Pre- and Post-Fit Sigmas

John Gipson

January 6, 1999

Abstract

This note describes the relation between pre- and post-fit sigmas.

1 Introduction

In least squares problems you start with a set of observations which you assume are described by a set of n parameters which you want to derive. One way of looking at this is that you have an n-parameter set of functions, and you want to find the function that matches the data best. Each observation has assigned to it an uncertainty σ_j, where the subscript j labels the observation. This uncertainty is the expected difference between the ideal observation in the absence of noise and the observation made in the real world. It is the scatter of an ensemble of identical observations. If your model is correct, this is also the expected difference between the true model, perhaps known only to God, and the observations. Once you estimate the parameters you have another model, hopefully close to the true model. The differences between the observations and the derived model are the post-fit residuals Δ_j. Naively you would think that the expected size of the post-fit residuals is the same as the uncertainty, explicitly:

$$\sqrt{\left\langle \Delta_j^2 \right\rangle} = \sigma_j \qquad \text{This is not correct!}$$

Here $\langle ... \rangle$.denotes mathematical expectation value. It turns out that this is not the case, and in fact one can show that:

$$\sqrt{\left\langle \Delta_j^2 \right\rangle} < \sigma_j$$

In words, the post-fit residuals will generally be smaller than the uncertainty of the measurement. Since the size of the post-fit residual of an individual point is often used to discard outliers, it is important that we able to calculate what the expected size is. I will refer to expectation value of the post-fit residual as the post-fit sigma, and I will denote it by $\widehat{\sigma}_j$:

$$\sqrt{\left\langle \Delta_j^2 \right\rangle} = \widehat{\sigma}_j$$

A concrete example may clarify the proceeding paragraph. Suppose that you expect that your data will lie along a straight line. This line is specified by two parameters, the offset and the slope. The expected deviation of an individual point from the "true" line is given by it's uncertainty, σ_j You can do a least squares fit to obtain a straight line through the points. This line will in general

be different from, although close to, the "true" line. Since this line is derived from the data, it seems plausible that residuals of individual data points will be smaller for it then for the true line. In fact, if you have just two points, the residuals will be 0.

In the section 2 we review the least squares technique. In section 3 we derive the formula for the post fit sigmas. In section 4 we apply these formulas to the familiar case of fitting data to a straight line.

All of this discussion ignores the effects of constraints, which is covered in another memo on "Constraints, Covariance and χ^2".

2 Lightning Review of Least Squares

In least squares problems you have a set of parameters A_a which you want to estimate from a set of observations O_j. You start by assuming that in the absence of noise, each observation O_j is given in terms of a known function $F_j(A)$:

$$O_j = F_j(A)$$

The next step is to linearize this problem around some value for the A's. This is done by expanding the $F(A)$ around a default for the A's.

$$\begin{aligned} O_j &\cong F_j(A_0) + \sum_a \frac{\partial F_j(A_0)}{\partial A_a} \delta A_a \qquad (1) \\ &= F_j(A_0) + \sum_a f_{j,a}(A_0) \delta A_a \end{aligned}$$

where

$$f_{j,a}(A_0) = \frac{\partial F_j(A_0)}{\partial A_a} \qquad (2)$$

The first term on the right hand side (RHS) of equation (1) is the calculated observation at the a priori values for the parameters. If the F_j are linear functions of the parameters, then this equality will be exact. (Even if they are not, the linear approximation is generally good if the a priori value for the A's is close to the true value, and the F's are smooth functions. If these conditions are not met, then it may be necessary to iterate the least squares process.) Before proceeding, it is useful to introduce some new notation. Let

$$O_{C,i} = O_j - F_j(A_0)$$

Then we can rewrite equation (1) in matrix notation as follows:

$$O_C = f(A_0)\delta A$$

In what follows we will usually suppress the dependence on A_0.

Our goal is to find the values for δA_a which make this equation hold. One way of doing this is to form the sum of the residuals:

$$\chi^2 = \sum_j \Delta_j^2 \qquad (3)$$

where

$$\Delta_j = [O_C - f\delta A_a]_j$$

By assumption, for noiseless data and with the correct choice of δA_a the residuals vanish. The general approach is to vary the δA_a to minimize χ^2.At this point we could proceed to derive the least squares equations for noiseless data, but will forgo doing so since we are interested in the case with noise.

In the real world data has noise, and even for the true values of δA_a we have:

$$O_{C,j} = [f\delta A]_j + \varepsilon_j \tag{4}$$

This equation will be true if we are correctly modeling the observations. If we are not, then this needs to be modified by an additional term. For example, if we are trying to modify data as straightline, but the data is really parabolic, then this equation is incorrect.

For any individual observation we may be able to adjust the parameters so that the left hand side is 0, but in general we won't be able to choose the parameters so that this is true for all observations simultaneously. Instead we attempt to minimize the residuals in some average sense which will be made clear momentarily. Let σ_j be the uncertainty in the j-th measurement:

$$\sigma_j^2 = \left\langle \varepsilon_j^2 \right\rangle$$

Then it is customary to replace equation (3) by the weighted sum:

$$\chi^2 = \sum_j \frac{\Delta_j^2}{\sigma_j^2}$$

The weighting factor insures that poor data is downweighted with respect to good data. Roughly speaking, if σ_j is twice as large as σ_k then we expect Δ_j to be twice as large Δ_k We derive the our estimate of δA_a by minimizing χ^2 with respect to them. This is done by setting the derivatives of χ^2 equal to 0:

$$\frac{\partial \chi^2}{\partial \delta A_a} = 0$$

Evaluating this is straightforward, and we obtain the ordinary least squares equations:

$$\sum_a \left[\sum_j f_{j,b}(A_0) f_{j,a}(A_0) \frac{1}{\sigma_j^2} \right] \delta \widehat{A}_a = \sum_j \frac{1}{\sigma_j^2} f_{j,b} O_{C,j}$$

Here the "hat" on δA indicates that it is an estimate, and not the true value, which is unknowable. This equation can be rewritten in matrix notation as:

$$\left[f^T W^{-1} f \right] \delta \widehat{A} = f^T W^{-1} O_C \tag{5}$$

Where

$$W = diag(\sigma_1^2, \sigma_2^2, \sigma_3^2, ...)$$

This is the covariance matrix of the observations

$$W_{i,j} = \langle \varepsilon_j \varepsilon_k \rangle \tag{6}$$

and I have assumed that the observations are uncorrelated, i.e., W is diagonal. If the observations are correlated then W will not be diagonal. Equation (5) will remain the correct least squares equation, where W is given by equation (6). The solution for the $\delta\widehat{A}_a$ is given by:

$$\delta\widehat{A} = \left[f^T W^{-1} f\right]^{-1} f^T W^{-1} O_C$$

This of course assumes that $\left[f^T W^{-1} f\right]$ is invertible.

3 Derivation of Post Fit Residuals and Expected χ^2

At this point we have derived a best fit model from the data. It is natural to ask what is the expected size of the residuals for this particular solution That is, given the $\delta\widehat{A}$ found by the above procedure, what is:

$$\left\langle \Delta_j^2 \right\rangle \equiv \widehat{\sigma}_j^2$$

for a given observation? As mentioned previously, naively one might expect that

$$\widehat{\sigma}_j^2 = \sigma_j^2$$

However, a simple counter-example shows that this can not be the case. If we have two points and are estimating a straight-line, then the post-fit residuals are 0, that is, $\widehat{\sigma}_j = 0$ even though the uncertainty of each observation is not. The reason our intuition is incorrect is that our estimates depend on the observations, and indeed are constructed from the observation.

As a first step in deriving $\widehat{\sigma}_j^2$ we rewrite the residuals entirely in terms of the observations:

$$\begin{aligned}
\Delta &= O_C - f\delta\widehat{A} \\
&= O_C - f\left[f^T W^{-1} f\right]^{-1} f^T W^{-1} O_C \\
&= \left(I - f\left[f^T W^{-1} f\right]^{-1} f^T W^{-1}\right) O_C
\end{aligned}$$

Equation (4) gives O_C in terms of the true parameters and the noise. Substituting this in we obtain:

$$\Delta = \left(I - f\left[f^T W^{-1} f\right]^{-1} f^T W^{-1}\right)(f\delta A + \varepsilon)$$

It is trivial to show that

$$\left(I - f\left[f^T W^{-1} f\right]^{-1} f^T W^{-1}\right) f = 0$$

and hence the $f\delta A$ term vanishes. We are left with:

$$\Delta = \left(I - f\left[f^T W^{-1} f\right]^{-1} f^T W^{-1}\right)\varepsilon$$

Hence it follows that

$$\langle \Delta_j \Delta_k \rangle = \sum_{p,q} \left(I - f\left[f^T W^{-1} f\right]^{-1} f^T W^{-1}\right)_{j,p} \langle \varepsilon_p \varepsilon_q \rangle \left(I - f\left[f^T W^{-1} f\right]^{-1} f^T W^{-1}\right)_{k,q}$$

However, $\langle \varepsilon_p \varepsilon_q \rangle$ is just the covariance of the observations, which was given in equation (6) above. This equation becomes:

$$\begin{aligned} \langle \Delta_j \Delta_k \rangle &= \left[\left(I - f \left[f^T W^{-1} f \right]^{-1} f^T W^{-1} \right) W \left(I - W^{-1} f \left[f^T W^{-1} f \right]^{-1} f^T \right) \right]_{j,k} \\ &= \left[W - f \left[f^T W^{-1} f \right]^{-1} f^T \right]_{j,k} \end{aligned} \tag{7}$$

For general j, k this equation gives the covariance of the residuals with each other. If we specialize to $j = k$ this equation gives the expected size of a residual for a given observation:

$$\left\langle \Delta_j^2 \right\rangle \equiv \widehat{\sigma}_j^2 = \sigma_j^2 - \left[f \left[f^T W^{-1} f \right]^{-1} f^T \right]_{jj}$$

Now the term $[]_{jj}$ is positive, which implies that the post-fit sigmas are always smaller than the prefit sigmas.

Before closing this section, it is interesting to derive one more relationship. Namely, what is the expected size of χ^2? By definition:

$$\chi^2 = \sum_j \frac{\Delta_j^2}{\sigma_j^2}$$

Which can be rewritten as:

$$\begin{aligned} \chi^2 &= \Delta^T W^{-1} \Delta \\ &= \mathrm{Trace}(W^{-1} \Delta \Delta^T) \end{aligned}$$

Hence

$$\begin{aligned} \left\langle \chi^2 \right\rangle &= \mathrm{Trace}(W^{-1} \left\langle \Delta \Delta^T \right\rangle) \\ &= \mathrm{Trace}(W^{-1} (W - f \left[f^T W^{-1} f \right]^{-1} f^T) \\ &= \mathrm{Trace} I_{num_obs} - \mathrm{Trace} I_{num_params} \\ &= num_obs - num_params \end{aligned}$$

In going from the first to the second line I have used equation (7). In going to the third line I have used the cyclic properties of traces. In the third line I_{num_obs} and I_{num_params} are unit vectors labeled by their dimension, and num_obs and num_params are the number of observations and parameters. This leads to the familiar result that the expected value for $\langle \chi^2 \rangle$ is the number of observations minus the number of parameters you are estimating.

4 Application to Straight Line

In this section we apply the results of the last section to a simple example: the estimation of a straight line by least squares. By assumption

$$O_j = A_0 + t_j A_1$$

Suppose that we have n equally spaced observations starting at 1, all with the same uncertainty σ. Then the matrix f takes the form:

$$f = \begin{pmatrix} 1 & 1 & .. & 1 & ... & 1 \\ 1 & 2 & .. & j & ... & n \end{pmatrix}$$

and

$$\begin{aligned} fWf^T &= \frac{1}{\sigma^2} \begin{pmatrix} \sum_{j=1}^{n} 1 & \sum_{j=1}^{n} j \\ \sum_{j=1}^{n} n & \sum_{j=1}^{n} n^2 \end{pmatrix} \\ &= \frac{1}{\sigma^2} \begin{pmatrix} n & \frac{n(n+1)}{2} \\ \frac{n(n+1)}{2} & \frac{n(n+1)(2n+1)}{6} \end{pmatrix} \end{aligned}$$

The inverse of the normal matrix is:

$$\left[fWf^T\right]^{-1} = \sigma^2 \frac{12}{n(n^2-1)} \begin{pmatrix} \frac{(n+1)(2n+1)}{6} & -\frac{(n+1)}{2} \\ -\frac{(n+1)}{2} & 1 \end{pmatrix}$$

It is straightforward to show that:

$$\begin{aligned} \widehat{\sigma}_j^2 &= \sigma^2 \left[1 - \frac{12}{n(n^2-1)} \begin{pmatrix} 1 & j \end{pmatrix} \begin{pmatrix} \frac{(n+1)(2n+1)}{6} & -\frac{(n+1)}{2} \\ -\frac{(n+1)}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ j \end{pmatrix} \right] \\ &= \sigma^2 \left[1 - \frac{12}{n(n^2-1)} \left\{ \left(\frac{n+1}{2} - j \right)^2 + \frac{n^2-1}{12} \right\} \right] \end{aligned}$$

Note that for $n = 2$, this vanishes at the end points. This is symmetric around reflection about the middle, namely under the substitution:

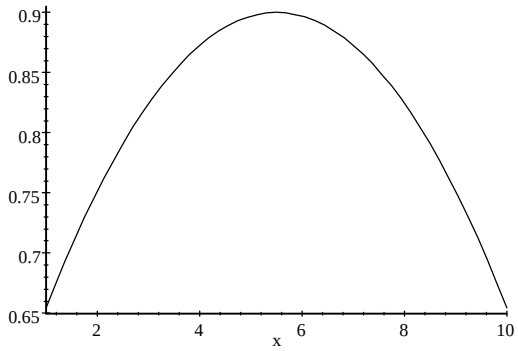
$$j \to n + 1 - j$$

For $n = 3$ we have

$$\begin{aligned} \widehat{\sigma}_1^2 &= \widehat{\sigma}_3^2 = \sigma^2 \frac{1}{6} \\ \widehat{\sigma}_2^2 &= \sigma^2 \frac{2}{3} \end{aligned}$$

All of the residuals are smaller than σ^2 but the effect is more pronounced at the end points. . You can show that in general

$$\begin{aligned} \widehat{\sigma}_1^2 &= \widehat{\sigma}_n^2 = \sigma^2 \frac{n-1}{n} \frac{n-2}{n+1} \\ \widehat{\sigma}_{n/2}^2 &= \sigma^2 \frac{n-1}{n} \end{aligned}$$

Since $\frac{n-2}{n+1} < 1$ the post fit residuals of the endpoints are always smaller than the middle points. This is the basis of the folklore that for the least squares fit for a straight-line, the end points matter

more than the middle As n gets large the term $\frac{n-2}{n+1}$ approaches 1, and the discrepancy between middle and end points diminishes. The figure below plots $\widehat{\sigma}_j^2/\sigma^2$ as a function of j for $n = 10$.