# Linear Regression

Prof. Alessandro Lucantonio

Aarhus University - Department of Mechanical and Production Engineering

?/?/2023

# Weight-Height example

Dataset: heights and weights of different people.

Task: build a model that predict the height given the weight.

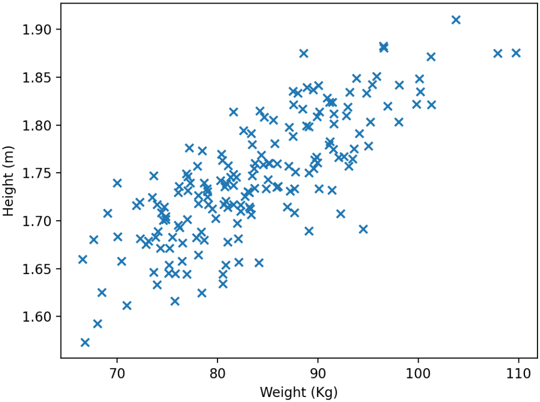


Figure: Data plot

# A solution - Linear regression model

Some remarks on data.

- Regression problem (continuous output).
- Data with different order of magnitude.

A possible solution to this problem is represented by **linear regression** (LR).

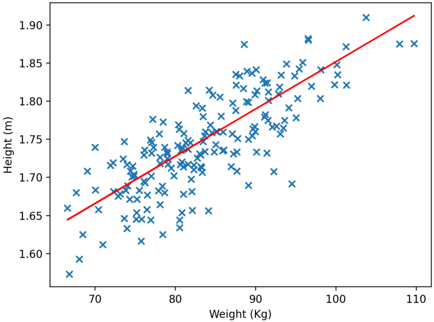


Figure: Trained model (in red)

# General ingredients

Notation:

- $x$: a data sample.
- $y$: the data target corresponding to $x$
- $N$: number of data.

Model/hypothesis: $h_{\boldsymbol{w}}(x) = w_1 x + w_0$, where $\boldsymbol{w} = [w_0, w_1]$ is the vector of parameter that has to be learned.
In our example, $x$ is the weight of a single sample and $h_{\boldsymbol{w}}(x)$ corresponds to the prediction of its height.

Usually the vector $\boldsymbol{w}$ is called **weights vector** and the set $\mathcal{H} := \{h_{\boldsymbol{w}} | \boldsymbol{w} \in \mathbb{R}^2\}$ is called **hypothesis space**.

How to learn $w$ from data?

# Mean squared error (MSE)

Given a training sample $x_i$ and a model $h_{\boldsymbol{w}}$ we can predict the target computing $h_{\boldsymbol{w}}(x_i)$. To evaluate how good is the prediction we compute the error $(h_{\boldsymbol{w}}(x_i) - y_i)^2$.

$(h_{\boldsymbol{w}}(x_i) - y_i)^2 \geq 0$ and $(h_{\boldsymbol{w}}(x_i) - y_i)^2 = 0$ if and only if $h_{\boldsymbol{w}}(x_i) = y_i$. The **mean squared error** (MSE) is:

$$E(\boldsymbol{w}) := \frac{1}{N} \sum_{i=1}^{N} (h_{\boldsymbol{w}}(x_i) - y_i)^2.$$

To find the best model we minimize the training error, hence in this case the MSE.

$$\boldsymbol{w} \in \arg\min_{\tilde{\boldsymbol{w}} \in \mathbb{R}^2} E(\tilde{\boldsymbol{w}}).$$

# $n$-dimensional LR

*Dataset samples.*

Previous case: $x \in \mathbb{R}, y \in \mathbb{R}$.

Now: $\boldsymbol{x} \in \mathbb{R}^n, y \in \mathbb{R}$.

Notation: $x_j^i$ is the $j$-th coordinate of the $i$-th sample.

*Hypothesis.*

Previous case:

$$h_w(x) = w_1 x + w_0,$$

where $w = [w_0, w_1]$.

Now:

$$\begin{aligned} h_{\boldsymbol{w}}(\boldsymbol{x}) &= w_n x_n + w_{n-1} x_{n-1} + \cdots + w_1 x_1 + w_0 \\ &= \sum_{i=0}^{n} w_i \tilde{x}_i = \boldsymbol{w}^T \tilde{\boldsymbol{x}}, \end{aligned}$$

where $\boldsymbol{w} = [w_0, \ldots, w_n]$ and $\tilde{\boldsymbol{x}} = [1, x_1, \ldots, x_n]$.

# n-dimensional LR

*MSE.*

Previous case:

$$E(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}(h_{\boldsymbol{w}}(x_i) - y_i)^2.$$

Now:

$$\begin{aligned} E(\boldsymbol{w}) &= \frac{1}{N}\sum_{i=1}^{N}(h_w(\boldsymbol{x}^i) - y^i)^2 \\ &= \frac{1}{N}(\mathrm{X}\boldsymbol{w} - \boldsymbol{y})^T(\mathrm{X}\boldsymbol{w} - \boldsymbol{y}) \\ &= \frac{1}{N}||\mathrm{X}\boldsymbol{w} - \boldsymbol{y}||^2 \end{aligned}$$

where

$$\mathrm{X} := \begin{bmatrix} \tilde{\boldsymbol{x}}^1 \\ \vdots \\ \tilde{\boldsymbol{x}}^N \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}.$$

# Spot the minimum - Gradient descent

How to find $\boldsymbol{w} \in \arg\min_{\tilde{\boldsymbol{w}} \in \mathbb{R}^2} E(\tilde{\boldsymbol{w}})$?
Main idea:

- Start with a random $\boldsymbol{w}^0$.
- For $j \geq 0$, update $\boldsymbol{w}^{j+1} := \boldsymbol{w}^j + \boldsymbol{d}^j$, where $\boldsymbol{d}^j$ is such that

$$E(\boldsymbol{w}^{j+1}) \leq E(\boldsymbol{w}^j)$$

Gradient descent: $\boldsymbol{d}^j = -\alpha \nabla E(\boldsymbol{w}^j)$. $\alpha$ is called **learning rate**.