



Guide to linear Least Squares curve fitting techniques

www.esdu.com

Endorsed by
The Royal Aeronautical Society

ESDU DATA ITEMS

Data Items provide validated information in engineering design and analysis for use by, or under the supervision of, professionally qualified engineers. The data are founded on an evaluation of all the relevant information, both published and unpublished, and are invariably supported by original work of ESDU staff engineers or consultants. The whole process is subject to independent review for which crucial support is provided by industrial companies, government research laboratories, universities and others from around the world through the participation of some of their leading experts on ESDU Technical Committees. This process ensures that the results of much valuable work (theoretical, experimental and operational), which may not be widely available or in a readily usable form, can be communicated concisely and accurately to the engineering community.

We are constantly striving to develop new work and review data already issued. Any comments arising out of your use of our data, or any suggestions for new topics or information that might lead to improvements, will help us to provide a better service.

THE PREPARATION OF THIS DATA ITEM

The work on this particular Data Item was monitored and guided by the Performance Committee. This Committee first met in 1946 and now has the following membership:

Chairman	
Mr G.J.R. Skillen	— Independent, UK
Vice-Chairman	
Mr K.J. Balkwill	— Independent, UK
Members	
Mr P.V. Aidala *	— Northrop Grumman ESID, Bethpage, New York, USA
Dr J.D. Avery	— MoD Boscombe Down, UK
Mr T.G.R. Bos *	— Dutch Airline Pilots Association, The Netherlands
Mr M. Broad	— Independent, UK
Dr G. van Es	— National Aerospace Laboratory NLR, The Netherlands
Mr N.J. Herniman	— Airbus Industrie, Toulouse, France
Mr M. Holcomb *	— Eclipse Aviation Corporation, Albuquerque, New Mexico, USA
Mr R.G. Humpston	— Rolls-Royce plc, Aero Division, Derby, UK
Mr D. Law	— Civil Aviation Authority, Safety Regulation Group, Gatwick
Mr R.J. Orłowski *	— Boeing Commercial Airplane Company, Seattle, Washington, USA
Dr P. Render	— University of Loughborough, UK
Dr A. Smith	— Air Systems Test & Evaluation, QinetiQ, MOD Boscombe Down, UK
Mr G.E. Smith	— Independent, UK
Mr A. Stanbrook	— Independent, UK
Mr R. Storey	— BAE SYSTEMS, Brough, UK
Mr S. Tsach *	— Israel Aircraft Industries, Jerusalem, Israel
Mr C.J. Turner *	— Boeing Commercial Airplane Company, Seattle, Washington, USA
Mr M. Wilson	— GAMTA Ltd, Aylesbury, UK

* Corresponding Member

The technical work involved in the assessment of the available information and the construction and subsequent development of the Data Item was carried out under contract to ESDU by a Member of the Performance Committee

Mr G.E. Smith.

The person with overall responsibility for the work in this subject area is Mr C.J. Loughton, Head of the Mechanical Motion and System Dynamics Group.

GUIDE TO LINEAR LEAST SQUARES CURVE FITTING TECHNIQUES

CONTENTS

	Page
1. NOTATION AND GLOSSARY	1
2. INTRODUCTION	4
3. LINEAR LEAST SQUARES METHOD OF CURVE FITTING	6
3.1 Basis of Method and its Application	6
3.2 Examples of Types of Function Used to Illustrate Curve Fitting	6
3.2.1 Polynomial function of single independent variable	7
3.2.2 Rational function	7
3.2.3 Peak function	8
3.2.4 Exponential function	9
3.3 Criteria for Quality of Fit of Function	9
3.4 Scaling of Data	10
3.5 Preparation of Data for Curve Fitting	11
4. UNDERFITTING, OVERFITTING AND SPARSE DATA SETS	12
4.1 Underfitting and Overfitting	12
4.2 Problems Arising with Sets of Sparse Data	16
5. EXAMPLES OF FITTING ARBITRARY EQUATIONS TO GIVEN DATA	19
5.1 Fuel Flow of Shaft Gas Turbines	19
5.2 Spray Patterns from Aircraft Tyres Running in Water	22
5.3 Induced Power Required by Helicopter	27
6. EXAMPLES OF FITTING EQUATIONS HAVING PHYSICALLY SIGNIFICANT TERMS	30
6.1 Drag and Lift Coefficients	30
6.1.1 Linear relationship between lift and drag coefficients	30
6.1.2 Nonlinear relationship between lift and drag coefficients	30
6.2 Compressibility Drag Rise of Large Transport Aircraft	32
6.3 Trigonometric Function Expressing Nonlinear Relationship	34
6.4 Nonlinear Equation of Circle and Circular Arc Data	35
6.5 Nonlinear Equation of Circle and Data for Complete Circle	38
6.6 Nonlinear Equation for Ellipse and Ellipse Data	39
7. EXAMPLES OF SPECIFIC FUNCTIONS TO FIT LOCALISED FEATURES OF CURVES	41
7.1 Curve Fit to Data with Distinct Peaks and Troughs	41
7.2 Curve Fit to Data with Well Separated Peaks	44
7.3 Curve Fits for Discontinuous Functions	46

8. REFERENCES	47
APPENDIX A – THE LINEAR LEAST SQUARES METHOD	48
A1. Derivation of Linear Least Squares Method of Curve Fitting	48
A2. Generalisation of Least Squares Method to Functions of Several Variables	50
APPENDIX B – ORTHOGONAL FUNCTIONS	51
B1. Application of Orthogonal Functions to Curve Fitting	51

GUIDE TO LINEAR LEAST SQUARES CURVE FITTING TECHNIQUES

1. NOTATION AND GLOSSARY

Some symbols are used in more than one context, with different meanings. Each definition is given within the section in which it occurs, but the majority of the symbols are defined here.

A, B	coefficients in exponential function (see Equation (3.9))
$[A]$	square symmetric matrix (see Appendix A)
B	variable to control sharpness of peak or trough (see Equation (3.8))
C_D	drag coefficient
C_{DM}	compressibility drag rise coefficient
C_{D0}	profile drag coefficient
C_L	lift coefficient
C_P	base pressure coefficient on afterbody
$C_1, C_2, \dots, C_p$	coefficients in functions (see Equation (3.1))
F_p	function producing unity peak at specified position on abscissa
F_{sx}, F_{sy}	scaling factors for transformed values of x and y
$F_1, F_2, \dots, F_p$	known functions of x (see Equation (3.1))
$f(x), g(x), h(x)$	general functions x
G	fuel flow of shaft gas turbine
G_n	non-zero value of integrals of Legendre polynomials
$H(x, y)$	generalised function of x and y
K	value of abscissa at which peak or trough has its maximum or minimum (see Equation (3.8))
K_D	lift factor
M	Mach number
M_D	value of M above which compressibility effects are significant
N	number of data points
P	axial load on struct
P_I	induced power required by helicopter
P_s	shaft power of gas turbine

Issued November 2006 – 53 pages

$P_1, P_2, P_3, \dots$	Legendre polynomials
p	number of terms in polynomial function and in numerator of rational function
q	number of terms in denominator of rational function
r	radius of circle
r	general variable
S	sum of weighted squares of residuals
V	forward airspeed of helicopter
V_G	groundspeed of tyre running in water
V_P	reference speed
$\bar{V}$	V/V_P
$\bar{V}_p$	$(\bar{V} - 1)/0.6$
$w_1, w_2, \dots, w_N$	weightings applied to square of residuals (see Appendix A)
x, y	independent and dependent variable
x_c, y_c	x and y values of centre of circles
x_0	offset value of x (see Equation (3.13))
γ_{Ds}	side spray elevation angle
Δx	increment in x
δ	ratio of ambient pressure to sea level standard value
δ	deflection
θ	ratio of ambient absolute temperature to sea level standard value
ϕ	phase angle
φ	orthogonal function

Subscripts

min	denotes minimum
$crit$	denotes critical value
L, U	denotes lower and upper
0	denotes initial value

Superscript

$'$	used with symbol to denote a transformed range
-----	--

Glossary

Asymptote	A straight or curved line that is closely approached by a plane curve so that the perpendicular distance between them decreases steadily to zero as the independent variable governing the curve approaches a critical value or tends to infinity.
Degree	The highest power in a polynomial is known as its degree.
Dependent Variable	The variable in a mathematical function whose value is entirely dependent on the chosen values of the independent variable(s) and on the function itself. (It is customary, whenever practical, for the dependent variable to appear on the left-hand-side of an equation and as the vertical axis (ordinate) on a simple graph.)
Discontinuity	A point on a curve where the function is not continuous, <i>i.e.</i> its value takes a jump.
Extrapolation	The process of estimating the value of a function beyond the range of values already known, by the extension of its curve.
Even Function	A function, $f(x)$, is even if $f(x) = f(-x)$.
Independent Variable	The variable(s) in a mathematical function whose values are considered to be selected independently, <i>i.e.</i> without reference to the function or any other variable(s). (It is customary, whenever practicable, for independent variables to appear on the right-hand-side of an equation and as the horizontal axis (abscissa) of a simple graph.)
Interpolation	The process of estimating the value of a function at a point within the range of values already known.
Least Squares	A criterion used to fit a function to a set of data (see Appendix A).
Number of terms (in a polynomial)	In this Item when referring to a polynomial, the number of terms minus one is synonymous with the degree of the polynomial, <i>e.g.</i> a <i>four-term</i> polynomial is <i>third-degree</i> , represented by terms up to and including a <i>third-order</i> term, <i>i.e.</i> a term involving x^3 .
Odd function	A function, $f(x)$, is odd if $f(x) = -f(-x)$.
Order	The order of a term within a polynomial describes the power to which the independent variable is raised, <i>e.g.</i> an equation of third-degree contains terms of first-, second- and third-order.
Overfitting	This is the term used when a multiple-term mathematical function fitted to a set of data tends to follow the irregularities resulting from possibly inaccurate individual points rather than the overall trends of the data.
Singularity	A point on an algebraic function where it is not differentiable although it is differentiable in the immediate neighbourhood of that point, <i>e.g.</i> the function $f(x) = 1/x$ on the real axis is infinity at the singularity, $x = 0$.
Underfitting	This is the term used when a multiple-term mathematical function fitted to a set of data has too few terms to enable it to follow satisfactorily the overall trends of the data.

2. INTRODUCTION

The purpose of this Item is to promote 'good practice' when choosing and fitting a function with a linear Least Squares error profile to represent a set of data. There are a host of spreadsheets, mathematical calculation software and dedicated plotting packages that enable the automatic fitting of a chosen function to a set of data. Care and judgement however must always be exercised when choosing the form of equation that is best suited to follow the overall trends of the data. The chosen equation can then be subsequently used to provide intermediate values between given data points and possibly also values outside the range of the data when required. Throughout this Item sketches are used to illustrate the relation between the set of data points to be fitted and the fitted equations.

The Least Squares criterion minimises the sum of the weighted[†] squares of the differences, or residuals, between the values of y , calculated at each of the values of x_i from the approximating function $f(x)$, and the corresponding data values y_i .

Curve fitting is a procedure for the fitting of algebraic functions to a set of known data points in order that the approximating function can be used in place of the data points for estimation of the dependent variable for all desired values of the independent variable. It is essentially a smoothing process in which the fitted curve follows the general trends of the data rather than passing through each individual point. It is not to be confused with the various types of interpolation procedure, such as splines, that produce curves constrained to pass through all the data points.

A companion Item, No. 01011^[1], presents a 'Compendium of curves and approximating algebraic functions' that may be used initially to choose the form of function required to represent a set of data. Once the function has been chosen then the Least Squares procedure is used to enable that choice of function to be ratified.

There are several situations in which it may be required to fit an equation to a set of data, some of which are as follows.

- An equation fitted to experimentally obtained data enables intermediate values to be calculated without the need for numerical interpolation of tabulated data. The function may also be required for obtaining values outside of the range of the data, within a limited and clearly defined range. Such an equation could be incorporated into a routine within a computer program.
- If physical considerations suggest that the data obey a particular form of equation the process of fitting that equation gives the values of its constants that provide the Least Squares fit to the data. This is often required for calibrating measuring instruments such as flow meters.
- With the use of a curve fitting procedure an approximating equation can be fitted to data calculated from a lengthy and complex numerical process. This equation could then be used elsewhere to shorten the calculation of further results.

Section 3 of this Item introduces the Least Squares method of curve fitting, the equations for which are derived in Appendix A. Four types of function, from the many available, are chosen to illustrate the curve fitting process; the power series polynomial function of a single independent variable, the logarithmic function, the rational function and the peak function. Section 3 also details five criteria for the quality of

[†] A weighting is the known, or in some way estimated, uncertainty of the value of a data point y , or y and x , compared with all other points in a data set, (x_i, y_i) , $i = 1$ to N . Thus a comparative uncertainty could be assigned to all values in a data set. An extreme example of a weighting, when a data point has been clearly identified as a 'rogue', and the wish is that this point does not form part of the fitting process, would be to assign it a weighting of zero. However extreme care should be exercised before data points are effectively eliminated from the fitting process.

fit of a function, introduces scaling of variables and presents notes relating to the preparation of data for curve fitting.

Section 4 presents examples to illustrate what is meant by underfitting and overfitting of functions to data sets and the fitting of functions to sparse data sets.

The curve fitting procedure is developed in further sections of this Item by illustration using a series of examples using data from aeronautical and other engineering applications. The examples include:

- the fitting of arbitrary equations to experimental data (**Section 5**),
- fitted equations that have physically significant terms (**Section 6**),
- specific functions to fit localised features of curves (**Section 7**).

Appendices A and B provide, respectively, a derivation of the Least Squares method of curve fitting and the application of orthogonal functions, in particular Legendre polynomials, to curve fitting.

3. LINEAR LEAST SQUARES METHOD OF CURVE FITTING

3.1 Basis of Method and its Application

The object of curve fitting is to find a function which defines a smooth approximating curve to a set of known data points, *e.g.* (x_i, y_i) , $i = 1$ to N . Here the dependent variable, y , is taken to be a function of a single independent variable, x . The values y_i may be measurements or observations made of a physical quantity at the value x_i , or previously calculated by a numerical procedure.

In linear curve fitting the approximating function, $f(x)$, is expressed as a linear combination of separate independent functions, each being multiplied by an arbitrary unknown numerical coefficient, thus

$$y \approx f(x) = C_1 F_1(x) + C_2 F_2(x) + C_3 F_3(x) + \dots + C_p F_p(x), \quad (3.1)$$

in which $F_1, F_2, \dots, F_p$ are known functions of x and $C_1, C_2, \dots, C_p$ are numerical coefficients whose values are to be found and which make $f(x)$ the best fit, in some defined way, to the data. The number of terms in Equation (3.1) is p and in this Item this is used to identify the polynomial. It is important to note that the approximating function can only be unique when the number of unknowns, p , is not larger than the number of known data points, N . The Least Squares algorithm of Appendix A identifies a unique fit to the data when $p \leq N$.

The overall requirement for the approximating function to be a good fit is that it should follow the general trends of the data, having a roughly similar shape to the set of data points when plotted. Thus the curve defined by the approximating function should pass as closely as possible to the majority of the data points. The closeness of fit, however, may be judged against more than one criterion.

The most commonly used criterion for the best fit and the one used in this Item is the Least Squares criterion in which the sum of the weighted squares of the differences, or residuals, between the values of y , calculated at each of the values of x_i from the approximating function $f(x)$, and the corresponding data values y_i , should be a minimum.

The Least Squares algorithm for calculating the values of the coefficients $C_1, C_2, \dots, C_p$ for a given set of data and a chosen function $f(x)$ is derived in Appendix A. In that analysis it is assumed that the dependent variable, y , is a function of a single independent variable, x . However the analysis may also be applied to a function of several variables $x_1, x_2, \dots, x_r$ and nothing in the analysis precludes the functions $F_1, F_2, \dots, F_p$ containing y itself. Further, the left-hand-side of Equation (3.1) can be a function of both y and x , and not just a function of y or x itself. Thus, Equation (3.1) may be written in the more general form given by

$$H(x_1, x_2, \dots, x_r, y) \approx C_1 F_1(x_1, x_2, \dots, x_r, y) + \dots + C_p F_p(x_1, x_2, \dots, x_r, y). \quad (3.2)$$

When Equation (3.2), the generalised form of Equation (3.1), is used, it is the sum of the weighted squares of the residuals calculated from the function H that is minimised, rather than the sum calculated from y itself. Equation (3.2) forms the basis for many different types of function which can be reduced to this general form and which are therefore amenable to the Least Squares method of curve fitting. Examples of types of function applicable to this method are given in Section 3.2.

3.2 Examples of Types of Function Used to Illustrate Curve Fitting

The examples of the functions included in this section are selected from those described in a compendium of curves given in Data Item No. 01011^[1] and are used in this Item to illustrate the curve fitting procedure.

The most common function used for curve fitting is the polynomial function which, together with the rational function, forms the basis of the fitting process described. With polynomial and rational functions in powers of x it is not necessary to include all the powers of x up to the highest used. If the curve is required to fit data that are known to be representable by an even function, *i.e.* $f(x) = f(-x)$, then it is only necessary to use even powers of x in the approximating function. In fact it may be detrimental to the accuracy of the fit if odd powers are included. Similarly, if data are known to be representable by an odd function, *i.e.* $f(x) = -f(-x)$, then only odd powers of x should be included. All of the examples featured in this Item, however, are fitted with polynomials containing all the powers of x up to the highest used, *i.e.* the degree of a polynomial is $p-1$ when the number of terms in the polynomial is p .

Other functions can be employed at the discretion of the user to suit a particular set of data. For example the powers of x in the polynomial and rational forms may be replaced by more complicated expressions such as those used in the orthogonal Legendre Polynomials. A series of exponential, trigonometric or hyperbolic functions may also be used. However, in general the use of over-complicated functions should be avoided as they can introduce extraneous features such as waves and discontinuities that are not inherent in the data.

3.2.1 Polynomial function of single independent variable

Polynomial functions in positive integer powers of a single independent variable, x , may be expressed in the form

$$y(x) = C_1 + C_2x + C_3x^2 + \dots + C_px^{p-1}, \quad (3.3)$$

which, when compared with Equation (3.2) gives

$$H(x, y) \equiv y; \quad F_1 \equiv 1; \quad F_2 \equiv x; \quad F_3 \equiv x^2; \quad \dots; \quad F_p \equiv x^{p-1}. \quad (3.4)$$

The Least Squares fit by polynomials with a small number of terms to 'well-behaved' data, *i.e.* data in which the range of slopes from data point to data point are not too large, are usually extremely good. However the behaviour of all polynomial functions of the type of Equations (3.1) or (3.3) are ultimately dominated at sufficiently high values of the independent variable, x , by the term having the highest power, and so all polynomials will approach infinity as x approaches infinity, positive or negative. Polynomials with terms of high-order tend to rapidly diverge from data to which they are fitted outside the range of the data and so extreme care should be exercised in allowing the prediction of extrapolated values, *i.e.* values outside of the range of the data. Polynomials with a large number of terms should only be used for fitting data in those rare cases when there is an overriding reason for the function used, say a knowledge of the physical behaviour of a property of the data, and when tight control is exercised over the range of the independent variable, x , used for estimation purposes.

3.2.2 Rational function

In its simplest form a rational function can be the ratio of two polynomials in powers of the independent variable, x , of number of terms p and q , given by

$$y(x) = \frac{C_1 + C_2x + C_3x^2 + \dots + C_px^{p-1}}{1 + C_{p+1}x + C_{p+2}x^2 + \dots + C_{p+q-1}x^{q-1}}. \quad (3.5)$$

This last equation may be expressed in a linear form similar to Equation (3.2) by cross-multiplying and rearranging to the form

$$y(x) = C_1 + C_2x + C_3x^2 + \dots + C_px^{p-1} - \left(C_{p+1}x y(x) + C_{p+2}x^2 y(x) + \dots + C_{p+q-1}x^{q-1}y(x) \right) . \quad (3.6)$$

Comparison of Equations (3.2) and (3.6) gives

$$\left. \begin{aligned} H(x, y) &\equiv y ; F_1 \equiv 1 ; F_2 \equiv x ; F_3 \equiv x^2 ; \dots ; F_p \equiv x^{p-1} \text{ and} \\ F_{p+1} &\equiv -x y(x) ; F_{p+2} \equiv -x^2 y(x) ; \dots ; F_{p+q-1} \equiv -x^{q-1}y(x) . \end{aligned} \right\} \quad (3.7)$$

In this linear form the functions $F_{p+1}, \dots, F_{p+q-1}$ are seen to be functions of y as well as x .

For very large values of x , positive or negative, the behaviour of the numerator and denominator of the rational function is dominated by the terms having the highest power of x . Thus for a rational function when $p = q$, its asymptotic behaviour, *i.e.* the constant value of the function that is approached at very large x (see Data Item No. 01011^[1]), is given by

$$\left(y(x)_{p=q} \right)_{x \rightarrow \pm\infty} \text{ approaches } \frac{C_p}{C_{p+q-1}} .$$

However a rational function with $p = q$ can also, in many cases, approach an asymptotic value at low to moderate values of x and this is a useful feature of such functions.

The rational function may also be used for cases when the data show singularities, such as when y tends to infinity at some point in the data range. Such behaviour can be reproduced by zero values in the denominator of a rational function but cannot be obtained from a polynomial in positive powers of x .

3.2.3 Peak function

The peak function may be expressed in the form

$$F_p(x) = \frac{C}{1 + B \left(\frac{x - K}{\Delta x} \right)^2} , \quad (3.8)$$

in which C is the magnitude of the peak value which is unknown and found by the Least Squares procedure, K is the value of the independent variable, x , at which the peak occurs and B controls the sharpness of the peak in terms of the interval, Δx , the higher the value of B the quicker the function decreases in value from its peak value. The values of B , K and Δx are known values, *i.e.* chosen, independently of the Least Squares procedure, to fit the data. A negative value for C inverts the peak into a trough, *i.e.* the minimum value of the function is C . The function asymptotically approaches zero at both positive and negative x .

Peak functions may be used in isolation but are best used in conjunction with other functions that can follow the background trends of the data.

3.2.4 Exponential function

An exponential function may be expressed in the form

$$y(x) = Ax^B. \quad (3.9)$$

This is an example of a function that is nonlinear with regard to its unknown coefficients but can be easily transformed to a linear relationship. Taking natural logarithms (the logarithm to any base could alternatively be used) of both sides of Equation (3.9) results in a linear form of the equation given by

$$\ln y(x) = \ln A + B \ln x. \quad (3.10)$$

Comparison of Equations (3.2) and (3.10) gives

$$\left. \begin{aligned} H(x, y) &\equiv \ln y(x); \quad F_1 \equiv 1; \quad F_2 \equiv \ln x \text{ and} \\ C_1 &\equiv \ln A; \quad C_2 \equiv B. \end{aligned} \right\} \quad (3.11)$$

In this example $H(x, y)$ is a function of y rather than y itself.

The values of the coefficients C_1 and C_2 which give the Least Squares fit to the data points are found from the procedure of Appendix A. The required values of the constants A and B are then found from Equation (3.6). It is noted, though, as described in Section 3.1, that it is the sum of the residuals of $\ln y(x)$ that is minimised and not the sum of the original function y .

3.3 Criteria for Quality of Fit of Function

Criteria for the quality of fit of a function to a set of data are as follows.

- Once an approximating function to a set of data is chosen, the Least Squares fit to the data is obtained when the sum of the weighted squares of the differences between the fitted function and the data to be fitted, *i.e.* the sum of the residuals, S , is a minimum. This Least Squares procedure for a given function, derived in Appendix A, is interpreted in this Item as the "best" fit.
- The curve generated by a chosen function should follow the general trends of the plotted data points smoothly, without introducing features such as kinks or waves between data points.
- Even if a curve passes through all the data points with $S = 0$, this does not automatically mean that the quality of fit is acceptable. For example, a polynomial with a high number of terms, p , fitted to a data set of size N , in which $N = p$, would pass exactly through all the data points and so give the sum of the residuals, S , of zero, but would undoubtedly not be a satisfactory fit at intermediate values between the given data points. This is because the curve would resemble a series of waves passing through the points and oscillating between them. This phenomena is known as overfitting (see the example in Section 4.1).
- The degree of overfitting could be limited by introducing a further criterion whereby the length of each arc of the curve between adjacent points is minimised, provided that no kinks or other discontinuities are introduced. No quantitative formulation of such a criterion has yet been found however, and recourse must therefore be made to a qualitative judgement of any curve derived from the Least Squares method.

- It is desirable, though by no means essential, that the function should permit a limited degree of extrapolation beyond the range of the data without diverging too abruptly from the general trend. This tends to occur with polynomials in powers of the independent variable which have many terms, but is less likely with rational functions.

3.4 Scaling of Data

In some instances the original data may have large numerical values or, once the fitting process has been applied to a set of data, there may be relatively large differences in magnitude between the values of the coefficients, $C_1, C_2, \dots, C_p$, of the fitting polynomial. In either case, when all the data are applied to the functions, $F_1, F_2, \dots, F_p$, in the curve fitting process some of the elements in the matrix $[A]$ in Equation (A1.7) may become extremely large. This can lead to numerical overflows and loss of significant figures in the computation of the required coefficients, $C_1, C_2, \dots, C_p$. In such cases it may be beneficial to scale the independent variable, x , in order to prevent these rounding errors. Conversely, the data values may be numerically small, in which case similar errors result.

In order to keep the range of magnitudes of the coefficients of the fitting polynomial as small as possible the data values actually used in the computations should be of the order of unity. This may be achieved by dividing, or scaling, the data by a suitable reference value, such as by the greatest value in the data set itself or else by an arbitrary number.

The procedure for scaling is as follows. Suppose that the values of the independent variable, x , lie within the range x_L to x_U , i.e. $x_L < x < x_U$, and x is to be transformed to another variable x' whose values lie within the range x'_L to x'_U , i.e. $x'_L < x' < x'_U$, then a linear transformation is given by

$$x' = \frac{(x - x_0)}{F_{sx}}, \quad (3.12)$$

in which x_0 is an offset value of x and F_{sx} is a scaling factor. The required values of x_0 and F_{sx} to give the transformation are

$$\left. \begin{aligned} x_0 &= \left(\frac{x_L x'_U - x_U x'_L}{x'_U - x'_L} \right) \text{ and} \\ F_{sx} &= \left(\frac{x_U - x_L}{x'_U - x'_L} \right) . \end{aligned} \right\} \quad (3.13)$$

As an example consider that a set of values has an independent variable, x , in the range $55 < x < 125$, which is to be transformed into x' in the range $0 < x' < 1$, then Equations (3.13) give

$$\begin{aligned} x_0 &= (55 \times 1 - 125 \times 0) / (1 - 0) = 55 \text{ and} \\ F_{sx} &= (125 - 55) / (1 - 0) = 70 . \end{aligned}$$

Substitution of these values into Equation (3.12) gives the required transformation $x' = (x - 55)/70$.

It is sometimes advantageous to transform a given set of data into another set lying within the range -1 to $+1$, such as when using Legendre Polynomials (see Appendix B). To transform the data set defined by the range, $55 < x < 125$, into the range, $-1 < x' < 1$, then use of the procedure of Equations (3.12) and (3.13) leads to the transformation $x' = (x - 90)/35$.

Note that it is not necessary for the upper and lower ends of a range of data to correspond exactly with the highest and lowest numerical values of the data. It is often more convenient to choose round numbers for the ends of a range which encloses the highest and lowest of the data values.

In some cases it may be necessary to transform also the dependent variable y , say into the variable y' , having a more convenient range. This may be carried out using the linear transformation $y' = (y - y_0)/F_{sy}$, in which the values of y_0 and F_{sy} are given by equations similar to Equations (3.13). The transformed data pairs, x' and y' , are used in the curve fitting procedure to obtain the values of the coefficients, $C_1, C_2, \dots, C_p$. Then to evaluate the values of y the reverse transformation is used, i.e.

$$y = y'F_{sy} + y_0.$$

3.5 Preparation of Data for Curve Fitting

The first step should always be to plot the raw data using suitable axes, and observe the trends. A smooth freehand curve drawn by eye through the points may be of help in establishing any trends. However this may not be possible if the data show considerable scatter and appear as a cloud of points, rather than falling reasonably well onto a single line. Assuming that the data are reliable and that the scatter is not due to gross observation or instrument errors, a possible cause of scatter is that the readings were obtained under different ambient conditions such as pressure or temperature.

Attempts should be made to collapse the data points onto a single curve by re-plotting them against a suitable dimensionless parameter. Alternatively the raw data might be brought onto a single line by adjusting their values with non-dimensional parameters suggested by physical considerations, combined with dimensional analysis. It may also be advantageous to scale the data numerically in order to reduce the possibility of overflows or underflows in the computations and subsequent loss of significant figures.

The final step in preparing the data, if a level of comparative uncertainty is known, is to assign a weighting to each of the data pairs in a set of data, (x_i, y_i) , $i = 1$ to N , as outlined in Appendix A. However, the choice of weighting depends entirely on known levels of comparative uncertainty in the data and no general guidance is given in this Item. Accordingly, in the numerical examples in this Item all the weightings are taken as unity.

4. UNDERFITTING, OVERFITTING AND SPARSE DATA SETS

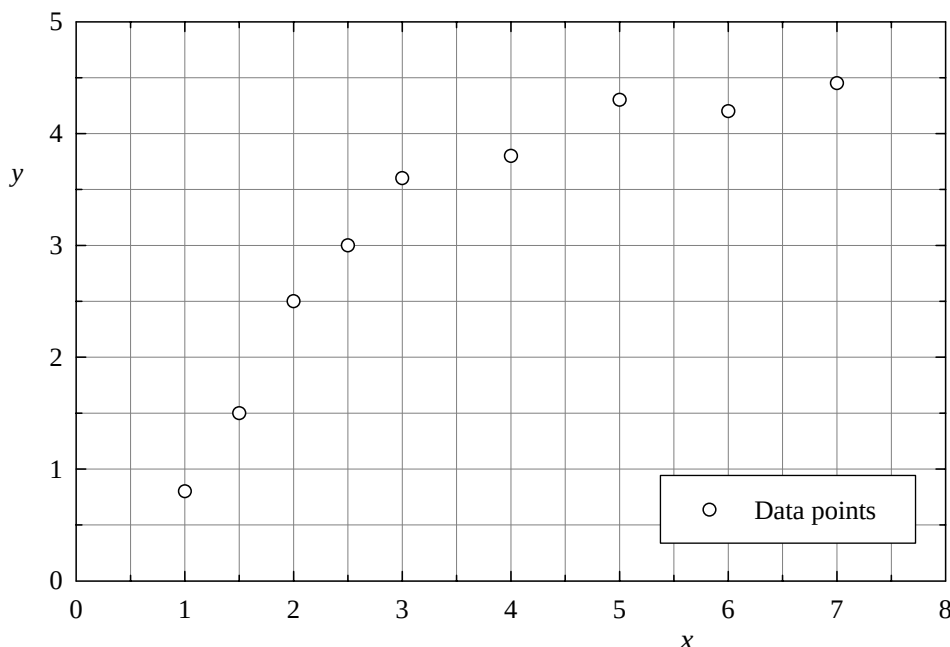
In this Item the terms ‘underfitting’ and ‘overfitting’ refer to cases when the function chosen to represent a set of data is judged to be inadequate in terms of the Least Squares fit that results. The former term is used when a multi-term function has too few terms to enable it to follow adequately the trends of the data. The latter is used when the chosen multi-term function, even though it is considered to have a sufficiency in the number of terms included in the function, still does not adequately represent a fit to the data. An overfitted function tends to follow the individual points in a data set rather than the overall trends. It is stressed that an inadequate fit of a function to a set of data is entirely subjective, both in terms of what is required as a representation of the data set and the level of differences between the points and function that is deemed to be acceptable.

Overfitting should not be confused with the problems associated with the fitting of sparse data sets. An overriding requirement of any Least Squares fit is that the number of data points, as a minimum, must equal the number of unknown coefficients in a chosen function.

Sections 4.1 and 4.2 describe the phenomena of underfitting, overfitting and the problems associated with sparse data sets through examples in which a set of data points have a degree of scatter about a envisaged smooth curve fit to the data.

4.1 Underfitting and Overfitting

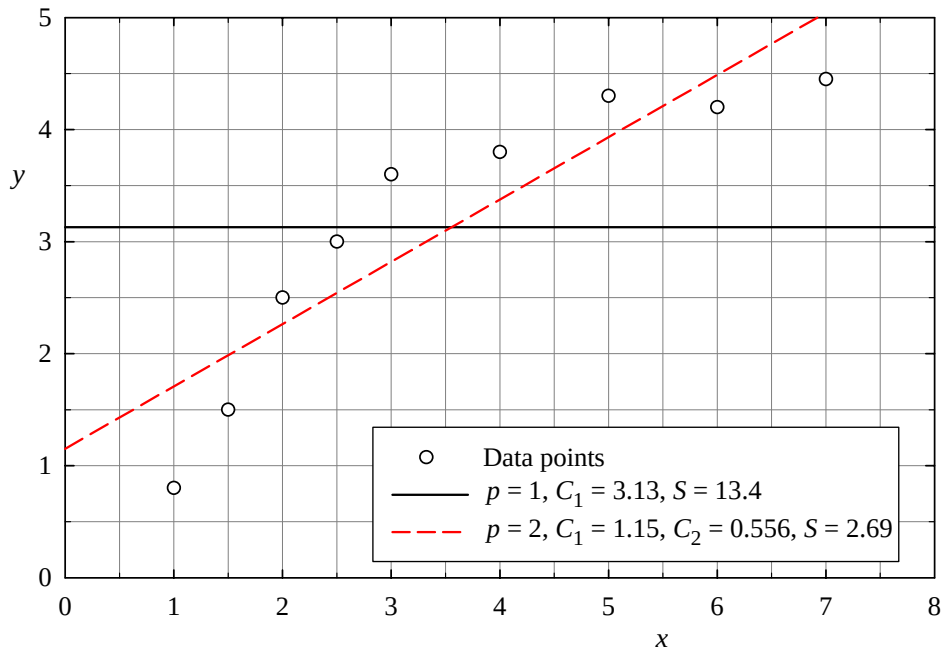
Consider the data set shown in Sketch 4.1. The trend of the data is seen to lie on a curved line increasing in value, y , as the independent variable, x , increases.



Sketch 4.1 Data with assumed measurement error

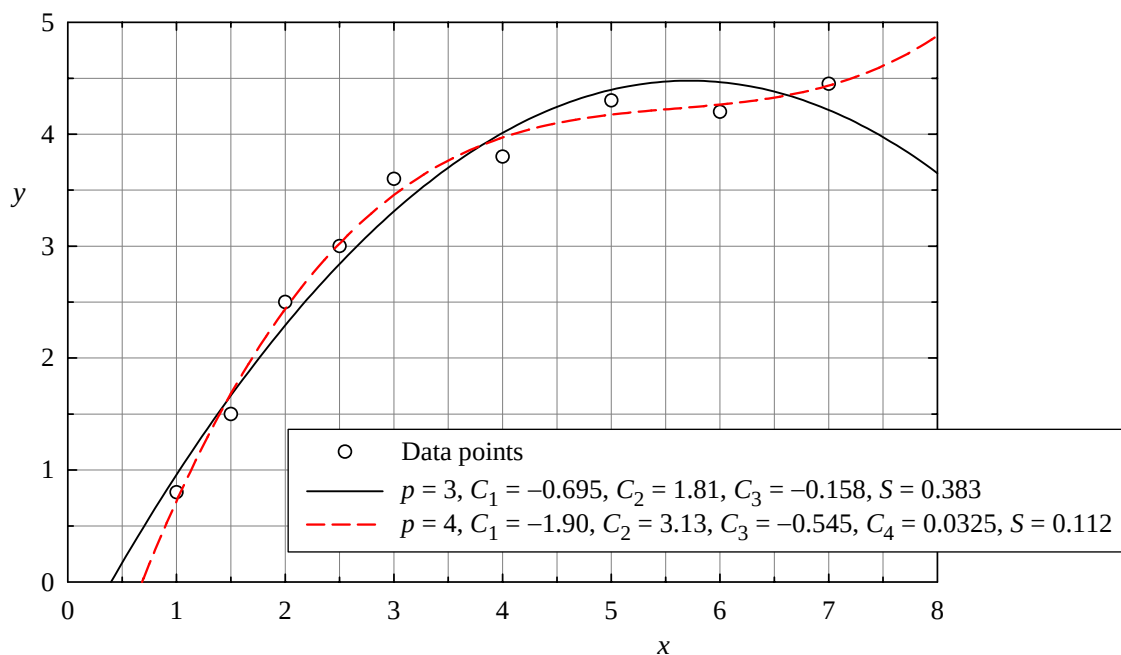
The fitting process is assessed by fitting a series of polynomials in powers of the independent variable of the form of Equation (3.3).

Fitting a line of zero slope, $y = C_1$ ($p = 1$), and one of constant slope, $y = C_1 + C_2x$ ($p = 2$), to the data results in the lines shown in Sketch 4.2. Either of these line fits could be judged as adequate by the intended user, but in terms of following the apparent trend of the data the fits are inadequate. If the fits are deemed unacceptable then this illustrates underfitting, which occurs whenever the fitted equation is judged to have too few terms to enable it to follow adequately the underlying curvature of the data.



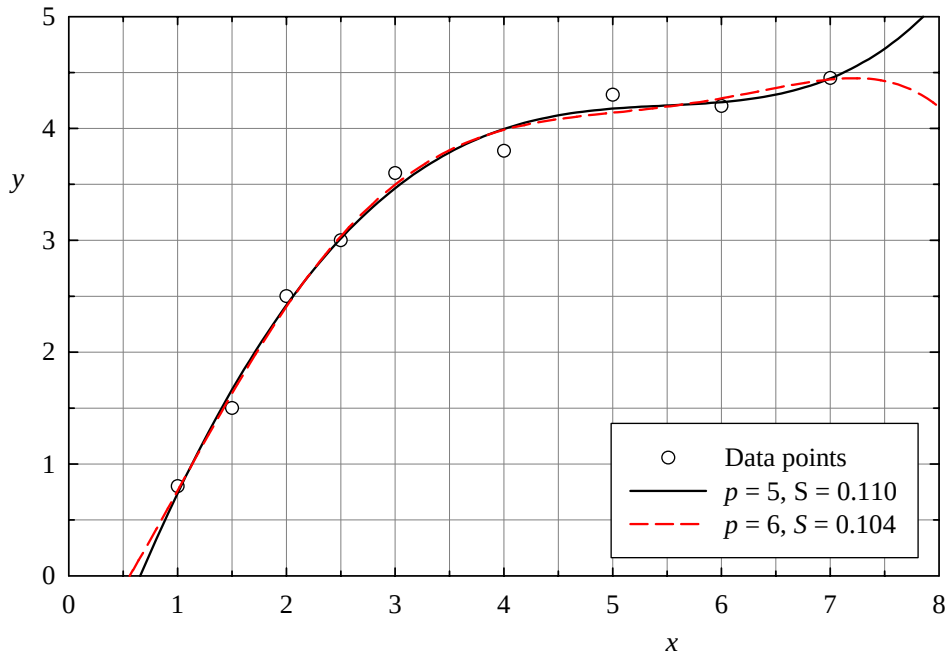
Sketch 4.2 Data with zero slope and constant slope polynomial fits

Improved fits are seen to be obtained with polynomials having three ($p = 3$) and four terms ($p = 4$), shown in Sketch 4.3, both of which have the capability of following the curvature of the data. The four-term polynomial is seen to be superior since it has the ability to, and attempts to, follow a possible point of inflexion at $x \approx 6$. The sum of the squared residuals, S , for the four-term polynomial compared with the three-term polynomial has reduced to 0.112 from 0.383.



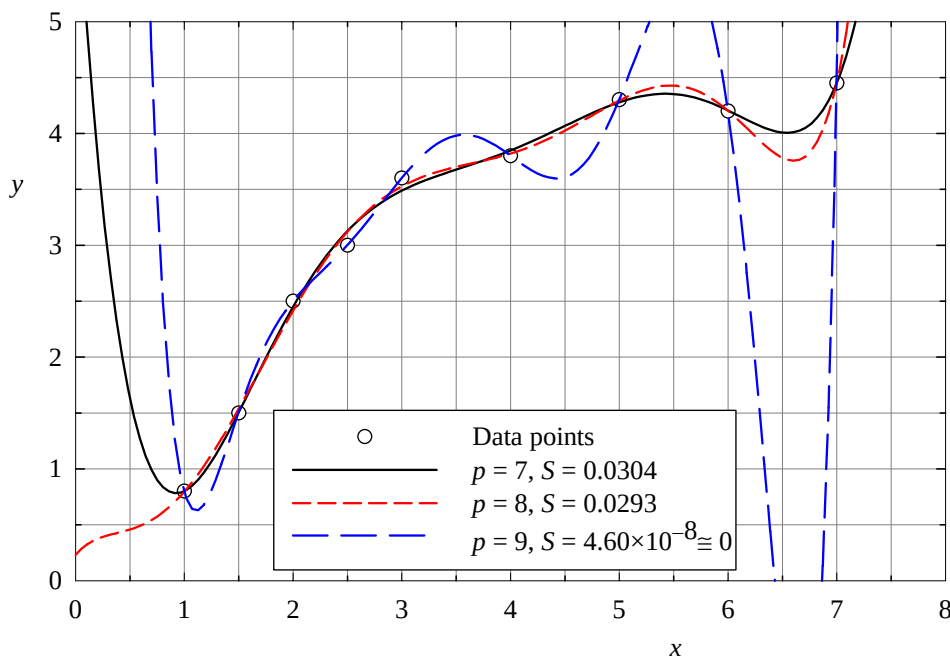
Sketch 4.3 Data with three-term and four-term polynomial fits

A further increase in the number of terms of the polynomial does improve fit from that of the four-term, but only marginally. Sketch 4.4 shows the fits obtained with the five-term and six-term polynomials, for which the values of S have decreased to 0.110 and 0.104, respectively.



Sketch 4.4 Data with five- and six-term polynomial fits

In Sketch 4.5 plots of seven-, eight- and nine-term polynomial fits are shown. The plots are superior to the earlier curves in the sense that they lie on average closer to the points as verified by the values of S obtained, 0.0304, 0.0293 and 4.6×10^{-8} , respectively. However, the curves are obviously not satisfactory and are seen to oscillate between the points, more and more violently as the number of terms is increased. In fact the nine-term polynomial between the last two points is seen to descend to a value below $y = 0$.



Sketch 4.5 Data with seven-, eight- and nine-term polynomial fits illustrating overfitting