Computational Mathematics/Information Technology

Dr Oliver Kerr

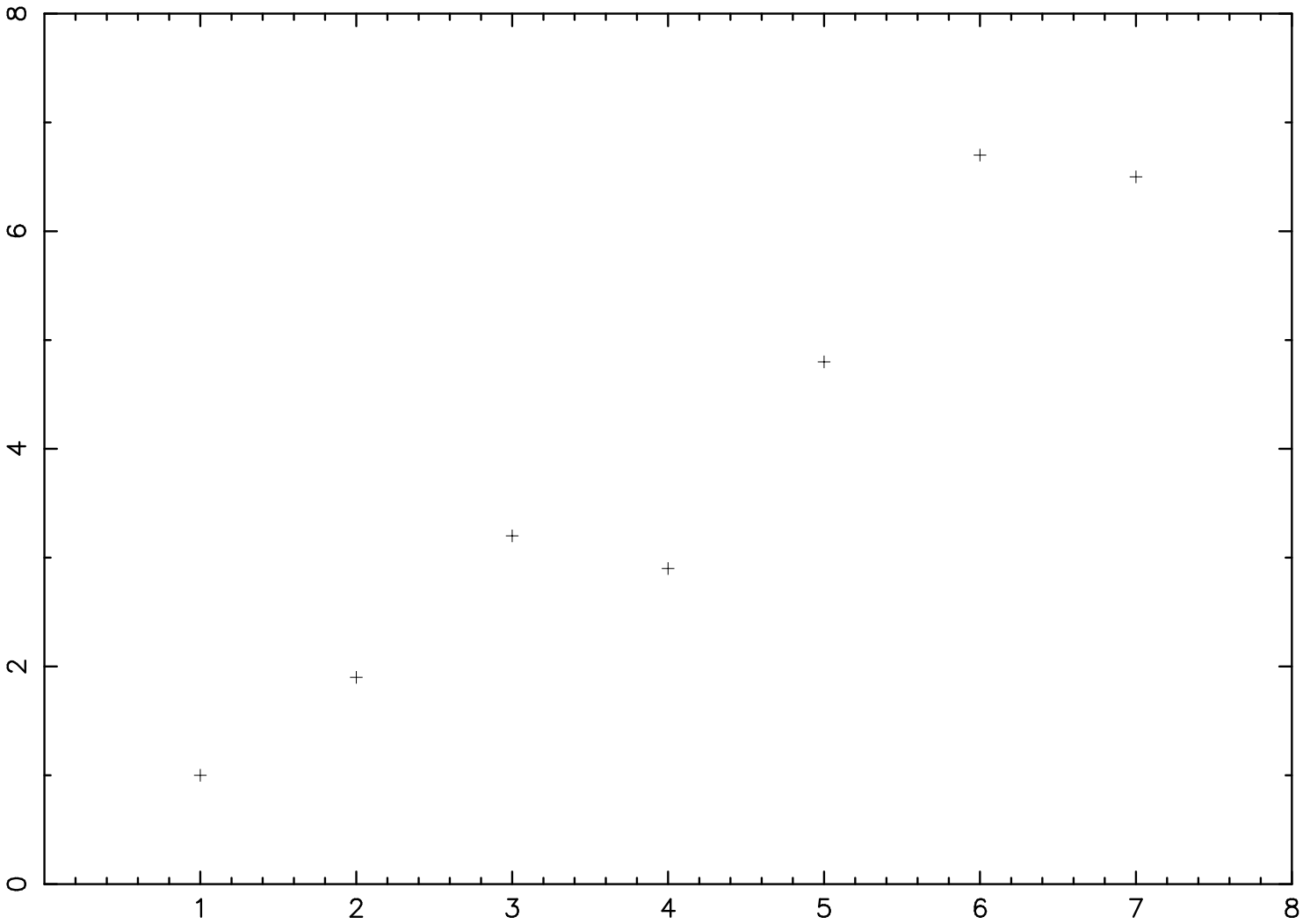
2009–10

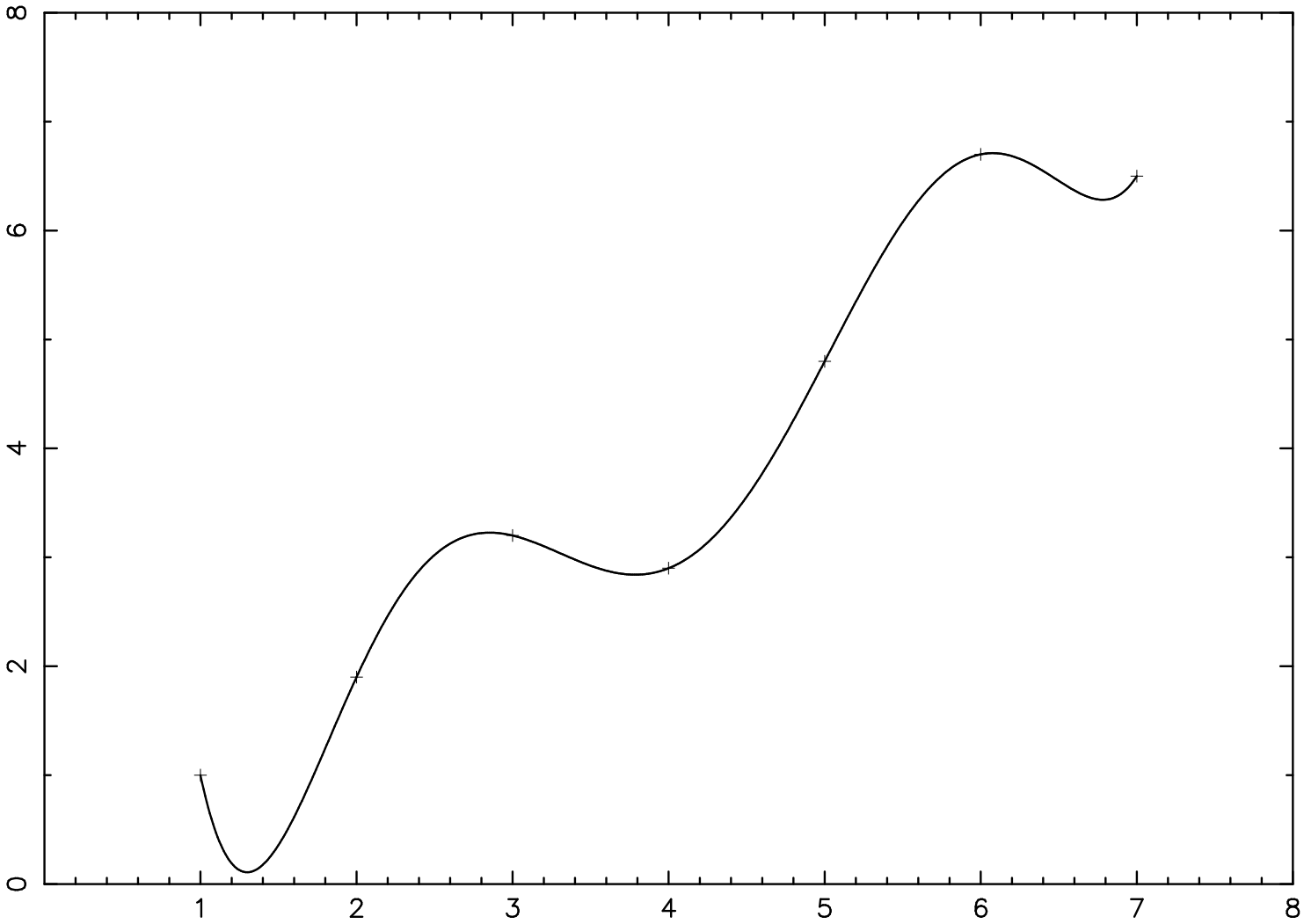
Not all data points are accurate...

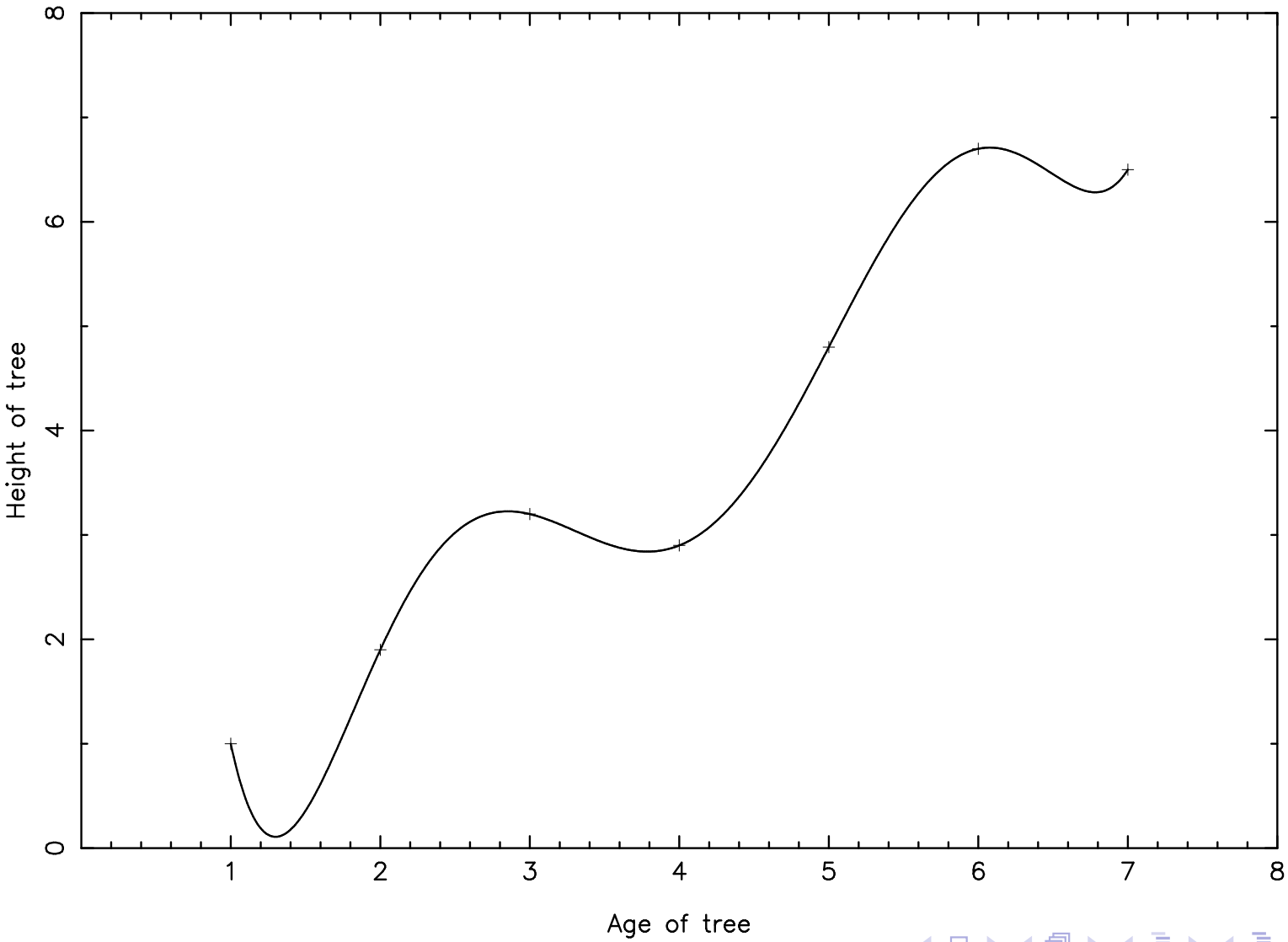
We have seen several methods for fitting curves through data points:

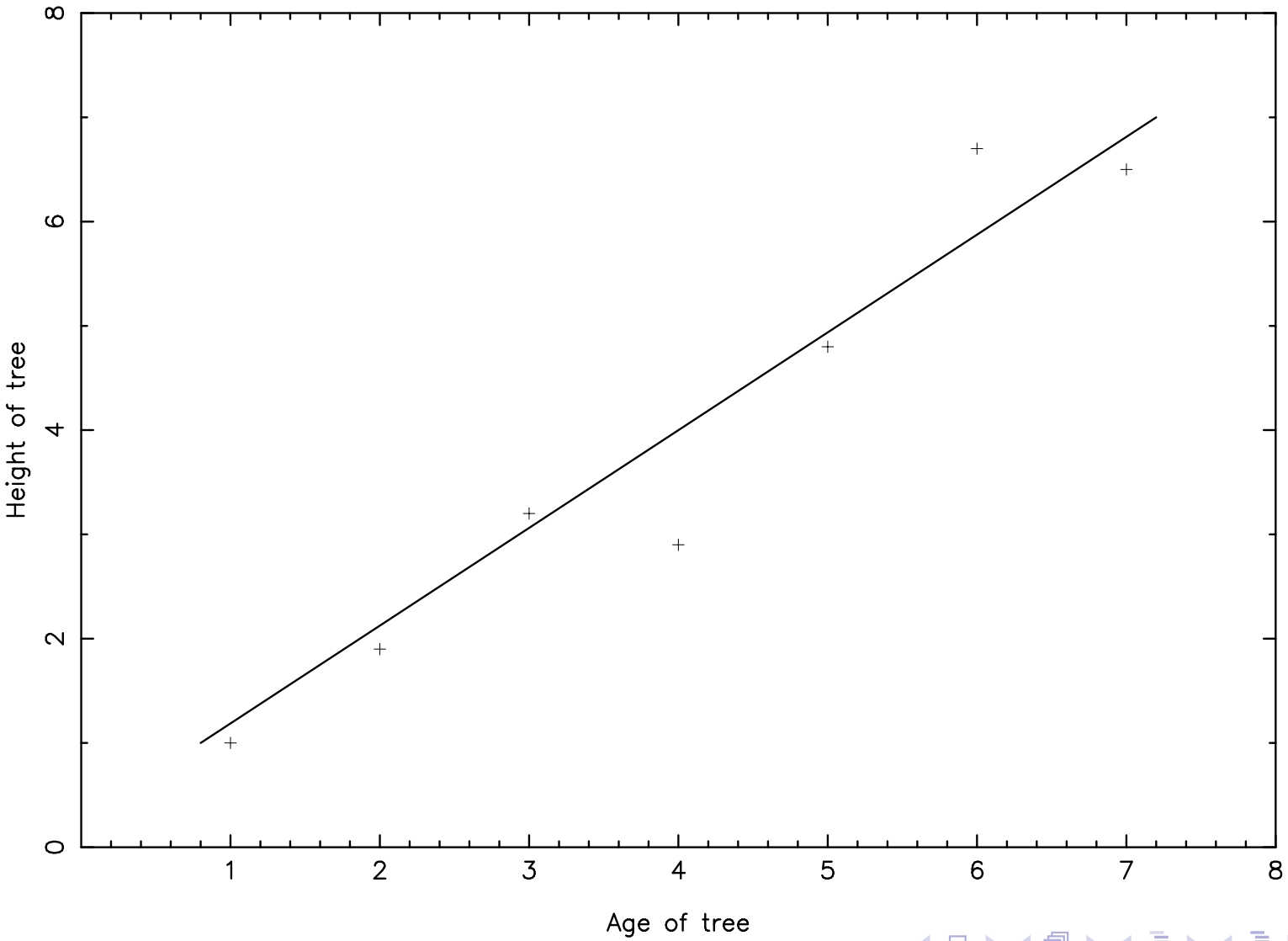
- Polynomial fitting
- Linear splines
- Cubic splines

All these curves pass through the data points.


0
2
4
6
8
1
2
3
4
5
6
7
8


0
2
4
6
8
1
2
3
4
5
6
7
8


Height of tree
Age of tree
0
2
4
6
8
1
2
3
4
5
6
7
8


Height of tree
Age of tree
0
2
4
6
8
1
2
3
4
5
6
7
8

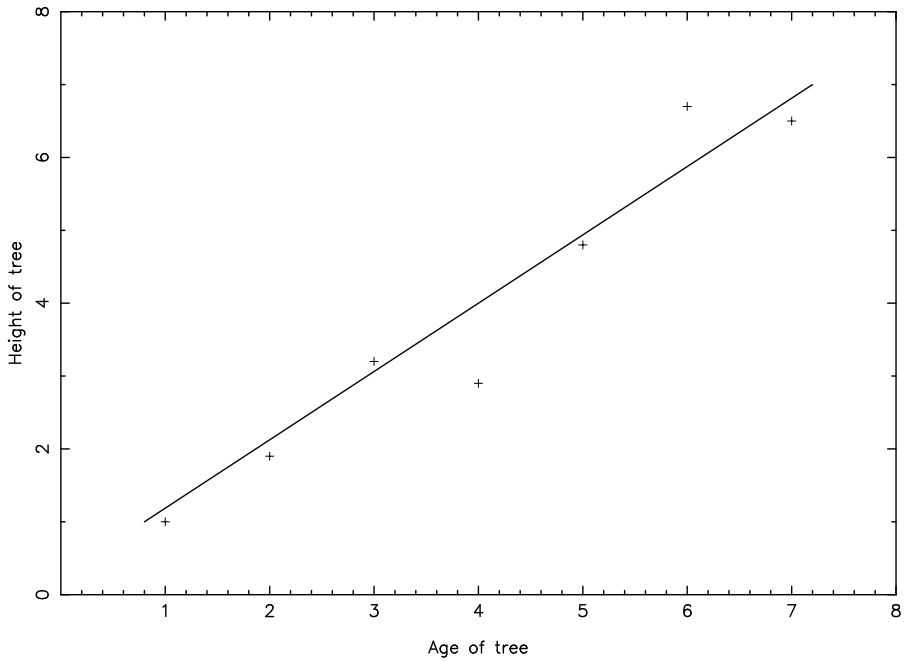
Least squares fitting

Not all data is perfect. There may be a degree of statistical scatter — either through errors in measurement or through inherent variation in what is being measured.

We will first look at fitting curves through such random data. We consider fitting a straight line to a set of data — this is sometimes referred to as regression.

We will try to construct a straight line that in some sense gets close to as many points as possible whilst at the same time indicating the general trend of the data.

The problem therefore is to specify mathematically, in a unique fashion, a way of constructing such a line from the given data.



Here we have an example. It is not a very good fit for all the data points, but to make the line fit these points would make it worse for others.

How go we know a line is a good fit?

The way we will select a good fit is to use the **least square criteria for fitting a straight line**.

- We will plot the **independent variable** along the x-axis, and the **dependent variable** along the y-axis.

How go we know a line is a good fit?

The way we will select a good fit is to use the **least square criteria for fitting a straight line**.

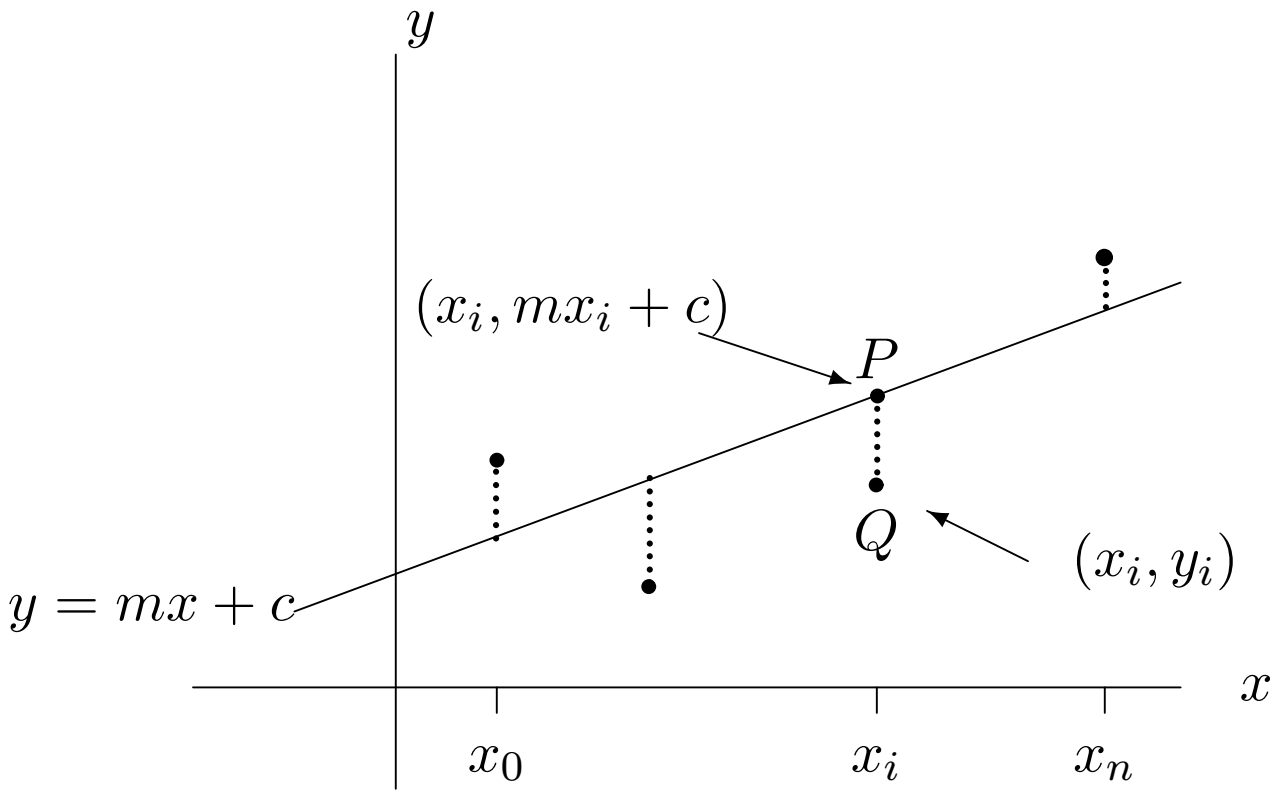
- We will plot the **independent variable** along the x-axis, and the **dependent variable** along the y-axis.
  Are cancer rates affected by smoking?
  Or, is smoking affected by cancer rates?

How go we know a line is a good fit?

The way we will select a good fit is to use the **least square criteria for fitting a straight line**.

- We will plot the **independent variable** along the x-axis, and the **dependent variable** along the y-axis.
  Are cancer rates affected by smoking?
  Or, is smoking affected by cancer rates?
- The least square criteria for fitting a straight line to a set of data is given as:

> The line is drawn such that the sum of the squares of the vertical distances of each data point from the line is a minimum.



If the point Q has coordinates (x_i, y_i) then the point vertically above it, P, lying on the line $y = mx + c$ has coordinates $(x_i, mx_i + c)$. The length of PQ is $|(mx_i + c) - y_i|$.

If the data points are $\{(x_0, y_0),\ (x_1, y_1),\ \ldots\ , (x_N, y_N)\}$ and the equation of the line is given by $y = mx + c$ then the sum of the squares of the distances is given by

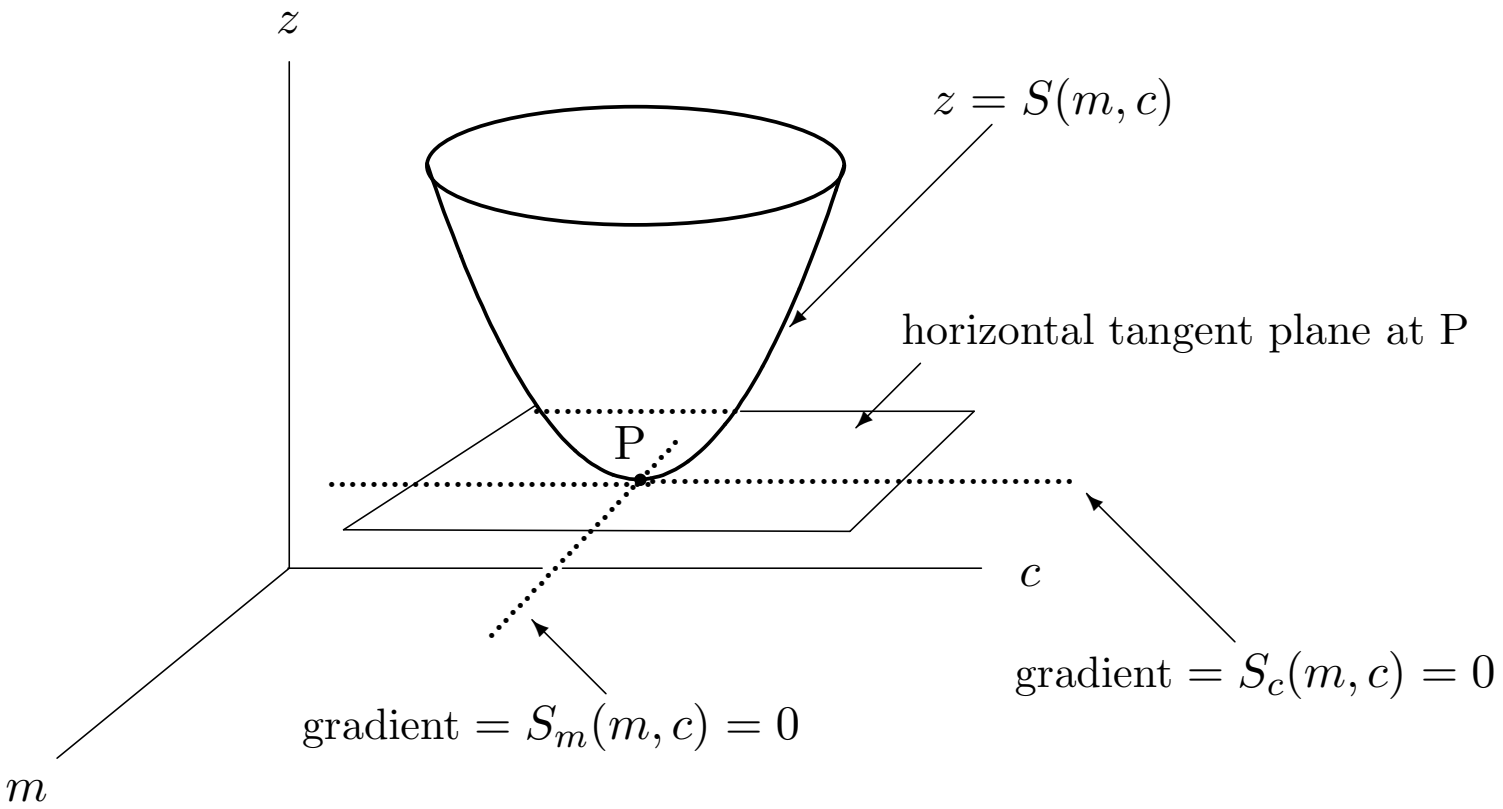
$$S = \sum_{i=0}^{N} (mx_i + c - y_i)^2$$

We want to find m and c such that this is a minimum.

The shortest approach is to use calculus for two variables. For this purpose S is considered to be a function of two variables m and c, so we write

$$S(m,c) = \sum_{i=0}^{N} (mx_i + c - y_i)^2$$

- Since $S(m,c)$ is the sum of squares it will be positive.
- Since it only depends quadratically on its two variables it will be in the shape of an open bowl.



This surface will have one point where the tangent plane is horizontal. This will be the minimum.

The point we identify as a possible extremum will be a minimum. It will not be a maximum, or indeed the equivalent of a stationary point of inflection in two dimensions.

How to deal with these other cases for general functions of two variables is left to your Calculus course in Part 2.

We are going to look for the point where

$$S_m(m,c) = \frac{\partial S}{\partial m} = 0 \qquad \text{and} \qquad S_c(m,c) = \frac{\partial S}{\partial c} = 0$$

This is equivalent to the case in one dimension where turning points of $f(x)$ are located by solving $f'(x) = 0$.

Since

$$S(m,c) = \sum_{i=0}^{N} (mx_i + c - y_i)^2$$

we have

$$S_m(m,c) = \sum_{i=0}^{N} 2(mx_i + c - y_i)x_i = 0$$

$$\text{or} \qquad m\sum_{i=0}^{N} x_i^2 + c\sum_{i=0}^{N} x_i - \sum_{i=0}^{N} x_i y_i = 0$$

$$S_c(m,c) = \sum_{i=0}^{N} 2(mx_i + c - y_i) = 0$$

$$\text{or} \qquad m\sum_{i=0}^{N} x_i + \sum_{i=0}^{N} c - \sum_{i=0}^{N} y_i = 0$$

If we divide the last of these

$$m\sum_{i=0}^{N} x_i + \sum_{i=0}^{N} c - \sum_{i=0}^{N} y_i = 0$$

by $N+1$ (the number of data points), we get

$$m\sum_{i=0}^{N} \frac{x_i}{N+1} + \sum_{i=0}^{N} \frac{c}{N+1} - \sum_{i=0}^{N} \frac{y_i}{N+1} = 0$$

But

$$\sum_{i=0}^{N} \frac{x_i}{N+1} = \overline{x}, \qquad \sum_{i=0}^{N} \frac{y_i}{N+1} = \overline{y}, \qquad \sum_{i=0}^{N} \frac{c}{N+1} = c$$

where $\overline{x}$ is the average value of the x_i and $\overline{y}$ is the average value of the y_i. So

$$m\overline{x} + c - \overline{y} = 0$$

Similarly

$$m\sum_{i=0}^{N} x_i^2 + c\sum_{i=0}^{N} x_i - \sum_{i=0}^{N} x_i y_i = 0$$

gives

$$m\overline{x^2} + c\overline{x} - \overline{xy} = 0$$

where

$$\overline{x^2} = \sum_{i=0}^{N} \frac{x_i^2}{(N+1)} \qquad \text{and} \qquad \overline{xy} = \sum_{i=0}^{N} \frac{x_i y_i}{(N+1)}$$

We have two equations for m and c

$$m\overline{x} + c - \overline{y} = 0 \qquad \text{and} \qquad m\overline{x^2} + c\overline{x} - \overline{xy} = 0$$

which we solve simultaneously to get

$$m = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \qquad \left[= \frac{\text{covariance of } x \text{ and } y}{\text{variance of } x} \right]$$

You then find

$$c = \overline{y} - m\overline{x}$$

Using Excel

That is the theory, but we will let Excel do the work for us!

Excel can cope with problems where each y_i depends on several variables, but we will stick to one for the moment.

We will 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 (for now) a single column of the x variables.

The other two arguments (which you can usually ignore) are:

- *c constant* Optional parameter set to be **true** or **false**. If **true** or omitted then c is calculated normally as above. If **false** then c is set to zero and the m parameters are found such that $y = mx$ is the best least squares fit. Useful if you know the line should pass through the origin (e.g., tree height is zero when its age is zero).
- *statistics* Optional parameter set to be **true** or **false**. If **false** or omitted then LINEST returns only c and the m values. If **true** then LINEST also returns a set of regression statistics that enable us to assess the goodness of the approximation.

LINEST is an **array function** and will return the values m and c together, along with statistics which can be used to judge how good a fit has been achieved.

With the optional parameters omitted LINEST places in a preselected row and in the following order the values $\{m,\ c\}$.

As with any array function to implement LINEST you need to first highlight the required space, in this case 2 cells in any vacant row, and finally, after entering the y and x data ranges into LINEST, and use Ctrl-Shift-Enter to obtain the results.

Example: Find the least squares linear fit to the following data points:

$x =$	0.0	0.1	0.2	0.3	0.4
$y =$	0.000	0.100	0.199	0.296	0.389

$x =$	0.5	0.6	0.7	0.8	0.9
$y =$	0.479	0.565	0.644	0.717	0.7 83

Now, over to Excel...