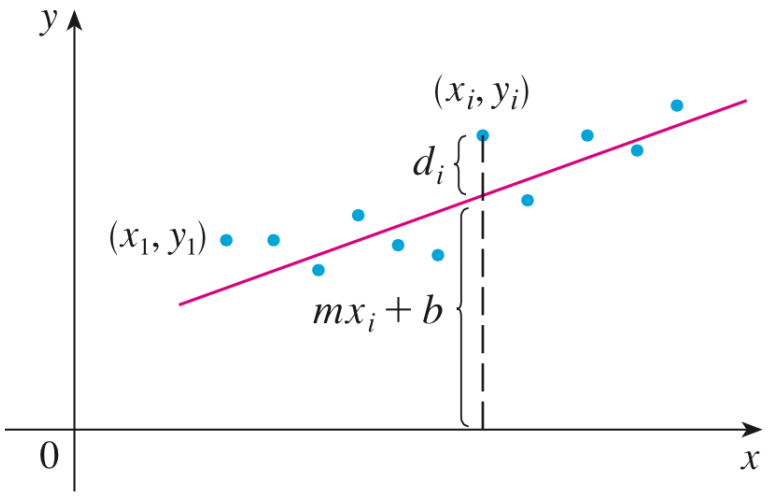
Objectives

☐ I know that finding a best-fit line/curve is an optimization problem.

In some sense, we've been using optimization all along in this class. The process of fitting data to a line (or curve) is exactly an optimization process known as the **Method of Least Squares**. In the example below, we explain the set up and its connection with the equation

$$A^T A\vec{x} = A^T\vec{b}$$

Example 24.1: (How We Fit Data to a Line.) Suppose you have a set of points $(x_1, y_1), \ldots (x_n, y_n)$ and you believe these points follow a linear relationship $y = ax + b$. How can we find the best values for the coefficients a and b so that it fits the data? We try to minimize the distance between each observed value, y_i, and the predicted value based on the line, $ax_i + b$. We can define these distances as $d_i = y_i - (ax_i + b)$.



The **method of least squares** is what we use to fit a line to data. It turns out to be an optimization problem. That is, we try to pick a and b so that

$$S = d_1^2 + d_2^2 + \ldots + d_n^2$$

is minimized!

Show that S is minimized when

$$a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i$$

and

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i.$$

Thus, the best fit line is found by solving these two equations with two unknowns, a and b.

Solution 24.2: Based on the description, the problem is to minimize

$$S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

We, therefore, want to find values of a and b so that

$$\nabla S = \begin{pmatrix} S_a \\ S_b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

So let's calculate each partial derivative.

$$\begin{aligned} 0 = S_a &= \left[\sum_{i=1}^{n} (y_i - ax_i - b)^2 \right]_a \\ &= \sum_{i=1}^{n} 2(y_i - ax_i - b) \cdot (-x_i) \\ &= -2 \sum_{i=1}^{n} x_i (y_i - ax_i - b) \end{aligned}$$

So from $0 = S_a$, we get

$$0 = -2 \sum_{i=1}^{n} x_i (y_i - ax_i - b) \implies a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i.$$

Now let's calculate the remaining partial derivative.

$$\begin{aligned} 0 = S_b &= \left[\sum_{i=1}^{n} (y_i - ax_i - b)^2 \right]_b \\ &= \sum_{i=1}^{n} 2(y_i - ax_i - b) \cdot (-1) \\ &= -2 \sum_{i=1}^{n} (y_i - ax_i - b) \end{aligned}$$

So from $0 = S_b$, we get

$$0 = -2 \sum_{i=1}^{n} (y_i - ax_i - b) \implies a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} 1 = \sum_{i=1}^{n} y_i$$

Notice that $\sum_{i=1}^{n} 1 = 1 + 1 + \ldots + 1 = n$. Hence, we get

$$a \sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i$$

So we have proved the claim.

Summary of Ideas

- The method of least squares, which is what we use when we find best-fit curves, is an optimization problem.