0 0 1 0 0 2 0 2 4 1 1

UCRL 8523

UNIVERSITY OF CALIFORNIA

Radiation Laboratory

A PRACTICAL GUIDE TO THE METHOD OF LEAST SQUARES

BERKELEY, CALIFORNIA

For Reference

Not to be taken from this room

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

0 0 1 0 0 2 0 2 4 1 2

UCRL-8523
Physics and Mathematics

UNIVERSITY OF CALIFORNIA
Radiation Laboratory
Berkeley, California

Contract No. W-7405-eng-48

A PRACTICAL GUIDE TO THE METHOD OF LEAST SQUARES

Peter Cziffra and Michael J. Moravscik

October 17, 1958

Printed for the U.S. Atomic Energy Commission

Printed in USA. Price 75 cents. Available from the
Office of Technical Services
U. S. Department of Commerce
Washington 25, D. C.

A PRACTICAL GUIDE TO THE METHOD OF LEAST SQUARES

Peter Cziffra

Radiation Laboratory
University of California
Berkeley, California

and

Michael J. Moravcsik

Radiation Laboratory
University of California
Livermore, California

October 17, 1958

ABSTRACT

This report gives a short and practical summary of curve fitting by the method of least squares. The purpose is to list and roughly justify the formulae used in finding the best fitting curve, the errors, and the quantities which describe the goodness of the fit.

0 0 1 0 0 2 0 2 4 1 3

A PRACTICAL GUIDE TO THE METHOD OF LEAST SQUARES*

Peter Cziffra

Radiation Laboratory
University of California
Berkeley, California

and

Michael J. Moravcsik

Radiation Laboratory
University of California
Livermore, California

October 17, 1958

INTRODUCTION

This report gives a short and practical summary of curve fitting by the method of least squares. The purpose is to list and roughly justify the formulae used in finding the best-fitting curve, the errors, and the quantities which describe the goodness of the fit. The contents are therefore well known to the statisticians, but apparently very poorly known to physicists, although the problem in question arises very often in everyday analysis of data. For more details as well as for a mathematically more satisfactory treatment the reader is referred to textbooks on statistics. This report, however, is self-contained, and explicitly utilitarian in tone.

1. NOTATION AND DEFINITIONS

Experimental Point:

$$y(x_i) \equiv y_i , \qquad i = 1 \ldots n . \tag{1.1}$$

This is usually the average of a series of measurements of y at the point x_i.

Uncertainty (experimental) in y_i:

$$\zeta_i \equiv \zeta(x_i). \tag{1.2}$$

This is the experimentally determined estimate of the standard deviation. We assume throughout that x_i is measured exactly, or at least with an error which is negligible compared to that of y_i.

We approximate the experimental points by the series of degree n:

$$f(x_i) = \sum_{k=0}^{n} a_k \phi_k(x_i). \tag{1.3}$$

*This work was done under the auspices of the U.S. Atomic Energy Commission.

The functions $\phi_k(x)$ may be any set of linearly independent functions of x. Most simply we may set $\phi_k(x) = x^k$, which would give us a polynomial of degree n. Other possible choices are, for instance, Legendre polynomials, Tschebycheff polynomials, $\phi_k(x) = \sin kx$, etc

Residual for this function:

$$r(x_i) \equiv r_i = y_i - f(x_i) = y_i - \sum_{k=0}^{n} a_k \phi_k(x_i) . \tag{1.4}$$

Weights:

$$w_i = \frac{1}{\zeta_i^2} \geq 0. \tag{1.5}$$

The "real" or "true" values of the y_i are denoted by

$$Y(x_i) \equiv Y_i . \tag{1.6}$$

The Y_i are the exact unknowable values which the y_i would approach if an infinite number of measurements could be made of them.

Standard deviation (true) or variance:

$$\sigma(x_i) \equiv \sigma_i . \tag{1.7}$$

The series of degree n approximating the $Y(x_i)$'s:

$$F(x_i) = \sum_{k=0}^{n} A_k \phi_k(x_i) \tag{1.8}$$

Residual for this series:

$$R(x_i) \equiv R_i = Y_i - F(x_i) . \tag{1.9}$$

"True" error in the y_i's:

$$E(x_i) \equiv E_i = Y_i - y_i . \tag{1.10}$$

We define for the "experimental" approximating function

$$q = \sum_{i=1}^{N} w_i \left[y_i - f(x_i)\right]^2 = \sum_{i=1}^{N} w_i r_i^2 . \tag{1.11}$$

We define for the "true" approximating function

$$Q = \sum_{i=1}^{N} w_i \left[Y_i - F(x_i)\right]^2 = \sum_{i=1}^{N} w_i R^2(x_i) . \tag{1.12}$$

We will set

$$\frac{\sigma_i^2}{\zeta_i^2} = \rho^2$$

for all i, where

$$\rho^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{\sigma_i^2}{\zeta_i^2} \tag{1.13}$$

"Experimental" quantities are lower-case:

$$y_i,\ x_i,\ \zeta_i,\ f(x_i),\ a_k\ r(x_i),\ w_i,\ q.$$

"True" quantities are upper-case (except σ_i):

$$Y_i,\ F(x_i),\ A_k,\ R(x_i),\ Q,\ E(x_i).$$

We wish to reassure the reader that the "true" quantities are used merely for mathematical purposes, and that a knowledge of them is not necessary for the use of the results of this article.

2. STATEMENT OF THE PROBLEM

The basic problem is as follows. For each abscissa x_i $(i = 1 \ldots N)$ we are given an experimental value of the ordinate y_i, with experimental uncertainty ζ_i. We wish to construct a curve of a given type, i.e., with a given $\phi_k(x)$ or of a given degree, in such a way that this curve most closely approximates the data.

The problem is solved as follows. We assume that the measured values y_i have a Gaussian (normal) distribution around the "true" value Y_i, with standard deviation σ_i. Thus the probability that for an experimentally determined y_i the "true" value Y_i lies between Y_i and dY_i is

$$P(Y_i)dY_i = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{(Y_i - y_i)^2}{2\sigma_i^2}\right] dY_i\,. \tag{2.1}$$

Thus the best (most probable) guess of the Y_i, $i = 1 \ldots N$ is obtained when

$$P = \prod_{i=1}^{N} P(Y_i) = \prod_{i=1}^{N} \left[\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(Y_i - y_i)^2}{2\sigma_i^2}\right)\right]$$

$$= \prod_{i=1}^{N} \left(\frac{1}{\sigma_i \sqrt{2\pi}}\right) \exp\left[-\sum_{j=1}^{N} \frac{(Y_j - y_j)^2}{2\sigma_j^2}\right], \tag{2.2}$$

is a maximum, or--what amounts to the same thing--when

$$\sum_{i=1}^{N} \frac{(Y_i - y_i)^2}{\sigma_i^2} \tag{2.3}$$

is a minimum. Now, we wish to determine the a_k's in Eq. (1.3) so that the $f(x_i)$ is the most probable estimate of the Y_i's. From (2.3) this will occur for

$$q = \sum_{i=1}^{N} w_i \, (y_i - f(x_i))^2 \tag{2.4}$$

a minimum. Here we have used ζ_i ($w_i = \zeta_i^2$) as the best available estimate of σ_i. Equation (2.4) is Legendre's principle of least squares.

3. OBTAINING THE COEFFICIENTS OF f(x)

We want to minimize

$$q \equiv \sum_{i=1}^{N} \left(\frac{\sum_{k=0}^{N} a_k \, \phi_k \, (x_i) - y_i}{\zeta_i} \right)^2 \tag{3.1}$$

or we want to satisfy

$$\frac{\partial q}{\partial a_\ell} = 2 \sum_{i=1}^{N} \left(\frac{\sum_{k=0}^{N} a_k \, \phi \, (x_i) - y_i}{\zeta_i} \right) \frac{\phi_\ell (x_i)}{\zeta_i} = 0 \tag{3.2}$$

for all $\ell = 0 \ldots n$.

Thus we get a set of linear equations

$$\sum_{k=0}^{n} a_k \left(\sum_{i=1}^{N} \frac{\phi_k(x_i) \, \phi_\ell \, (x_i)}{\zeta_i^2} \right) = \sum_{i=1}^{N} \frac{y_i \, \phi_\ell (x_i)}{\zeta_i^2} \quad \ell = 0 \ldots n. \tag{3.3}$$

These are called the <u>normal equations</u> for the least-square-fitting procedure. These equations can also be written as a matrix equation,

$$h \, a = g. \tag{3.4}$$

where

$$h_{k\ell} = \sum_{i=1}^{N} \frac{\phi_k(x_i)\,\phi_\ell(x_i)}{\zeta_i^{\,2}} \; ; \quad g_\ell = \sum_{i=1}^{N} \frac{y_i\,\phi_\ell(x_i)}{\zeta_i^{\,2}} \; . \tag{3.5}$$

The solution can therefore be written as

$$a_k = \sum_{\ell=0}^{n} (h^{-1})_{k\ell}\, g_\ell = \sum_{\ell=0}^{n} (h^{-1})_{k\ell} \sum_{i=1}^{N} \frac{y_i\,\phi_\ell(x_i)}{\zeta_i^{\,2}} \; . \tag{3.6}$$

The matrix h is symmetric. If the $\phi_\ell(x_i)$'s are orthogonal, h is a diagonal matrix.

4. ERRORS IN THE COEFFICIENTS

The "true" error in the y_i is given by $E_i = Y_i - y_i$. From Eq. (2.1) the probability that E_i lies between E_i and $E_i + dE_i$ is given by

$$P(E_i)dE_i = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{E_i^{\,2}}{2\sigma_i^{\,2}}\right] dE_i \; . \tag{4.1}$$

Thus, if we make a very large number of measurements of y_i we find, for the average value of $E_i^{\,2}$,

$$\overline{E_i^{\,2}} = \frac{1}{\sigma_i\sqrt{2\pi}} \int_{-\infty}^{\infty} E_i^{\,2} \exp\left[-\frac{E_i^{\,2}}{2\sigma_i^{\,2}}\right] dE_i = \sigma_i^{\,2} \; , \tag{4.2}$$

and for $i \neq j$,

$$\overline{E_i E_j} = \int_{-\infty}^{\infty} \frac{dE_i}{\sigma_i\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{dE_j}{\sigma_j\sqrt{2\pi}}\; E_i\,E_j \exp\left[-\frac{1}{2}\left(\frac{E_i^{\,2}}{\sigma_i^{\,2}} + \frac{E_j^{\,2}}{\sigma_j^{\,2}}\right)\right] = 0. \tag{4.3}$$

The bar over the variables denotes an average over a very large number of measurements.

Let the "true" series fitting the "true" data Y_i be given by

$$F(x_i) = \sum_{k=0}^{\infty} A_k\,\phi_k(x_i).$$

We can then define the uncertainty in the coefficient of f(x) to be

$$\left[\overline{(a_k - A_k)^2}\right]^{1/2} . \tag{4.4}$$

Now, from Eq. (3.6),

$$a_k = \sum_{\ell=0}^{n} h_{k\ell}^{-1} \sum_{i=1}^{N} \frac{y_i \, \phi_\ell(x_i)}{\zeta_i^2} .$$

Since h^{-1} does not depend on the y_i, being dependent only on the x_i and ζ_i, we find

$$A_k = \sum_{\ell=0}^{n} h_{k\ell}^{-1} \sum_{i=1}^{N} \frac{Y_i \, \phi_\ell(x_i)}{\zeta_i^2} . \tag{4.5}$$

Hence

$$(a_k - A_k)^2 = \left[\sum_{\ell=0}^{n} h_{k\ell}^{-1} \sum_{i=1}^{N} \frac{\phi_\ell(x_i)}{\zeta_i^2} E_i\right]^2 \tag{4.6}$$

Let us denote temporarily

$$\Phi_k(x_i) = \sum_{\ell=0}^{n} h_{k\ell}^{-1} \phi_\ell(x_i) \tag{4.7}$$

Then we have

$$(a_k - A_k)^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{\zeta_i^2} \frac{1}{\zeta_j^2} \Phi_k(x_i) \Phi_k(x_j) E_i E_j . \tag{4.8}$$

Taking the average of both sides, using (4.2) and (4.3) gives us

$$\overline{(a_k - A_k)^2} = \sum_{i=1}^{N} \left(\frac{\Phi_k^2(x_i)}{\zeta_i^2}\right)\left(\frac{\sigma_i^2}{\zeta_i^2}\right) . \tag{4.9}$$

The ζ_i are supposed to be the best possible experimental estimates of the σ_i. It is therefore not unreasonable to assume that σ_i^2/ζ_i^2 does not vary much with i, and to replace it by its average value

$$\rho^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{\sigma_i^2}{\zeta_i^2} .$$

We can thus write (4.7)

$$\overline{(a_k - A_k)^2} = \rho^2 \sum_{i=1}^{N} \left(\frac{\Phi_k^{\,2}(x_i)}{\zeta_i^{\,2}} \right). \tag{4.10}$$

Now

$$\sum_{i=1}^{N} \frac{\Phi_k^{\,2}(x_i)}{\zeta_i^{\,2}} = \sum_{i=1}^{N} \frac{1}{\zeta_i^{\,2}} \sum_{\ell=0}^{n} h_{\ell k}^{-1} \phi_\ell(x_i) \sum_{m=0}^{n} h_{mk}^{-1} \phi_m(x_i)$$

$$= \sum_{\ell=0}^{n} h_{\ell k}^{-1} \sum_{m=0}^{n} h_{mk}^{-1} \sum_{\ell=1}^{N} \frac{1}{\zeta_i^{\,2}} \phi_\ell(x_i) \phi_m(x_i)$$

$$= \sum_{\ell=0}^{n} h_{\ell k}^{-1} \sum_{m=0}^{n} h_{mk}^{-1} h_{\ell m} \qquad \text{by (3.5)}$$

$$= \sum_{\ell=0}^{n} h_{\ell k}^{-1} \delta_{\ell k} = h_{kk}^{-1}, \tag{4.11}$$

so that

$$\overline{(a_k - A_k)^2} = \rho^2 h_{kk}^{-1}. \tag{4.12}$$

This is as far as we can go without further assumptions, since ρ^2 is not a directly measurable quantity.

In order to estimate ρ we proceed as follows:

$$E_i = Y_i - y_i,$$

and therefore

$$E_i + r_i = F(x_i) - f(x_i) + R_i,$$

where

$$r_i = y_i - f(x_i); \qquad R_i = Y_i - F(x_i);$$

thus

$$E_i + r_i = \sum_{k=0}^{n} (A_k - a_k)\, \phi_k(x_i) + R_i$$

$$= \sum_{k=0}^{n} \phi_k(x_i) \sum_{j=1}^{N} \frac{\Phi_k(x_j)}{\zeta_j^{\,2}} E_j + R_i \,. \tag{4.13}$$

Multiplying both sides by $r_i/\zeta_r^{\,2}$ and summing over i, we get

$$\sum_{i=1}^{N} \frac{1}{\zeta_i^{\,2}} r_i E_i + \sum_{i=1}^{N} \frac{1}{\zeta_i^{\,2}} r_i^{\,2} =$$

$$\sum_{i=1}^{N} \sum_{k=0}^{n} \sum_{j=1}^{N} \frac{1}{\zeta_i^{\,2}} \frac{1}{\zeta_j^{\,2}} r_i\, \phi_k(x_i)\, \Phi_k(x_j)\, E_j + \sum_{i=1}^{N} \frac{r_i R_i}{\zeta_i^{\,2}} \tag{4.14}$$

From Eq. (3.2) we see immediately

$$\sum_{i=1}^{N} \frac{1}{\zeta_i^{\,2}} r_i\, \phi_k(x_i) = 0 \,, \tag{4.15}$$

whereupon (4.14) becomes

$$\sum_{i=1}^{N} \frac{1}{\zeta_i^{\,2}} r_i E_i + \sum_{i=1}^{N} \frac{1}{\zeta_i^{\,2}} r_i^{\,2} = \sum_{i=1}^{N} \frac{r_i R_i}{\zeta_i^{\,2}} \,. \tag{4.16}$$

Multiplying Eq. (4.13) by $E_i/\zeta_i^{\,2}$ and summing over i, we get

$$\sum_{i=1}^{N} \frac{E_i^2}{\zeta_i^2} + \sum_{i=1}^{w} \frac{r_i E_i}{\zeta_i^2} = \sum_{k=0}^{n} \sum_{j=0}^{n} \sum_{i=0}^{N} \phi_i(x_i)\, \Phi_k(x_j) \frac{E_i E_j}{\zeta_j^2 \zeta_i^2}$$

$$+ \sum_{i=1}^{N} \frac{E_i R_i}{\zeta_i^2} . \tag{4.17}$$

Eliminating $\sum \frac{1}{\zeta_i^2} r_i E_i$ between (4.16) and (4.17),

$$\sum_{i=1}^{N} \frac{E_i^2}{\zeta_i^2} + \sum_{i=1}^{N} \frac{r_i R_i}{\zeta_i^2} - \sum_{i=1}^{N} \frac{1}{\zeta_i^2} r_i^2 =$$

$$\sum_{k=0}^{n} \sum_{j=1}^{N} \sum_{i=1}^{N} \phi_k(x_i)\, \Phi_k(x_j) \frac{E_i E_j}{\zeta_i^2 \zeta_j^2} + \sum_{i=1}^{N} \frac{E_i R_i}{\zeta_i^2} . \tag{4.18}$$

Let us now take the average of both sides of Eq. (4.18) on the basis of the normal distribution of the y_i. We note that $\overline{y_i} = Y_i$, so that $\overline{r_i} = R_i$, and $\overline{E_i} = 0$. Using these results as well as (4.2) and (4.3), we get for Eq. (4.18),

$$\sum_{i=1}^{N} \frac{\sigma_i^2}{\zeta_i^2} + \sum_{i=1}^{N} \frac{R_i^2}{\zeta_i^2} - \overline{\sum \frac{1}{\zeta_i^2} r_i^2} = \sum_{k=0}^{n} \sum_{i=1}^{N} \frac{\phi_k(x_i)\Phi_k(x_i)}{\zeta_i^2} \frac{\sigma_i^2}{\zeta_i^2} \tag{4.19}$$

From the definition of ρ^2 we have $\sum_{i=1}^{N} \frac{E_i^2}{\zeta_i^2} = N \rho^2$, and setting $\frac{\sigma_i^2}{\zeta_i^2} = \rho^2$ on the left-hand side of (4.19), we get

$$N \rho^2 + \sum_{i=1}^{N} \frac{R_i^2}{\zeta_i^2} - \overline{\sum_{i=1}^{N} \frac{1}{\zeta_i^2} r_i^2} = \rho^2 \sum_{k=0}^{n} \sum_{i=1}^{N} \frac{\phi_k(x_i)\Phi_k(x_i)}{\zeta_i^2} \tag{4.20}$$

However,

$$\sum_{i=1}^{N} \frac{\phi_k(x_i)\,\Phi_k(x_i)}{\zeta_i^{\,2}} = \sum_{i=1}^{N} \frac{\phi_k(x_i)}{\zeta_i^{\,2}} \sum_{\ell=0}^{n} h_{k\ell}^{\,-1}\,\phi_\ell(x_i)$$

$$= \sum_{\ell=0}^{n} h_{k\ell}^{\,-1}\,h_{k\ell} = 1$$

and

$$\sum_{k=0}^{n} \sum_{i=1}^{N} \frac{\phi_k(x_i)\,\Phi_k(x_i)}{\zeta_i^{\,2}} = \sum_{k=0}^{n} 1 = n + 1 ,$$

so that Eq. (4.20) becomes

$$\rho^2 = \frac{1}{N - n - 1} \left[\sum_{i=1}^{N} \overline{\frac{1}{\zeta_i^{\,2}}\, r_i^{\,2}} - \sum_{i=1}^{N} \frac{1}{\zeta_i^{\,2}}\, R_i^{\,2} \right] . \tag{4.21}$$

We may approximate the first term on the right by the actually observed residuals, in which case

$$\rho^2 \approx \frac{1}{N - n - 1} \left[q - \sum_{i=1}^{N} \frac{R_i^{\,2}}{\zeta_i^{\,2}} \right] . \tag{4.22}$$

If now $R_i = 0$ for all i -- that is, if there is an underlying physical law which obeys a series of degree n,

$$Y_i = \sum_{k=0}^{n} A_k\,\phi_k(x_i) \quad --$$

then $\rho^2 \approx q/(N - n - 1)$. In case $R_i \neq 0$ we see from Eq. (4.22) that $q/(N - n - 1) > \rho^2$; thus, in setting $\rho^2 = q/(N - n - 1)$ in Eq. (4.10) we are making a conservative estimate of ρ^2 and also in some measure taking into account the poorness of the fit. The error in the coefficients may consequently be written

$$(\Delta a_k)^2 = \overline{(a_k - A_k)^2} = \frac{h_{kk}^{-1} q}{N - n - 1} \tag{4.23}$$

$$= \frac{h_{kk}^{-1}}{N - n - 1} \sum_{i=1}^{N} \frac{1}{\zeta_i^2} r_i^2 .$$

5. ERROR AT ANY GIVEN POINT

The error at any given point is given by

$$\Delta^2(x) = [F(x) - \ell(x)]^2 = [\sum_{k=0}^{n} (A_k - a_k)\, \phi_k(x)]^2 =$$

$$= \sum_{k=0}^{n} \sum_{\ell=0}^{n} (A_k - a_k)(A_\ell - a_\ell)\, \phi_k(x)\, \phi_\ell(x) \tag{5.1}$$

Using Eq. (4.4) and (4.5), we can write this as

$$\Delta^2(x) = \sum_{k=0}^{n} \sum_{\ell=0}^{n} \sum_{i=1}^{N} \frac{1}{\zeta_i^2} \frac{1}{\zeta_i^2} E(x_i)\, E(x_i)\, \Phi_k(x_i)\, \Phi_\ell(x_i)\, \phi_k(x)\, \phi_\ell(x). \tag{5.2}$$

Again assuming, as in Section 4, the randomness of the distribution in y_i, we get

$$\Delta^2(x) = \rho^2 \sum_{k=0}^{n} \sum_{\ell=0}^{n} \phi_k(x)\, \phi_\ell(x) \sum_{i=1}^{N} \frac{1}{\zeta_i^2} \Phi_k(x_i)\, \Phi_\ell(x_i) . \tag{5.3}$$

But

$$\sum_{i=1}^{N} \frac{1}{\zeta_i^2} \Phi_k(x_i)\, \Phi_\ell(x_i) = \sum_{i=1}^{N} \frac{1}{\zeta_i^2} \sum_{p=0}^{n} h_{pk}^{-1} \phi_p(x_i) \sum_{q=0}^{n} h_{q\ell} \phi_q(x_i) =$$

$$= \sum_{i=1}^{N} \sum_{p=0}^{n} \sum_{q=0}^{n} \frac{1}{\zeta_i^2} \phi_p(x_i)\, \phi_q(x_i)\, h_{pk}^{-1}\, h_{q\ell}^{-1} =$$

$$= \sum_{p=0}^{n} \sum_{q=0}^{n} h_{pq} h_{pk}^{-1} h_{q\ell}^{-1} = \sum_{q=0}^{n} \delta_{p\ell} h_{\ell k}^{-1} = h_{\ell k}^{-1} , \tag{5.4}$$

so that

$$\Delta^2(x) = \rho^2 \sum_{k=0}^{n} \sum_{\ell=0}^{n} h_{\ell k}^{-1} \phi_k(x) \phi_\ell(x) . \tag{5.5}$$

Again, assuming that

(a) the "true" function $F(x)$ is an ideally good approximation to Y_i, i.e. $R_i \approx 0$, and
(b) the sum of the squares of the average values of the weighted residuals can be approximated by the sum of the squares of the observed values of the residuals, we get

$$\Delta^2(x) = \frac{1}{N - n - 1} \left(\sum_{i=1}^{N} \frac{r^2(x_i)}{\zeta_i^2} \right) \left[\sum_{k=0}^{n} \sum_{\ell=0}^{n} h_{\ell k}^{-1} \phi_k(x) \phi_\ell(x) \right] . \tag{5.6}$$

6. SOME PROPERTIES OF THE ERRORS

A. The Average Error of the Fitting Function at the Points of Observation

If we set $x = x_i$ in Eq. (5.5), divide by ζ_i^2 and sum over i, we obtain

$$\sum_{i=1}^{N} \frac{\Delta^2(x_i)}{\zeta_i^2} = \rho^2 \sum_{i=1}^{N} \frac{1}{\zeta_i^2} \sum_{k=0}^{n} \sum_{\ell=0}^{n} h_{\ell k}^{-1} \phi_k(x_i) \phi_\ell(x_i)$$

$$= \rho^2 \sum_{k=0}^{n} \sum_{\ell=0}^{n} h_{\ell k}^{-1} h_{\ell k} = (n + 1) \rho^2 , \tag{6.1}$$

and therefore

$$\left(\frac{\Delta^2(x_i)}{\zeta_i^2}\right)_{av} = \frac{n+1}{N}\,\rho^2 = \frac{n+1}{N}\,\frac{1}{N-n-1}\sum_{i=1}^{N}\frac{r^2(x_i)}{\zeta_i^2}, \tag{6.2}$$

where the av refers to averaging over the points of observation. The extreme right-hand side holds under the conditions that make Eq. (4.26) and Eq. (5.6) possible. Thus if ρ^2 is unity (see section 7), the average error of the fitting curve is always less than the error of the observation at that point, since $n + 1 \leq N$ for a meaningful fit.

B. The Average Error over an Interval

This average is given by

$$\overline{\Delta^2(x)} = \frac{1}{x_1 - x_2}\int_{x_1}^{x_2}\Delta^2(x)\,dx = \frac{1}{N-n}\left(\sum_{i=1}^{n}\frac{r^2(x_i)}{\zeta_i^2}\right)\frac{1}{x_2 - x_1}$$

$$\sum_{k=0}^{n}\sum_{\ell=0}^{n} h_{\ell k}^{-1}\int_{x_1}^{x_2}\phi_k(x)\,\phi_\ell(x)\,dx, \tag{6.3}$$

which in case of polnomial functions can be written as

$$\overline{\Delta^2(x)} = \frac{1}{N-n-1}\left(\sum_{i=1}^{N}\frac{r^2(x_i)}{\zeta_i^2}\right)\frac{1}{x_2 - x_1}\sum_{k=0}^{n}\sum_{\ell=0}^{n}\frac{h_{\ell k}^{-1}}{\ell+k+1}\left(x_2^{\ell+k+1} - x_1^{\ell+k+1}\right). \tag{6.4}$$

7. GOODNESS-OF-FIT CRITERIA

The error in the fitting series, as given by Eq. (5.6) is composed of two factors, One of them, namely

$$\sum_{k=0}^{n} \sum_{\ell=0}^{n} h_{k\ell}^{-1} \phi_k(x) \phi_\ell(x), \tag{7.1}$$

gives the contribution of the statistical errors of the experiment. This expression is independent of y_i and is determined only by the experimental errors ζ_i and the abscissae and basic functions we use for fitting.

The other factor in Eq. (5.6), namely

$$\frac{1}{N-n-1} \sum_{i=1}^{N} \frac{r^2(x_i)}{\zeta_i^2} \equiv \frac{1}{N-n-1} q, \tag{7.2}$$

which, once we have assumed that the "true" values Y_i are ideally fitted by the "true" function $F(x_i)$, gives an indication of the goodness of the set of data under consideration. In particular, it can tell us whether the set of data is an "unlikely" one or not. If it is a very unlikely set, one might suspect some systematic error in the measurement.

In practice, however, often it is not known for certain whether $F(x_i)$ indeed gives an ideal fit to Y_i. If not, and the minimum value of Q is not 0, this will also show up in Eq. (7.2). Thus it is in fact difficult to tell whether this is the case or whether there is a systematic error in the experiment.

The goodness-of-fit criteria are concerned with the quantity in Eq. (7.2). We discuss two tests for such fit, the so-called chi-square test and the so-called F test.

A. The Chi-Square (χ^2) Test

In Section 4 (Eq. (4.19)), we showed that for y_i which are distributed normally about the Y_i, then

$$\overline{\sum_{i=1}^{N} \frac{1}{\zeta_i^2}} = (N-n-1)\rho^2 + \sum_{i=1}^{N} \frac{1}{\zeta_i^2} R_i^2. \tag{7.3}$$

If the ζ_i are good estimates of the σ_i, then $\rho^2 \approx 1$. If in addition the $R_i \approx 0$, which would be the case for a good fit, then we would expect

$$q = \sum_{i=1}^{N} \frac{1}{\zeta_i^2} r_i^2 \approx N - n - 1 . \tag{7.4}$$

For a poor fit, $R_i^2 > 0$, so that we would expect $q > N - n - 1$.

It can be shown that for normally distributed y_i, with $R_i = 0$, the statistic q obeys what is known as a chi-square (χ^2) distribution, that is, the probability that q lies between q and q + dq is given by

$$P_M(q)dq = \frac{1}{2^{M/2} \Gamma (M/2)} e^{-q/2} q^{(M-2)/2} dq, \tag{7.5}$$

where $M = N - n - 1$ is the number of degrees of freedom of the distribution. A simple integration shows

$$\bar{q} = \int_0^\infty q P_M(q)dq = M , \tag{7.6}$$

in agreement with Eq. (7.4). In order to find the degree of the series which best fits the data, one calculates q/M for n = 1, 2, 3, . . . , etc. In general, q/M as a function of n will first decrease, then level off to a plateau, and finally perhaps slowly increase. The best series is thus the one with the smallest n at which the plateau begins.

Sometimes the plateau will not occur at $q/M \approx 1$, but at some higher value. If this value is so high as to be improbable it may be concluded either that the R_i are not small, or that there is some internal inconsistency in the data. The probability p that q be larger than a given value q_o is

$$\int_{q_o}^{\infty} P_M(q')dq' = p ; \qquad (0 < p < 1). \tag{7.7}$$

A brief table of p, q_o and M is given in the Appendix.

Another thing that can lead to values of q/M differing significantly from one is a poor estimate of the uncertainty ζ_i. For instance, unwarranted rejection of data would make ζ_i too small, resulting in a large ρ^2. On the other hand, estimates of ζ_i which are too conservative would make ρ^2 small, so that for $R_i \approx 0$, q/M would be considerably less than one.

Supplementing the χ^2 test one can also use the F test, which is discussed next.

B. The F Test

Strictly speaking the F test was designed to solve the following problem. Suppose we are given a set of normally distributed variables y_i, of which it is known that the true values Y_i obey a relation of the form

$$Y_i = \sum_{k=0}^{n} A_k \phi_k(x_i) ,$$

where n is unknown. For a given n, what is the probability, on the basis of the available data, that $A_n = 0$? In most practical cases we do not know whether the Y_i's obey a relation of the above form; as a matter of fact, in many cases (e.g., scattering cross-sections in nuclear physics) the underlying physical law may be an infinite series. However, even though the test is not rigorously applicable it may still be used to indicate the degree of the series that best fits the data.

Let the data be fitted with a series of degree n and with a series of degree n-1; let q_n and q_{n-1} be the respective weighted sums of the residuals. We form the statistic

$$S_M = \frac{M}{q_n} (q_{n-1} - q_n) , \qquad M = N - n - 1 . \tag{7.8}$$

Under the conditions stated above q_n obeys a χ^2 distribution with M degrees of freedom; similarly $q_{n-1} - q_n$ obeys a χ^2 distribution with one degree of freedom. Now the distribution of the ratio of two χ^2 variables divided by their degrees of freedom is called a Fisher $F(m_1, m_2)$ distribution, where m_1 and m_2 are the degrees of freedom of the variables in the numerator and denominator, respectively. In our case $m_1=1$, $m_2=M$, so that S_M has an F(1, M) distribution. If $P_M(F)$ is the F(1, M) distribution, then the probability α for

$S_M \geq F_\alpha(M)$ is given by

$$\alpha = \int_{F_\alpha(M)}^{\infty} P_M(F)\, dF . \tag{7.9}$$

The F test now states that for $S_M \geq F_\alpha(M)$ one may assume $A_n \neq 0$ with a probability α of being wrong in this assumption. Thus if S_M corresponds to an $F_\alpha(M)$ with $\alpha = 0.75$ there is a 75% chance of being right if we assume $A_n = 0$. A table of $F_\alpha(M)$ is given in the Appendix.

The F test must always be used in concert with the χ^2 test. For suppose the F test indicates that with a high probability $A_n = 0$, it is still possible that A_{n+1}, A_{n+2}, ... are not zero. However, if in that case we terminate the series at n-1 the χ^2 test indicates that a good fit has not yet been obtained.

8. EXAMPLE

Consider the set of numbers given at the left. By use of an IBM 650 computer these were fitted with the function

$$f(\chi) = \sum_{k=0}^{n} a_k \chi^k$$

for n = 1, 2, 3, 4, 5. The results are shown below.

χ	y	ζ
.12	3.85	.09
.56	9.42	.15
.83	12.90	.42
1.36	17.36	.42
1.48	19.31	.23
1.73	22.73	.27
2.20	32.89	.36
2.57	44.51	.83
2.83	53.01	.52
3.01	62.09	.61
3.32	81.00	.93
3.62	102.11	.86
3.90	124.00	.71

n	1	2	3	4	5
M	11	10	9	8	7
a_0	- 0.5567	4.8593	1.9142	1.8899	1.9802
a_1	18.9928	- 1.5784	17.0535	17.2574	16.2962
a_2	----	7.3924	- 8.0616	- 8.3539	- 6.2119
a_3	----	----	3.0132	3.1472	1.4636
a_4	----	----	----	- 0.01877	0.5136
a_5	----	----	----	----	- 0.0580
q_n	11148.407	1471.809	13.142	13.097	12.838
q_n/M	1004.01	147.1809	1.4002	1.6371	1.8340
S_M	65.7462	1041.741	0.02748	0.14122	----

M = N - n - 1, N = 13

We see immediately that the χ^2 test selects n = 3 as giving the best fit. The F test corroborates this selection, since $S_M = 0.02748$ for M = 9 corresponds, according to the table in the appendix, to $\alpha \approx 0.85$. Thus in terminating the series at n = 3 we have approximately an 85% chance of being right.

APPENDIX

Distribution of χ^2

Values of q_o for given M and p

$$\int_{q_o}^{\infty} p_M(\chi^2)\, d\chi^2 = p$$

M \ p	.01	.05	.10	.20	.50	.80	.90	.95	.99
1	6.63	3.84	2.71	1.64	0.455	0.064	0.016	0.004	0.0002
3	11.34	7.81	6.25	4.64	2.37	1.00	0.584	0.352	0.115
5	15.09	11.07	9.24	7.29	4.35	2.34	1.61	1.14	0.554
7	18.47	14.07	12.02	9.80	6.35	3.82	2.83	2.17	1.24
9	21.67	16.92	14.68	12.24	8.34	5.38	4.17	3.32	2.09
10	23.21	18.31	15.99	13.44	9.34	6.18	4.86	3.94	2.56
15	30.58	25.00	22.31	19.31	14.34	10.31	8.55	7.26	5.23
20	37.57	31.41	28.41	25.04	19.34	14.58	12.44	10.85	8.26
25	44.31	37.65	34.38	30.67	24.34	18.94	16.47	14.61	11.52
30	50.89	43.77	40.26	36.25	29.34	23.36	20.60	18.49	14.95

Distribution of F:

Values of $F_\alpha(M)$ for given M and α

$$\int_{F_\alpha(M)}^{\infty} p_M(F)\, dF = \alpha$$

α \ M	5	10	15	20	30	40	60	120	∞
.999	1.73×10^{-6}	1.65×10^{-6}	1.62×10^{-6}	1.61×10^{-6}	1.60×10^{-6}	1.59×10^{-6}	1.58×10^{-6}	1.58×10^{-6}	1.57×10^{-6}
.99	1.75×10^{-6}	1.65×10^{-4}	1.62×10^{-4}	1.61×10^{-4}	1.60×10^{-4}	1.59×10^{-4}	1.58×10^{-4}	1.58×10^{-4}	1.57×10^{-4}
.95	4.35×10^{-3}	4.13×10^{-3}	4.07×10^{-3}	4.03×10^{-3}	4.00×10^{-3}	3.98×10^{-3}	3.96×10^{-3}	3.95×10^{-3}	3.93×10^{-3}
.90	0.0175	0.0166	0.0163	0.0162	0.0160	0.0160	0.0159	0.0158	0.0158
.80	0.0713	0.0676	0.0666	0.0660	0.0655	0.0650	0.0645	0.0645	0.0640
.75	0.113	0.107	0.105	0.104	0.103	0.103	0.1025	0.102	0.1015
.50	0.523	0.490	0.477	0.472	0.466	0.464	0.461	0.458	0.454
.25	1.69	1.49	1.43	1.40	1.38	1.36	1.35	1.34	1.32
.10	4.06	3.29	3.07	2.97	2.88	2.84	2.79	2.75	2.71
.05	6.61	4.96	4.54	4.35	4.17	4.08	4.00	3.92	3.84
.01	16.26	10.04	8.68	8.10	7.56	7.31	7.08	6.35	6.63
.005	22.78	12.83	10.80	9.94	9.18	8.83	8.49	8.18	7.88
.001	47.18	21.04	16.59	14.82	13.29	12.61	11.97	11.38	10.38

REFERENCES

1. Sir Ronald Aylmer Fisher, Statistical Methods for Research Workers, 12 ed. rev. (Hafner, New York, 1954).

2. Henry Berthold Mann, Analysis and Design of Experiments; Analysis of Variance and Analysis of Variance Design (Dover, New York, 1949).

3. Francis B. Hildebrand, Introduction to Numerical Analysis (McGraw-Hill, New York, 1956), especially Chapter 7.

4. Harald Cramé, Mathematical Methods of Statistics, (Princeton University Press, Princeton, 1946).

5. S.S. Wilks, Mathematical Statistics (Princeton University Press, Princeton, 1943).

6. E.S. Pearson and H. O. Hartley (Eds.) Biometrika Tables for Statisticians. (Cambridge University Press, Cambridge, England 1954).

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission to the extent that such employee or contractor prepares, handles or distributes, or provides access to, any information pursuant to his employment or contract with the Commission.