



Why

What is the best linear predictor if we choose according to a squared loss function.

Definition

Suppose we have a paired dataset of records with inputs in \mathbf{R}^d and outputs in \mathbf{R} . A *least squares linear predictor* or *linear least squares predictor* is a linear transformation $f : \mathbf{R}^d \rightarrow \mathbf{R}$ which minimizes

$$\frac{1}{n} \sum_{i=1}^n (f(x^i) - y^i)^2.$$

over a dataset of pairs $(x^1, y^1), \dots, (x^n, y^n)$, and we want to select the predictor f to minimize. The space of linear functions is in one-to-one correspondence with vectors in \mathbf{R}^d . So we want to find $\theta \in \mathbf{R}^d$ to minimize

$$\frac{1}{n} \|X\theta - y\|^2.$$

Proposition 1. *There exists a unique linear least squares predictor and its parameters are given by $(X^\top X)^{-1} X^\top y$.*

