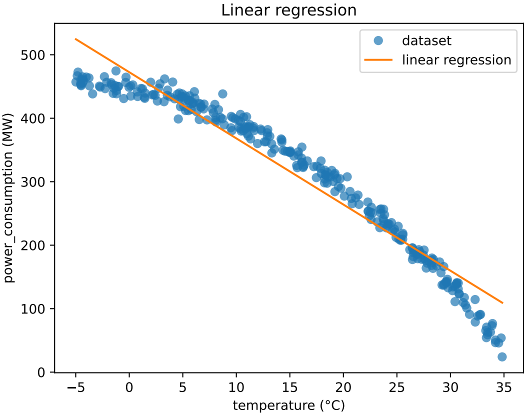
Solution to the linear regression in dimension 1



Linear regression

Formalization :

- input space (temperature) : $\mathcal{X} = \mathbb{R}$
- output space (power consumption) : $\mathcal{Y} = \mathbb{R}$
- dataset : $D_n = \{(x_1, y_1), \dots, (x_n, y_n), i \in [1, n]\}$.

When doing linear regression, our estimator is of the form :

$$h(x) = \theta x + b \tag{1}$$

with $\theta \in \mathbb{R}$, $b \in \mathbb{R}$.

Empirical risk minimization

$$R_n(\theta, b) = \frac{1}{n}\sum_{i=1}^{n}(\theta x_i + b - y_i)^2 \tag{2}$$

We want to find θ and b such that $R_n(\theta, b)$ has the **smallest possible value**.

Since $(\theta, b) \mapsto R_n(\theta, b)$ is convex, finding its global minimum is equivalent to finding the points where the gradient cancels, which means that the partial derivatives with respect to both θ and b are equal to 0.

Derivatives

We note that if we look for the (θ, b) values that cancel the gradient, we can remove the $\frac{1}{n}$ factor without changing the result.

$$\begin{aligned}
\frac{\partial R_n}{\partial \theta}(\theta, b) &= \sum_{i=1}^{n} 2(\theta x_i + b - y_i)x_i \\
&= 2[\theta \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i]
\end{aligned} \tag{3}$$

$$\begin{aligned}
\frac{\partial R_n}{\partial b}(\theta, b) &= \sum_{i=1}^{n} 2(\theta x_i + b - y_i) \\
&= 2[\theta \sum_{i=1}^{n} x_i + nb - \sum_{i=1}^{n} y_i]
\end{aligned} \tag{4}$$

Hence we have a system of 2 equations with 2 unknowns (dropping the θ^* notation)

$$\theta \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0 \tag{5}$$

$$\theta \sum_{i=1}^{n} x_i + nb - \sum_{i=1}^{n} y_i = 0 \tag{6}$$

Which means

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - \theta \sum_{i=1}^{n} x_i) \tag{7}$$

$$\theta \sum_{i=1}^{n} x_i^2 + \frac{1}{n}(\sum_{i=1}^{n} y_i - \theta \sum_{i=1}^{n} x_i) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0 \tag{8}$$

Finally :

$$\theta(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2) + \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i = 0 \qquad (9)$$

or

$$\theta^* = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}[\sum_{i=1}^{n} x_i]^2} \qquad (10)$$