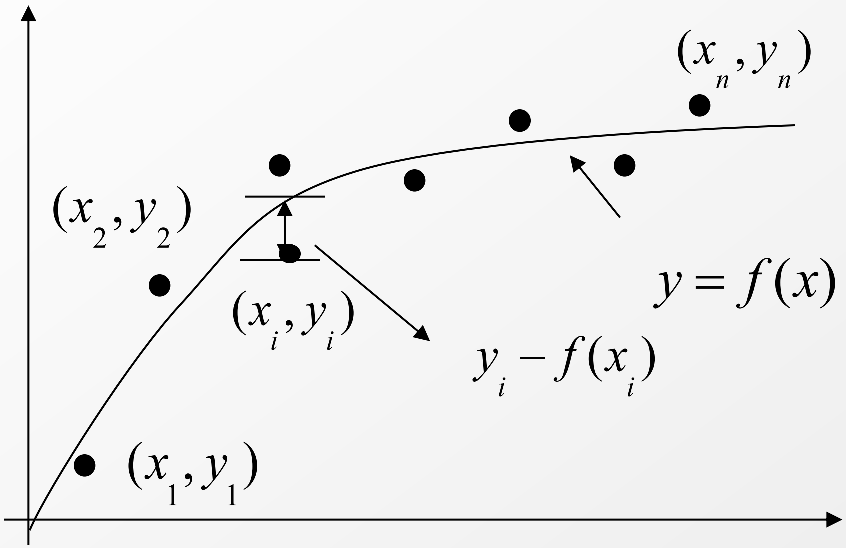
Z kvadratickej interpolácie máme hodnotu 392.19

SATURATION GROWTH MODEL

Zadanie regresie

Majme zadaných n bodov $(x_1,y_1), (x_2,y_2),\dots,(x_n,y_n)$ a hľadáme funkciu typu $$y = \frac{ax}{b+x}$$

Platí pritom $$\lim_{x\to\infty} y \frac{ax}{b+x} = \lim_{x\to\infty} \frac{a}{\frac{b}{x}+1} = a$$ a pre x=0 je y=0



transformácia

$$y = \frac{ax}{b+x} \quad \Rightarrow \quad \frac{1}{y} = \frac{b+x}{ax} = \frac{b}{a}\frac{1}{x} + \frac{1}{a}$$

$$z = a_0 + a_1 w$$

A ďalej možem použiť lineárnu regresiu...

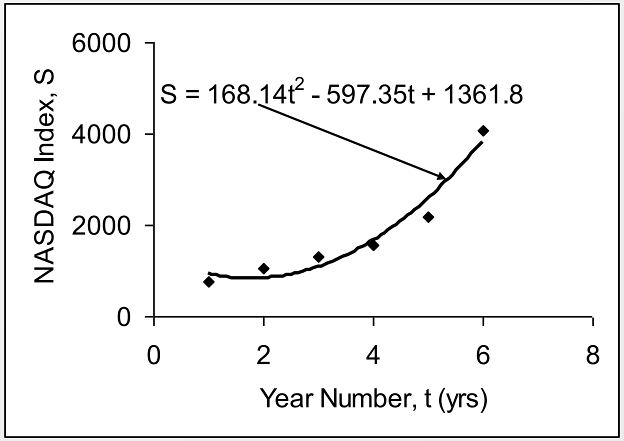
$$a = \frac{1}{a_0} \qquad \frac{b}{a} = a_1 \Rightarrow b = a_1 a \Rightarrow b = \frac{a_1}{a_0}$$

ZÁVER

Zhrnutie

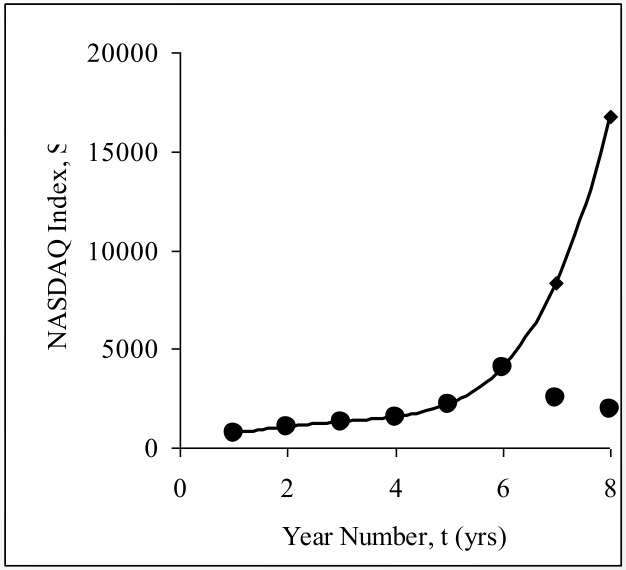
- Uviedli sme niekoľko príkladov pre regresiu funkcie a vhodnú transformáciu funkcií.
- Je viacero metódna riešenie komplexnejších funkcií. Napríklad hľadanie koeficientov pomocou náhodných malých zmien a opakovanom vyhodnocovaní minimalizačného kritéria.
- Treba byť napriek tomu opatrní

Year Number ( )	NASDAQ Index ( )
1 (1994)	752
2 (1995)	1052
3 (1996)	1291
4 (1997)	1570
5 (1998)	2193
6 (1999)	4069



Zhrnutie

Year Number ( )	NASDAQ Index ( )	Predicted Index	Absolute Relative True Error (%)
1 (1994)	752	933	24
2 (1995)	1052	840	20
3 (1996)	1291	1083	16
4 (1997)	1570	1663	6
5 (1998)	2193	2579	18
6 (1999)	4069	3831	6
7 (2000)	2471	5419	119
8 (2001)	1951	7344	276



THE END