The Statistical Underpinning of Least-Squares

Probability for a Set of Measured Values

Consider the probability of obtaining a single measurement from a normal pdf.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right] \tag{1}$$

Now write the product representing the probability of observing *N* values from *the same pdf*.

$$\begin{aligned} p_{total} &= p(x_1)\, p(x_2)\, p(x_3) \cdots p(x_N) \\ p_{total} &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x_i-\mu)^2}{\sigma^2} \right] \end{aligned} \tag{2}$$

Expand the product and collect terms. The pre-exponential factor has no subscripted terms, thus is the same term multiplied by itself *N* times. This is accomplished by simply raising it to a power. The product of the exponential terms can be simplified by adding the exponentials together, i.e. $e^a \times e^b = e^{a+b}$. The resultant sum is indicated by the operator, Σ.

$$p_{total} = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^N \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i-\mu)^2 \right] \tag{3}$$

The Method of Maximum Likelihood (MML)

What estimate of a pdf parameter will maximize the probability of observing the measured set of values? In this example we will estimate μ. Since μ doesn't appear in the pre-exponential term, that term can be ignored. Because the term within the exponential is negative, minimizing it (obtaining the least squares) will maximize the exponential. Start by taking the partial derivative of the term within the exponential with respect to the estimated parameter (denoted $\hat{\mu}$) and setting it equal to zero.

$$\frac{\partial}{\partial \hat{\mu}} = \frac{2}{2\sigma} \sum_{i=1}^{N} (x_i - \hat{\mu}) \tag{4}$$

Now evaluate the summation and set the derivative to zero.

$$\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i - \frac{1}{\sigma^2} N\hat{\mu} = 0$$
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

The result is the arithmetic average. That is, the arithmetic average is the estimate of μ that maximizes the probability of observing the measured set of data, $x_1 \ldots x_N$.

MML for the Linear Function $y = a_0 + a_1 x$

Let a_0 and a_1 be the "true" coefficients that are to be estimated by the method of maximum likelihood using N experimental pairs of x,y-values, (x_i , y_i). Assume that the measured value of y has error described by a normal pdf and that the magnitude of this error, σ, can vary with x. Also assume that x has no error. For any one chosen value of $x = x_i$, the probability that some particular value, y_i , would be measured is,

$$p(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\sigma_i^2}} = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y_i-a_0-a_1x_i)^2}{\sigma_i^2}} \tag{6}$$

The method of maximum likelihood will be applied to the total probability of N measurements in the belief that this procedure will provide the most probable estimate of the coefficients.

$$p_{total} = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y_i-\hat{a}_0-\hat{a}_1x_i)^2}{\sigma_i^2}} = \left[\prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}}\right]\left[e^{-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i-\hat{a}_0-\hat{a}_1x_i)^2}{\sigma_i^2}}\right] \tag{7}$$

Only the rightmost term within brackets contains $\hat{a}_0$ and $\hat{a}_1$. As a result, only this term is important in maximizing p_{total} with respect to the estimates of $\hat{a}_0$ and $\hat{a}_1$. Again, maximization is achieved by minimizing the summation within the exponent. This particular summation is so common it is given its own name, chi-square, and symbol, χ^2.

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{(y_i - \hat{a}_0 - \hat{a}_1 x_i)^2}{\sigma_i^2} \tag{8}$$

The method of maximum likelihood requires that chi-square be minimized with respect to the two coefficients. This is done by taking the partial derivatives and setting the result to zero.

$$\frac{\partial \chi^2}{\partial \hat{a}_0} = -2\sum \frac{\left(y_i - \hat{a}_0 - \hat{a}_1 x_i\right)}{\sigma_i^2} = 0 \qquad \sum \frac{y_i}{\sigma_i^2} = \left(\sum \frac{1}{\sigma_i^2}\right)\hat{a}_0 + \left(\sum \frac{x_i}{\sigma_i^2}\right)\hat{a}_1$$

$$\frac{\partial \chi^2}{\partial \hat{a}_1} = -2\sum \frac{x_i\left(y_i - \hat{a}_0 - \hat{a}_1 x_i\right)}{\sigma_i^2} = 0 \qquad \sum \frac{x_i y_i}{\sigma_i^2} = \left(\sum \frac{x_i}{\sigma_i^2}\right)\hat{a}_0 + \left(\sum \frac{x_i^2}{\sigma_i^2}\right)\hat{a}_1 \tag{9}$$

The simultaneous equations on the right can be solved several ways. For two equations and two unknowns I find solving determinants easier than matrix algebra.

$$\hat{a}_0 = \frac{\begin{vmatrix} \sum \frac{y_i}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i y_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{vmatrix}}{\Delta} = \frac{\sum \frac{y_i}{\sigma_i^2}\sum \frac{x_i^2}{\sigma_i^2} - \sum \frac{x_i y_i}{\sigma_i^2}\sum \frac{x_i}{\sigma_i^2}}{\Delta}$$

$$\hat{a}_1 = \frac{\begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{y_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i y_i}{\sigma_i^2} \end{vmatrix}}{\Delta} = \frac{\sum \frac{1}{\sigma_i^2}\sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2}\sum \frac{y_i}{\sigma_i^2}}{\Delta} \tag{10}$$

$$\Delta = \begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{vmatrix} = \sum \frac{1}{\sigma_i^2}\sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2$$

Note that Δ has no y-terms, thus is free of error.

For complicated expressions like these, I like to check my work by using a units analysis. To this end, the σ^2 terms have units of y^2, thus Δ has units of x^2/y^4. As a result,

a_0 has units of $(x^2y/y^4)/(x^2/y^4) = y$
a_1 has units of $(xy/y^4)/(x^2/y^4) = y/x$

When a_1 is multiplied by x the result has units of y. Thus, the left and right sides of $y = a_0 + a_1x$ both have the same units!

MML for an Un-weighted Least-Squares

An un-weighted least-squares is used when all y-values have the same standard deviation. Thus, all $\sigma_i = \sigma$.

$$\Delta \equiv N\sum x^2 - \left(\sum x\right)^2$$
$$\hat{a}_0 = \frac{\sum y \sum x^2 - \sum xy \sum x}{\Delta} \tag{11}$$
$$\hat{a}_1 = \frac{N\sum xy - \sum x \sum y}{\Delta}$$

where simplification relies on,

$$\sum_{i=1}^{N} \frac{1}{\sigma^2} = N\left(\frac{1}{\sigma^2}\right)$$

Note that the units of Δ are now x^2, again giving a_0 units of y and a_1 units of y/x.

Errors in the Least-Squares Coefficients

This derivation is quite complex and is presented for the sole purpose of demonstrating that the error equations used in straight-line fits can be derived using nothing more than the propagation of variance.

A detailed derivation will only be given for the error in the intercept. The error in the slope is left as an exercise for the reader. Use the form of a_0 where each individual y_i has a unique σ_i, that is, a weighted least-squares.

$$a_0 = \frac{\sum_i \frac{y_i}{\sigma_i^2} \sum_i \frac{x_i^2}{\sigma_i^2} - \sum_i \frac{x_i y_i}{\sigma_i^2} \sum_i \frac{x_i}{\sigma_i^2}}{\Delta} \tag{12}$$

Note that the only terms having errors are those with y. The denominator, Δ, has no y-terms. Thus, the equation is linear in y and the variance of the intercept can be determined by simple propagation,

$$\sigma_{a_0}^2 = \sum_j \left(\frac{\partial a_0}{\partial y_j}\right)^2 \sigma_j^2 \tag{13}$$

where two different indices will be used, i and j. The index i appears within the summations used to define a_0 and a_1. The index j is used to keep track of the particular y-value used in the

partial derivative. Both indices run over the range of 1 to *N*. The index *j* will eventually go away!

All the partial derivatives in Eq. 13 have the same functional form, differing only in the **subscript *j***. Note that the summations over y_i and $x_i y_i$ in Eq. 12 only have one y_j. All of the other terms in the summation have different subscripts, $i \neq j$, thus their partial derivative is zero.

$$\frac{\partial a_0}{\partial y_j} = \frac{1}{\Delta}\left[\frac{1}{\sigma_j^2}\sum_i \frac{x_i^2}{\sigma_i^2} - \frac{x_j}{\sigma_j^2}\sum_i \frac{x_i}{\sigma_i^2}\right] \tag{14}$$

Now square the partial derivative so it can be substituted into the equation for the propagation of variance.

$$\left(\frac{\partial a_0}{\partial y_j}\right)^2 = \frac{1}{\Delta^2}\left[\frac{1}{\sigma_j^4}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)^2 - \frac{2x_j}{\sigma_j^4}\sum_i \frac{x_i^2}{\sigma_i^2}\sum_i \frac{x_i}{\sigma_i^2} + \frac{x_j^2}{\sigma_j^4}\left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2\right] \tag{15}$$

Multiply the square of the partial derivative by the variance of y_i , σ_i^2.

$$\left(\frac{\partial a_0}{\partial y_j}\right)^2 \sigma_j^2 = \frac{1}{\Delta^2}\left[\frac{\sigma_j^2}{\sigma_j^4}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)^2 - \frac{2x_j\sigma_j^2}{\sigma_j^4}\sum_i \frac{x_i^2}{\sigma_i^2}\sum_i \frac{x_i}{\sigma_i^2} + \frac{x_j^2\sigma_j^2}{\sigma_j^4}\left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2\right] \tag{16}$$

Each of these *j* partial derivatives represent one term in the summation of Eq. 13. Finally, sum all the *j* terms of Eq. 13. The above equation has three separate terms. Apply the summation over each of them.

$$\sigma_{a_0}^2 = \frac{1}{\Delta^2}\left[\sum_j \frac{\sigma_j^2}{\sigma_j^4}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)^2 - 2\sum_j \frac{x_j\sigma_j^2}{\sigma_j^4}\sum_i \frac{x_i^2}{\sigma_i^2}\sum_i \frac{x_i}{\sigma_i^2} + \sum_j \frac{x_j^2\sigma_j^2}{\sigma_j^4}\left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2\right] \tag{17}$$

At this point the j-index has served its purpose in identifying partial derivatives with respect to specific *y* values. All summations over *j* can now be converted into **summations over *i***.

$$\sigma_{a_0}^2 = \frac{1}{\Delta^2}\left[\sum_i \frac{1}{\sigma_i^2}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)^2 - 2\sum_i \frac{x_i}{\sigma_i^2}\sum_i \frac{x_i^2}{\sigma_i^2}\sum_i \frac{x_i}{\sigma_i^2} + \sum_i \frac{x_i^2}{\sigma_i^2}\left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2\right] \tag{18}$$

Add the last two terms of Eq. 18 together.

$$\sigma_{a_0}^2 = \frac{1}{\Delta^2}\left[\sum_i \frac{1}{\sigma_i^2}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)^2 - \sum_i \frac{x_i^2}{\sigma_i^2}\left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2\right] \tag{19}$$

Finally, pull out the common term within the square brackets.

$$\sigma_{a_0}^2 = \frac{1}{\Delta^2}\left[\sum_i \frac{x_i^2}{\sigma_i^2}\left\{\sum_i \frac{1}{\sigma_i^2}\sum_i \frac{x_i^2}{\sigma_i^2} - \left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2\right\}\right] \tag{20}$$

Referring to Eq. 10, the term within the curly braces is seen to be Δ. Make this substitution and simplify.

$$\sigma_{a_0}^2 = \frac{1}{\Delta^2}\left[\sum_i \frac{x_i^2}{\sigma_i^2}\Delta\right] = \frac{1}{\Delta}\sum_i \frac{x_i^2}{\sigma_i^2} \tag{21}$$

QED.

In an equally laborious manner it can be shown that the variance of the slope is given by the following.

$$\sigma_{a_1}^2 = \frac{1}{\Delta}\sum_i \frac{1}{\sigma_i^2} \tag{22}$$

Equations for coefficient variance with an un-weighted least-squares can be obtained by replacing all σ_j with σ. When doing this remember that Δ has σ terms that will be eliminated!

$$\sigma_{a_0}^2 = \frac{\sigma^2}{\Delta}\sum_i x_i^2 \qquad\qquad \sigma_{a_1}^2 = \frac{N\sigma^2}{\Delta} \tag{13}$$

In order to evaluate these equations you need to have an estimate of σ^2. The estimate can be obtained by computing the deviations between the measured *y* and the value computed with the coefficients. Since two parameters have been computed from the data set, the degrees of freedom are *N* - 2.

$$s^2 = \frac{1}{N-2}\sum_i \left(y_i - \hat{a}_0 - \hat{a}_1 x_i\right)^2 \tag{14}$$