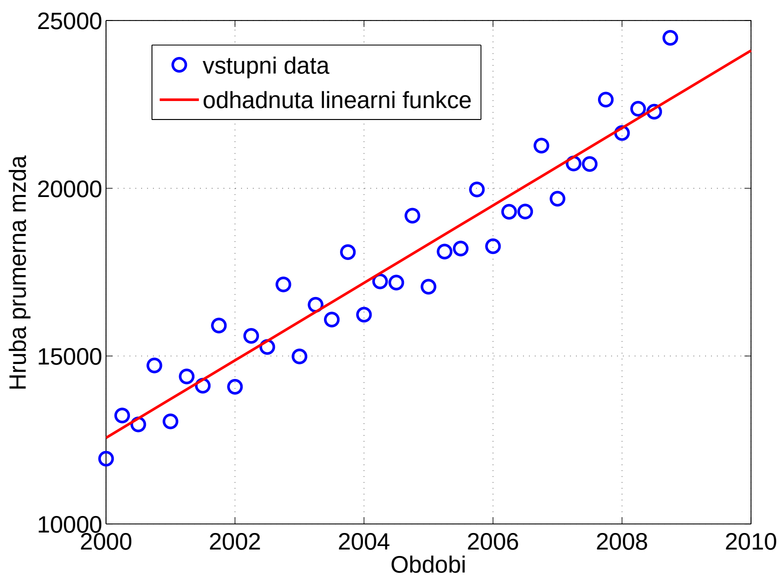
Cvičení z optimalizace:
Metoda nejmenších čtverců

Vojtěch Franc, ZS2014

Úloha 1: Predikce průměrné hrubé mzdy

Graf 1 ukazuje vývoj průměrné hrubé mzdy (PHM) v České republice v období od roku 2000 do roku 2008 [1]. Pro ilustraci jsou některé hodnoty z grafu 1 uvedeny v tabulce 1.



Obrázek 1: Modré body odpovídají průměrné hrubé mzdě měřené ve všech kvartálech let 2000 až 2008. Červená přímka je odhadnuta z dat metodou nejmenších čtverců.

Naším cílem bude predikovat hodnotu PHM ve druhém kvartálu roku 2009, tj. v období pro něž nemáte údaje o PHM k dispozici. Úlohu vyřešíme tak, že nejprve

[1] Data byla stažena ze stránek Českého statistického úřadu `www.czso.cz/`

Mzda $M(t)$ [Kč]	11,941	13,227	12,963	14717	...	22,282	24,448
Období t [rok]	2000.00	2000.25	2000.50	2000.75	...	2008.50	2008.75

Tabulka 1: Vybrané hodnoty z časové řady popisující vývoj PHM v čase. Časový údaj je ve formátu $t = \text{rok} + (\text{kvartál} - 1)/4$, kde rok $\in \{2000, \dots, 2008\}$ a kvartál $\in \{1, 2, 3, 4\}$.

nalezneme funkci, která co nejlépe odpovídá zadaným údajům o PHD. Nalezenou funkci následně použijeme pro odhad PHD v požadované době. Z obrázku 1 je vidět, že trend PHM v závislosti na čase je téměř lineární. Tudíž budeme hledat lineární funkci

$$\hat{M}(t) = x_0 + x_1 t\,, \tag{1}$$

kde $\hat{M}(t)$ je odhad HPM v čase t a $x_0, x_1 \in \mathbb{R}$ jsou neznámé parametry. Parametry (x_0, x_1) určíme z dat metodou nejmenších čtverců. Naše data (viz. graf 1 a tabulka 1) je množina $\{(M(t_1), t_1), \ldots, (M(t_n), t_n)\}$ obsahují n dvojic (HPM, čas). Parametry (x_0, x_1) nalezneme tak, aby součet kvadrátů odchylek skutečné a odhadnuté mzdy byl v zadaných bodech minimální. To znamená, že budeme řešit minimalizační úlohu

$$(x_0^*, x_1^*) = \operatorname*{argmin}_{x_0 \in \mathbb{R}, x_1 \in \mathbb{R}} F(x_0, x_1) \tag{2}$$

kde

$$F(x_0, x_1) = \sum_{i=1}^{n} \left(\hat{M}(t_i) - M(t_i) \right)^2 = \sum_{i=1}^{n} (x_0 + x_1 t_i - M(t_i))^2 \ .$$

Vaším úkolem bude převést úlohu (2) na problém řešení soustavy lineárních rovnic $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ pomocí metody nejmenších čtverců, tj. na úlohu

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^2} \|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|^2 \,. \tag{3}$$

Úkoly k vypracování

1. Stáhněte si textový soubor `http://cmp.felk.cvut.cz/cmp/courses/OPT/cviceni/06/mzdy.txt`. Nahrajte data do Matlabu pomocí příkazů:

```
Data = load('mzdy.txt','-ascii');
Rok = Data(:,1);
Mzdy = Data(:,2);
```

2. Převeďte úlohu (2) na úlohu (3). Napište přesnou formu matice $\mathbf{A}$ a vektoru $\boldsymbol{b}$ vystupujících v úloze (3).

   Řešení (1 bod):

$$\mathbf{A} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} M(t_1) \\ M(t_2) \\ \vdots \\ M(t_m) \end{bmatrix}$$

3. Použijte metodu nejmenších čtverců k odhadu parametrů lineární funkce (1) z dat nahraných v bodě 1. Do jednoho grafu zobrazte vstupní data a odhadnutou funkci.

   Řešení (1 bod): Obrázek 1.

4. Použijte odhadnutou funkci k předpovědi HPM ve druhém kvartálu roku 2009. Pro porovnání, skutečná hodnota HPM v tomto kvartálu byla 23,664 Kč.

   Řešení (1 bod): Odhadnutá mzda ve 2. kvartálu 2009 je 23236.620292 .

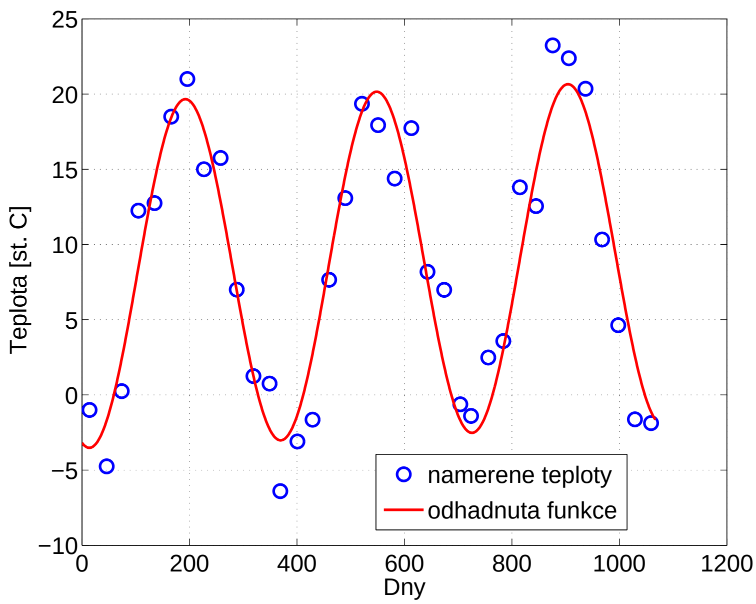
   Mnoho studentů místo toho počítá odhad pro 1. kvartál (=22948.112698) nebo 3. kvartál (=23525.127885).

5. Pro odhadnuté parametry spočtěte hodnotu kriteriální funkce $F(x_0, x_1)$.

   Řešení (1 bod): $F(x_0^*, x_1^*) = 24,800,910.248219$.

Interpolace denní teploty ve Svatoňovicích[2]

Představte si, že pracujete na meteorologické stanice ve Svatoňovicích (`http://www.meteosvatonovice.unas.cz`). Od Vašeho kolegy meteorologa jste získali měření průměrné denní teploty od roku 2005 do roku 2007. Váš kolega neprováděl měření každý den, ale jen jednou za měsíc. Konkrétně měřil teplotu pokaždé 15. dne v měsíci. Naměřené teploty jsou vyneseny v grafu 2 a vybrané hodnoty jsou pro ilustraci uvedeny v tabulce 2.



Obrázek 2: Modré body značí průměrné denní teploty měřené na meteorologické stanici ve Svatoňovicích mezi roky 2005 až 2007. Měření prováděno vždy 15. dne v měsíci. Den 0 odpovídá datu 1.1.2005. Červená křivka je odhadnuta z naměřených hodnot metodou nejmenších čtverců.

Teplota $T(t)$ [st. Celsia]	-1.00	-4.75	0.25	12.25	...	-1.63	-1.88
Období t [den]	14	46	74	105	...	1029	1059

Tabulka 2: Vybrané hodnoty z časové řady popisující vývoj průměrné denní teploty měřené na meteorologické stanici ve Svatoňovicích mezi roky 2005 až 2007. Měření prováděno vždy 15. dne v měsíci. Den 0 odpovídá datu 1.1.2005.

Vaším úkolem je odhadnout, jaká byla teplota i v ostatních dnech, pro něž Vám kolega měření nedodal. Tuto úlohu vyřešíte tak, že opět proložíte (interpolujete) množinu naměřených teplot funkcí, kterou pak lze použít k dopočítaní hodnot v bodech, kde měření nemáme k dispozici.

Z grafu 2 je jasně vidět, že závislost teploty na čase není v žádném případě lineární jako tomu bylo v předešlé úloze. Naopak, tato závislost je silně nelineární a je vidět, že se teplota mění cyklicky s periodou jednoho roku (tj. každých 365 dní).

[2]Myšlenka příkladu převzata ze Steve Schaffer: Case-study Least-Square solutions `http://infohost.nmt.edu/~schaffer/www_prevclasses/M254/14_LeastSqs/`

Jako mnohem rozumnější se tedy například jeví funkce

$$\hat{T}(t) = x_0 + x_1 t + x_2 \sin(\omega t) + x_3 \cos(\omega t) \ , \tag{4}$$

kde $\hat{T}(t)$ je odhadovaná teplota v čase t, $\omega = \frac{2\pi}{365}$ a $x_0, \ldots, x_3 \in \mathbb{R}$ jsou neznámé parametry. Proč je vhodné použít zrovna funkci tohoto tvaru ukážeme, respektive Vy si dokážete, v úloze popsané dále. Pro teď berme tvar funkce jako daný a zabývejme se pouze nalezením neznámých parametrů $(x_0, \ldots, x_3)$. Budeme opět hledat takové parametry, při nichž bude funkce $\hat{T}(t)$ co nejvíce odpovídat naměřeným hodnotám $\{(T(t_1), t_1), \ldots, (T(t_n), t_n)\}$ (viz. graf 2 a tabulka 2). A opět použijeme metodu nejmenších čtverců, tj. chce nalézt parametry $(x_0, \ldots, x_3)$, které řeší úlohu

$$(x_0^*, \ldots, x_3^*) = \operatorname*{argmin}_{x_0 \in \mathbb{R}, \ldots, x_3 \in \mathbb{R}} F(x_0, \ldots, x_3) \tag{5}$$

kde

$$\begin{aligned} F(x_0, \ldots, x_3) &= \sum_{i=1}^{n} \left(\hat{T}(t_i) - T(t_i) \right)^2 \\ &= \sum_{i=1}^{n} \left(x_0 + x_1 t_i + x_2 \sin(\omega t_i) + x_3 \cos(\omega t_i) - T(t_i) \right)^2 \ . \end{aligned}$$

Jak je vidět, i tento problém je ekvivalentní s úlohou řešení soustavy lineárních rovnic $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ metodou nejmenších čtverců, tj. lze převést na úlohu

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^4} \|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|^2 \ . \tag{6}$$

Úkoly k vypracování

1. Stáhněte si textový soubor `http://cmp.felk.cvut.cz/cmp/courses/OPT/cviceni/06/teplota.txt`. Nahrajte data do Matlabu pomocí příkazů:

```
Data = load('teplota.txt','-ascii');
Den = Data(:,1);
Teplota = Data(:,2);
```

2. Převeďte úlohu (5) na úlohu (6). Napište přesnou formu matice $\mathbf{A}$ a vektoru $\boldsymbol{b}$ vystupujících v úloze (6).

   Řešení (1 bod):

$$\mathbf{A} = \begin{bmatrix} 1 & t_1 & \sin(\omega t_1) & \cos(\omega t_1) \\ 1 & t_2 & \sin(\omega t_2) & \cos(\omega t_2) \\ \vdots & & & \\ 1 & t_m & \sin(\omega t_m) & \cos(\omega t_m) \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} T(t_1) \\ T(t_2) \\ \vdots \\ T(t_m) \end{bmatrix}$$

3. Použijte metodu nejmenších čtverců k odhadu parametrů funkce (4) z dat nahraných v bodě 1. Do jednoho grafu zobrazte vstupní data a odhadnutou funkci.

   Řešení (1 bod): Obrázek 2.

4. Pro odhadnuté parametry spočtěte hodnotu kriteriální funkce $F(x_0, \ldots, x_3)$.

   Řešení (1 bod): $F(x_0^*, \ldots, x_3^*) = 388.640151$.

5. Z grafu 2 je vidět, že závislost naměřených teplot zhruba odpovídá sinusoidě superponované na lineární funkci

$$\hat{G}(t) = y_0 + y_1 t + A \sin(\omega t + \phi) \ .$$

Lineární funkce $y_0 + y_1 t$ modeluje sklon sinusoidy daný např. globálním oteplováním. Perioda sinusoidy je očividně 365 dní, tj. $\omega = \frac{2\pi}{365}$, zatímco její amplituda A a fáze ϕ jsou neznámé. Neznámé parametry jsou tedy čísla $y_0, y_1, A \in \mathbb{R}$ a $\phi \in (0, 2\pi]$. Metodu (lineárních) nejmenších čtverců popsanou výše nelze pro takto definovanou funkci použít, protože hodnota odhadované funkce závisí na parametru ϕ nelineárně. My jsme namísto funkce $\hat{G}(t)$, použili funkci $\hat{T}(t)$, definovanou rovnicí (4), která závisí na všech svých parametrech lineárně. Fitování funkce $\hat{T}(t)$ lze ospravedlnit tím, že pro každou čtveřici (y_0, y_1, A, ϕ) existuje čtveřice $(x_0, \ldots, x_3)$ taková, že obě funkce jsou shodné, tj. že platí $\hat{T}(t) = \hat{G}(t)$, $\forall t \in \mathbb{R}$. Vaším úkolem je toto tvrzení dokázat. Nápověda: vzpomeňte si, jak lze zapsat $\sin(\alpha + \beta)$.

Řešení (3 body):

$$\begin{aligned}
\hat{G}(t) &= y_0 + y_1 t + A \sin(\omega t + \phi) \\
&= y_0 + y_1 t + A \sin(\omega t) \cos(\phi) + A \cos(\omega t) \sin(\phi)
\end{aligned}$$

po dosazení $x_0 = y_0$, $x_1 = y_1$, $x_2 = A\cos(\phi)$, $x_3 = A\sin(\phi)$ do (4) vidíme, že $\hat{T}(t) = \hat{G}(t)$, $\forall t \in \mathbb{R}$.