

5 Linear Regression

In linear regression, we have a real valued measurement Y of a signal \mathbf{x} (potentially a vector) that we want to measure, potentially corrupted by noise. Linear regression is ubiquitous as a component of several algorithms, but a commonplace standalone (and important) example is fMRI signals. We will assume that the measurements are linear (*i.e.*, the signal \mathbf{x} is transformed by linear operations, which is equivalent to multiplying by a matrix in general).

This is quite a vast topic in itself, and this module covers what is known as Ordinary Least Squares. In this section, we focus on the "frequentist" view that we tackle with the *Maximum Likelihood* (ML) principle. You can contrast this with the *Bayesian* approach, which will be an juxtaposition on Maximum Likelihood approach (see Ridge Regression).

5.1 Frequentist: Maximum Likelihood setup

Let $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$ be an arbitrary $k \times 1$ vector (called the the *model*, which we want to estimate). We linearly *measure* \mathbf{x} n times, and assume

$$y_i = [b_{i1} \ \cdots \ b_{ik}] \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} + Z_i, \quad 1 \leq i \leq n$$

where Z_i is a normal random variable (we will assume mean 0 and variance σ^2). The b_{i1}, \dots, b_{ik} , $1 \leq i \leq n$ are fixed numbers that we control, and are called the *measurements*. The output y is called the *target*. The variables Z_i , $1 \leq i \leq n$ are independent.

The n equations above can be written in one compact matrix form,

$$Y = BX + Z,$$

where Y is a vector of the n targets, B is the $n \times k$ measurement matrix, whose i 'th row is $[b_{i1} \ \cdots \ b_{ik}]$ from the equation above, and Z is the vector of the n Gaussian random variables. The columns of B are often called *features* or *attributes* (we will primarily use the word features).

A subtle point to note here is that we may not truly believe that the target y is linearly related to \mathbf{x} , but may choose to use this approach anyway. For example, in the Boston housing example attached to this module, Y the median price of a house in a neighborhood is predicted using features related

to the neighborhood, such as the average age of houses therein, average acreage of plots, zoning district, etc. We don't truly believe the median value of the house in the i 'th neighborhood, y_i , is a linear function of the features b_{i1}, \dots, b_{i13} measured in that neighborhood. But we use it anyway, because linear methods have a lot going for them as we will see.

Given our observations of the target Y , measurements B and given any choice of \mathbf{x} , we have

$$Z = Y - B\mathbf{x}.$$

Recall that we use $\|Z\|$ for the Euclidean length or ℓ_2 norm of the vector Z , and $\|Z\|^2 = Z^T Z$. See the module on norms.

For this choice of \mathbf{x} then, the likelihood $L(Z|\mathbf{x})$ of the corresponding Z is

$$\begin{aligned} L(Z|\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}Z^T Z\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|Z\|^2\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|Y - B\mathbf{x}\|^2\right). \end{aligned}$$

The Maximum Likelihood principle asks us to choose the value of \mathbf{x} that maximizes the likelihood $L(Z|\mathbf{x})$. From the above series of equalities, maximizing the likelihood is equivalent to minimizing $\|Y - B\mathbf{x}\|^2$, or minimizing the length of the vector $Y - B\mathbf{x}$.

This is the Ordinary Least Squares (OLS) formulation. We choose the vector \mathbf{x}_{OLS} satisfying

$$\mathbf{x}_{OLS} = \arg \min_{\mathbf{x} \in \mathbb{R}^k} \|Y - B\mathbf{x}\|^2.$$