17. Line Fitting

A very common problem in data analysis is looking for relationships between different parameters and fitting lines or surfaces to data. The simplest example is fitting a straight line and we will discuss that here – it is also covered in Chapter 4 of Pal Wessel’s notes.

Least-squares straight-line fitting

The process of fitting a straight line is one of the simplest examples of an inverse problem. For \( n \) pairs of data points \( X_i, Y_i, i = 1, 2, \ldots, n \) and fit the data with a relationship

\[
y_i = a + bX_i
\]

where \( y_i \) is the predicted value of \( Y_i \). We can use the \( \chi^2 \) statistic to measure the misfit

\[
\chi^2(a,b) = \sum_{i=1}^{n} \left( \frac{Y_i - a - bX_i}{\sigma_i} \right)^2
\]

where \( \sigma_i \) are the uncertainties (or estimates of the uncertainties which can be set to unity in the absence of better knowledge) in the \( Y \) values.

Our goal is to find the values of \( a \) and \( b \) that minimize \( \chi^2 \). To do this we find take the partial derivates of \( \chi^2 \) with respect to \( a \) and \( b \) and solve for the values of \( a \) and \( b \) at which they are both zero

\[
\frac{\partial \chi^2}{\partial a} = -2\sum_{i=1}^{n} \left( \frac{Y_i - a - bX_i}{\sigma_i^2} \right) = 0
\]

\[
\frac{\partial \chi^2}{\partial b} = -2\sum_{i=1}^{n} X_i \left( \frac{Y_i - a - bX_i}{\sigma_i^2} \right) = 0
\]

If we define

\[
S = \sum_{i=1}^{n} \frac{1}{\sigma_i}, \quad S_X = \sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}, \quad S_Y = \sum_{i=1}^{n} \frac{Y_i}{\sigma_i^2}, \quad S_{XX} = \sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2}, \quad S_{XY} = \sum_{i=1}^{n} \frac{X_iY_i}{\sigma_i^2}
\]

then equation (7) reduces to

\[
aS + bS_X = S_Y
\]

\[
aS_X + bS_{XX} = S_{XY}
\]

The solution is

\[
a = \frac{S_{XX}S_Y - S_XS_{XY}}{\Delta}
\]

\[
b = \frac{SS_{XY} - S_XS_Y}{\Delta}
\]

with

\[
\Delta = SS_{XX} - S_X^2
\]

We can also estimate the uncertainties in \( a \) and \( b \). To do this we sum the variance in \( a \) and \( b \) resulting from the variance in each of the \( Y \) values. This can be written mathematically
\[ s_a^2 = \sum_{i=1}^{n} \sigma_i^2 \left( \frac{\partial a}{\partial Y_i} \right)^2 \]
\[ s_b^2 = \sum_{i=1}^{n} \sigma_i^2 \left( \frac{\partial b}{\partial Y_i} \right)^2 \]  
\[ \text{After substituting derivatives obtained from equation (17-6) and a fair amount of manipulation we get} \]
\[ s_a^2 = \frac{S_{X\chi}}{\Delta} \]
\[ s_b^2 = \frac{S}{\Delta} \]  
\[ \text{We can also estimate the covariance of the uncertainties in } a \text{ and } b \]
\[ s_{ab}^2 = \sum_{i=1}^{n} \sigma_i^2 \left( \frac{\partial a}{\partial Y_i} \right) \left( \frac{\partial b}{\partial Y_i} \right) = -\frac{S_{X}}{\Delta} \]  
\[ \text{Our estimate of the correlation coefficient between } a \text{ and } b, \text{ becomes} \]
\[ r = \frac{s_{ab}}{s_a s_b} = \frac{-S_{X}}{\sqrt{SS_{XX}}} . \]  
\[ \text{If we assume our estimates of the uncertainty in } Y \text{ are correct, we can check if the fit is adequate (significant) at the } \alpha \text{ level by comparing our value of } \chi^2 \text{ to the critical } \chi_{\alpha}^2 \text{ for } n-2 \text{ degrees of freedom. Provided it does not exceed this value then the data is fit adequately by the straight line.} \]
\[ \text{We can test the significance of the correlation of } x \text{ and } y, \text{ by applying the } t \text{-statistic with } n-2 \text{ degrees of freedom to determine whether the slope and our estimate of its uncertainty are significantly different from 0} \]
\[ t = \frac{(b - 0)}{s_b} \]  
\[ \text{We can write the 95\% confidence limits for } b \text{ as} \]
\[ b \pm t_{0.025} s_b \]  
\[ \text{If these limits enclose zero we cannot be confident that } x \text{ and } y \text{ are correlated at the 95\% level.} \]
\[ \text{If we do not know the uncertainty of our data but know that the straight-line model is correct, then we can initially assume an uncertainty of 1 for the purpose of getting a straight line fit and then estimate it from the residuals according to} \]
\[ s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - a - bX_i)^2 \]  
\[ \text{We can use } s^2 \text{ in place of the population variance } \sigma^2 \text{ to estimate the slope uncertainties but this will lead to a underestimate of the uncertainty for small } n \text{ because there is additional uncertainty arising from using an estimate of } \sigma^2 \text{ and not its true value.} \]
(Note that Paul Wessel uses $\sigma_b$ etc instead of $s_b$ in section 4.1 but this is not consistent with the notation he introduces in Chapter 1 and that we have used since $s_b$ is clearly an estimate based on a limited sample of points not the entire population)

**Line fitting with errors in x and y**

It is important to note that equation (17-2) assumes that our determinations of x have no uncertainty. In some instances this is a good assumption — for example our determinations of time or spatial coordinate will often have negligible uncertainty. For other instances it is a poor approximation — for example if we plot the concentration of two dissolved chemicals in seawater or two trace elements in a rock, they may both have similar analytical errors. If we have errors in both variables then a better measure of misfit is given by

$$E = \sum_{i=1}^{n} \left[ \left( \frac{y_i - Y_i}{\sigma_{y,i}} \right)^2 + \left( \frac{x_i - X_i}{\sigma_{x,i}} \right)^2 \right]$$

(17-15)

where $X_i$ and $Y_i$ are the observed data and $x_i$ and $y_i$ are the modeled values that are required to lie on a straight line

$$y_i = a + bx_i$$

(17-16)

Our goal is to find the values of $a$ and $b$ that minimize $E$. To do this we use the method of Lagrange Multipliers. We can write equation (17-16) as

$$f_i = a + bx_i - y_i = 0$$

(17-17)

and since the $f_i$ values are constrained to be zero we can write equation (17-15) as

$$E = \sum_{i=1}^{n} \left[ \left( \frac{y_i - Y_i}{\sigma_{y,i}} \right)^2 + \left( \frac{x_i - X_i}{\sigma_{x,i}} \right)^2 + 2\lambda_i f_i \right]$$

(17-18)

where the $2\lambda_i$ values are unknown constant Lagrange multipliers and the factor of 2 is for algebraic convenience.

We now set the partial derivatives of $E$ to zero to find the values that give a minimum

$$\frac{\partial E}{\partial x_i} = \frac{\partial E}{\partial y_i} = \frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0$$

Now if we make the assumption that all the sigma values are equal to unity this gives

$$\frac{\partial E}{\partial x_i} = \frac{\partial}{\partial x_i} \left( x_i - X_i \right)^2 + \frac{\partial}{\partial x_i} \left( \lambda_i bx_i \right) = 2 \left( x_i - X_i \right) + 2b\lambda_i = 0$$

(17-19)

$$\frac{\partial E}{\partial y_i} = \frac{\partial}{\partial y_i} \left( y_i - Y_i \right)^2 - \frac{\partial}{\partial y_i} \left( \lambda_i y_i \right) = 2 \left( y_i - Y_i \right) - 2\lambda_i = 0$$

(17-20)

$$\frac{\partial E}{\partial a} = 2 \sum_{i=1}^{n} \left[ \frac{\partial}{\partial a} \left( \lambda_i a \right) \right] = 2 \sum_{i=1}^{n} \lambda_i = 0$$

(17-21)

$$\frac{\partial E}{\partial b} = 2 \sum_{i=1}^{n} \left[ \frac{\partial}{\partial b} \left( \lambda_i bx_i \right) \right] = 2 \sum_{i=1}^{n} \lambda_i x_i = 0$$

(17-22)

From equations (17-19) and (17-20) we can write
\[ x_i = X_i - b\lambda_i \]
\[ y_i = Y_i + \lambda_i \]  \hspace{1cm} (17-23)

Substituting for \( x_i \) and \( y_i \) in equation (17-16) yields
\[ Y_i + \lambda_i = a + b\left(X_i - b\lambda_i \right) = a + bX_i - b^2\lambda_i \]  \hspace{1cm} (17-24)

Solving for \( \lambda_i \)
\[ \lambda_i = \frac{a + bX_i - Y_i}{1 + b^2} \]  \hspace{1cm} (17-25)

Substituting for \( \lambda_i \) into equation (17-21) yields
\[ \sum_{i=1}^{n} \frac{a + bX_i - Y_i}{1 + b^2} = 0 \]  \hspace{1cm} (17-26)

Substituting for \( \lambda_i \) from equations (17-25) and for \( x_i \) from (17-23) into equation (17-22) yields
\[ \sum_{i=1}^{n} \left( \frac{a + bX_i - Y_i}{1 + b^2} \right) \left( X_i - b\lambda_i \right) = \sum_{i=1}^{n} \frac{aX_i + bX_i^2 - Y_iX_i}{1 + b^2} - \sum_{i=1}^{n} b \left( \frac{a + bX_i - Y_i}{1 + b^2} \right)^2 = 0 \]  \hspace{1cm} (17-27)

We now have reduced the \( n + 2 \) equations for \( a \), \( b \) and \( \lambda_i \) to 2 equations (17-26 and 17-27) for \( a \) and \( b \). Since the denominator in equation (17-26) cannot reduce to zero, we can write
\[ \sum_{i=1}^{n} a = \sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i \]
\[ \Rightarrow a = \overline{Y} - b\overline{X} \]  \hspace{1cm} (17-28)

where \( \overline{X} \) and \( \overline{Y} \) are the mean values of the data. We can substitute equation (17-28) into equation (17-27), multiply by \((1 + b^2)^2\), and use the variables \( U_i = X_i - \overline{X} \) and \( V_i = Y_i - \overline{Y} \) and after a few lines of manipulation get
\[ \sum_{i=1}^{n} \left[ b^2U_iV_i + b(U_i^2 - V_i^2) \right] - U_iV_i = 0 . \]  \hspace{1cm} (17-29)

This has the solution
\[ b = -\frac{\sum_{i=1}^{n} V_i^2 - \sum_{i=1}^{n} U_i^2 \pm \sqrt{\left( \sum_{i=1}^{n} U_i^2 - \sum_{i=1}^{n} V_i^2 \right)^2 + 4 \left( \sum_{i=1}^{n} U_iV_i \right)^2}}{2\sum_{i=1}^{n} U_iV_i} \]  \hspace{1cm} (17-30)

There are two solutions for \( b \) (each with a corresponding value of \( a \) from equation 17-28), one that minimizes \( E \) and a second that gives a perpendicular line that maximizes \( E \).

**Robust Line Fitting**

In a least squares line in which we assume all the data have the same uncertainty we seek to minimize
Minimize \( a, b \)  
\[ E = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 = \sum_{i=1}^{n} r_i^2 \]  
(17-31)

This process is sensitive to outliers, particularly so when the outliers lie near the lower or upper limits of the range of \( x_i \). The breakdown point for the least squares line fit (L_2 regression) is \( 1/n \). We can overcome this problem to a small extent by minimizing the sum of the absolute misfits (L_1 regression)

Minimize \( a, b \)  
\[ E = \sum_{i=1}^{n} |r_i| \]  
(17-32)

but the L_1 norm also has a breakdown point of \( 1/n \).

A robust approach with a breakdown point of \( \frac{1}{2} \) is to minimize the median misfit.

Minimize \( a, b \)  
\[ \text{median} |r| = \text{median} |Y_i - a - bX_i| \]  
(17-33)

This is equivalent to finding the narrowest strip that encloses half the points. The only way to do this, is by a systematic search through different values of \( b \). For each value of \( b \) we calculate \( Y_i - bX_i \), and then find the value of \( a \) that minimizes the median of \( |Y_i - bX_i - a| \). One then chooses the \( a \) and \( b \) values that gives the minimum median among all the values of \( b \) analyzed.

One can use this robust statistical method to find and eliminate outliers

\[
\frac{|Y_i - a - bX_i|}{\text{Median} |Y_i - a - bX_i|} > z_{\text{cut}}
\]  
(17-34)

where a value of \( z_{\text{cut}} = 4.45 \) is equivalent to 3 standard deviations for a normal distribution. Once the outliers are eliminated, one can apply the least squares line fitting approach.