Metoda nejmenších čtverců

Zápisky z internetu, poznámek z předmětu Praktická astrofyzika a skript

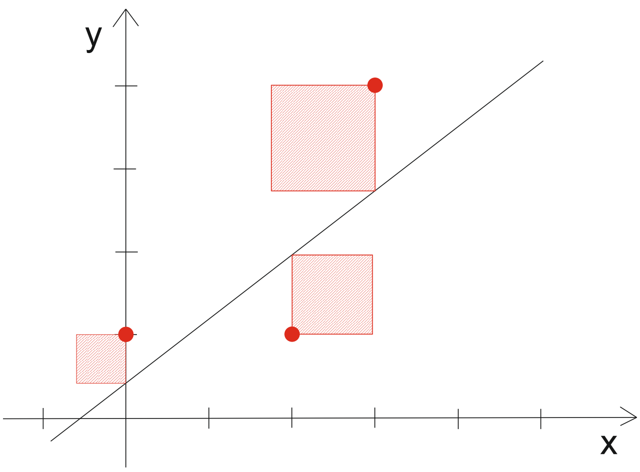
Poslední úprava: 9. června 2015

KAPITOLA 1

METODA NEJMENŠÍCH ČTVERCŮ

Metoda nejmenších čtverců je regresní analýza. Snažíme se najít ten nejlepší sklon přímky a výšky průsečíku s osou y, tak aby plocha čtverců byla co nejmenší. V tomto případě se optimalizují dva parametry. Čtverce znázorňují kvadráty odchylek naměřených hodnot od "modelových".

Například měřením jsme získali tři body $(0; 1)$, $(2; 1)$ a $(3; 4)$ a budeme hledat lineární funkci, která bude nejlépe popisovat jejich průběh. Zkusíme si to nakreslit a přibližně proložit lineární funkcí.



Obrázek 1.1: Vynesené body a odhad lineární funkce.

Toto je však pouhý odhad, potřebujeme znát přesnou hodnotu sklonu přímky a její průsečík s osou y. To zjistíme z těchto vztahů:

$$a \cdot \sum_{i=1}^{N} x_i^2 + b \cdot \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i \cdot y_i \quad , \tag{1.1}$$

$$a \cdot \sum_{i=1}^{N} x_i + b \cdot n = \sum_{i=1}^{N} y_i \quad , \tag{1.2}$$

kde a je sklon přímky, b je průsečík s osou y a n je počet bodů v grafu. Vyjádřím si nejdříve hodnoty pro jednotlivé sumy:

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 0 + 2 + 3 = 5 \ , \tag{1.3}$$

$$\sum_{i=1}^{3} x_i^2 = x_1^2 + x_2^2 + x_3^2 = 0 + 4 + 9 = 13 \ , \tag{1.4}$$

$$\sum_{i=1}^{3} y_i = y_1 + y_2 + y_3 = 1 + 1 + 4 = 6 \ , \tag{1.5}$$

$$\sum_{i=1}^{3} x_i \cdot y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 = 0 \cdot 1 + 2 \cdot 1 + 3 \cdot 4 = 14 \ . \tag{1.6}$$

Do rovnic 1.1 dosadíme hodnoty pro jednotlivé sumy a získáme:

$$13 \cdot a + 5 \cdot b = 14 \ , \tag{1.7}$$

$$5 \cdot a + 3 \cdot b = 6 \ . \tag{1.8}$$

Z těchto dvou lineárních rovnic získáme funkci pro lineární fit ($y = a \cdot x + b$) v tomto tvaru:

$$y = \frac{6}{7} \cdot x + \frac{4}{7} \ . \tag{1.9}$$

Lineární fit

Tuto metodu využijeme ve chvíli, když chceme aby se průběh dané funkce co nejvíce přiblížil naměřeným bodům $[x_i; y_i]$, k tomu slouží měřítko:

$$S_{(a;b)} = \sum_{i=1}^{N} \left(f_{(x_i)} - y_i \right)^2 \ , \tag{1.10}$$

kde i zastupuje počet měření. Pro lineární fit z předchozího příkladu pak platí:

$$\begin{aligned} S_{(a;b)} = \sum_{i=1}^{N} (a \cdot x_i + b - y_i)^2 = (a \cdot x_1 + b - y_1)^2 + \\ + (a \cdot x_2 + b - y_2)^2 + (a \cdot x_3 + b - y_3)^2 \ . \end{aligned} \tag{1.11}$$

Pro minimalizaci odchylky, je nutné nalézt minimum této funkce. Jednoduše výraz zderivujeme podle parametrů a a b:

$$\frac{\partial S}{\partial a} = 2 \cdot (a \cdot x_1 + b - y_1) \cdot x_1 + \\ + (a \cdot x_2 + b - y_2) \cdot x_2 + (a \cdot x_3 + b - y_3) \cdot x_3 \quad , \tag{1.12}$$

$$\frac{\partial S}{\partial b} = 2 \cdot (a \cdot x_1 + b - y_1) + (a \cdot x_2 + b - y_2) + (a \cdot x_3 + b - y_3) \quad . \tag{1.13}$$

Stačí provést drobné úpravy a získáme:

$$\frac{\partial S}{\partial a} = 2 \cdot a \cdot \left(x_1^2 + x_2^2 + x_3^2\right) + 2 \cdot b \cdot (x_1 + x_2 + x_3) - \tag{1.14}$$

$$-2 \cdot (x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3) \quad , \tag{1.15}$$

$$\frac{\partial S}{\partial b} = 2 \cdot a \cdot (x_1 + x_2 + x_3) + 2 \cdot 3 \cdot b - 2 \cdot (y_1 + y_2 + y_3) \quad . \tag{1.16}$$

Nyní stačí pouze položit rovno nule a získáme dvě rovnice:

$$2 \cdot a \cdot \left(x_1^2 + x_2^2 + x_3^2\right) + 2 \cdot b \cdot (x_1 + x_2 + x_3) = \tag{1.17}$$

$$= 2 \cdot (x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3) \quad , \tag{1.18}$$

$$2 \cdot a \cdot (x_1 + x_2 + x_3) + 2 \cdot 3 \cdot b = 2 \cdot (y_1 + y_2 + y_3) \quad . \tag{1.19}$$

Tyto rovnice nyní můžeme vynásobit 1/2 a součty v závorkách nahradit sumami:

$$a \cdot \sum_{i=1}^{3} x_i^2 + b \cdot \sum_{i=1}^{3} x_i = \sum_{i=1}^{3} x_i \cdot y_i \quad , \tag{1.20}$$

$$a \cdot \sum_{i=1}^{3} x_i + 3 \cdot b = \sum_{i=1}^{3} y_i \quad . \tag{1.21}$$

Číslo 3 zastupuje počet měření. Pro výpočet parametrů lineárního fitu máme tedy tyto dva vztahy:

$$a \cdot \sum_{i=1}^{N} x_i^2 + b \cdot \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i \cdot y_i \quad , \tag{1.22}$$

$$a \cdot \sum_{i=1}^{N} x_i + n \cdot b = \sum_{i=1}^{N} y_i \quad . \tag{1.23}$$

Dále si můžeme odvodit vztah pro a a b, který bude platit pro libovolný počet měření při použití lineární regrese. Na začátek vyjdeme ze vzorce 1.23, ze kterého si vyjádříme b:

$$b = \frac{\sum_{i=1}^{N} y_i - a \cdot \sum_{i=1}^{N} x_i}{n} \quad . \tag{1.24}$$

Tento vztah si nyní dosadíme do rovnice 1.22 a budeme vyjadřovat parametr a :

$$a \cdot \sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} x_i \cdot \frac{\sum_{i=1}^{N} y_i}{n} - a \cdot \sum_{i=1}^{N} x_i \cdot \frac{\sum_{i=1}^{N} x_i}{n} = \sum_{i=1}^{N} x_i \cdot y_i \ , \tag{1.25}$$

celý vztah nyní vynásobíme n:

$$a \cdot n \cdot \sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i - a \cdot \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} x_i = n \cdot \sum_{i=1}^{N} x_i \cdot y_i \tag{1.26}$$

a z tohoto výrazu si vyjádříme a

$$a = \frac{n \cdot \sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{n \cdot \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i} \ . \tag{1.27}$$

Máme tedy výraz pro sklon a a průsečík osou y daný parametrem b, který si můžeme upravit:

$$b = \frac{\sum_{i=1}^{N} y_i - a \cdot \sum_{i=1}^{N} x_i}{n} = \hat{y} - a \cdot \hat{x} \ , \tag{1.28}$$

kde $\hat{y}$ a $\hat{x}$ jsou aritmetické průměry těchto hodnot.

Polynom druhého a vyššího stupně

Pro odvození vztahů pro polynom použijeme obdobný způsob, pro jednoduchost použiji polynom druhého stupně. Budeme počítat pouze pro dvě měření:

$$S_{(a;b)} = \sum_{i=1}^{N} \left(f_{(x_i)} - y_i\right)^2 = \sum_{i=1}^{N} \left(a \cdot x_i^2 + b \cdot x_i + c\right)^2 \quad , \tag{1.29}$$

opět zderivujeme tento výraz podle a, b a c:

$$\frac{\partial S}{\partial a} = 2 \cdot x_1^2 \cdot \left(a \cdot x_1^2 + b \cdot x_1 + c - y_1\right) + 2 \cdot x_2^2 \cdot \left(a \cdot x_2^2 + b \cdot x_2 + c - y_2\right) = \tag{1.30}$$

$$= 2 \cdot a \cdot \left(x_1^4 + x_2^4\right) + 2 \cdot b \cdot \left(x_1^3 + x_2^3\right) + 2 \cdot c \cdot \left(x_1^2 + x_2^2\right) - 2 \cdot \left(x_1^2 \cdot y_1 + x_2^2 \cdot y_2\right) \quad . \tag{1.31}$$

$$\frac{\partial S}{\partial b} = 2 \cdot x_1 \cdot \left(a \cdot x_1^2 + b \cdot x_1 + c - y_1\right) + 2 \cdot x_2 \cdot \left(a \cdot x_2^2 + b \cdot x_2 + c - y_2\right) = \tag{1.32}$$

$$= 2 \cdot a \cdot \left(x_1^3 + x_2^3\right) + 2 \cdot b \cdot \left(x_1^2 + x_2^2\right) + 2 \cdot c \cdot (x_1 + x_2) - 2 \cdot (x_1 \cdot y_1 + x_2 \cdot y_2) \quad . \tag{1.33}$$

$$\frac{\partial S}{\partial c} = 2 \cdot \left(a \cdot x_1^2 + b \cdot x_1 + c - y_1\right) + 2 \cdot \left(a \cdot x_2^2 + b \cdot x_2 + c - y_2\right) = \tag{1.34}$$

$$= 2 \cdot a \cdot \left(x_1^2 + x_2^2\right) + 2 \cdot b \cdot (x_1 + x_2) + 2 \cdot c - 2 \cdot (y_1 + y_2) \quad . \tag{1.35}$$

Dvojka ve výrazu $2 \cdot c$ zastupuje počet měření. Nyní opět stačí výsledky položit rovny nule a vynásobit 1/2.

$$a \cdot \left(x_1^4 + x_2^4\right) + b \cdot \left(x_1^3 + x_2^3\right) + c \cdot \left(x_1^2 + x_2^2\right) = \left(x_1^2 \cdot y_1 + x_2^2 \cdot y_2\right) \tag{1.36}$$

$$a \cdot \left(x_1^3 + x_2^3\right) + b \cdot \left(x_1^2 + x_2^2\right) + c \cdot (x_1 + x_2) = (x_1 \cdot y_1 + x_2 \cdot y_2) \tag{1.37}$$

$$a \cdot \left(x_1^2 + x_2^2\right) + b \cdot (x_1 + x_2) + c = (y_1 + y_2) \quad . \tag{1.38}$$

Tyto vztahy lze pomocí sumy přepsat pro libovolný počet měření:

$$a \cdot \sum_{i=1}^{N} x_i^4 + b \cdot \sum_{i=1}^{N} x_i^3 + c \cdot \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i^2 \cdot y_i \tag{1.39}$$

$$a \cdot \sum_{i=1}^{N} x_i^3 + b \cdot \sum_{i=1}^{N} x_i^2 + c \cdot \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i \cdot y_i \tag{1.40}$$

$$a \cdot \sum_{i=1}^{N} x_i^2 + b \cdot \sum_{i=1}^{N} x_i + n \cdot c = \sum_{i=1}^{N} y_i \quad . \tag{1.41}$$

Tuto soustavu tří rovnic (kvartické , kubické a kvadratické) lze přepsat do matice:

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^4 & \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 \\ \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i & n \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i^2 \cdot y_i \\ \sum_{i=1}^{N} x_i \cdot y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix} \quad . \tag{1.42}$$

Tento výsledek je nyní možné přepsat do tvaru, pro předem neurčený stupeň polynomu:

$$y = a_s \cdot x^s + \cdots + a_1 \cdot x^1 + a_0 \tag{1.43}$$

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^{2s} & \ldots & \sum_{i=1}^{N} x_i^{s+1} & \sum_{i=1}^{N} x_i^s \\ \vdots & \ddots & \vdots & \vdots \\ \sum_{i=1}^{N} x_i^{s+1} & & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^s & & \sum_{i=1}^{N} x_i & n \end{pmatrix} \cdot \begin{pmatrix} a_s \\ \vdots \\ a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i^s \cdot y_i \\ \vdots \\ \sum_{i=1}^{N} x_i \cdot y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix} \tag{1.44}$$

Použitím metody nejmenších čtverců předpokládáme, že drtivá většina odchylek patří do normálového rozdělení (Gaussovka [zvonovka]). Když si vykreslíme histogram, tak musím mít tvar normálového rozdělení.