Computational Mathematics/Information Technology

Dr Oliver Kerr

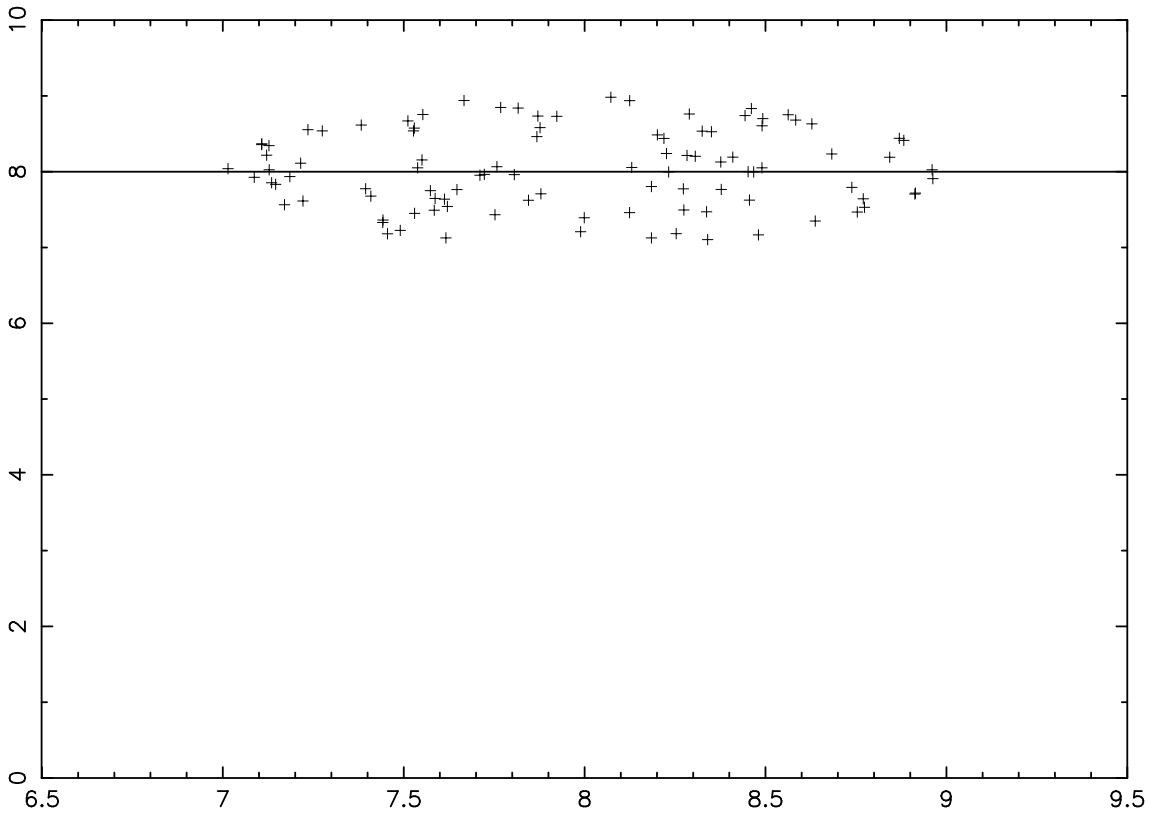
2009–10

Progress Test Results

- Lowest mark: 0%
- Highest mark: 100%
- Average mark: 61.9%
- 15% got less than 40%

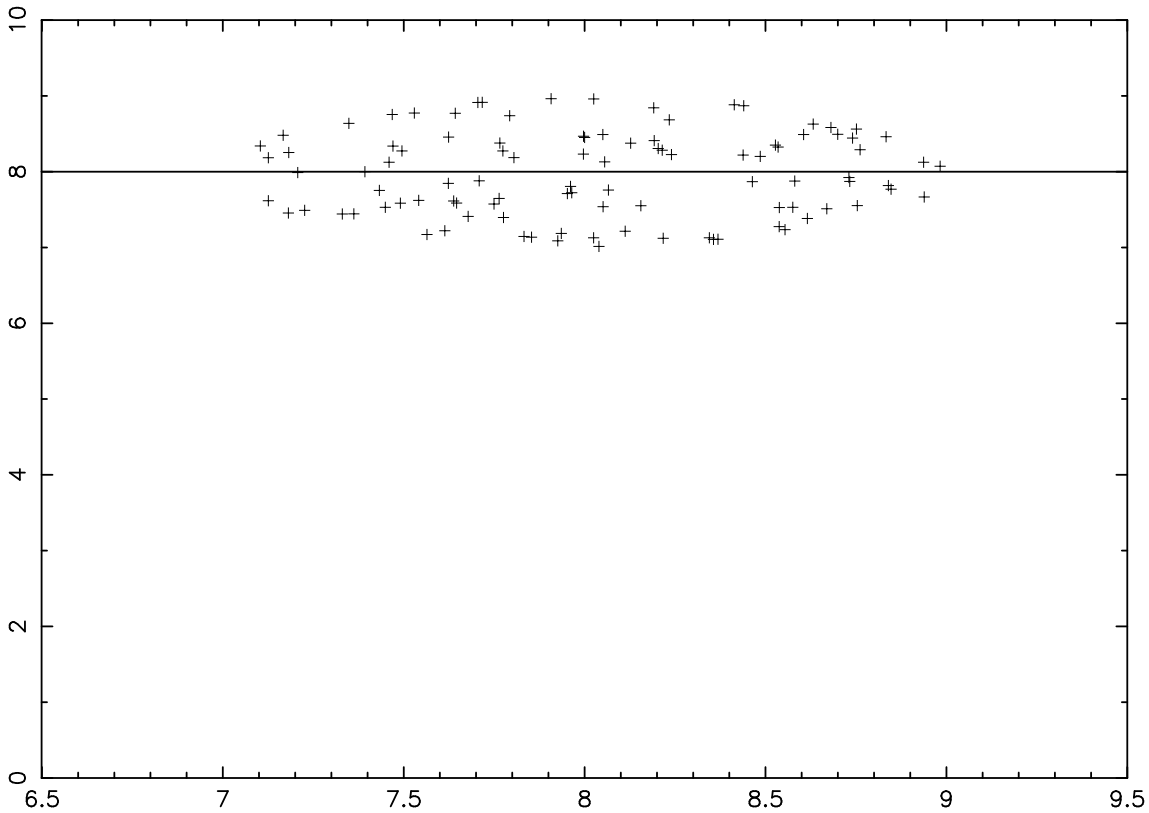
- Bottom 10% in range 0%–36%
- Bottom 25% in range 0%–46%
- Median mark 62%
- Top 25% in range 82%–100%
- Top 10% in range 90%–100%

Warning about independent variables



Some data, and the line of best fit.

Warning about independent variables

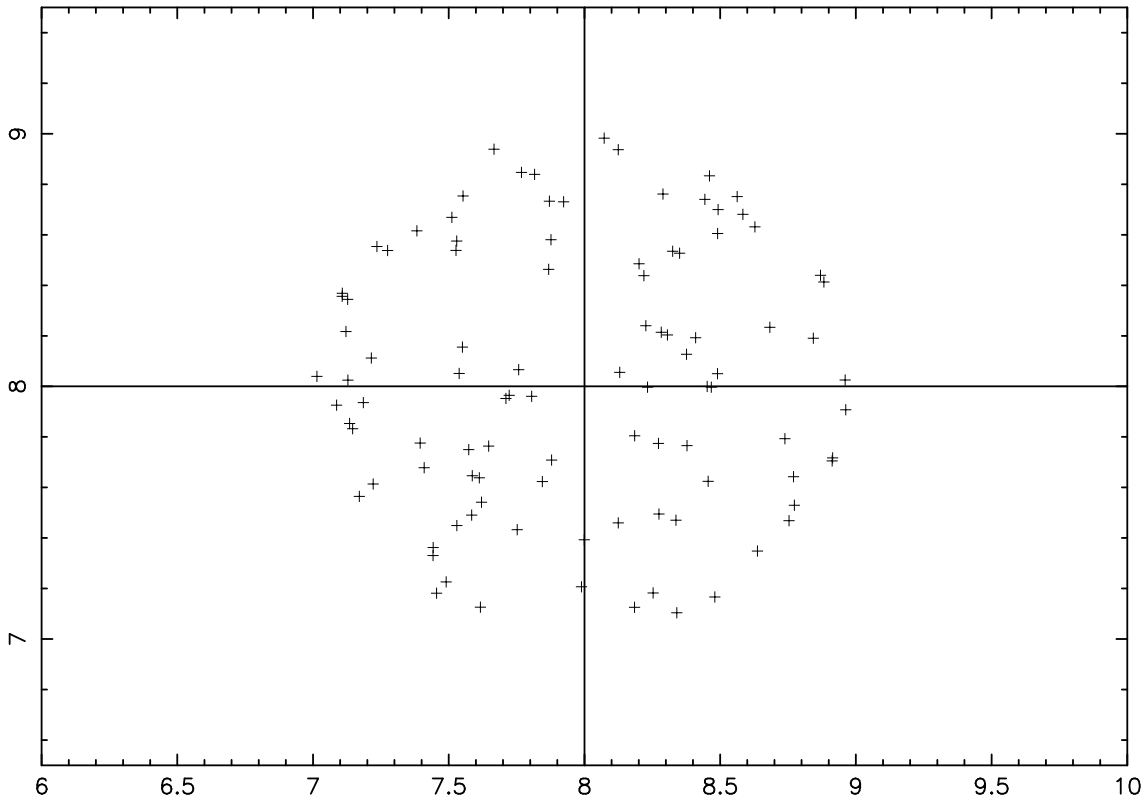


Another set of data, and the line of best fit.

What is the problem?

What is the problem?

Both sets of data were the same, but the axes were swapped



Several independent variables

Sometimes you measurements will depend on more than one independent variable. For example:

Life expectancy depends on:

Several independent variables

Sometimes you measurements will depend on more than one independent variable. For example:

Life expectancy depends on:

- Wealth (well off people live 7 year longer on average than poor people)

Several independent variables

Sometimes you measurements will depend on more than one independent variable. For example:

Life expectancy depends on:

- Wealth (well off people live 7 year longer on average than poor people)
- Weight

Several independent variables

Sometimes you measurements will depend on more than one independent variable. For example:

Life expectancy depends on:

- Wealth (well off people live 7 year longer on average than poor people)
- Weight
- Height

Several independent variables

Sometimes you measurements will depend on more than one independent variable. For example:

Life expectancy depends on:

- Wealth (well off people live 7 year longer on average than poor people)
- Weight
- Height
- Smoking

Several independent variables

Sometimes you measurements will depend on more than one independent variable. For example:

Life expectancy depends on:

- Wealth (well off people live 7 year longer on average than poor people)
- Weight
- Height
- Smoking
- Drinking

Several independent variables

Sometimes you measurements will depend on more than one independent variable. For example:

Life expectancy depends on:

- Wealth (well off people live 7 year longer on average than poor people)
- Weight
- Height
- Smoking
- Drinking
- ...

We will make a natural extension of the linear problem to several variables. We assume that the dependent variable y depends on several variables $\{x_1, x_2, \ldots, x_k\}$ and that we are given a set of data points each of the form $(x_1, x_2, \ldots, x_k, y)$.

We will look for the linear function

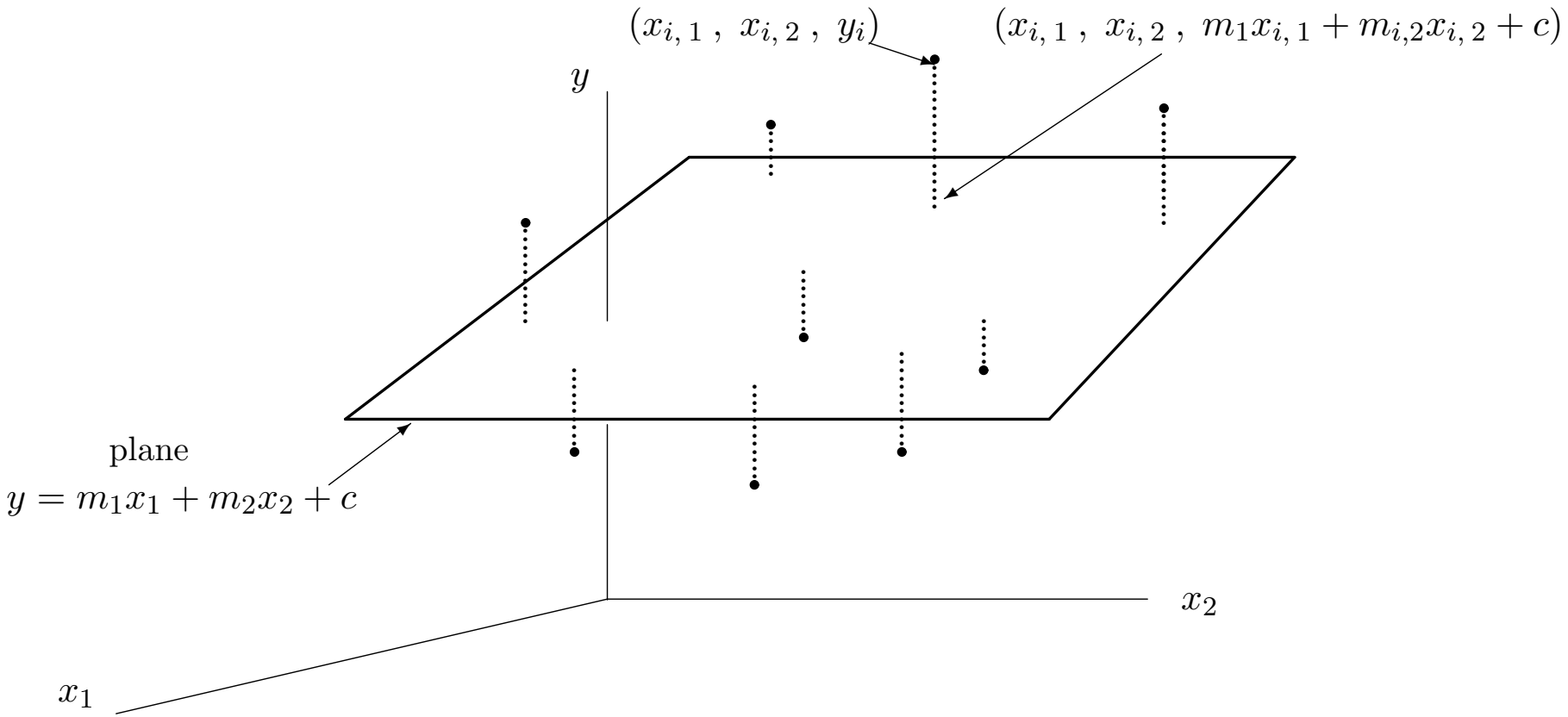
$$y = m_1 x_1 + m_2 x_2 + \cdots + m_k x_k + c$$

that best fits the data in the sense of least squares.

If y depends on two variables, x_1 and x_2, the data points are of the form (x_1, x_2, y) and geometrically are represented by points in three dimensional space.

In this case, with y depending on two variables, we minimise the sum of the squares of the vertical distances of the data points from the plane

$$y = m_1 x_1 + m_2 x_2 + c$$


$(x_{i,1}\ ,\ x_{i,2}\ ,\ y_i)$
$(x_{i,1}\ ,\ x_{i,2}\ ,\ m_1x_{i,1} + m_{i,2}x_{i,2} + c)$
y
plane
$y = m_1x_1 + m_2x_2 + c$
x_2
x_1

When y depends on two variables, it is necessary to label each data point as well as each independent variable, thus we introduce the index i. This can be done in different ways. For example in the figure

$$(x_{i,1}, x_{i,2}, y_i)$$

Here we will use

$$({}^{i}x_1, \ {}^{i}x_2, \ {}^{i}y)$$

In either case the index i will run from 0 to N.

The least squares criteria now requires the determination of m_1, m_2 and c such that:

$$S = \sum_{i=0}^{N} \left[(m_1 \, {}^{i}x_1 + m_2 \, {}^{i}x_2 + c) - {}^{i}y \right]^2$$

Is a minimum

For the general case where data points of the form $\{x_1, x_2 \ldots x_k,\ y\}$, the general ith data point is denoted as

$$({}^{i}x_1,\ {}^{i}x_2, \ldots {}^{i}x_k,\ {}^{i}y\}$$

and the least square criteria for fitting the "surface"

$$y = m_1 x_1 + m_2 x_2 \cdots + m_k x_k + c$$

means finding $\{m_1,\ m_2 \ldots m_k,\ c\}$ such that:

$$S = \sum_{i=0}^{N} \left[(m_1\, {}^{i}x_1 + m_2\, {}^{i}x_2 + \cdots + m_k\, {}^{i}x_k + c) - {}^{i}y \right]^2$$

is a minimum.

To find the minimum solve the $k+1$ equations

$$\frac{\partial S}{\partial m_1} = 0, \frac{\partial S}{\partial m_2} = 0, \ \ldots, \ \frac{\partial S}{\partial m_k} = 0 \quad \text{and} \quad \frac{\partial S}{\partial c} = 0$$

As before, because S is simply the sum of squared terms, the solution to these equations will be a minimum.

We we will not (usually) have to carry out any of the above calculations by hand — we use Excel.

We will again use the Excel function **LINEST**

=LINEST(*known y values***,** *known x values***,** *c constant***,** *statistics***)**
The parameter of LINEST are as follows:

- *known y values* A range of cells identifying a single column of y values.
- *known x values* A range of cells **identifying k columns of the x variables**.

The other two arguments are as before (usually ignored).

Example: Given the following data obtain the best least squares fit of the form:

$$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + c$$

Hence obtain an approximation to y at $x_1 = x_2 = x_3 = 4$.

$x_1 =$	1.0	1.7	2.3	-0.5	2.3
$x_2 =$	2.1	1.0	1.7	0.3	0.5
$x_3 =$	1.0	2.1	3.0	3.1	3.6
$y =$	2.17	2.272	2.795	1.515	2.519

3.7	-3.0	2.5	2.0	1.6	3.0
0.6	-1.2	-0.2	2.0	3.1	-1.0
1.4	5.0	1.2	6.0	7.1	-3.0
2.419	0.611	1.911	3.169	3.356	0.950

Over to Excel...

This process gives:

$$y = (0.155)x_3 + (0.309)x_2 + (0.273)x_1 + 1.062$$

Thus the estimate of the value of y at $x_1 = x_2 = x_3 = 4$ is given by

$$y = (0.155)(4) + (0.309)(4) + (0.273)(4) + 1.062 = 3.948$$

If you didn't want the values of m_1, m_2, etc., you could use the Excel function **TREND** (see G. Bowtell's notes)

Goodness of fit

There several ways of assessing the goodness of fit of you estimated line, surface or function to your original data.

If the **statistics** option is set to be true LINEST returns statistical data that can be used for this purpose. Here two of the more straightforward indicators given by LINEST are considered:

- the coefficient of determination
- standard error of the y-estimate

Note: This discussion of the meaning of these statistical concepts is far from complete and intended to give a flavour of what is possible. For more detail you are advised to consult any standard statistics text, or you probability and statistics notes.

The **standard error** is related to the standard deviation of the difference between the y-data and the y-estimate.

Suppose we fitted the regression line $Y = mx + c$ to the data

$$\{(x_0, y_0),\ (x_1, y_1), \ldots, (x_n, y_n)\}$$

by the method of least squares. The **standard error** of the estimate Y is defined as

$$SE_y = \sqrt{\frac{\sum_{i=0}^{n} (y_i - Yi)^2}{(n+1)}}$$

where $Y_i = mx_i + c$.

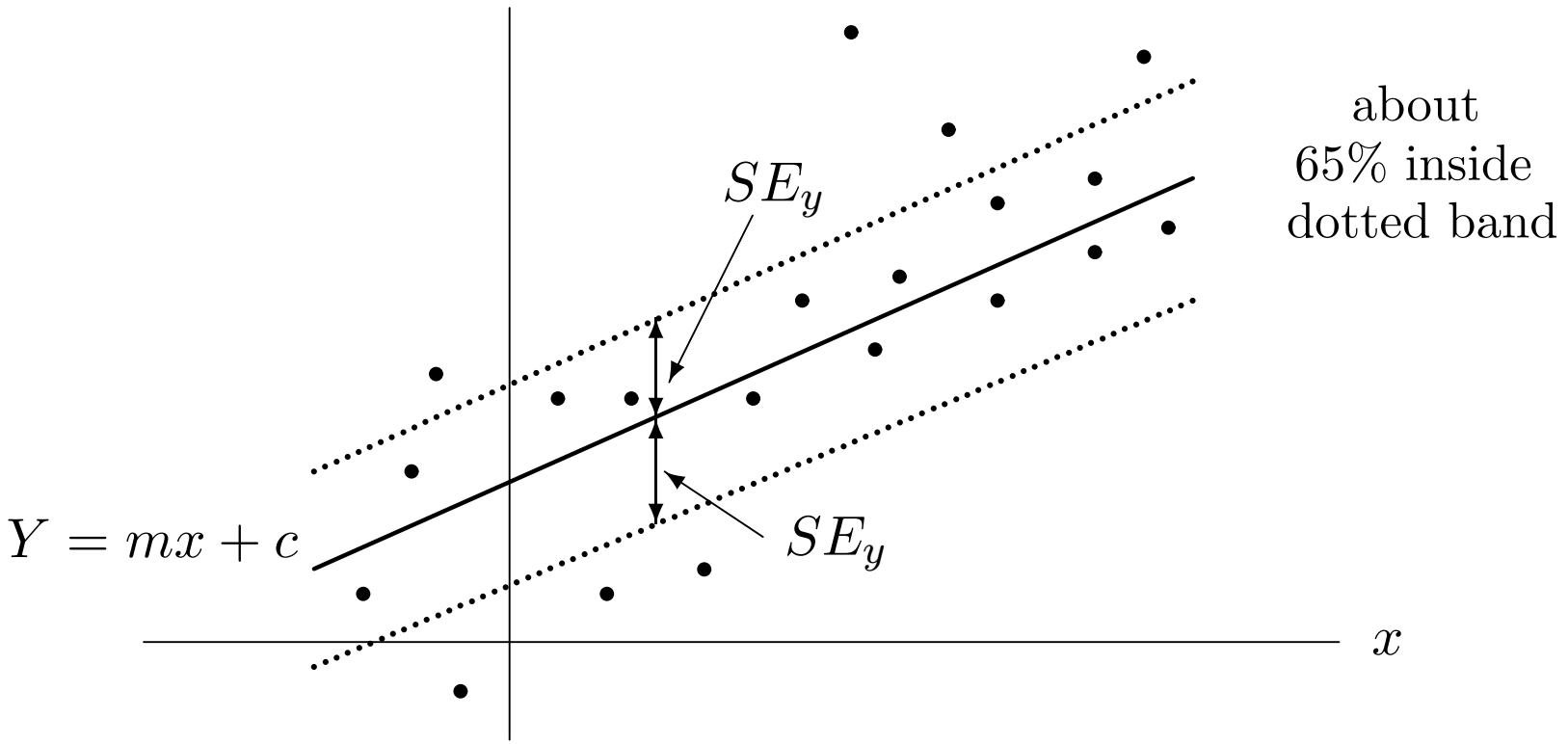
Recall that the straight line through the data was calculated such that

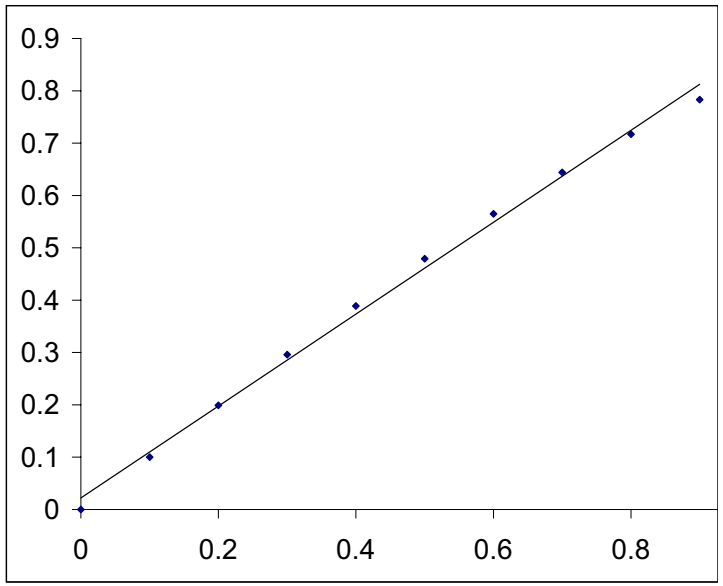
$$\sum_{i=0}^{n} (y_i - Yi)^2$$

was a minimum.

Another way of looking at finding a regression line is that it is the line that minimises the standard error, SE_y.

Statistically SE_y corresponds to an estimate of the standard deviation of the error variable $E_i = (y_i - Y_i)$. Indeed under certain assumptions concerning the population from which E_i is selected we can state that with increasing n we can expect about 65% of the data points to lie closer than the standard error from the estimated line of best fit. (See a statistics text for more details)


SE_y
SE_y
$Y = mx + c$
x
about
65% inside
dotted band


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8

Note:

- Calculating SE_y for a data set consisting of large values of y will yield a large value of SE_y
- calculate it for a data set consisting of small values of y it will yield a small value of SE_y.

So seeing the standard error by itself does not tell you if the fit is good or bad. We use the **the Coefficient of Determination**.