Least Squares in a Nutshell

INTRODUCTION

The method of least squares (LS) is the default data analysis tool in most of physical science. The statistical properties of *linear* least-squares estimators are well known.[1-3] These include, importantly: If the data are distributed independently and normally (*i.e.*, with the Gaussian distribution) about their true values, the estimators of the parameters will be *unbiased* and *minimum-variance* and will themselves be normally distributed about their true values, with variances that can be calculated directly from the *matrix of the normal equations*. The structure of the latter quantity depends only upon the distribution of values for the independent variable(s) and therefore can be calculated at the outset, permitting its use in *experimental design*.

Often the theoretical expressions that predict the relationships among the experimentally measured variables are not linear in the desired parameters, necessitating analysis by *nonlinear* least squares. Included among these cases are most of the situations in which the directly measured quantities are transformed into a form that yields a linear plot, $y' = A + B\,x'$, where y' and x' are the transformed variables, and where the linear parameters A and B may depend in various ways on the original parameters, a and b.

From a computational standpoint the chief differences between linear and nonlinear LS are the following: Linear LS equations are, in principle, solvable in a single step and yield a single, minimum-variance solution; nonlinear problems must be solved iteratively and may converge on multiple solutions, or may not converge at all.

THEORETICAL BACKGROUND

Linear Least Squares

Matrix Formulation. The least-squares equations are obtained by minimizing the sum of weighted squared residuals S,

$$S = \sum w_i \delta_i^2 , \tag{1}$$

with respect to a set of adjustable parameters β, where δ_i is the residual (observed–calculated mismatch) for the ith point and w_i is its weight. For the purpose of the matrix notation that follows, β is a column vector containing p elements, one for each adjustable parameter. Thus its transpose is a row vector: $\beta^{\mathbf{T}} = (\beta_1, \beta_2, \ldots \beta_p)$.[a] The solution to the minimization problem in the linear case is the set of p equations in p unknowns given in matrix form by

---

[a] The problem is a linear one if the measured values of the dependent variable (y) can be related to those of the independent variable(s) (x, u, ...) and the adjustable parameters through the matrix equation,[2]

$$\mathbf{y} = \mathbf{X}\,\beta + \delta , \tag{2}$$

$$\mathbf{A}\ \beta = \mathbf{z}, \tag{3}$$

The matrix **A** is the previously mentioned *matrix of the normal equations*.[b] Equations (3) are solved for the parameters β, *e.g.*,

$$\beta = \mathbf{A}^{-1}\ \mathbf{z}\ , \tag{4}$$

where $\mathbf{A}^{-1}$ represents the inverse of **A**. Knowledge of the parameters then permits calculation of the residuals δ from Eq. (2) and thence of S. Importantly, the variances in the parameters are the diagonal elements of the *variance-covariance matrix* **V**, which is proportional to $\mathbf{A}^{-1}$ (see below).

For these equations to make sense, it is essential that the measurements y_i be drawn from parent distributions of finite variance.[2] If additionally the data are *unbiased*, then the LS equations will yield unbiased estimates of the parameters β. If the data distributions are normal, then the parameter estimates will also be normally distributed. For these to be *minimum-variance* estimates as well, it is necessary that the weights be taken as proportional to the inverse variances,[1,2]

$$w_i \quad \sigma_i^{-2}\ . \tag{5}$$

Under these conditions, least squares is also a *maximum likelihood* method, which is reassuring, since maximum likelihood is the more fundamental statistical principle behind data analysis in physical science. Note that it is possible to have LS estimators that are unbiased but not minimum-variance, or minimum-variance but not unbiased, or even unbiased and minimum variance, but nonnormal.

If the parent distributions for the data are normal and the proportionality constant in Eq. (5) is taken as 1.00, then the quantity S is distributed as a χ^2 variate for $\nu = n - p$ degrees of freedom.[1–3] Correspondingly, the quantity S/ν follows the *reduced chi-square* (χ_ν^2) distribution, given by

$$P(z)\ dz = C\ z^{(\nu-2)/2} \exp(-\nu\ z/2)\ dz\ , \tag{6}$$

where $z = \chi_\nu^2$ and C is a normalization constant. It is useful to note that a χ^2 variate has a mean of ν and a variance of 2ν, which means that χ_ν^2 has a mean of unity and a variance of $2/\nu$. In the limit of large ν, $P(z)$ becomes Gaussian. [See Figure, next page.]

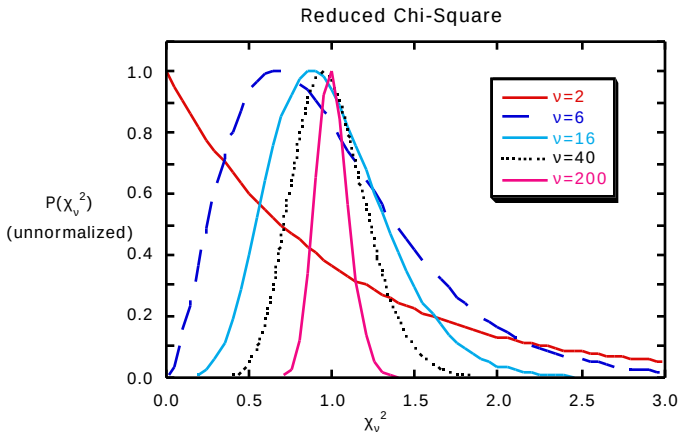
A Priori and a Posteriori Weighting. If the proportionality constant in Eq. (5) is taken as unity, then the proportionality constant connecting **V** and $\mathbf{A}^{-1}$ is likewise unity, giving

---

where **y** and δ are column vectors containing n elements (for the n measured values), and the *design matrix* **X** has n rows and p columns, and depends only on the values of the independent variable(s) (assumed to be error-free) and not on the parameters β or dependent variables **y**. For example, a fit to $y = ax + b/x^2 + c\ \ln(3u)$ qualifies as a linear fit, with two independent variables (x, u), three adjustable parameters (a, b, c), and **X** elements $X_{i1} = x_i$, $X_{i2} = x_i^{-2}$, $X_{i3} = \ln(3u_i)$. On the other hand, the fit becomes nonlinear if, for example, the first term is changed to x/a, or the third to $3\ln(cu)$. It also becomes nonlinear if one or more of the "independent" variables is not error-free, hence is treated (along with y) as a dependent variable.

[b] We worked out the form of the matrix **A** for several specific cases. For *any* linear problem it can be shown to be given by $\mathbf{A} = \mathbf{X^T\ W\ X}$, where the square weight matrix **W** is, in this case, diagonal, with n elements $W_{ii} = w_i$. The vector **z** can similarly be given as a simple matrix formula, $\mathbf{z} = \mathbf{X^T\ W\ y}$. And S is, in matrix form, $S = \delta^{\mathbf{T}}\ \mathbf{W}\ \delta$.

$$\mathbf{V} = \mathbf{A}^{-1} . \qquad (7)$$



If additionally, the parent data distributions are normal, the parameter distributions are also normal, as already noted. Then the confidence intervals for the parameters can be evaluated straightforwardly from the standard error function tables. For example, the 95% confidence interval on β_1 is $\pm 1.96\ V_{11}^{1/2}$. This is always the case in Monte Carlo calculations on linear fit models with normally distributed errors, since the σ_{yi} are set by the computationalist. Accordingly, in such calculations 95% of the estimates of β_1 are expected within $\pm 1.96\ V_{11}^{1/2}$ of the true value.

The use of Eq. (7) implies prior knowledge of the statistics of the y_i. Accordingly the weights obtained in this manner may be designated as *a priori* weights. Note that this *a priori* **V** is also *exact*, not an estimate. Unfortunately, from the experimental side we never have perfect *a priori* information about the statistics of the data. However, there are cases, especially with extensive computer logging of data, where the *a priori* information may be good enough to make Eq. (7) the proper choice and the resulting **V** *virtually* exact. A good example is data obtained using counting instruments, which often follow Poisson statistics closely, so that the variance in y_i (σ_{yi}^2) can be taken as y_i . (For large y_i Poisson data are also very nearly Gaussian.)

At the other extreme we have the situation where nothing is known in advance about the statistics of the y_i, except that we have good reason to believe the parent distributions all to have the same variance, independent of y_i. In this case the weights w_i can all be taken to be the same constant, which without loss of generality we can take to be 1.00. This is the case of unweighted least squares. The variance in y is then estimated from the fit itself, as

$$\sigma_y^2 \approx s_y^2 = \frac{\sum \delta_i^2}{n - p} = \frac{S}{\nu} , \qquad (8)$$

which is recognized as the usual expression for estimating a variance by sampling (except that the degrees-of-freedom correction is p rather than 1). The use of Eq. (8) represents an *a posteriori* assessment of the variance in y_i. The variance-covariance matrix now becomes

$$\mathbf{V} = \frac{S}{\nu}\,\mathbf{A}^{-1} . \qquad (9)$$

Under the same conditions as stated before Eq. (6), s_y^2 is distributed as a scaled χ^2 variate. This means, for example, if the s_y^2 values from a Monte Carlo treatment of unweighted LS are divided by the true value σ_y^2 used in the MC calculations, the resulting ratios are distributed in accord with Eq. (6) for χ_ν^2.

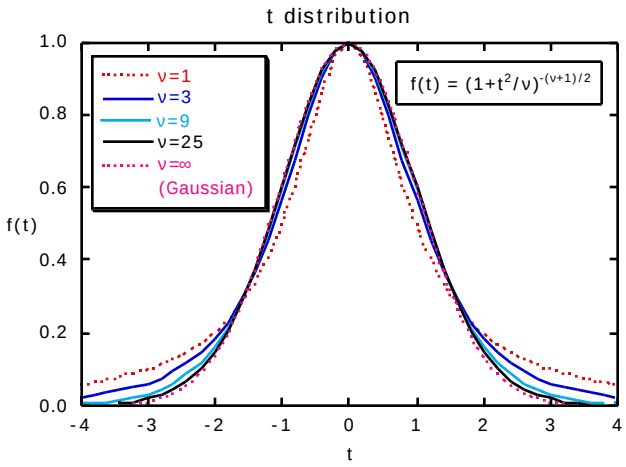
As already noted, the parameter standard errors are completely known at the outset in the case of *a priori* weighting. Even in the ignorance situation of unweighted regression, they are known to within the factor s_y, which of course must be assessed *a posteriori*. Because this is equivalent to an *a posteriori* assessment of the weights w_i (using $w_i = s_y^{-2}$), one might be tempted to determine the w_i in this manner from the fit, then call them "known" and use Eq. (7) for **V**. This is *not* proper. For the use of Eq. (7) the weights must truly be known *a priori*, hence independently from the fit; if they have to be assessed *a posteriori*, Eq. (9) is the proper expression for **V**. (Admittedly, the state of prior knowledge of the statistics of the y_i can become a judgment call.)

In the case of *a posteriori* assessment, the uncertainty in s_y does not greatly compromise the reliability of the parameter standard error estimates when the data set is large. For example, since the variance in χ_ν^2 is $^2/_\nu$, the relative standard deviation in s_y^2 is 0.1 when $\nu = 200$. This translates into a 5% relative standard deviation in s_y ($\sqrt{1}/_{(2\nu)}$) and hence also in all the parameter standard error estimates ($V_{ii}^{1/2}$). [Can you verify these statements using error propagation?]

What about the confidence limits on the parameters in the case of *a posteriori* assessment? The need to rely upon the fit itself to estimate s_y means the parameter errors are no longer exact but are uncertain estimates. Accordingly we must employ the *t*-distribution to assess the parameter confidence limits. Under the same conditions that yield a normal distribution for the parameters β and scaled χ^2 distributions for s_y^2 and for the V_{ii} from Eq. (9), the quantities $(\beta_i - \beta_{i,\text{true}})/V_{ii}^{1/2}$ belong to the *t*-distribution for ν degrees of freedom,[1] which is given by

$$f(t)\,dt = C'\,(1 + t^2/\nu)^{-(\nu+1)/2}\,dt\,, \qquad (10)$$

with C' another normalizing constant. For small ν the *t*-distribution is somewhat narrower in the peak than the Gaussian distribution, with more extensive tails. However the *t*-distribution converges on the unit-variance normal distribution in the limit of large ν, making the distinction between the two distributions unimportant for large data sets. [See figure.]



The structure of **V** is such that it scales with σ_y^2 and with $1/n$. The latter, for example, means that the parameter standard errors ($V_{ii}^{1/2}$) do correctly go as $n^{-1/2}$, all other things being equal. This means that they are to be interpreted in the same manner as the standard deviation in the mean in the case of a simple average. [As we showed in class, a fit of data to $y = a$ yields for σ_a the usual expression for the standard deviation in the mean.]

Intermediate Situations. Sometimes one has *a priori* information about the *relative* variation of σ_{yi} with y_i but not a good handle on the *absolute* σ_{yi}. For example, data might be read from a

logarithmic scale, or transformed in some way to simplify the LS analysis. As a specific example of the latter, data might be fitted to $y = ax + bx^2$ by first dividing by x to yield $y' \equiv y/x$, then fitting to $y' = a + bx$. If the original y_i have constant standard deviation σ_y, then simple error propagation (with x still treated as error-free) shows that the standard deviations in the y_i' values are σ_y/x_i, meaning the weights w_i are x_i^2.

Clearly a weighted fit is called for in the latter "straight-line" analysis; neglect of the now-unequal weighting of the dependent variable will fail to yield the desired minimum-variance estimates of the parameters. A check will show that the weighted fit to $y' = a + bx$ yields a set of equations (Eqs. 3) identical to those for the unweighted fit to $y = ax + bx^2$. [This was Problem 5 on p. 20 of the ClassPak.] Also, the results for both β and $\mathbf{V}$ (through Eq. 9) are independent of arbitrary scale factors in the weights. If the latter are taken as simply $w_i = x_i^2$, then S/ν (with S given by Eq. (1)) will be an estimate of σ_y^2. Since this is an *a posteriori* assessment, the *t*-distribution should be used to obtain the confidence limits on the parameters.

Another situation is the case where data come from two or more parent distributions of differing σ_y, but again known in only a relative sense. As before, the results of the calculations are independent of an arbitrary scale factor in the weights. However, to obtain meaningful estimates of the parent variances, it is customary to designate one subset as reference and assign $w_i = 1$ for these data, with all other weights taken as s_{ref}^2/s_i^2 (hence the need for knowledge of the relative precisions). Then the quantity S/ν ($= s_{ref}^2$) obtained from the fit is more properly referred to as the "estimated variance for data of unit weight," and the estimated variance for a general point in the data set is then s_{ref}^2/w_i.

A cautionary note is in order for users of commercial data analysis programs. Those programs which do provide estimates of the parameter errors do not always make clear which equation — (7) or (9) — is being used. For example the program KaleidaGraph uses (9) in unweighted fits to user-defined functions, but (7) in all weighted fits. This means that in cases like those just discussed, where the weights are known in only a relative sense, the user must scale the parameter error estimates by the factor $(S/\nu)^{1/2}$ to obtain the correct *a posteriori* values. [In the KaleidaGraph program the quantity S is called "Chisq" in the output box but is a scaled χ^2 variate except in cases where the input σ_i values are valid in an absolute sense.]

Thus, the explanation of the differences you obtained in the various fits of the PROB1 data set in the second problem set: The unweighted fit of the data to $y = a*x$ yielded $\sigma_a = 0.02883656$, while the weighted fit with all $\sigma_{yi} = 1.00$ yielded $\sigma_a = 0.01961161$. Both fits yielded $\chi^2 = 23.782247$ [since σ_{yi} is assumed to be 1.00 in an unweighted fit, and χ^2 is computed as $\Sigma\ (\delta_i/\sigma_i)^2$]. Thus, multiplication of the σ_a from the weighted fit by $(\chi^2/\nu)^{1/2}$ [$= (23.782247/11)^{1/2}$] yields the σ_a from the unweighted fit. Also, changing σ_{yi} to 1.50 (the original value) yields $\sigma_a = 0.0294174$, which is exactly 1.5×0.01961161 (the value obtained with $\sigma_{yi} = 1.00$); and χ^2 for $\sigma_{yi} = 1.50$ is 10.5699, which is smaller than 23.782247 by exactly $(1.5)^2$.

In summary, unbiased finite-variance data, properly weighted, yield unbiased, minimum-variance parameter estimates in a linear least-squares analysis. If the errors (σ_y) in the data are known, the exact parameter variances are obtained from the variance-covariance matrix **V**. Improper weighting biases the parameter error estimates but not the parameters themselves. It thus renders **V** useless for the parameter errors. Improper weighting also corrupts the expected chi-square distributions for S/ν and *t*-distributions for *a posteriori*-assessed parameters. These distributions and the normal parameter distributions also fail to hold when the errors in the data are not normal.

Nonlinear Least Squares

In nonlinear fitting the quantity minimized is again S, and the least-squares equations take a form very similar to Eq. (3) but must be solved iteratively. The search for the minimum in S can be carried out in a number of different ways;[3,4] but sufficiently near this minimum, the corrections $\Delta\beta$ to the current values β_0 of the parameters can be evaluated from[3,4]

$$\mathbf{A}\,\Delta\beta = \mathbf{z}\ , \tag{11}$$

leading to improved values,[c]

$$\beta_1 = \beta_0 + \Delta\beta\ . \tag{12}$$

In the case of a linear fit, starting with $\beta_0 = 0$, these relations yield for β_1 equations identical to Eqs. (3) and (4) for β. For nonlinear functions, multiple iterations are needed to achieve convergence.

Regardless of how convergence is achieved, the variance-covariance matrix is again given by Eq. (7) in the case of *a priori* weighting and Eq. (9) for *a posteriori* weighting. However, there is an important distinction between **V** in the general nonlinear case *vs*. the linear case: The matrix **A** now contains a dependence on the parameters. Thus even in the case of *a priori* weighting, **V** from Eq. (7) will vary from data set to data set.

Error Propagation

One property of linear fitting is particularly appealing: Provided the adjustable parameters are truly overdetermined by the data at hand, one is assured of a numerical solution to the problem. This does not hold for nonlinear fitting, in which a poor choice of initial parameter estimates β_0 can lead to divergence or slow convergence. Thus there are times when a transformation to a linear form is a practical convenience. However, in such cases one must take care in estimating the errors in the desired (nonlinear) parameters, because the simple rules for error propagation do not apply.

For example, suppose the data are to be fitted to $y = 1/a + bx/a$. This is a nonlinear fit, and if pursued as such will yield proper estimates of σ_a^2 and σ_b^2 as the appropriate diagonal elements of **V**.

---

[c] Here again $\mathbf{A} = \mathbf{X^T\ W\ X}$, but the elements of **X** are now given by $X_{ij} = (\partial F_i/\partial\beta_j)$, and are evaluated at x_i using the current values β_0 of the parameters. The function F expresses the relations among the variables and parameters in such a way that a perfect fit yields $F_i = 0$. For the commonly occurring case where y can be expressed as an explicit function of x, it takes the form, $F_i = y_{\mathrm{calc}}(x_i) - y_i = -\delta_i$. Also, **z** is different: $\mathbf{z} = \mathbf{X^T\ W}\ \delta$.

Alternatively one might choose to fit to $y = A + Bx$, which is linear and with all the usual assumptions will yield normally distributed estimates of A and B. But A and B are *correlated* parameters, so the calculation of the estimated error in any function f of A and B must employ the full expression,[2]

$$s_f^2 = \mathbf{g}^{\mathbf{T}}\,\mathbf{V}\,\mathbf{g}\ , \qquad (13)$$

in which the elements of $\mathbf{g}$ are $g_i = \partial f/\partial\beta_i$, and $\mathbf{V}$ is the variance-covariance matrix obtained from the linear fit to A and B. In this case a is a function of A alone ($a = 1/A$), and the usual rules of error propagation apply. However, b is a function of both A and B ($b = B/A$), so the full expression must be used. It can be shown that the estimates of σ_a^2 and σ_b^2 obtained from σ_A^2 and σ_B^2 using Eq. (13) are identical to those obtained directly from the nonlinear $\mathbf{V}$. Also, for any given data set, the nonlinear fit will yield a and b values identical to those obtained from the linear estimates of A and B. [Note: The usual simplified rules for error propagation are obtained from Eq. (13) by neglecting the off-diagonal elements of $\mathbf{V}$ — the covariances. You should be able to verify this.]

REFERENCES

1. Mood, A. M.; Graybill, F. A. *Introduction to the Theory of Statistics,* 2nd edit. **1963**, McGraw-Hill, New York.
2. Hamilton, W. C. *Statistics in Physical Science: Estimation, Hypothesis Testing, and Least Squares* **1964**, The Ronald Press Co., New York.
3. Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences* **1969**, McGraw-Hill, New York.
4. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes* **1986**, Cambridge Univ. Press, Cambridge, U. K.