Fitting and Error Estimation

1. Overview

Often, we take data where we expect our measurements to be a function of an independent variable, such as the case when we measure the position of a body in motion as a function of time. Usually, we know ahead of time (or can guess) the functional form, and it’s the parameters of this function that are of interest. This writeup gives you some mathematical tools you can use to determine the parameters (slope and intercept) of the straight line that best describes your data, and the errors on these parameters based on the errors of your individual measurements. This is called a linear fit. The full derivation given here is rather lengthy. Feel free to jump to the Summary at the end if you don’t care about where this comes from, but just want the formulas.

2. Least Squares Fitting

One of the most common methods of parameter determination is called “Least Squares” fitting. This is easy to understand: given the assumption that our data follows some functional form like \( F(x) = mx + b \), the best values of the parameters \( m \) and \( b \) occur when the deviations (distances) of all of the data points from the line are minimized. If our data points are sets of pairs \((x_i, y_i)\), then the deviation \( \delta \) between our hypothesis at each point \((mx_i + b)\) and what we measure \((y_i)\) is

\[
\delta_i = |mx_i + b - y_i|.
\]

Where does the “Squares” come in? You can see that the deviation \( \delta \) involves an absolute value. Mathematically, absolute values can be tricky to deal with, especially where taking derivatives is concerned. Instead, if we minimize \( \delta^2 \) over the whole data set of \( N \) points, we achieve the same basic goal without the problems associated with the absolute value:

\[
\delta^2 = \sum_{i=1}^{N} [(mx_i + b) - y_i]^2.
\]

The parameter \( \delta^2 \) is minimized when \( m \) and \( b \) give a line that is closest to passing through all of the data points.
3. Minimization to Find Fit Parameters

Let us now go through some calculations to find out what the minimum in $\delta^2$ is for our example. We have two parameters, $m$ and $b$, and $\delta^2$ must be at a minimum in its value when either one of them is varied, so we are looking for a common minimum. This happens when the derivative of $\delta^2$ with respect to either variable is zero:

$$\frac{\partial \delta^2}{\partial m} = 2 \sum_{i=1}^{N} [(mx_i + b) - y_i] \cdot x_i = 0$$

$$\frac{\partial \delta^2}{\partial b} = 2 \sum_{i=1}^{N} [(mx_i + b) - y_i] = 0$$

If we move some terms around, we find the following two equations:

$$b \sum_{i=1}^{N} x_i + m \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i$$

$$N \cdot b + m \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i.$$ 

As an aside, note that the second equation here can be rewritten as

$$b + \frac{m}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} \sum_{i=1}^{N} y_i; \quad b + m \bar{x} = \bar{y}, \quad (1)$$

Which can give us an easy way to get one of the variables $m$ or $b$ once we have the other.

For compactness, we’ll drop the explicit indication of sum indices for now, so $\sum_{i=1}^{N} x_i \equiv \sum x_i$, etc. We can write the above two equations in terms of a single matrix equation as follows:

$$\begin{bmatrix} \sum x_i \\ \sum x_i^2 \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} \sum x_i y_i \\ \sum y_i \end{bmatrix}.$$ 

Since we want to solve for $b$ and $m$, we multiply through by the inverse of the matrix on the left:

$$\begin{bmatrix} \sum x_i \\ \sum x_i^2 \end{bmatrix}^{-1} = \frac{1}{(\sum x_i)^2 - N \sum x_i^2} \begin{bmatrix} \sum x_i & -\sum x_i^2 \\ -N & \sum x_i \end{bmatrix}.$$
which yields

\[
\begin{bmatrix}
  b \\
  m
\end{bmatrix} = \frac{1}{(\sum x_i)^2 - N \sum x_i^2} \left[ \begin{array}{c}
  \sum x_i - \sum x_i^2 \\
  -N \sum x_i 
\end{array} \right] \times \left[ \begin{array}{c}
  \sum x_i y_i \\
  \sum y_i
\end{array} \right]
\]

\[
= \frac{1}{(\sum x_i)^2 - N \sum x_i^2} \left[ \begin{array}{c}
  \sum x_i \sum x_i y_i - \sum x_i^2 \sum y_i \\
  -N \sum x_i y_i + \sum x_i \sum y_i
\end{array} \right],
\]

or,

\[
b = \frac{\sum x_i \sum x_i y_i - \sum x_i^2 \sum y_i}{(\sum x_i)^2 - N \sum x_i^2}
\]

\[
= \frac{1}{N} \times \frac{\sum x_i \sum x_i y_i - \sum x_i^2 \sum y_i}{(\sum x_i)^2 - N \sum x_i^2} = \frac{\bar{x} \sum x_i y_i - \bar{y} \sum x_i^2}{N \bar{x}^2 - \sum x_i^2}
\]

\[
m = \frac{\sum x_i \sum y_i - N \sum x_i y_i}{(\sum x_i)^2 - N \sum x_i^2}
\]

\[
= \frac{1}{N} \times \frac{\sum x_i \sum y_i - N \sum x_i y_i}{(\sum x_i)^2 - N \sum x_i^2} = \frac{N \bar{x} \bar{y} - \sum x_i y_i}{N \bar{x}^2 - \sum x_i^2}.
\]

This is somewhat messy, although it really only requires the evaluation of four quantities, namely \(\bar{x}, \bar{y}, \sum x_i^2,\) and \(\sum x_i y_i,\) which is pretty straightforward.

### 4. Error Estimation

We want, of course, the errors on \(b\) and \(m\) as well. Working under the assumption that the error in the dependent variable \(y_i\) dominates (usually a good working hypothesis), we can use the propagation of errors formula to figure out the errors on the slope and the intercept. Going back to our slightly messy expressions for \(b\) and \(m\) (Equations 2 and 4), we can take the derivatives with respect to a single variable \(y_j\) (this treats all of the \(y_j\) as independent sources of error).

\[
\frac{\partial b}{\partial y_j} = \frac{\sum x_i^2 - x_j \sum x_i}{N \sum x_i^2 - (\sum x_i)^2}
\]

\[
\frac{\partial m}{\partial y_j} = \frac{N x_j - \sum x_i}{N \sum x_i^2 - (\sum x_i)^2}.
\]
Again, in order to use the error formula, which in this case is given by

\[ \sigma_F = \sqrt{\left( \frac{\partial F}{\partial y_1} \right)^2 \sigma_{y_1}^2 + \left( \frac{\partial F}{\partial y_2} \right)^2 \sigma_{y_2}^2 + \ldots + \left( \frac{\partial F}{\partial y_N} \right)^2 \sigma_{y_N}^2} \]

where \( F = b \) or \( m \), we need to have an estimate for the errors \( \sigma_{y_j} \). In general, the errors arise from both instrumental (systematic) error and statistical fluctuations in the measured values, and these errors will need to be combined to produce the total error \( \sigma_{y_j} \). As before, you can use the sample standard deviation to estimate the statistical error:

\[ \sigma_y = s_y = \sqrt{\frac{1}{N - 2} \sum_{i=1}^{N} [y_i - (mx_i + b)]^2}. \]

Note that it’s \( N - 2 \) in this case because we’ve already used up two free parameters in finding \( b \) and \( m \).

To make things easier to write, let’s define

\[ \Delta \equiv N \sum x_i^2 - \left( \sum x_i \right)^2. \]

If we make the assumption that all of the errors \( \sigma_{y_j} \) are the same (and equal to just \( \sigma \)), then we can write the error on the intercept \( b \) as

\[
\sigma_b = \sqrt{\sum_{j=1}^{N} \left[ \sum x_i^2 - x_j \sum x_i \right]^2 \frac{\sigma^2}{\Delta^2}} = \sqrt{\sum_{j=1}^{N} \left[ (\sum x_i^2)^2 - 2x_j \sum x_i \sum x_i^2 + x_j^2 (\sum x_i)^2 \right] \frac{\sigma^2}{\Delta^2}}
\]

\[
= \sqrt{[N (\sum x_i^2)]^2 - 2 (\sum x_i)^2 \sum x_i^2 + \sum x_i^2 (\sum x_i)^2 \frac{\sigma^2}{\Delta^2}}
\]

\[
= \frac{\sigma}{\Delta} \sqrt{\sum x_i^2 [N (\sum x_i^2) - 2 (\sum x_i)^2 + (\sum x_i)^2]} = \frac{\sigma}{\Delta} \sqrt{\sum x_i^2 [N (\sum x_i^2) - (\sum x_i)^2]}
\]

\[
= \frac{\sigma}{\Delta} \sqrt{\Delta \sum x_i^2} = \sigma \sqrt{\frac{\sum x_i^2}{\Delta}}.
\]
The error on the slope \( m \) can be figured out in a similar manner:

\[
\sigma_m = \sqrt{\sum_{j=1}^{N} \left[ Nx_j - \sum x_i \right]^2 \frac{\sigma^2}{\Delta^2}} = \sqrt{\sum_{j=1}^{N} \left[ N^2 x_j^2 - 2N x_j \sum x_i + (\sum x_i)^2 \right] \frac{\sigma^2}{\Delta^2}}
\]

\[
= \sqrt{N^2 (\sum x_i^2) - 2N (\sum x_i)^2 + N (\sum x_i)^2} \frac{\sigma^2}{\Delta^2}
\]

\[
= \frac{\sigma}{\Delta} \sqrt{N \left[ N (\sum x_i^2) - 2 (\sum x_i)^2 + (\sum x_i)^2 \right]} = \frac{\sigma}{\Delta} \sqrt{N \left[ N (\sum x_i^2) - (\sum x_i)^2 \right]}
\]

\[
= \frac{\sigma}{\Delta} \sqrt{N \Delta} = \sigma \sqrt{\frac{N}{\Delta}}.
\]

5. Note on Uncertainties

As mentioned above, the treatment used to derive the uncertainty on \( m \) and \( b \) assumes that each point \((x_i, y_i)\) we’re fitting is independent of all the other points. If this is not true—for example, if there is some source of error that affects more than one point at a time—then the above formulas don’t apply. Consider the examples below:

Suppose we are so excited about least squares fitting that we decide to redo the “Drops in a Bucket” lab using fitting. One way to do this would be to measure the mass of different numbers of drops of water and make a plot of mass versus number of drops. If you fit a straight line to this plot, the slope of the line would be the mass of one drop. To begin, we put 10 drops in the beaker and measure its mass. Then we add another 10 drops (for a total of 20) and measure again. We continue to add another 10 drops and measure until we get to a total of 100 drops. We use this data to make a plot of mass versus number of drops and extract the mass of a drop from the slope of the linear fit. (Do you know what the value of the intercept represents?) However, when we go to calculate the uncertainty on this estimate we realize we have a problem. Because each data point represents the effect of adding another 10 drops to the same beaker we were filling, the data points are not independent. A fluctuation in the mass of the first 10 drops, also affects the results of subsequent measurements. The measurements are not independent, and the error estimation formulas do not apply.

If we instead empty the beaker between measurements—so that we put 10 drops in, measure, empty, then put 20 drops in, measure, empty, etc.—then the measurements are independent and we can use the error estimation formulas from above.
6. Summary

For a set of \( N \) data points \((x_i, y_i)\), the best fit of a line \( y = mx + b \) through the points is parametrized by

\[
\begin{align*}
b &= \frac{\bar{x} \sum x_i y_i - \bar{y} \sum x_i^2}{N \bar{x}^2 - \sum x_i^2} \\
m &= \frac{N \bar{x} \bar{y} - \sum x_i y_i}{N \bar{x}^2 - \sum x_i^2}
\end{align*}
\]  

where all of the sums run from \( i = 1 \) to \( N \). Just like we used before, the errors on the slope and intercept depend on the errors of the individual points. The error on an individual point \((\sigma_y)\) can be computed using the sample standard deviation,

\[
\sigma_y = s_y = \sqrt{\frac{1}{N-2} \sum_{i=1}^{N} [y_i - (mx_i + b)]^2}.
\]

Note that it’s \( N-2 \) in this case because we’ve already used up two free parameters in finding \( b \) and \( m \). Assuming all of the errors to be equal \((= \sigma)\), the errors on \( b \) and \( m \) are given by

\[
\begin{align*}
\sigma_b &= \sigma \sqrt{\frac{\sum x_i^2}{\Delta}} \\
\sigma_m &= \sigma \sqrt{\frac{N}{\Delta}},
\end{align*}
\]

where

\[
\Delta \equiv N \sum x_i^2 - \left( \sum x_i \right)^2.
\]

Don’t forget that the above error estimation formulas can only be applied if the data points are independent of one another.

Bibliography


Acknowledgements

These notes were adapted from notes written by Mike Hildreth for this class in 2005.