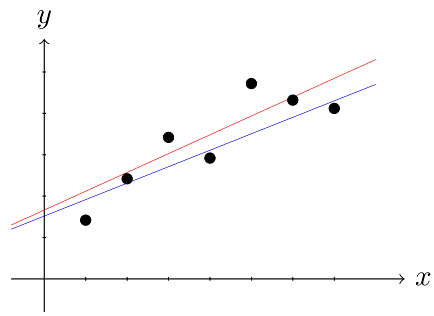
Suppose we collect some sample data.

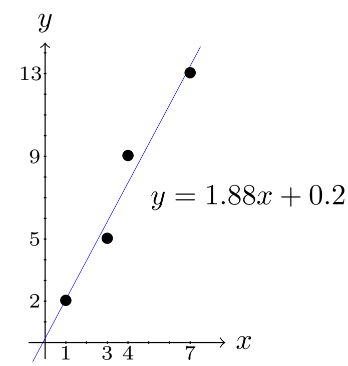
We believe the connection between the inputs and outputs is linear.

How do we find the slope and y-intercept of the best line? (Various ways to determine "best".)



We make the assumption that the inputs are known.

Find the line of best fit that goes through the points $(1, 2)$, $(3, 5)$, $(4, 9)$, and $(7, 13)$.



We seek a line of the form $y = mx + b$, where m and b are determined by the inconsistent system

$$\begin{aligned} 2 &= m + b; \\ 5 &= 3m + b; \\ 9 &= 4m + b; \\ 13 &= 7m + b; \end{aligned} \qquad \begin{aligned} b + m &= 2; \\ b + 3m &= 5; \\ b + 4m &= 9; \\ b + 7m &= 13; \end{aligned} \qquad \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ 9 \\ 13 \end{bmatrix} \qquad A\mathbf{x} = \mathbf{b}$$

$$A^T A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 7 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \end{bmatrix} = \begin{bmatrix} 4 & 15 \\ 15 & 75 \end{bmatrix} \quad \text{and} \quad A^T\mathbf{b} = \begin{bmatrix} 29 \\ 144 \end{bmatrix} \quad \text{then solve} \quad A^T A\mathbf{x} = A^T\mathbf{b}$$

$$x_1 = \frac{\begin{vmatrix} 29 & 15 \\ 144 & 75 \end{vmatrix}}{\begin{vmatrix} 4 & 15 \\ 15 & 75 \end{vmatrix}} = \frac{15(145 - 144)}{15(20 - 15)} = \frac{1}{5}; \qquad x_2 = \frac{\begin{vmatrix} 4 & 29 \\ 15 & 144 \end{vmatrix}}{\begin{vmatrix} 4 & 15 \\ 15 & 75 \end{vmatrix}} = \frac{576 - 435}{75(4 - 3)} = \frac{141}{75} = \frac{47}{25}$$

The line of best fit is $y = \dfrac{47}{25}x + \dfrac{1}{5} = 1.88x + 0.2$.

Consider the general case for linear problems of this type.

$$\begin{aligned} b + x_1 m &= y_1 \\ b + x_2 m &= y_2 \\ &\vdots \\ b + x_n m &= y_n \end{aligned} \qquad \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad A\mathbf{x} = \mathbf{b}$$

$$A^T A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}, \quad A^T\mathbf{b} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$b = \frac{\begin{vmatrix} \sum_{i=1}^n y_i & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i & \sum_{i=1}^n x_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix}} = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}, \quad m = \frac{\begin{vmatrix} n & \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix}} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

Let $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$. We claim the line of best fit passes through the point $(\overline{x}, \overline{y})$.

The projection of $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ onto the column space of $\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ is the vector $\begin{bmatrix} mx_1 + b \\ mx_2 + b \\ \vdots \\ mx_n + b \end{bmatrix}$.

It follows that $\begin{bmatrix} y_1 - mx_1 - b \\ y_2 - mx_2 - b \\ \vdots \\ y_n - mx_n - b \end{bmatrix}$ is in the orthogonal complement of $\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$.

In particular, it is orthogonal to the first column consisting of all 1's. It follows that

$$0 = \sum_{i=1}^n (y_i - mx_i - b) = \sum_{i=1}^n y_i - m\sum_{i=1}^n x_i - \sum_{i=1}^n b = n\overline{y} - mn\overline{x} - nb = n(\overline{y} - m\overline{x} - b)$$

It follows that $\overline{y} = m\overline{x} + b$.

Hence, if we can find m some easier way, then b can be determined by the above equation.

Note that the matrices $\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ and $\begin{bmatrix} 1 & x_1 - \overline{x} \\ 1 & x_2 - \overline{x} \\ \vdots & \vdots \\ 1 & x_n - \overline{x} \end{bmatrix}$ have the same column space.

The second matrix has orthogonal columns.

The projection of $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ onto $\begin{bmatrix} 1 & x_1 - \overline{x} \\ 1 & x_2 - \overline{x} \\ \vdots & \vdots \\ 1 & x_n - \overline{x} \end{bmatrix}$ is

$$\frac{y_1 + y_2 + \cdots + y_n}{1 + 1 + \cdots + 1} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \frac{y_1(x_1 - \overline{x}) + y_2(x_2 - \overline{x}) + \cdots + y_n(x_n - \overline{x})}{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2} \begin{bmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{bmatrix}$$

$$\big(\overline{y} - m\overline{x}\big) \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + m \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{where } m = \frac{\sum_{i=1}^{n} y_i(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \qquad b = \overline{y} - m\overline{x}$$

We have used the fact that $\sum_{i=1}^{n} \overline{y}(x_i - \overline{x}) = \overline{y} \sum_{i=1}^{n} (x_i - \overline{x}) = \overline{y}\big(n\overline{x} - n\overline{x}\big) = 0.$

Find the line of best fit that goes through the points $(1, 2)$, $(3, 5)$, $(4, 9)$, and $(7, 13)$.

$$\overline{x} = \tfrac{15}{4}, \ \overline{y} = \tfrac{29}{4} \qquad \frac{1}{4}\left(\begin{bmatrix} 4 \\ 12 \\ 16 \\ 28 \end{bmatrix} - \begin{bmatrix} 15 \\ 15 \\ 15 \\ 15 \end{bmatrix} \right) = \frac{1}{4} \begin{bmatrix} -11 \\ -3 \\ 1 \\ 13 \end{bmatrix}, \qquad \begin{bmatrix} 2 \\ 5 \\ 9 \\ 13 \end{bmatrix}$$

$$m = \frac{\frac{1}{4}(-22 - 15 + 9 + 169)}{\frac{1}{16}(121 + 9 + 1 + 169)} = \frac{4 \cdot 141}{300} = \frac{141}{75} = \frac{47}{25}, \qquad b = \frac{29}{4} - \frac{47}{25} \cdot \frac{15}{4} = \frac{145}{20} - \frac{141}{20} = \frac{1}{5}$$

We have found two expressions for the slope m.

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}\big((x_i-\overline{x})(y_i-\overline{y})\big)}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

For the numerator, we find that

$$\sum_{i=1}^{n}\big((x_i-\overline{x})(y_i-\overline{y})\big) = \sum_{i=1}^{n}\big(x_iy_i - \overline{x}y_i - \overline{y}x_i + \overline{x}\,\overline{y}\big) = \sum_{i=1}^{n} x_iy_i - n\overline{x}\,\overline{y} - n\overline{y}\,\overline{x} + n\overline{x}\,\overline{y}$$

and thus

$$n\sum_{i=1}^{n}\big((x_i-\overline{x})(y_i-\overline{y})\big) = n\sum_{i=1}^{n} x_iy_i - (n\overline{x})(n\overline{y}) = n\sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

For the denominator, we find that

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum_{i=1}^{n}\big(x_i^2 - 2x_i\overline{x} + \overline{x}^2\big) = \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\overline{x}(n\overline{x}) + n\overline{x}^2$$

and thus

$$n\sum_{i=1}^{n}(x_i-\overline{x})^2 = n\sum_{i=1}^{n} x_i^2 - n^2\overline{x}^2 = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$$

CAUTION: These formulas are only for the line of best fit, not other curves.

Consider the quadratic equation $y = a + bx + cx^2$. The matrix form would be

$$\begin{aligned} a + bx_1 + cx_1^2 &= y_1 \\ a + bx_2 + cx_2^2 &= y_2 \\ &\vdots \\ a + bx_n + cx_n^2 &= y_n \end{aligned} \qquad \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad A\mathbf{x} = \mathbf{b}$$

If we seek a plane of the form $z = a + bx + cy$, then we might have something like

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 1 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 7 \\ 9 \\ 12 \\ 10 \\ 17 \\ 16 \end{bmatrix}$$