Problem 1: Can you fit a line?

Problem Statement

Consider N data points $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, N$ obtained from an experiment , where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. Let $N > n+1$. The goal of linear regression is to find the "best" linear fit; i.e. find $\mathbf{c} \in \mathbb{R}^n$, $d \in \mathbb{R}$ s.t.:

$$y_i \approx \mathbf{c}^T \mathbf{x}_i + d, \text{ for } i = 1, \ldots, N$$

a

Suppose each measurement is corrupted by independent Gaussian noise with identical variances and 0 mean. Find $\mathbf{A} \in \mathbb{R}^{N\times(n+1)}$ in terms of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ s.t. that the maximum likelihood estimate of $\hat{\mathbf{c}}$, $\hat{d}$ is

$$\begin{pmatrix} \hat{\mathbf{c}} \\ \hat{d} \end{pmatrix} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1} \mathbf{A}^T y, \text{ where } \mathbf{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

b

Consider the case with $n = 1$; i.e. $x_1, \ldots, x_N$ are scalars. Finding the best linear fit amounts to finding $(\hat{c}, \hat{d})$ tht minimizes:

$$J(c, d) := \sum_{i=1}^{N} (y_i - cx_i - d)^2$$

Compute a stationary point $(\hat{c}, \hat{d})$ by setting the derivative of J w.r.t c and d to zero.

c

Use the second derivative tests to conclude that your $(\hat{c}, \hat{d})$ computed in (b) is indeed a local minimizer of J. Can you conclude from this second derivative test alone that it is a local minimizer?

d

Consider the following (x, y) pairs:

- $(1.00, 1.10)$
- $(1.50, 1.62)$
- $(2.00, 1.98)$
- $(2.57, 2.37)$
- $(3.00, 3.23)$
- $(3.50, 3.69)$
- $(4.00, 3.97)$

Draw a scatter plot of these points; then, plot the best linear fit to these points using your formula in part (b).

Solution

a

b