§4.4 Data fitting by a straight line   Lecture 27

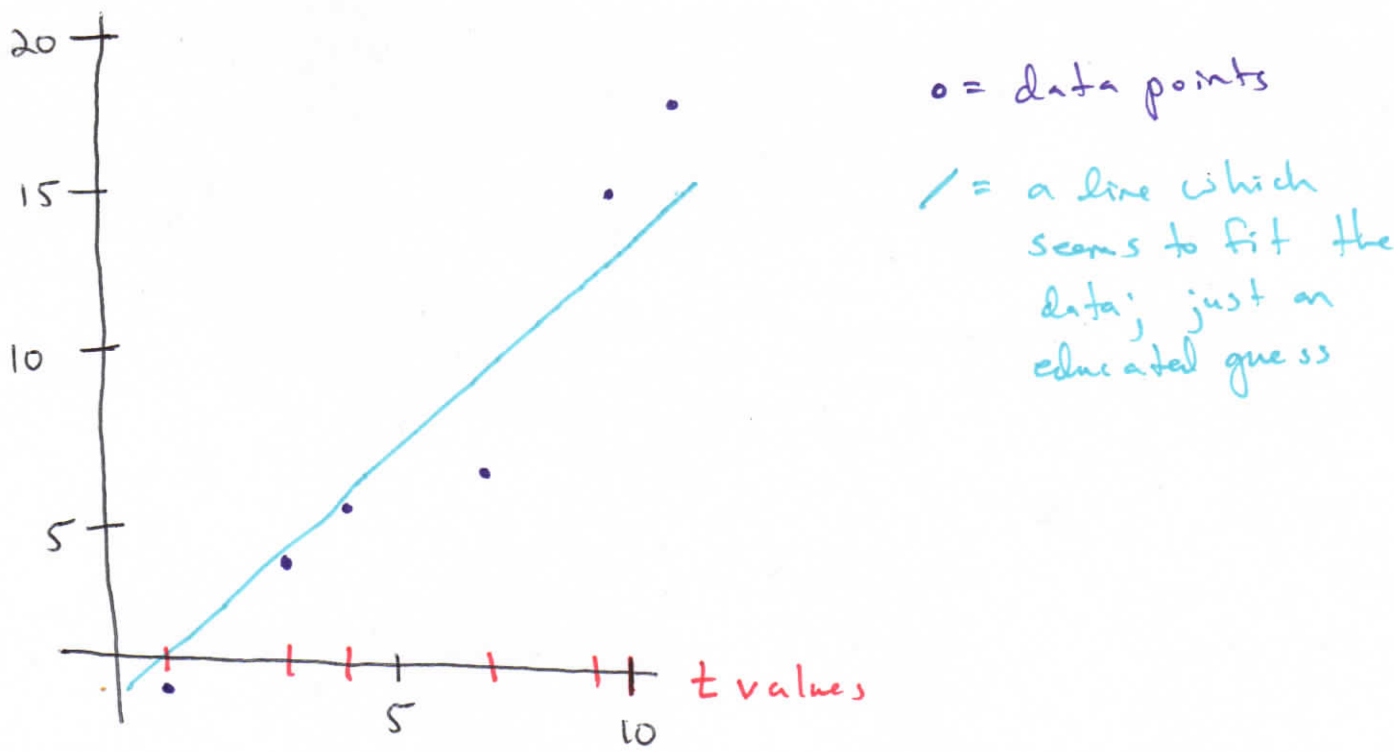
Suppose we are given data points with an input variable t and an output variable y. For example, t may be time and y is distance traveled. How do we use these data points to predict the output variable y for other input values of t? Moreover, how do we do this effectively in the presence of experimental error?

<u>Case 1</u>: The data is "roughly linear"

<u>ex</u> Suppose we have the following data points

t_i	1	3	4	7	9	10
y_i	-1	4	6	9	16	18

We plot these data points below



To make this precise, we do the following. Let

$$y(t) = \alpha_0 + \alpha_1 t$$

be a line, where will attempt to choose the coefficients α_0 and α_1 as well as possible.

Let $(t_1, y_1), (t_2, y_2), \ldots, (t_m, y_m)$ be our measured data points. Let

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

be the vector of measured values, and set

$$A = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix}, \quad x = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}.$$

Then

$$Ax = \begin{pmatrix} \alpha_0 + \alpha_1 t_1 \\ \alpha_0 + \alpha_1 t_2 \\ \vdots \\ \alpha_0 + \alpha_1 t_m \end{pmatrix} = \begin{pmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_m) \end{pmatrix}.$$

The error in the known values ~~win~~ compared to the predicted values is

$$\begin{pmatrix} y(t_1) - y_1 \\ y(t_2) - y_2 \\ \vdots \\ y(t_m) - y_m \end{pmatrix} = Ax - \vec{y}.$$

We will choose our line of best fit according to which line minimizes the norm of the error, $\|Ax - \vec{y}\|$.

Thus, our coefficients $x = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}$ are given by the least squares solution to $Ax = \vec{y}$.

By section 4.3, we conclude that we find our coefficients x by solving the normal equations

$$A^T A x = A^T y.$$

ex (cont.) Using the data points from before, we form

$$A = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \\ 1 & t_5 \\ 1 & t_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \\ 1 & 9 \\ 1 & 10 \end{pmatrix}, \quad \vec{y} = \begin{pmatrix} -1 \\ 4 \\ 6 \\ 9 \\ 16 \\ 18 \end{pmatrix}.$$

We compute

$$A^T A = \begin{pmatrix} 6 & 34 \\ 34 & 256 \end{pmatrix}, \quad A^T y = \begin{pmatrix} 52 \\ 422 \end{pmatrix}.$$

Solving the normal equation

$$A^T A x = A^T y$$

yields

$$\alpha_0 \approx -2.7263 \qquad \alpha_1 \approx 2.0105$$

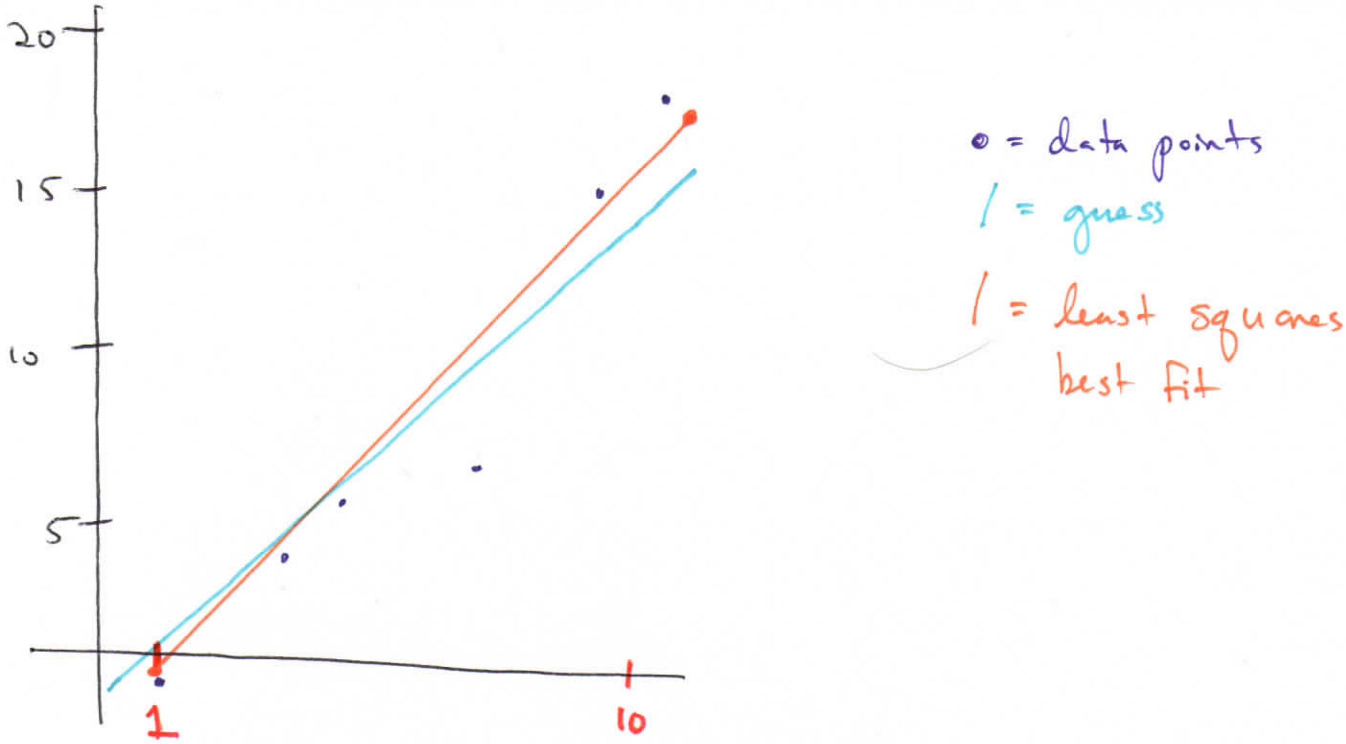
So our line of best fit is

$$y(t) = -2.7263 + 2.0105t.$$

Let's visually compare this with the line we tried to guess. We find two points

$$y(1) = -.7158 \qquad y(10) = 17.3789$$

and plot



We see that our guess is fairly close with the line of best fit. However, our line of best fit also helps us to see that the point (7, 9) is most likely an outlier or subject to high experimental error