

## B Linear Regressions

---

---

### Is your data linear?

So you want to fit a straight line to a set of measurements... First, make sure you really want to do this. That is, see if you can convince yourself that a plot of your data, a series of  $(x, y)$  pairs, is compatible with a linear model

$$y = mx + b \quad (1)$$

where  $m$  is the slope, and  $b$  is the  $y$  intercept. Fig. 1 shows a plot of a set of data that does not appear to be linear. We could apply a linear fit to this data, but it is unclear what the results would mean.

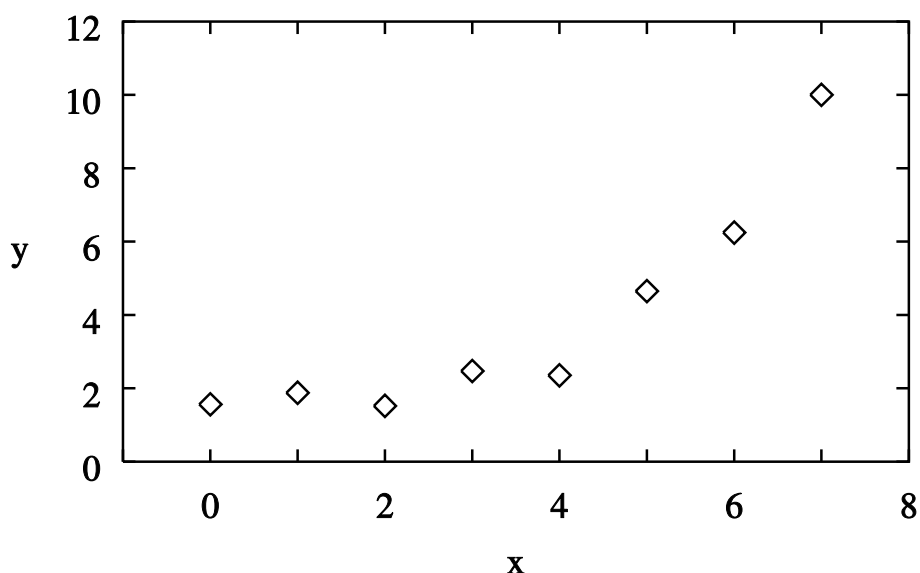


Figure 1: An example of a set of data which does not appear to be compatible with a linear model. Instead, the slope of the graph appears to increase with  $x$ .

A sample set of data compatible with a linear model is given in Table 1 and plotted in Fig. 2. Note that none of the data points fall along the straight line shown in the figure. Measurements compatible with a linear model generally do not all fall exactly on a single straight line, because measurements include uncertainties. Although the data do not fall on a single line, the diamonds in Fig. 2 appear to be scattered randomly about a common line.

### The Method of Least Squares

Once you are convinced that your data are compatible with a linear model, it is reasonable to apply the **method of least squares**, also called a **linear regression**, to your data.

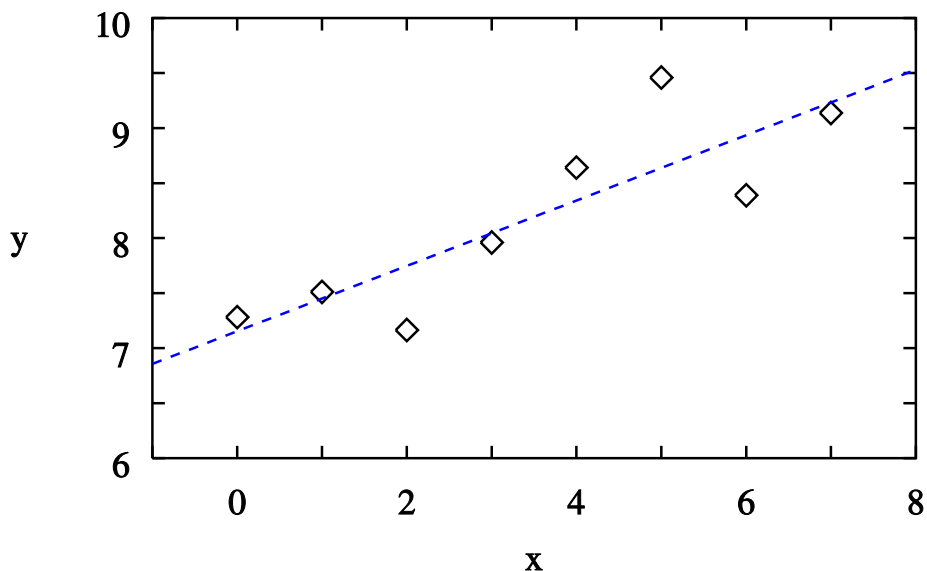


Figure 2: The data (diamonds) and the best linear fit (dashed line).

This is a method of finding the slope and intercept, with associated uncertainties, of the line giving the “best fit” to your data. The method is implemented as a function in the **Excel** spreadsheet and as an analysis tool in **Logger Pro**. Detailed instructions for using them are given below.

A derivation of the method of least squares is beyond the scope of this course,<sup>1</sup> and we won’t need to use the associated equations, since the method is automated in the software tools we use. Instead of working through a derivation, we will consider the the general idea behind the process. In Fig. 3, the solid bars show the differences

$$y_i - (mx_i + b) \quad (2)$$

between the data and the fit function (Eq. 1). The method of least squares gives the line which minimizes the sum of the squares of these differences,

$$\sum_i [y_i - (mx_i + b)]^2 \quad (3)$$

These are the “least squares” referred to in the name of the method. Try sketching a different line on Fig. 3, and add your own “difference bars,” and you should be able to convince yourself that they would give a larger value of the sum in Eq. 3.

The method of least squares is analogous to calculating the mean and the associated standard deviation of a set of measurements of a single quantity. The uncertainties in the measurements are assumed to be random, and the resulting slope and intercept are estimates

---

<sup>1</sup>However, a first course in calculus qualifies you to follow the mathematics. See, for example, Hugh D. Young, *Statistical Treatment of Experimental Data*, McGraw-Hill (1962).

x values	y values
0	7.28
1	7.51
2	7.16
3	7.96
4	8.64
5	9.46
6	8.39
7	9.14

Table 1: Data which appear to be compatible with a linear model.

of the most probable true values. The dashed line shown in Fig. 2 is the result of a least squares fit to the data in Table 1. The fit results are

$$\begin{aligned}
 m &= 0.297619 \\
 \sigma_m &= 0.075643 \\
 b &= 7.150833 \\
 \sigma_b &= 0.316438 \\
 \sigma_y &= 0.490223 \\
 r^2 &= 0.720676
 \end{aligned}$$

where  $m$  is the slope,  $b$  is the  $y$  intercept,  $\sigma_y$  is an estimate of the uncertainty of individual  $y$  measurements called the **mean square error** or the **standard deviation of the  $y$  estimate** and  $r$  is the **correlation coefficient**.

## Interpreting the Fit Results

- $\sigma_m$  and  $\sigma_b$  The uncertainties  $\sigma_m$  and  $\sigma_b$  in the slope and intercept are standard deviations. Hence, with a large number of measurements, the fit values of the slope and intercept fall within  $\sigma_m$  and  $\sigma_b$  of the true values with a probability of 68%, and it is 95% probable that the true values fall within  $2\sigma_m$  and  $2\sigma_b$ . The shaded region in Fig. 4 reflects the uncertainty ranges  $\pm\sigma_m$  and  $\pm\sigma_b$ .
- $\sigma_y$  The mean square error is calculated via

$$\sigma_y^2 = \frac{1}{N-2} \sum_i [y_i - (mx_i + b)]^2 \quad (4)$$

and is analogous to the standard deviation (squared) of a distribution of repeated measurements. Error bars equal to  $\sigma_y$  are shown in Fig. 4. Note that not all of the measurements agree with the best fit line within uncertainty. By the definition of the standard deviation, we only expect 68% of the measurements to come within  $\sigma_y$  of the best fit line.

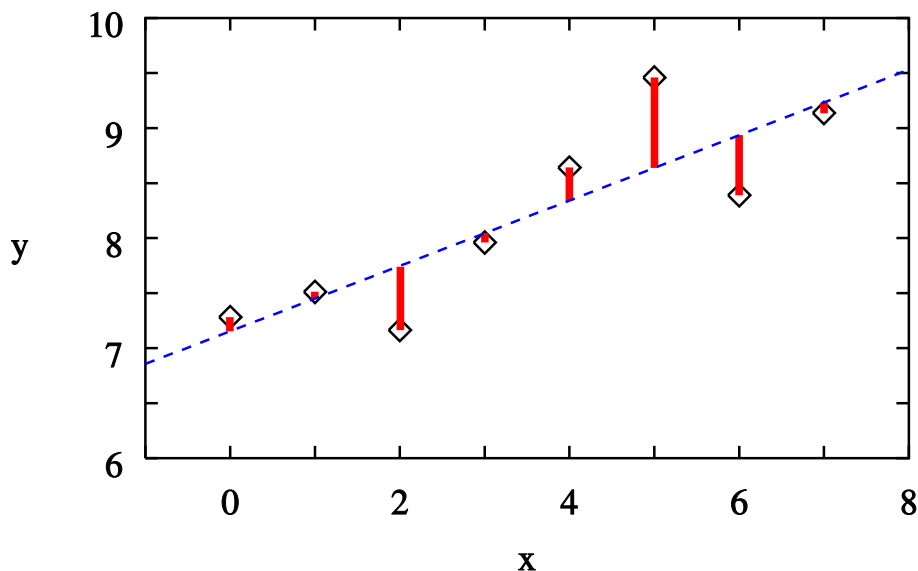


Figure 3: The solid vertical bars are the differences  $y_i - (mx_i + b)$  between the data and the linear fit. The sum of the squares of the deviations are minimized by the fitting procedure.

- $r^2$

Finally, the correlation coefficient  $r$  is a measure of the degree of correlation between  $y$  and  $x$  values. Typically, we are interested in  $r^2$ , called the **coefficient of determination**. It falls in the range  $0 \leq r^2 \leq 1$ , and describes the fraction of the variation in  $y$  values explained by the linear model. It follows that the quantity  $1 - r^2$  is the fraction of the variation we can attribute to the uncertainties in the measurements, provided the differences between the data and the linear model are purely random. In the example given here, about 72% of the variation in the data is explained by the linear model and about 28% of the variation is random. If the data falls precisely on the fit line,  $\sigma_y = 0$ , and  $r^2$  is exactly 1. If the data is completely uncorrelated,  $r^2$  is close to zero. There will generally be a clear correlation between measurements and models in the laboratory work for this course, so we will not be particularly concerned with  $r^2$  values. If you are given  $r$  instead of  $r^2$ , the range of possible values is  $-1 \leq r \leq 1$ . Negative  $r$  values correspond to negative correlations (negative slopes).

*Caution!* The correlation coefficient will not help you to identify data that is incompatible with a linear model. In fact, the least squares fit to the data shown in Fig. 1 is  $r^2 = 0.769507534$ , which indicates a stronger correlation than that of the linear data of Table 1 and Fig. 2.

*Note :* If we were working with actual data, the slope and intercept would have physical meaning, and we would report  $m = 0.30 \pm 0.08$  and  $b = 7.2 \pm 0.3$  with appropriate units. We would also report the error estimate for individual  $y$  measurements as  $\pm 0.5$  ( $\pm \sigma_y$ ).

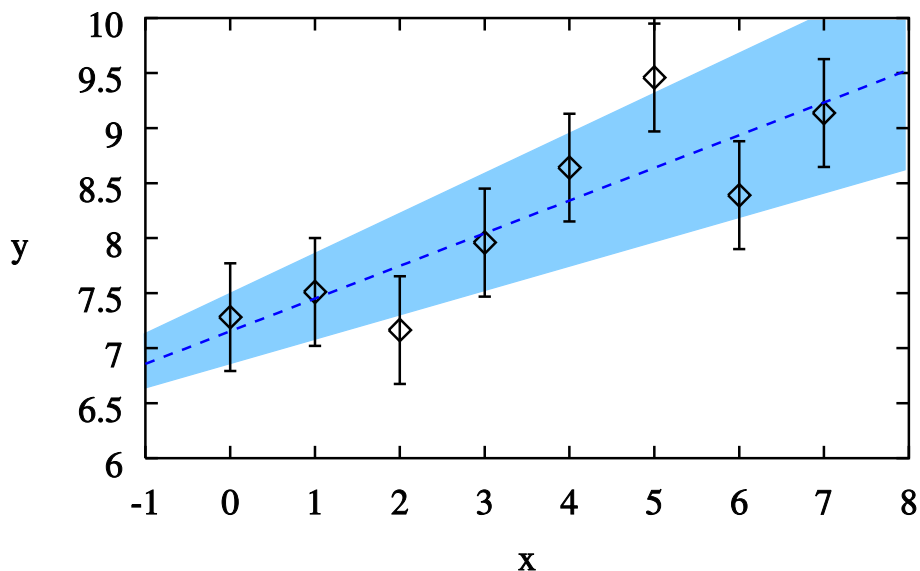



Figure 4: Error bars on the individual  $y$  measurements are  $\sigma_y$ . The shaded region reflects the uncertainty ranges  $\pm\sigma_m$  and  $\pm\sigma_b$  of the best fit slope and intercept.

## Zero Intercept

In some cases, you may have a reason (a theoretical model, e.g.) to test whether your data are consistent with a linear model with a  $y$  intercept  $b$  of zero. If a linear fit to your data yields a standard deviation of the  $y$  intercept  $\sigma_b$  greater than  $b$  itself, then you may conclude that your data are consistent with a zero  $y$  intercept.

If your theoretical model predicts a zero  $y$  intercept, *and* you find that a linear fit yields an intercept consistent within uncertainty with zero, you may want to perform a fit in which you fix the value of  $b$  at zero in order to find a best value of the slope compatible with your model. Instructions for performing linear fits with  $b$  fixed at zero with **Logger Pro** and **Excel** are included below.

## Fitting with Logger Pro

- The data you wish to fit must be in the **Logger Pro** data table and displayed in the graph window.
- If you want to fit a subset of the data shown on the graph, select a rectangular area in the graph window containing the points you want to fit by dragging the mouse.
- Click on the Curve Fit button . If you have saved several runs, select the one you'd like to fit. Select the **Linear** form ( $mx + b$ ), and click on **Try Fit** and **Ok**.

- The labels that **Logger Pro** gives to the fit results relate to our notation as

$$\begin{aligned} \text{m (Slope)} &: m \pm \sigma_m \\ \text{b (Y - Intercept)} &: b \pm \sigma_b \\ \text{Correlation} &: r \\ \text{RMSE} &: \sigma_y \end{aligned}$$

Note that you are given  $r$  instead of  $r^2$ .

- To fix the intercept ( $b$ ) to zero, after you click on **Analyze -> Automatic Curve Fit ...**, choose the **Proportional** form ( $Ax$ ) instead of the **Linear** form ( $mx + b$ ).

## Fitting with Excel

- First, enter your data into the spreadsheet. The data should be arranged in two columns,  $x$  values in one column and  $y$  values in another.
- Select a group of 6 cells, two columns wide by three rows high, for the fit results.
- In the formula bar type

`=linest(<y-cells>, <x-cells>, true, true)`

where `<y-cells>` is the range of cells containing your  $y$  values and `<x-cells>` is the range of cells containing your  $x$  values. The two “true” entries tell the `linest` function to allow the intercept  $b$  to be non-zero and to report fit statistics  $\sigma_m$ ,  $\sigma_b$ ,  $r^2$ , and  $\sigma_y$ .

- Hit **<CTRL> <SHIFT> <ENTER>**, and fit results will appear in the six cells as follows

$m$	$b$
$\sigma_m$	$\sigma_b$
$r^2$	$\sigma_y$

- To fix the intercept ( $b$ ) to zero, set the first logical argument of the `linest()` function to `false`. That is, in the formula bar, type

`=linest(<y-cells>, <x-cells>, false, true)`