Notes on uncertainties in extrapolating a multiple regression

G. W. Hammett, et.al.
Princeton University Plasma Physics Laboratory
P.O. Box 451, Princeton, NJ 08543 USA

Abstract

Here is the abstract.

1 First section

A general linear regression equation can be written in the form:

$$y = \sum_i a_i x_i = \vec{a} \cdot \vec{x}$$

[Bevington tends to separate out a constant term, but a formula of the form $y = a_0 + a_1 x$ can always be written in this form by defining $x_0 = 1$, and I think this form is more compact and simpler.] A convenient way to derive the error propagation formulas is to assume that each regression coefficient $a_i = \bar{a}_i + \delta a_i$, where $\bar{a}_i$ is the true value of a_i, and δa_i is a random variable that represents the uncertainty in a_i. δa_i has a mean of 0 and a variance of σ_{ai}^2. I.e. upon ensemble averaging we have

$$\langle a_i \rangle = \bar{a}_i$$

and

$$\langle (a_i - \bar{a}_i)^2 \rangle = \langle (\delta a_i)^2 \rangle = \sigma_{ai}^2$$

And of course the mean value of y is the trivial result

$$\langle y \rangle = \bar{y} = \sum_i \bar{a}_i x_i$$

Given a multiple regression formula, what is the uncertainty in the predicted y (which might be the H-mode power threshold) for a new set of parameters $\vec{x}$ (which might represent ITER for example, or which might represent C-MOD if we are trying to predict C-MOD from the rest of the database)? The uncertainty in y is the square root of

$$\sigma_y^2 = \langle (y - \bar{y})^2 \rangle = \langle (\sum_i \delta a_i x_i)^2 \rangle$$

$$= \langle (\sum_i \delta a_i x_i)(\sum_j \delta a_j x_j) \rangle$$

Or

$$\sigma_y^2 = \sum_i \sum_j x_i \langle \delta a_i \delta a_j \rangle x_j$$
$$= \sum_i \sum_j x_i \sigma_{a,i,j}^2 x_j$$
$$= \vec{x} \cdot \vec{\vec{\sigma}}_a^2 \cdot \vec{x} \qquad (1)$$

And the important point to remember is that the error in the i'th coefficient may be correlated with the error in the j'th coefficient so in general one has to keep this full matrix. To check this result, we note that in the 2-D limit, this reproduces one of Bevington's summary formulas for error propagation. On p. 64 of the original edition of his book, at the end of chapter 4 on propagation of errors, he notes that for the formula $x = au \pm bv$, where a and b are fixed constants but u and v have uncertainties, then the uncertainty in x is given by

$$\sigma_x^2 = a^2 \sigma_u^2 + b^2 \sigma_v^2 + 2ab\sigma_{uv}^2$$

This is equivalent to Eq. (1). [I am using Bevington's early Fortran edition, where Bevington is sole author. There is a later edition that has programs in Basic instead of Fortran I think, and I think it is written posthumously with a co-author.]

To make sure we all understand, I will use some standard notation for linear regression as given in a tutorial by Otto J.W.F. Kardaun and Andreas Kus, "Basic Probability Theory and Statistics for Experimental Plasmas Physics" (IPP 5/68, September 1996, Max-Planck-Institut Fur Plasmaphysik). Starting with Eq. 3.3 on p. 60, a standard set of data for regression can be written as

$$\vec{Y} = \vec{\vec{X}} \vec{\alpha} + \vec{E}$$

where Y_i is the i'th observation of the dependent variable, X_{ij} is the set of j independent variables for the i'th observation (and the first independent variable is always 1 to represent the constant offset term), and E_i is the error on the i'th observation. The standard linear regression formula for the estimate of the value of the coefficients $\vec{\alpha}$ that minimizes the RMS error is given by Kardaun Eq. 3.13:

$$\hat{\vec{\alpha}} = (\vec{\vec{X}}^t \vec{\vec{X}})^{-1} \vec{\vec{X}}^t \vec{Y}$$

and the covariance matrix for $\vec{\alpha}$, which gives the uncertainties in the α_i's and the correlations between those uncertainties, is given by Kardaun Eq. 3.15:

$$\vec{\vec{\mathrm{Var}}}(\hat{\vec{\alpha}}) = (\vec{\vec{X}}^t \vec{\vec{X}})^{-1} \sigma^2$$

where the standard assumption is that σ^2 is estimated from the errors as $\sigma^2 = \sum_i E_i^2/(N - p - 1)$, where N is the number of independent observations and (p+1) is the number of fit coefficients α_j. This variance matrix is what I define as $\vec{\vec{\sigma}}^2{}_a$ in Eq. (1) (the $\vec{a}$ I used in deriving Eq. (1) is the same as Kardaun's $\vec{\alpha}$).

Kardaun's tutorial show various limiting cases of this formula, and his Fig 3.3 illustrates the main point that the uncertainty is smallest if you evaluate x in the middle of the data base, but increases as you extrapolate. A trivial limit to check (which I think is Kardaun's Case 1 on p. 66) is the 0-d case, where there is a single unknown parameter $\alpha_1 = \alpha$. In this case $X_{i1} = 1$, and one finds that

$$\vec{\vec{X}}^t \vec{\vec{X}} = N$$

so $Var(\alpha) = \sigma^2/N$. This is the standard result that if you have N observations, the uncertainty in the mean is less than the scatter of the N observations by a factor of $1/\sqrt{N}$.

To make sure you understand these results, another check is in the 1-D case of simple least-squares fit to a straight line. The result is that for an equation of the form $y = a + bx$, the uncertainty in a predicted $\hat{y}$ when extrapolated to $\hat{x}$, is the square root of

$$\sigma_{\hat{y}}^2 = \frac{\sigma^2}{N} \left[1 + \lambda^2 \right] \tag{2}$$

where

$$\lambda = \frac{\hat{x} - \bar{x}}{\sigma_x}$$

is the distance being extrapolated from the center of the database, in units of standard deviations over which x has been varied. I.e., $\bar{x} = \sum_i x_i/N$ and $\sigma_x^2 = \sum_i (x_i - \bar{x})^2/N$. One can get this same result from Bevington's summary formulas at the end of his Chapter 6 (on least squares fit to a straight line) by combining his formulas for σ_a^2 and σ_b^2 in the appropriate way, for the case where the x variable has been redefined so that $\bar{x} = 0$. To get the full formula of Eq. (2), one would need to generalize Bevington's calculation to include the cross-correlation between the errors in a and b (in general the cross-correlation σ_{ab}^2 is nonzero unless $\bar{x} = 0$, and Bevington neglected to write down this cross-correlation formula for the 1-D case). For multiple regression, he does give the covariance matrix, but again in a special form separating out the a_0 term. I think the above formulas are in a simpler form.