

Matematika 1

pro PEF PaE

14. Metoda nejmenších čtverců

Přemysl Jedlička

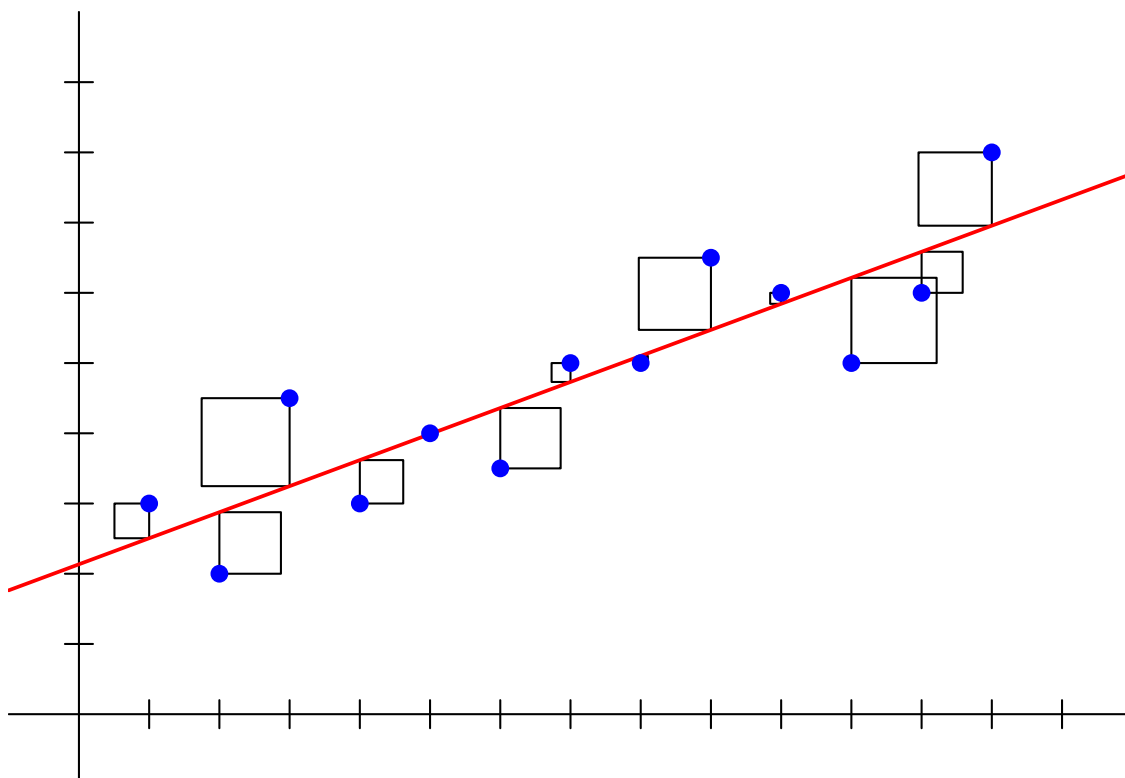
Katedra matematiky, TF ČZU

Metoda nejmenších čtverců

Lineární regrese

2 / 10

Metoda nejmenších čtverců



Odvození lineární regrese

V tabulce máme hodnoty

x	x_1	x_2	x_3	$\cdots$	x_{m-1}	x_m
y	y_1	y_2	y_3	$\cdots$	x_{m-1}	y_m

Chceme nalézt lineární funkci $f(x) = a + b \cdot x$ takovou, že součet plochy čtverců bude minimální. Funkce plochy je

$$S(a, b) = (a + b \cdot x_1 - y_1)^2 + (a + b \cdot x_2 - y_2)^2 + \cdots + (a + b \cdot x_m - y_m)^2,$$

což je funkce dvou proměnných a , b a my hledáme její globální minimum. Což jsou jediné body, kde jsou parciální derivace nulové.

$$\frac{\partial S}{\partial a} = 2(a + b \cdot x_1 - y_1) + 2(a + b \cdot x_2 - y_2) + \cdots + 2(a + b \cdot x_m - y_m),$$

$$\frac{\partial S}{\partial b} = 2(a + b \cdot x_1 - y_1) \cdot x_1 + \cdots + 2(a + b \cdot x_m - y_m) \cdot x_m.$$

Odvození lineární regrese

Dostaneme soustavu dvou rovnic o dvou neznámých

$$2(a + b \cdot x_1 - y_1) + 2(a + b \cdot x_2 - y_2) + \cdots + 2(a + b \cdot x_m - y_m) = 0,$$

$$2(a + b \cdot x_1 - y_1) \cdot x_1 + \cdots + 2(a + b \cdot x_m - y_m) \cdot x_m = 0.$$

Po drobných úpravách vyjde

$$m \cdot a + b \cdot (x_1 + \cdots + x_m) = y_1 + \cdots + y_m,$$

$$a \cdot (x_1 + \cdots + x_m) + b \cdot (x_1^2 + \cdots + x_m^2) = x_1 y_1 + \cdots + x_m y_m,$$

což se dá zapsat v lepším tvaru jako

$$m \cdot a + b \cdot \sum_{i=1}^m x_i = \sum_{i=1}^m y_i,$$

$$a \cdot \sum_{i=1}^m x_i + b \cdot \sum_{i=1}^m x_i^2 = \sum_{i=1}^m x_i y_i.$$

Použití lineární regrese

Úloha

Nalezněte přímku, která nejlépe odpovídá naměřeným hodnotám ve smyslu metody nejmenších čtverců.

x	1	2	3	4	5	6	7	8	9	10	11	12	13
y	3	2	4,5	3	4	3,5	5	5	6,5	6	5	6	8

Řešení: Zapišeme hodnoty do tabulky:

i	x_i	y_i	x_i^2	$x_i y_i$
1	1	3	1	3
2	2	2	4	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
13	13	8	169	104
$m = 13$	$\sum_{i=1}^{13} x_i = 91$	$\sum_{i=1}^{13} y_i = 61,5$	$\sum_{i=1}^{13} x_i^2 = 819$	$\sum_{i=1}^{13} x_i y_i = 498$

Použití lineární regrese

Dostaneme dvě rovnice o dvou neznámých

$$13 \cdot a + 91 \cdot b = 61,5$$

$$91 \cdot a + 819 \cdot b = 498$$

Z první rovnice vyjádříme $a = \frac{61,5 - 91 \cdot b}{13}$ a dosadíme do druhé rovnice. Vyjde nám řešení

$$b = \frac{135}{364} \quad \text{a} \quad a = \frac{111}{52}.$$

Přímka má rovnici $y = \frac{135}{364} \cdot x + \frac{111}{52}$.

Kvadratická regrese

Vzorec pro nalezení funkce $f(x) = a + b \cdot x + c \cdot x^2$, která nejlépe vyhovuje naměřeným hodnotám ve smyslu metody nejmenších čtverců:

$$\begin{aligned} m \cdot a + b \cdot \sum_{i=1}^m x_i + c \cdot \sum_{i=1}^m x_i^2 &= \sum_{i=1}^m y_i \\ a \cdot \sum_{i=1}^m x_i + b \cdot \sum_{i=1}^m x_i^2 + c \cdot \sum_{i=1}^m x_i^3 &= \sum_{i=1}^m x_i \cdot y_i \\ a \cdot \sum_{i=1}^m x_i^2 + b \cdot \sum_{i=1}^m x_i^3 + c \cdot \sum_{i=1}^m x_i^4 &= \sum_{i=1}^m x_i^2 \cdot y_i \end{aligned}$$

Polynomická regrese

Vzorec pro nalezení funkce $f(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \dots + a_n x^n$ podle metody nejmenších čtverců:

$$\begin{aligned} m \cdot a_0 + a_1 \cdot \sum_{i=1}^m x_i + \dots + a_n \cdot \sum_{i=1}^m x_i^n &= \sum_{i=1}^m y_i \\ a_0 \cdot \sum_{i=1}^m x_i + a_1 \cdot \sum_{i=1}^m x_i^2 + \dots + a_n \cdot \sum_{i=1}^m x_i^{n+1} &= \sum_{i=1}^m x_i \cdot y_i \\ &\vdots \\ a_0 \cdot \sum_{i=1}^m x_i^n + a_1 \cdot \sum_{i=1}^m x_i^{n+1} + \dots + a_n \cdot \sum_{i=1}^m x_i^{2n} &= \sum_{i=1}^m x_i^n \cdot y_i \end{aligned}$$

Neboli, k -tá rovnice zní $\sum_{j=0}^n a_j \cdot \sum_{i=1}^m x_i^{j+k} = \sum_{i=1}^m x_i^k \cdot y_i$, pro $0 \leq k \leq n$.

Exponenciální regrese

Chceme najít nejlepší funkci ve tvaru

$$y = a \cdot e^{bx}.$$

Když tuto funkci zlogaritmujeme, dostaneme

$$\ln y = \ln a + b \cdot x.$$

Označíme $Y = \ln y$ a $A = \ln a$ a provádíme lineární regresi pro

$$Y = A + b \cdot x.$$

Úloha

V roce 1600 bylo na zemi 545 mil. lidí, v roce 1700 bylo na zemi 679 mil. lidí, v roce 1800 bylo na zemi 910 mil. lidí a v roce 1900 to bylo 1,6 mld. lidí. Kolik by bylo na zemi lidí v roce 1950, nebýt světových válek?

Řešení exponenciální regrese

Údaje dosadíme do tabulky:

i	t_i	y_i	$Y_i = \ln y_i$	t_i^2	$t_i Y_i$
1	1600	545	6,301	2,56 mil.	10081
2	1700	679	6,520	2,89 mil.	11085
3	1800	910	6,813	3,24 mil.	12264
4	1900	1600	7,378	3,61 mil.	14017
	$\sum t_i = 7000$		$\sum Y_i = 27,012$	$\sum t_i^2 = 12,3$	$\sum t_i Y_i = 47447$

Z lineární regrese dostaneme $b = 3,52$ tisícín a $A = 0,593$, tedy $a = 1,81$.

Vzorec pro růst populace tedy je $1,81 \cdot e^{0,00352 \cdot t}$, po dosazení $t = 1950$ dostaneme výsledek 1731.