



## Why

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

## Definition

Let  $X \in \mathbf{R}^{n \times d}$  and  $y \in \mathbf{R}^d$ . In other words, we have a paired dataset of records with inputs in  $\mathbf{R}^d$  (the rows of  $X$ ) and outputs in  $\mathbf{R}$  (the elements of  $y$ ).

A *least squares linear predictor* or *linear least squares predictor* is a linear transformation  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  (the field is  $\mathbf{R}$ ) which minimizes

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

over the dataset of pairs  $(x^1, y_1), \dots, (x^n, y_n) \in \mathbf{R}^d \times \mathbf{R}$  where  $(x^i)^\top$  is the  $i$ th row of  $X$  for  $i = 1, \dots, n$ . The set of linear functions from  $\mathbf{R}^d$  to  $\mathbf{R}$  is in one-to-one correspondence with  $\mathbf{R}^d$ . So we want to find  $\theta \in \mathbf{R}^d$  to minimize

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

## Solution

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

---

<sup>1</sup>Future editions will include an account.



