

## Linear regression problem

Let  $y, x_1, x_2, \dots, x_n$  be variables. We believe that dependent variable  $y$  could be approximated by linear combination of independent variables  $x_1, x_2, \dots, x_n$ . That is,

$$y \approx a_1 x_1 + a_2 x_2 + \dots + a_n x_n.$$

The problem is to find coefficients  $a_1, a_2, \dots, a_n$ . In order to find them we conduct  $m$  experiments (samples) and get following dataset of observations

$$\begin{cases} y_1 \approx a_1 x_{11} + \dots + a_n x_{n1} \\ y_2 \approx a_1 x_{12} + \dots + a_n x_{n2} \\ \vdots \\ y_m \approx a_1 x_{1m} + \dots + a_n x_{nm} \end{cases}$$

where  $x_{1i}, x_{2i}, \dots, x_{ni}, y_i$  is a result of  $i$ -th experiment. Clearly, the problem could be equivalent to matrix equation of the form

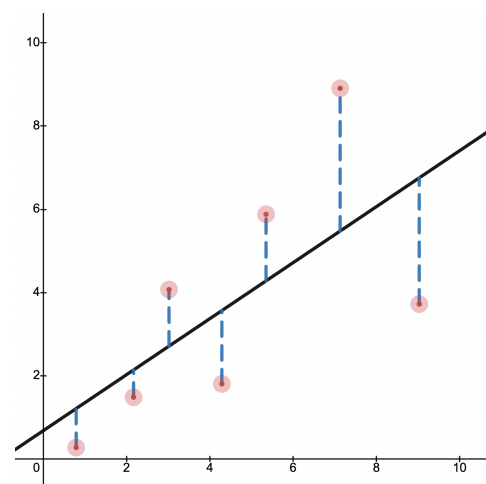
$$X\vec{a} \approx \vec{y},$$

where

$$X = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{nm} \end{bmatrix}, \quad \vec{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

The solution of the problem is least square solution (pseudosolution)

$$\vec{a} = X^+ \vec{y}.$$



In linear regression, the observations (red) are assumed to be the result of random deviations (blue) from an underlying relationship (black) between a dependent variable ( $y$ ) and an independent variable ( $x$ ).

### Problem 1. Linear regression

Find linear approximation of temperature using following dataset of temperature observations. Find temperature on day 4.

#	Day	Temperature
1	1	19°
2	2	18°
3	2	20°
4	3	15°

**Solution:** Let us denote day and temperature of the row  $i$  by  $x_i$  and  $y_i$  respectively. We assume that temperature ( $y(x)$ ) could be approximated by linear model

$$y(x) = ax + b \in \mathbb{R},$$

where  $x \in \mathbb{R}$  is number of the day and  $a, b \in \mathbb{R}$  are unknown coefficients. This means that we need to find pseudosolution of the following system

$$X\vec{a} = \vec{y},$$

where

$$X = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} 19 \\ 18 \\ 20 \\ 15 \end{bmatrix}, \quad \vec{a} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

So we need to evaluate the following  $\vec{a} = X^+ \vec{y}$ . Since  $X$  is full column rank, we get

$$\begin{aligned} X^+ &= (X^* X)^{-1} X^* = \left( 2 \begin{bmatrix} 9 & 4 \\ 4 & 2 \end{bmatrix} \right)^{-1} X^* \\ &= \frac{1}{2} \frac{1}{2} \begin{bmatrix} 2 & -4 \\ -4 & 9 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \\ &= \frac{1}{4} \begin{bmatrix} -2 & 0 & 0 & 2 \\ 5 & 1 & 1 & -3 \end{bmatrix} \end{aligned}$$

Hence, we get

$$\vec{a} = X^+ \vec{y} = \frac{1}{4} \begin{bmatrix} -2 & 0 & 0 & 2 \\ 5 & 1 & 1 & -3 \end{bmatrix} \begin{bmatrix} 19 \\ 18 \\ 20 \\ 15 \end{bmatrix} = \begin{bmatrix} -2 \\ 22 \end{bmatrix}.$$

This means that linear model  $y(x)$  is

$$y(x) = -2x + 22.$$

According to linear model  $y(x)$ , on day 4 temperature will be  $y(4) = 14^\circ$ . ■