University of Alabama
Department of Physics and Astronomy

LeClair Spring 2014

Linear regression: the best fit line

Contents

1 What is a "best-fit" line?

Let's say we have a collection of data (x_i, y_i) we believe to have a linear relationship, namely $y = mx + b$. What line best fits our data?

Previously, we considered a collection of repeated measurements x_i, with n measurements. The average of this collection is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

We also wanted to know something about the dispersion of these measurements about the average. The procedure was to take every data point x_i, calculate its deviation from the mean $(x_i - \overline{x})$, and square the result so all the deviations were positive. We then defined the standard deviation σ of our collection as

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

What this did for our collection of data points x_i scattered around the mean $\overline{x}$ was to find the constant value (horizontal line) that described our data with the minimum squared deviation.

If our data follows a linear relationship, what we would like to do is find the line that passes through our data set with the minimum squared deviation about that line, rather than about a constant. We will imagine that the parameter we control (the independent variable) are the x_i, and for each x_i, we measure a response y_i (the dependent variable). In this case, the x_i are nominally known exactly, and we want to find the linear function $y = mx_i + b$ that best describes our measured y_i. If our linear function predicts values of

$y_{\text{pred}} = mx_i + b$, we could, in analogy with the standard deviation, define an error ϵ that describes how much squared deviation there is between our prediction and the measured data:

$$\epsilon^2 = \sum_{i=1}^{n} (y_i - y_{\text{pred}}) = \sum_{i=1}^{n} (y_i - mx_I - b)^2 = \sum_{i=1}^{n} \left(m^2 x_i^2 + 2mbx_i - 2mx_iy_i + b^2 - 2by_i + y_i^2\right) \tag{3}$$

What we want is the line that gives the minimum ϵ^2 for our data. Given our two adjustable parameters m and b, this requires

$$\frac{d\epsilon^2}{dm} = 0 = 2m\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_iy_i \tag{4}$$

$$\frac{d\epsilon^2}{db} = 0 = 2m\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n} b - 2\sum_{i=1}^{n} y_i \tag{5}$$

Clearly, $\sum_{i=1}^{n} b = nb$. Rearranging the two equations, we find

$$m = \frac{\sum_{i=1}^{n} x_iy_i - b\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \tag{6}$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_I - m\sum_{i=1}^{n} x_i\right) \tag{7}$$

Equivalently, these two relationships could be stated

$$\sum_{i=1}^{n} x_iy_i = m\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i \tag{8}$$

$$\sum_{i=1}^{n} y_i = m\sum_{i=1}^{n} x_i + nb \tag{9}$$

Notice how these two equations resemble the linear relationship, but for the whole data set. We now have 2 equations with 2 unknowns, namely, m and b. Plugging the equation for b into the equation for m, we can find the slope of our best-fit line m in terms of sums of our experimental data and the number of points alone:

$$m = \frac{n\sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\left(\sum_{i=1}^{n} x_iy_i\right) - n\overline{x}\,\overline{y}}{\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2} \tag{10}$$

Note that we divided everything by n and used the definitions $\overline{x} = (\sum_{i=1}^{n} x_i)/n$ and $\overline{y} = (\sum_{i=1}^{n} x_i)/n$. The intercept is then found easily if we already have the slope:[i]

[i] We should note that one can also calculate the uncertainty or *confidence interval* for the slope and intercept, a measure

$$b = \text{trendline intercept} = \overline{y} - m\,\overline{x} \tag{11}$$

A neat fact is that trend line must pass through the dataset average point $(\overline{x}, \overline{y})$. Of course, these parameters m and b describing the trend line have uncertainties, which can be calculated.

2 Uncertainty in the fit parameters

Of course, our data will not follow a linear relationship perfectly, and there will be uncertainty in the measured values y_i. Following our definition of standard deviation, a good estimate of the uncertainty σ_y would be

$$\sigma_y = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - mx_i - b\right)^2} \tag{12}$$

Here the factor $1/(n-2)$ is a bit mysterious, and we will not dwell on it, but for large n it is a minor point. One can think that because with only two points we could always find a perfect line, we must really only consider the subsequent $n-2$ points in finding the best fit line. From this uncertainty, one can derive the uncertainty in the fit coefficients:

$$\sigma_b = \sigma_y\sqrt{\frac{\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}} \tag{13}$$

$$\sigma_m = \sigma_y\sqrt{\frac{n}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}} \tag{14}$$

Thus, we could report the fit parameters as $b \pm \sigma_b$ and $m \pm \sigma_m$ to provide a confidence interval.

We note a few limitations here. First, this does not consider uncertainty in the x_i - they are assumed to be a perfectly known independent variable. Second, this does not handle the case of *weighted* fitting. In practice, if the uncertainty in the y values are known, we would like to weight points with lower uncertainty more heavily, and deemphasize points with large uncertainty. We provide these formulas in the next section.

3 Weighted least squares fitting

If we expect the y_i to fall on a straight line $y = mx + b$ and the measured y_i have known uncertainties σ_i, we can introduce a *weight* w for each point:

of how reliable the extracted trendline parameters are. This is a little more difficult and nuanced, and we will not dwell on it right now. Be aware that the trendline parameters are merely *best fit* values – in any particular case, however, even the 'best fit' can be pretty terrible! The correlation coefficient, discussed below and reported by most programs Excel or OriginLab, is a simple way to judge the quality of the trendline in describing your data. Ideally, the slope and intercept should be reported with uncertainty margins, however.

$$w_i = \frac{1}{\sigma_i^2} \tag{15}$$

This has the desired property that both positive and negative uncertainties weight the data in the same way (as in our calculation of standard deviation), and that more uncertain points have less influence over the fit parameters. In this case, we can determine the fit coefficients from

$$b = \frac{\sum_{i=1}^{n} wx_i^2 \sum_{i=1}^{n} wy_i - \sum_{i=1}^{n} wx_i \sum_{i=1}^{n} wx_i y_i}{\sum_{i=1}^{n} w \sum_{i=1}^{n} wx_i^2 - \left(\sum_{i=1}^{n} wx_i\right)^2} \tag{16}$$

$$m = \frac{\sum_{i=1}^{n} w \sum_{i=1}^{n} wx_i y_i - \sum_{i=1}^{n} wx_i \sum_{i=1}^{n} wy_i}{\sum_{i=1}^{n} w \sum_{i=1}^{n} wx_i^2 - \left(\sum_{i=1}^{n} wx_i\right)^2} \tag{17}$$

The uncertainty in the fit parameters becomes

$$\sigma_b = \sqrt{\frac{\sum_{i=1}^{n} wx_i^2}{\sum_{i=1}^{n} w \sum_{i=1}^{n} wx_i^2 - \left(\sum_{i=1}^{n} wx_i\right)^2}} \tag{18}$$

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{n} w}{\sum_{i=1}^{n} w \sum_{i=1}^{n} wx_i^2 - \left(\sum_{i=1}^{n} wx_i\right)^2}} \tag{19}$$

Note that this case still does not handle uncertainty in the x_i, we still assume the x_i are exactly controlled independent variables.

Programs like OriginLab will let you specify that the y error values are to be used as weights in fitting. Programs like Excel will require you to perform this calculation manually using the formulas above. The primary point is that if your data has uncertainty, and the uncertainty is *not* the same for all points, a more accurate estimate of the best-fit line should incorporate the uncertainties.

4 Correlation coefficient

We can also define (but will not derive) a "goodness of fit" parameter r. If $r = 0$, there is no correlation between x and y – total randomness. If $r = -1$, the data are perfectly *negatively* correlated – a line with negative slope. If $r = +1$, the data is perfectly *positively* correlated – a line with positive slope. The closer $|r|$ is to 1, the better the correlation, while a small value of $|r|$ near zero indicates poor correlation. We would guess that our r should be positive, and close to 1 based on the plot above. Without derivation, we'll simply quote how we calculate r:

$$\text{quality of fit} = r = \frac{n\left(\sum\limits_{i=1}^{i=n} x_i y_i\right) - \left(\sum\limits_{i=1}^{i=n} x_i\right)\left(\sum\limits_{i=1}^{i=n} y_i\right)}{\sqrt{n\left(\sum\limits_{i=1}^{i=n} x_i^2\right) - \left(\sum\limits_{i=1}^{i=n} x_i\right)^2}\sqrt{n\left(\sum\limits_{i=1}^{i=n} y_i^2\right) - \left(\sum\limits_{i=1}^{i=n} y_i\right)^2}} \tag{20}$$

It is a bit fearsome-looking, and we will use a computer to calculate it automatically in general, but it just involves sums of our data that we've already done to find the best-fit line.

5 Linearization of a non-linear relationship

Many experiments we will perform will not result in nicely linear data. As a simple example you have all seen before, consider free-fall motion. In this case, we know that the vertical position versus time for an object dropped from rest at a height x_o is given by

$$x = \frac{1}{2}a_o t^2 + x_o = -\frac{1}{2}gt^2 + x_o \tag{21}$$

Here the second form explicitly assumes that $a_o = -g$ with the upward direction defined as positive y. How can we make this into a linear relationship so we can make use of the regression analysis we just learned? We can start by noticing that we *would* have a straight line relationship if we had t instead of t^2 in the equation above. If that is the case, all we need to do is plot x versus t^2 instead of x versus t, or in other words we can change the dependent variable from t to t^2. Mathematically, we can do that by making the replacement $u = t^2$:

$$x = \frac{1}{2}a_o u + x_o = -\frac{1}{2}gu + x_o \tag{22}$$

Now we have a straight line of slope $-g/2$ and intercept x_o, and we can perform our regression analysis on the substituted data $x(u) = y(t^2)$, plotting x versus t^2 and finding the appropriate trendline. For rather subtle statistical reasons, this procedure is *not* generally as accurate as if the data were linear in the first place, but it is typically a very good way to estimate the best-fit parameters for a non-linear relationship.

As another example, let's say we had an exponential relationship, like this:

$$x = x_o e^{-at} \tag{23}$$

In this case, we could linearize the equation by taking the natural logarithm of each side and re-arranging:

$$\ln x = -at + \ln x_o \tag{24}$$

Now if we plot $\ln y$ versus t we should get a straight line of slope $-a$ and intercept $\ln x_o$. Not every equation can be linearized in a simple way, of course, but more often than not it is the case. There are subtle problems in handling uncertainties correctly this way if the uncertainties are not the same for all y_i, but if the variation in uncertainties is small or the uncertainties are unknown, linearization is an unambiguous and simple way to get reasonable estimates of fit parameters.