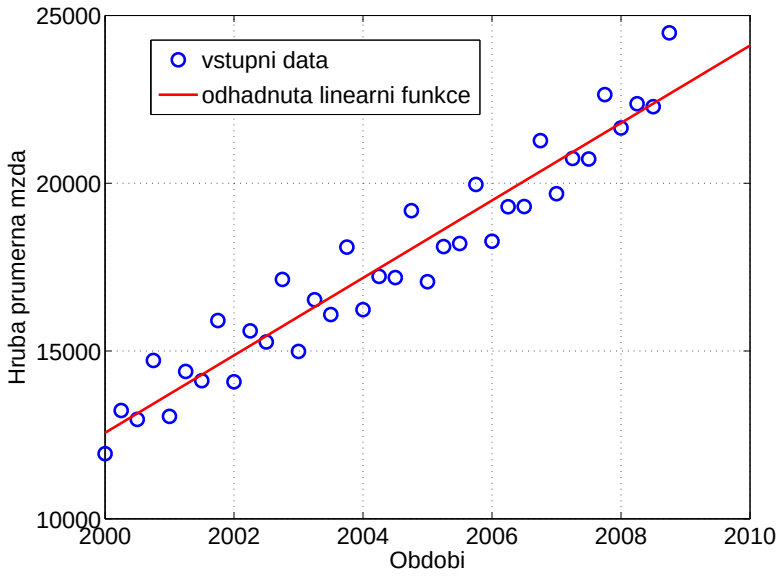
Labs from Optimization:
Least squares method

Ivana Duricova, ZS2015

Excercise 1: Forecast of average gross wage

Graph 1 shows the development of average gross wage (AWG) in Czech Republic in the period from year 2000 to 2008 [1]. To illustrate some values from chart 1 are shown in table 1.



Obrázek 1: The blue points correspond to average gross wage measured in all quarters from year 2000 to 2008. Red line is estimated from data using least squares method.

Our goal will be to predict value of AWG in the second quarter of the year 2009, ie. period for which you have no available data. To solve this excercise we have to

[1] Data were downloaded from Czech statistical office webpage `www.czso.cz/`

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

Tabulka 1: Selected values from the time series describing development of AWG at a time. The time is formated as $t = \text{year} + (\text{quarter} - 1)/4$, where year $\in \{2000, \ldots, 2008\}$ and quarter $\in \{1, 2, 3, 4\}$.

firstly find the function that best matches the specified data of AWG. After locating function than we used to estimate the AWG in the required time. From figure 1 it can be seen that the trend of AWG versus time period is almost linear. Thus we seek a linear function

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

where $\hat{M}(t)$ is estimated AWG in time t and $x_0, x_1 \in \mathbb{R}$ are unknown parameters. Parameters (x_0, x_1) determinated from the data least squares method. Our data (see graph 1 and table 1) is a set $\{(M(t_1), t_1), \ldots, (M(t_n), t_n)\}$ containing n pairs (AWG, time). Parameters (x_0, x_1) can be found so that the sum of squared deviations of the actual and estimated wage was minimum in specified points. This means that we will solve the minimization task

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

where

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

Your task will be to transfer the task (2) to problem solving systems of linear equations $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ using the least squares method, it means formulate it as excercise

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

Tasks which need to be solved

1. Download the text file `http://cmp.felk.cvut.cz/cmp/courses/OPT/cviceni/06/mzdy.txt`. Upload the data to Matlab using the command:

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

2. Convert the task (2) to task (3). Define the exact form of a matrix $\mathbf{A}$ and vector $\boldsymbol{b}$ performs in task(3).

3. Use the least squares method to estimate the parameters of the linear function (1) from data recorded in paragraph 1. Within one graph to display input data and estimated function.

4. Use the estimated function to predict the AWG in the second quarter of year 2009. For comparison, the actual value of AWG in that quarter was 23,664 Kč.

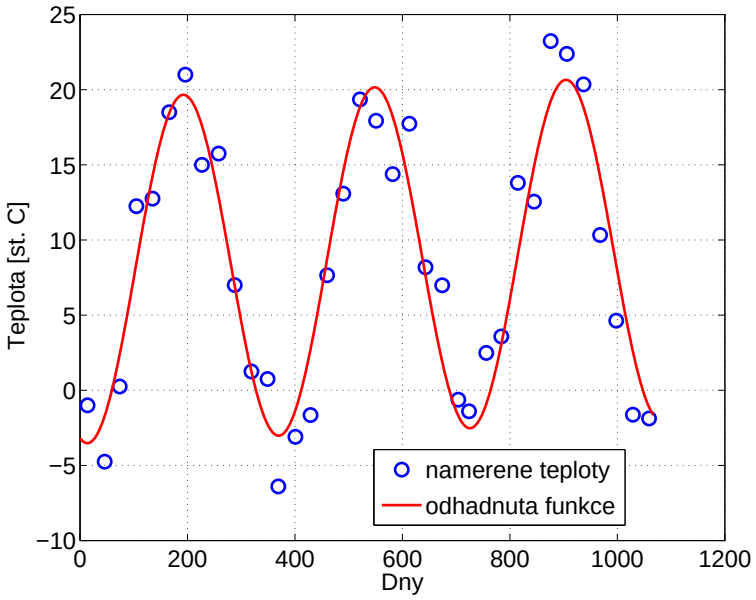
5. For estimated parameters compute the value of objective function $F(x_0, x_1)$.

Excercise 2: Interpolation daytime temperature in Svaton[2]

Imagine that you are working on a weather station in Svaton (`http://www.meteosvatonovice.unas.cz`). From your fellow meteorologist you get measurements of the average

[2]The idea of the example is taken from Steve Schaffer: Case-study Least-Square solutions `http://infohost.nmt.edu/~schaffer/www_prevclasses/M254/14_LeastSqs/`

daily temperature from 2005 to 2007. Your colleague did not carry out measurements every day, but only once a month. Specifically, his temperature every 15th of the month. Measured temperatures are plotted on the graph 2 and selected values are illustrated in table 2.



Obrázek 2: Blue dots indicate the average daily temperature measured at the weather station in Svaton from 2005 to 2007. Measurements performed always the 15th of the month. Day 0 corresponds to the date 1.1.2005. The red curve is estimated from the measured values by least square.

Temperature $T(t)$ [st. Celsius]	-1.00	-4.75	0.25	12.25	...	-1.63	-1.88
Period t [day]	14	46	74	105	...	1029	1059

Tabulka 2: Selected values of the time series describing the evolution of the average daily temperature measured at the weather station Svaton between years 2005 and 2007. Measurements performed always the 15.th of the month. Day 0 corresponds to the date 1.1.2005.

Your task is to estimate what the temperature was in the other day for which your colleague have no measurement. To solve this task so that once again space (interpolated) set of measured temperatures of function, which can than be used for recalculating values at the points where measurements are available.

The graph 2 is clearly seen that the temperature versus time is certainly not linear as in the previous task. Conversely, this dependency is strongly nonlinear and it is seen that the temperature varies cyclically with period of one year (ie. every 365 days). As more reasonable seem to be used function

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

where $\hat{T}(t)$ is estimated temperature in time t, $\omega = \frac{2\pi}{365}$ and $x_0, \ldots, x_3 \in \mathbb{R}$ are unknown parameters. Why should we use exactly the function of the form as it is here you will find out later - in task descripted below. For now, let's take the

shape of function as it is and let's deal with just finding the unknown parameters $(x_0, \ldots, x_3)$. We will again seek such parameters, where the function $\hat{T}(t)$ as much as possible match the measured values $\{(T(t_1), t_1), \ldots, (T(t_n), t_n)\}$ (see chart 2 and table 2). We will use again least squares method, it means that we want to find parameters $(x_0, \ldots, x_3)$, which solve the task

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

where

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

As you can see, even this problem is equivalent to the task of solving a system of linear equations $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$ by least squares method, ie. can be converted to the task

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

Tasks which need to be solved

1. Download the text file `http://cmp.felk.cvut.cz/cmp/courses/OPT/cviceni/06/teplota.txt`. Upload the data to Matlab with using the command:

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

2. Convert the task (5) to task (6). Define the exact form of matricx $\mathbf{A}$ and vector $\boldsymbol{b}$ performers in the task (6).

3. Use the least squares method to estimate the parameters of the function (4) of data recorded in paragraph 1. Within one graph to display input data and estimated function.

4. To compute the value of the parameters estimated cost function $F(x_0, \ldots, x_3)$.

5. In chart 2 is seen that the dependence of the measured temperatures corresponds to a sine wave superimposed on the linear function

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

   Linear function $y_0 + y_1 t$ models such as the slope of the sinusoid for example global warning. The period of the sine wave is apparently 365 days, ie. $\omega = \frac{2\pi}{365}$, while its amptitude A and phase ϕ are unknown. Unknown parameters are therefore the number $y_0, y_1, A \in \mathbb{R}$ and $\phi \in (0, 2\pi]$. Method (linear) least squares described above can not be defined for duch a feature, because the estimated value of the function depends on the parameter ϕ nonlinearly. We use this function $\hat{G}(t)$, instead of this function $\hat{T}(t)$, defined by the equation (4), which depends on all its parameters linearly. Fitting function $\hat{T}(t)$ can be justified by the fact that for each foursome (y_0, y_1, A, ϕ) exists foursome $(x_0, \ldots, x_3)$ such that both functions are identical, ie. that pays $\hat{T}(t) = \hat{G}(t)$, $\forall t \in \mathbb{R}$. Your task is to prove this assertion. Help: remember how to write $\sin(\alpha + \beta)$.