Tina Memo No. 2008-005
Internal Memo

Tutorial: Least-Squares Fitting

P. A. Bromiley

Last updated
06 / 06 / 2008

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Tutorial: Least-Squares Fitting

P. A. Bromiley
Imaging Science and Biomedical Engineering Division
Medical School, University of Manchester
Manchester, M13 9PT, UK
paul.bromiley@man.ac.uk

Abstract

Least-squares fitting, first developed by Carl Friedrich Gauss, is arguably the most widely used technique in statistical data analysis. It provides a method through which the parameters of a model can be optimised in order to obtain the best fit to a data set through the minimisation of the squared differences between the model and the data. This tutorial document describes the closely associated methods of least-squares and weighted least-squares (χ^2) fitting. We derive least-squares estimators both as Maximum-Likelihood (ML) estimators and as Best Linear Unbiased Estimators (BLUE), reconciling the two treatments in the conclusion. We then use the method to derive estimators for the parameters of some simple models, including the straight-line fit, together with the standard errors on the estimated parameters. We conclude with some general observations on how the lessons learned from the specific case of least-squares fitting can inform our understanding of machine vision and medical image analysis algorithms in general.

1 The Method of Least Squares

The method of least squares has, at various times, been ascribed to a number of different authors, notably Gauss and Laplace. Plackett [**?**, **?**] describes the historical development of the method, and was responsible for establishing that the fundamental results are due to Gauss. It is arguably the most commonly used statistical estimation procedure, being almost ubiquitous in science and engineering, and is usually one of the first to be learned. It provides a procedure through which a model can be fitted to a set of measurements by minimising the squared differences between the measurements and the model prediction,s with respect to the parameters of the model, in order to obtain the optimal parameters. However, this apparently simple yet powerful procedure depends on a set of assumptions that are much less well known than the method itself, leading to invalid applications.

There are two, notably different, approaches to justifying the least-squares fitting procedure, differing in their assumptions. The (arguably) simpler and easier to interpret approach is to assume that the errors on the measurements are described by a normal distribution, in which case the least-squares estimators can be derived using maximum likelihood. However, it is also possible to derive least-squares estimators as those that, amongst all unbiased, linear estimators for linear models, have the lowest variance. We initially describe both derivations without comment, and reconcile them in the conclusion.

1.1 Derivation as a Maximum Likelihood Estimator

Suppose that you have a set of n data (x_i, y_i) where $i = 1...n$, to which you wish to fit some model $f(x)$. An implicit assumption is made at this stage that the model is a correct description of the physical process that generates the data (the consequences of using an incorrect model are described later). Noise will have been added to the data during the acquisition process, so that

$$y_i = f(x_i) + \eta_i$$

where η_i is the noise on the ith data point. Therefore, the residuals r_i generated by subtracting the model prediction at x_i from the measured y_i are given by

$$r_i = y_i - f(x_i) = \eta_i$$

We now need to apply this model to make statements regarding the degree of conformity of data. In the strictest sense, a distinction must be made between probabilities, which are defined over a range, and probability densities, which are not i.e. the probability $P(r_i|\theta)$ that the residual r_i will lie within the range $r \pm \Delta/2$ is given by

$$P(r_i|\theta) = \int_{r-\Delta/2}^{r+\Delta/2} p(r_i|\theta)\ dr\ =\ p(r_i|\theta)\Delta$$

where $p(r_i|\theta)$ is the probability density distribution of the residuals[1]. The aim of least-squares fitting is to find the set of model parameters that maximise the probability that the model could have generated the data. The probability density over r_i for a Gaussian distribution is given by

$$p(r_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp[-\frac{(y_i - f(x_i))^2)}{2\sigma_i^2}]$$

where θ is the vector of model parameters and σ_i is the standard deviation of the noise on the ith data point. Assuming that the noise on each data point is uncorrelated, the probability of the whole set of residuals is given by the product of the probabilities for each residual i.e. they consist only of noise. Assuming that the noise is described by a normal distribution, the probability density $p(r_i|\theta)$. for r_i is given by

$$p(r|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp[-\frac{(y_i - f(x_i))^2)}{2\sigma_i^2}]$$

We can maximise this quantity, or **likelihood**, with respect to the model parameters θ (see NAT comments). Note that we are considering $p(r|\theta)$ as function of the θ, not as a function of the x_i. Furthermore, we are interested only in the location of the maximum of the function, rather than in its absolute value. Therefore, we can apply any monotonic transformation to it, since such transformations will change the absolute value of the function but not the location of its optimum. One such transformation is taking the logarithm: this simplifies the expression considerably[2]. We can now define the likelihood L according to

$$lnL = \sum_{i=1}^{n} -\frac{[y_i - f(x_i)]^2}{2\sigma_i^2}$$

In order to maximise the likelihood i.e. to obtain the model parameters that maximise the probability that the data could have been generated by the model, we minimise (due to the minus sign in the above equation) the sum of the squares of the differences between the model and the data, weighting each term in the sum by the square of the standard deviation of the error on that data point. The negative log-likelihood is referred to as the χ^2 function

$$\chi^2 = \sum_{i=1}^{n} \frac{[y_i - f(x_i)]^2}{\sigma_i^2}$$

Furthermore, if the standard deviation of the noise is independent of x, then $\sigma_i = \sigma$, and this can also be removed as a constant factor, leaving[3]

$$lnL = \sum_{i=1}^{n} -[y_i - f(x_i)]^2 = \sum_{i=1}^{n} -r_i^2$$

Due to the minus sign, maximisation of the likelihood is achieved by minimising the squares of the residuals, and the method is therefore known as "least-squares fitting". Since the process is a maximisation of the likelihood, it falls into a class of methods known as maximum-likelihood estimators.

1.2 Linear and Non-linear Least Squares, and the Matrix Formulation

Least-squares fitting problems can be divided into two categories: linear and non-linear. A least-squares problem is said to be linear when the model can be expressed as a linear combination of its parameters i.e.

$$f(x|\mathbf{a}) = \sum_{r=1}^{n_p} c_r(x) a_r$$

where $\mathbf{a}$ is the vector of n_p model parameters and the coefficients c_r are either constants or functions of the x_i. Note that it is linearity in the $\mathbf{a}$, not in x, that matters. The χ^2 then becomes

$$\chi^2 = \sum_{i=1}^{n} \frac{[y_i - \sum_{r=1}^{n_p} c_r(x) a_r]^2}{\sigma_i^2}$$

[1] The notational convention is adopted that capital P refers to a probability, whereas lower-case p refers to a probability density.

[2] Taking the logarithm also has the desirable property that it turns the product into a sum, thus avoiding the need to take the product of many small numbers, and so avoids loss of accuracy due to the limits imposed by machine precision.

[3] This property is referred to as homoscedasticity, and spaces in which it is true as homoscedastic spaces; however, due to the obscurity of this term we have referred to such spaces in other documents on this web site as "equal variance spaces".

The estimators of the model parameters are derived by maximising the χ^2 w.r.t. the parameters; this can be achieved by differentiating w.r.t. each parameter,

$$\frac{d\chi^2}{da_r} = \sum_{i=1}^{n} c_r(x_i) \frac{[y_i - \sum_{s=1}^{n_p} c_s(x) a_s]^2}{\sigma_i^2}$$

and setting the result equal to zero. At the point where the differential is set equal to zero we are implicitly substituting the optimal value of the parameter into the equation i.e. replacing the parameter as a variable with the maximum-likelihood estimator of the parameter. Therefore, we replace the symbol for the parameter a_s with the symbol for the estimator of the parameter $\hat{a_s}$ at this point

$$\sum_{i=1}^{n} c_r(x_i) \frac{[y_i - \sum_{s=1}^{n_p} c_s(x) \hat{a}_s]^2}{\sigma_i^2} = 0$$

This generates a system of n_p simultaneous equations with r unknowns (the c_r). Solving this system of equations generates the least-squares estimators for the model parameters.

The above equations can be expressed in a more concise form using matrix notation, avoiding the summation signs and indices. The χ^2 becomes

$$\chi^2 = (\mathbf{y}^T - \mathbf{f}^T)\mathbf{V^{-1}}(\mathbf{y} - \mathbf{f})$$

We adopt a notational convention in which bold lower-case variables represent vectors and bold upper-case variables represent matrices, so $\mathbf{y}$ is the vector of y_i, $\mathbf{f}$ is the vector of model predictions at each x_i, and $\mathbf{V}$ is the covariance matrix of the measurements i.e. for independent measurements, the diagonal elements of this matrix are the squares of the standard deviations of the errors on each data point. Introducing another matrix $\mathbf{C}$, where $C_{ir} = c_r(x_i)$, we have

$$\mathbf{f} = \mathbf{Ca}$$

$$\chi^2 = (\mathbf{y}^T - \mathbf{a}^T\mathbf{C}^T)\mathbf{V^{-1}}(\mathbf{y} - \mathbf{Ca})$$

so the normal equations become

$$\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C}\hat{\mathbf{a}} = \mathbf{C}^T\mathbf{V}^{-1}\mathbf{y}$$

Note that the vector $\mathbf{y}$ is of length n, the vector $\mathbf{a}$ is of length n_p, $\mathbf{V}$ is an $n \times n$ square matrix, and $\mathbf{C}$ is rectangular with n rows and n_p columns. In general n will be much greater than n_p (unless $n_p \leq n$ there is no unique solution to the normal equations) and so a factor of $\mathbf{C}^T$ cannot be cancelled from this equation as $\mathbf{C}$ does not have an inverse, and so the weighted least-squares estimate of $\hat{\mathbf{a}}$ is

$$\hat{\mathbf{a}} = (\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{V}^{-1}\mathbf{y}$$

If the errors on the data are equal, then the $\mathbf{V}$ can be cancelled to give the normal equations for the least-squares fit

$$\hat{\mathbf{a}} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}$$

Note that these equations are deceptively simple: only in the most straightforward cases can least-squares or weighted least-squares fitting be implemented by evaluating them directly. For all but the simplest models the computational effort required to compute the matrix inversion grows rapidly, and numerical problems with accuracy and rounding rapidly get out of control, particularly if some of the model parameters are weakly constrained by the data. In general, unless the fitting problem involves only proportional, linear (for which the equations for the estimators are given in the next section) or quadratic models, it is better to rely upon packages provided by software libraries, which generally apply techniques such as SVD in order to avoid the explicit matrix inversion.

Given the above forms for the normal equations, the errors on the estimators can be obtained directly. The estimators of the parameters $\hat{\mathbf{a}}$ are obtained through multiplying the $\mathbf{y}$ by a matrix $(\mathbf{C}^T\mathbf{V}(\mathbf{y})^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{V}(\mathbf{y})^{-1}$. In the present case only the $\mathbf{y}$ have errors: the $\mathbf{x}$ (and thus the matrix $(\mathbf{C}^T\mathbf{V}(\mathbf{y})^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{V}(\mathbf{y})^{-1}$) are not random variates. The matrix form of the propagation of errors formula (see Appendix 1) states that the variance transforms using the same matrix i.e. if

$$\hat{\mathbf{a}} = \mathbf{My}$$

then

$$\mathbf{V}(\hat{\mathbf{a}}) = \mathbf{MV}(\mathbf{y})\mathbf{M}^T$$

where

$$\mathbf{M} = (\mathbf{C}^T\mathbf{V}(\mathbf{y})^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{V}(\mathbf{y})^{-1}$$

Since $\mathbf{V}^T = \mathbf{V}$ and $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ this reduces to

$$\mathbf{V}(\hat{\mathbf{a}}) = (\mathbf{C}^T\mathbf{V}(\mathbf{y})^{-1}\mathbf{C})^{-1}$$

In cases where the model cannot be expressed as a linear combination of the parameters, the least-squares problem is said to be non-linear. There is no closed-form solution in such cases. An iterative approach must then be adopted in which the solution is found by repeatedly evaluating linear approximations around the current estimates of the parameters, using much the same procedure as described above. However, the process becomes sensitive to the initial estimates of the model parameters, and so a good initial "guess" for the parameters is required.

1.3 Derivation of Least-Squares Estimators as Best Linear Unbiased Estimators (BLUE)

Linear least-squares fitting can also be derived, under a slightly different set of assumptions, by finding the linear, unbiased estimators of the parameters of the linear model that have the lowest variance. First, assume that the model is linear (again, in its parameters rather than in x), so that

$$\mathbf{y} = \mathbf{Ca} + \eta$$

Next, assume that the errors on each measurement are uncorrelated

$$cov(\eta_i, \eta_j) = 0$$

that the error distribution is symmetric

$$< \eta >= 0$$

and that the errors are homoscedastic (i.e. we are working in an equal-variance space)

$$V(\eta) = \sigma^2\mathbf{I}$$

Let $\hat{\mathbf{T}} = \mathbf{t}^T\mathbf{y}$ be any linear function of the observations that provides an unbiased estimate of $\alpha\mathbf{a}$, where α is an arbitrary vector of constants. Since $\hat{\mathbf{T}}$ is unbiased it follows that

$$< \hat{\mathbf{T}} >=< \mathbf{t}^T\mathbf{y} >= \mathbf{t}^T < \mathbf{y} >= \mathbf{t}^T\mathbf{Ca}$$

and

$$< \hat{\mathbf{T}} >= \alpha\mathbf{a}$$

so

$$\mathbf{t}^T\mathbf{C} = \alpha \quad or \quad \alpha^T = \mathbf{C}^T\mathbf{t} \tag{1}$$

We now find the variance of $\hat{\mathbf{T}}$

$$\begin{aligned} V(\mathbf{t}^T\mathbf{y}) &=< (\mathbf{t}^T\mathbf{y})(\mathbf{y}^T\mathbf{t}) > \\ &= \mathbf{t}^T V(\mathbf{y})\mathbf{t} \\ &= \mathbf{t}^T\sigma^2\mathbf{I}\mathbf{t} \\ &= \mathbf{t}^T\mathbf{t}\sigma^2 \end{aligned} \tag{2}$$

We wish to find an estimator that, with the constraint that it is unbiased, has the lowest possible variance. Therefore, we minimise the variance (Eq. 2) subject to unbiasedness (Eq. 1) using a vector of Lagrange multipliers λ. Let

$$Q = \mathbf{t}^T\mathbf{t} + 2\lambda^T(\alpha^T - \mathbf{C}^T\mathbf{t})$$

We adopt the usual approach to minimisation of differentiating this with respect to $\mathbf{t}$, setting the result equal to zero, and solving for $\mathbf{t}$. So

$$\frac{dQ}{d\mathbf{t}} = 2\mathbf{t} - 2\mathbf{C}\lambda$$

giving

$$\mathbf{t} = \mathbf{C}\lambda$$

Premultiplying this by $\mathbf{C}^T$ gives

$$\mathbf{C}^T\mathbf{t} = \mathbf{C}^T\mathbf{C}\lambda$$

but from Eq. 1 $\alpha^T = \mathbf{C}^T\mathbf{t}$ so

$$\alpha^T = \mathbf{C}^T\mathbf{C}\lambda \tag{3}$$

Now

$$\hat{\mathbf{T}} = \mathbf{t}^T\mathbf{y} = \lambda^T\mathbf{C}^T\mathbf{y} \tag{4}$$

Solving Eq. 3 for λ and substituting into Eq. 4 gives

$$\begin{aligned}\hat{\mathbf{T}} &= [(\mathbf{C}^T\mathbf{C})^{-1}\alpha^T]^T\mathbf{C}^T\mathbf{y} \\ &= \alpha(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}\end{aligned}$$

Now, since $< \hat{\mathbf{T}} >= \mathbf{T} = \alpha\mathbf{a}$, $\hat{\mathbf{T}}$ must have the form $\alpha\hat{\mathbf{a}}$, and so

$$\alpha\hat{\mathbf{a}} = \alpha(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}$$

or

$$\hat{\mathbf{a}} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}$$

This is simply the expression for the normal equations given above, and so the unbiased linear estimator of the linear model that has the lowest variances is the least-squares estimator, and due to this we refer to it as the Best Linear Unbiased Estimator (BLUE). The result is known as the Gauss-Markov Theorem.

Two further results follow immediately. First, taking the expectation value of the least-squares estimator gives

$$\begin{aligned}< \hat{\mathbf{a}} > &=< (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y} > \\ &=< (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T(\mathbf{C}\mathbf{a} + \eta) > \\ &= \mathbf{a} + (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T < \eta >\end{aligned}$$

and, since under the above assumptions the expectation value of the noise is zero,

$$< \hat{\mathbf{a}} >= \mathbf{a}$$

The expectation value of the estimator is equal to the value of the parameter being estimated, so the least-squares estimator is unbiased.

Furthermore, the variance of the estimator is defined as

$$V(\hat{\mathbf{a}}) =< (\hat{\mathbf{a}} - \mathbf{a})(\hat{\mathbf{a}} - \mathbf{a})^T >$$

but

$$\begin{aligned}\hat{\mathbf{a}} - \mathbf{a} &= (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y} - \mathbf{a} \\ &= (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T(\mathbf{C}\mathbf{a} + \eta) - \mathbf{a} \\ &= (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\eta\end{aligned}$$

So,

$$\begin{aligned}V(\hat{\mathbf{a}}) &=< (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\eta\eta^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1} > \\ &= (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T < \eta\eta^T > \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1} \\ &= (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T(\sigma^2\mathbf{I})\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1} \\ &= \sigma^2(\mathbf{C}^T\mathbf{C})^{-1}\end{aligned}$$

thus reproducing the result obtained from the derivation using maximum likelihood.

2 Simple Examples of Linear Least-Squares Fitting

In this section, the estimators for each model parameter for some commonly occurring models are derived. The general process is the same in each case; we find the optimum of the χ^2 function w.r.t. each of the model parameters by differentiating w.r.t. that parameter, setting the result equal to zero, and solving for the parameter. The over-bar symbol is used to represent the mean of a set of values e.g.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The errors on the estimators are also derived. Therefore, the procedure in each case is to write the estimator as a function of the y_i, differentiate w.r.t. the y_i, and then sum the squares of the differentials multiplied by the squares of the standard deviations. This procedure gives the variance of the error on the estimator: the square-root of this quantity is referred to as the standard error on the estimator.

2.1 Fitting $y = mx$

Situations in which y is directly proportional to x are reasonably common in physical laws; linear elasticity in the form of Hooke's Law provides one example. In this case, the model is

$$y = mx$$

where m is called the constant of proportionality (Young's modulus in the case of Hooke's Law). The χ^2 is therefore

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - mx_i)^2}{\sigma_i^2}$$

or, if σ_i is constant,

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - mx_i)^2$$

Differentiating w.r.t. m and setting the result equal to zero gives

$$\frac{d\chi^2}{dm} = \frac{1}{\sigma^2} \sum_{i=1}^{n} -2x_i(y_i - \hat{m}x_i) = 0$$

and therefore

$$\sum_{i=1}^{n} x_i y_i - \hat{m}x_i^2 = 0$$

so

$$\frac{\bar{xy}}{n} = \frac{\hat{m}\bar{x^2}}{n}$$

and

$$\hat{m} = \frac{\bar{xy}}{\bar{x^2}}$$

In order to find the error on the estimator $\hat{m}$, we write it as a sum over y_i

$$\hat{m} = \sum_{i=1}^{n} \frac{x_i}{n\bar{x^2}} y_i$$

and then differentiate w.r.t. y_i: only one term in the sum contains y_i (the others contain y_1, y_2,...,y_{i-1},y_{i+1},...,y_n), so

$$\frac{d\hat{m}}{dy_i} = \frac{x_i}{n\bar{x^2}}$$

Applying the equation of error propagation therefore gives

$$\sigma_{\hat{m}}^2 = \sum_{i=1}^{n} (\frac{x_i}{n\bar{x^2}})^2 \sigma^2$$

and so

$$\sigma_{\hat{m}}^2 = \frac{\sigma^2}{n\bar{x^2}}$$

Taking the square-root of this quantity provides the standard error on $\hat{m}$.

2.2 Fitting $y = mx + c$

The second example we provide is the case of fitting a straight line, in which the model is

$$y = mx + c$$

where m is the gradient and c is the intercept. The χ^2 is

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - (mx_i + c))^2}{\sigma_i^2}$$

In this case, we differentiate w.r.t. both m and c, set the results equal to zero, and solve the resultant pair of simultaneous equations. Differentiating gives

$$\frac{d\chi^2}{dc} = \sum_{i=1}^{n} -2(y_i - \hat{m}x_i - \hat{c}) = 0 \tag{5}$$

and

$$\frac{d\chi^2}{dc} = \sum_{i=1}^{n} -2x_i(y_i - \hat{m}x_i - \hat{c}) = 0 \tag{6}$$

From Eq.5 we have

$$\sum_{i=1}^{n} -2(y_i - \hat{m}x_i - \hat{c}) = 0 = \sum_{i=1}^{n} y_i - \hat{m}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \hat{c} = \bar{y} - \hat{m}\bar{x} - \hat{c} \tag{7}$$

Eq.6 provides a similar function

$$\bar{xy} - \hat{m}\bar{x^2} - \hat{c}\bar{x} = 0$$

so

$$\frac{\bar{xy} - \hat{m}\bar{x^2}}{\bar{x}} = \hat{c}$$

Substituting this into Eq.7 gives

$$\hat{m} = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2}$$

Substituting this into Eq.7 gives

$$\hat{c} = \frac{\bar{y}\bar{x^2} - \bar{xy}\bar{x}}{\bar{x^2} - \bar{x}^2}$$

Again, we can write the estimator of $\hat{m}$ as a function of y_i

$$\hat{m} = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{n(\bar{x^2} - \bar{x}^2)} y_i$$

differentiate

$$\frac{d\hat{m}}{dy_i} = \frac{x_i - \bar{x}}{n(\bar{x^2} - \bar{x}^2)}$$

and apply the propagation of errors formula to obtain the error on $\hat{m}$

$$\sigma^2_{\hat{m}} = \frac{\sigma^2}{n(\bar{x^2} - \bar{x}^2)}$$

A similar procedure gives the error on $\hat{c}$ as

$$\sigma^2_{\hat{c}} = \frac{\sigma^2\bar{x^2}}{n(\bar{x^2} - \bar{x}^2)}$$

Again, taking the square-roots of these formulae provides the standard errors on $\hat{m}$ and $\hat{c}$.

3 Estimating Data Errors from the Fit Itself

The results of applying error propagation in the above examples gives the standard errors on the estimators of each model parameter in terms of the standard deviation of the noise on the data. In some cases this standard deviation may be known from the calibration of the measurement equipment or from error propagation in previous algorithmic stages. If it is not known then we can estimate it from the fitting process itself: the residuals around the fitted model provide us with a sample drawn from the noise distribution, from which we can calculate the standard deviation. However, the error propagation formulae require the standard deviation of the parent distribution that generated the noise: applying the usual formula

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

to the residuals gives us the standard deviation of the sample of noise drawn from the parent distribution, unless there are an infinite number of data points. This estimate of the standard deviation of the parent distribution generated by the above equation is biased: fortunately, it is possible to evaluate the magnitude of this bias, and it turns out that the standard deviation generated by the above equation is exactly $(n - n_p)/n$ too low. Knowing this, we can apply a correction leading to the formula

$$\sigma^2 = \frac{1}{n - n_p} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Many fitting packages will provide error estimates generated using this equation. However, it should be noted that its use prevents any independent estimate of the goodness-of-fit from the χ^2.

4 Conclusions

Least-squares fitting can be derived from two different sets of assumptions. Assuming that

- the model is an accurate description of the data generation process;
- the errors on the data are uncorrelated;
- the noise generation process has a normal distribution;

least-squares estimators emerge as the maximum-likelihood estimators of the model parameters. In this case it is also easy to interpret the meaning of the fitted model parameters: they are the parameters that give the highest probability that the data could have been generated by the model. However, it is also possible to prove that linear least-squares estimators are BLUE under an alternative set of assumptions i.e. that

- the model is an accurate description of the data generation process;
- the errors on the data are uncorrelated;
- the distribution of the errors is symmetric;
- the errors are homoscedastic (i.e. have equal variances regardless of where in the space the measurement is made).

These assumptions, whilst requiring homoscedasticity and a symmetrical distribution for the errors, do not assume a particular distribution i.e. a normal distribution, and so are less strict. It may therefore appear that the case in which least-squares estimators are also maximum-likelihood estimators is a special case of a more general approach. Indeed, this is the interpretation which many statistical textbooks prefer to put forward, in order that the use of least-squares can be justifed in the maximum number of applications.

There are two points to make regarding this issue. First, if the distribution of the errors is not normal, then the least-squares estimators only have the lowest variances amongst the **linear** estimators. There will, in general, be maximum-likelihood estimators that are non-linear and have lower variances. Second, the proof that least-squares estimators are BLUE relies on taking expectation values, and so is an asymptotic property (i.e. only guaranteed in the limit of infinite amounts of data). The assumption that the error distribution is symmetric is particularly sensitive; even when this is true, the residuals (i.e. the sample taken from the error distribution) may be non-symmetric i.e. there may be more data points on one side of the mean than on the other[4]. Such considerations have led to the development of "robust" estimation procedures, which are less sensitive to outliers. As a consequence, we can conclude that simply choosing to use least-squares based upon the BLUE interpretation is likely to lead to inferior results in practical applications. We certainly can not conclude that it is likely to be the best approach. Using the BLUE analysis to support such an argument amounts to over-interpretation. An approach based upon quantitative probability, which matches the assumed distributions to those practically observed, must do better. Unfortunately, such over interpretation of specific mathematical analyses is often encountered when looking for arguments to support many common statistical measures. Scientists therefore need to be more generally aware of this issue.

[4]The influence this has on the fit will be inversely proportional to the amount of data. In addition, the probability of having data points with large residuals, which therefore have a large effect on the fit, and which are not balanced by a similar data point on the opposite side of the mean, will increase as the relative size of the tails on the distribution increases. Since the Gaussian is the most compact distribution, the probability of having such outliers will increase the further the distribution departs from normality.

Another point of interest is that of homoscedaticity. Why is it needed in the BLUE analysis but not the conventional derivation from Gausian distrbutions? In fact, we can argue that a correctly motivated derivation of likelihood would also have required the property of homoscedasticity (see NAT comments).

Whichever justification for the use of least-squares estimators is applied, it is important to ensure that the assumptions are met. In particular, the model must be a complete description of the physical process that generates the data, such that subtracting the model predictions from the data leaves only noise. If this assumption is not met, then the residuals will contain an x-variate dependent structure. This will exert a biasing effect on the fit i.e. introduce systematic errors. The same is true of non-symmetric error distributions. Finally, for best use, the error distribution must be both normal **and** homoscedastic. Due to these considerations, it is not sufficient to simply quote the fitted model parameters when applying the procedure; they should be accompanied by some measure of the goodness-of-fit, in order to demonstrate that the model genuinely fits the data. The most powerful goodness-of-fit tests require independent estimates of the errors, and are precluded if the errors must be estimated from the residuals. Therefore, it is also good practice to provide estimates of the errors on the fitted parameters, in order to avoid constraining the statistical procedures available for further analysis of the results.

All statistical analysis procedures can, at some level, be considered as model fitting processes, of which least-squares fitting is a simple example. Therefore, the conclusions drawn from this analysis of least-squares can provide a toolkit with which to approach the consideration of any statistical procedure. In particular, we can pose two questions.

- Does the model really fit the data?
- Are the errors on the results quantified?

These are particularly relevant in the case of algorithms developed in machine vision and medical image analysis since the answers are, with surprising frequency, "no". If this is the case, we are justified in being sceptical about the value of the proposed algorithm.

5 Appendix 1: Error Propagation

Suppose we have a set of m functions $f_1, f_2, f_3, ..., f_m$ of n different variables $x_1, x_2, x_3, ..., x_n$. If the x_i have errors associated with them, then so do the f_k. Furthermore, since the f_k share the x_i they will be correlated even if the x_i are not. The variances on the f_k are given by

$$V(f_k) =< f_k^2 > - < f_k >^2$$

The f_k can be expanded in a Taylor series about the mean to give

$$f_k \approx f_k(\mu_1, \mu_2, ...) + (\frac{\delta f_k}{\delta x_1})(x_1 - \mu_1) + (\frac{\delta f_k}{\delta x_2})(x_2 - \mu_2)$$

Inserting this into the formula for the variance gives

$$V(f_k) = (\frac{\delta f_k}{\delta x_1})^2 < (x_1 - \mu_1)^2 > + ... + 2(\frac{\delta f_k}{\delta x_2}) < (x_1 - \mu_1)(x_2 - \mu_2) > ...$$

$$= \sum_i (\frac{\delta f_k}{\delta x_i})^2 V(x_i) + \sum_i \sum_{j \neq i} (\frac{\delta f_k}{\delta x_i})(\frac{\delta f_k}{\delta x_j}) cov(x_i, x_j)$$

This is a generalised form of the standard law of combination of errors; for example, with a single function of independent variables (where the covariances are all zero) it simplifies to the more familiar

$$\sigma_f^2 = \sum_i (\frac{\delta f}{\delta x_i})^2 \sigma_{x_i}^2$$

The covariances between the f_k can be found in the same way

$$< f_k f_l > - < f_k >< f_l > \approx < (x_1 - \mu_1)(x_1 - \mu_1) > (\frac{\delta f_k}{\delta x_1})(\frac{\delta f_l}{\delta x_1}) + ... + < (x_1 - \mu_1)(x_2 - \mu_2) > (\frac{\delta f_k}{\delta x_1})(\frac{\delta f_l}{\delta x_2}) + ...$$

which can be expressed in summation notation as

$$cov(f_k, f_l) = \sum_i \sum_j (\frac{\delta f_k}{\delta x_i})(\frac{\delta f_l}{\delta x_j}) cov(x_i, x_j)$$

So, if we define a matrix $\mathbf{G}$ such that

$$g_k i = \frac{\delta f_k}{\delta x_i}$$

then the law of combination of errors can be expressed in its most generalised, matrix form as

$$\mathbf{V_f} = \mathbf{G V_x G}^T$$

Since it is based on a Taylor expansion, this is an approximation to the true error. The accuracy is dependent on the rate of change of the derivatives around the point of expansion; it is reasonably accurate if the derivatives do not change much over a few standard deviations.

Comments from Neil Thacker

The conventional definition for a likelihood is based upon the use of probability densities such as in the derivation above.

$$p(r|\theta) = \prod_{i=1}^{n} p(r_i|\theta) \qquad (a)$$

Fisher's stated aim was to define likelihood such that **"the ratio of likelihoods for two sets of parameters should tell us the ratio of the number of times that the data would have been generated by the two models"**. It was also Fishers intention to define a unique estimation process. It is commonly known that a likelihood is not a probability. There appear to two reasons for this, as will be discussed below.

point 1

Beware, the conventional use of likelihood (as equation (a)) is not guaranteed to satisfy Fisher's requirements. In particular, in situations where the probability densities cannot be provided by theory, but must be sampled from data, we will get different sample (residual) distributions and consequently different (non-unique) likelihood results for non-linear transformations of the measurements. We can understand this problem better if we take a closer look at the steps leading up to (a).

Fisher's requirements are true for likelihood if we can write

$$lnL = ln[P(r|\theta)] \; + \; const$$

where

$$P(r|\theta) = \prod_{i=1}^{n} \int_{r_i-\Delta_i/2}^{r_i+\Delta_i/2} p(r_i|\theta)\; dr \;\approx \prod_{i=1}^{n} p(r_i|\theta)\Delta_i \qquad (b)$$

Consequently, his aims can be met by defining likelihood according to equation (b), provided we can define the Δ_i terms consistently. The most obvious definition would seem to be the one which links Δ to the measurement accuracy, thereby defining the probability in accordance with the evidence. It can be immediately observed that **fixed** Δ_i factors will cancel in any ratio of $P(r|\theta)$ (θ being the model parameters). Cancellation of Δ_i's is guaranteed for any transformation of r (which applies consistently to all data on a one off basis), provided we make appropriate modifications to the associated densities $p(r_i|\theta)$.

An associated property of impirical approaches is that we often have to make assumptions in order to estimate distributions from the available samples. Clearly, we can only estimate a single consistent distribution (true for all data) for homoscedastic spaces. This mitigates against observing any modifications, and we will need a lot more data and Bland-Altman plots in order to understand the statistics. This comment is made all the more interesting by the observation that the proof of BLUE for least squares makes an explicit homoscedastic assumption while conventional likelihood does not.

point 2

Another thing which is often said regarding likelihood, is that $p(r|\theta)$ is not a probability as it defines a function over θ, which does not satisfy the axioms of probability. Clearly therefore, likelihood terms can never be justfied as a substitute for $P(r|\theta)$ [5]. However, taken at face value, such a statement might also imply that equation (b) in point 1 (which is also a function of the parameters) could not be a probability either.

In fact these considerations appear when attempting to justify comparison of likelihoods. Then we would want to be sure that any terms we compare during optimisation are consistently defined according to Kolmogorov's axioms.

[5] This last statement appears to flatly contradict popular practice.

However, as Fisher's intent is to make predictions regarding the number of times that data would be generated, likelihood is therefore based on a 'frequentist' definition of probability (see *Tina Memo 2007-008*). Provided we take steps to maintain a quantitative link between data and probability we can do more with the resulting likelihoods. In particular, we can assess the relative merits of parameter estimates on the basis of absolute quantities. We do not then need to rely upon an axiomatic definition of probability to believe that a comparison is meaningful. In this respect it is enough that $P(r|\theta)$ is consistent with the axioms of probability when considered over all possible measurements, rather than parameters. Equation (b) is a perfectly good expression of frequentist probability, and under some circumstances likelihood will be equivalent.

There is no such justification however under a strict Bayesian interpretation, for use of either L or (it would seem) $P(r|\theta)$ [6].

summary

It seems that if we look at Fisher's motivation for suggesting likelihood, there is a more general form which relates directly to probability underlying it. Unfortunately, functions based upon (a) can fall short of the intended goals. There was however, always enough in the original work to undo these shortcomings, if one knew where to look.

[6] It is beyond the scope of this document to explain the logical mess anyone will get into if they want to justify Bayesian estimation.