Linear least squares (mathematics)

In statistics and mathematics, **linear least squares** is an approach to fitting a mathematical or statistical model to data in cases where the idealized value provided by the model for any data point is expressed linearly in terms of the unknown parameters of the model. The resulting fitted model can be used to summarize the data, to predict unobserved values from the same system, and to understand the mechanisms that may underlie the system.

Mathematically, linear least squares is the problem of approximately solving an overdetermined system of linear equations, where the best approximation is defined as that which minimizes the sum of squared differences between the data values and their corresponding modeled values. The approach is called *linear* least squares since the assumed function is linear in the parameters to be estimated. Linear least squares problems are convex and have a closed-form solution that is unique, provided that the number of data points used for fitting equals or exceeds the number of unknown parameters, except in special degenerate situations. In contrast, non-linear least squares problems generally must be solved by an iterative procedure, and the problems can be non-convex with multiple optima for the objective function. If prior distributions are available, then even an underdetermined system can be solved using the Bayesian MMSE estimator.

In statistics, linear least squares problems correspond to a particularly important type of statistical model called linear regression which arises as a particular form of regression analysis. One basic form of such a model is an ordinary least squares model. The present article concentrates on the mathematical aspects of linear least squares problems, with discussion of the formulation and interpretation of statistical regression models and statistical inferences related to these being dealt with in the articles just mentioned. See Outline of regression analysis for an outline of the topic.

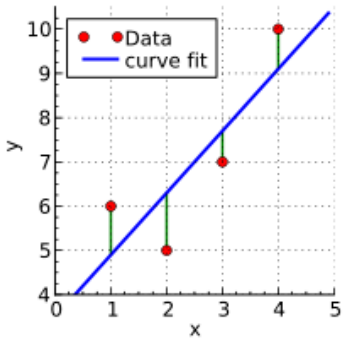
Contents

Example

As a result of an experiment, four (x, y) data points were obtained, $(1, 6)$, $(2, 5)$, $(3, 7)$, and $(4, 10)$ (shown in red in the diagram on the right). We hope to find a line $y = \beta_1 + \beta_2 x$ that best fits these four points. In other words, we would like to find the numbers β_1 and β_2 that approximately solve the overdetermined linear system

$$\begin{aligned}
\beta_1 + 1\beta_2 &= 6 \\
\beta_1 + 2\beta_2 &= 5 \\
\beta_1 + 3\beta_2 &= 7 \\
\beta_1 + 4\beta_2 &= 10
\end{aligned}$$

of four equations in two unknowns in some "best" sense.



A plot of the data points (in red), the least squares line of best fit (in blue), and the residuals (in green).

The "error", at each point, between the curve fit and the data is the difference between the right- and left-hand sides of the equations above. The least squares approach to solving this problem is to try to make the sum of the squares of these errors as small as possible; that is, to find the minimum of the function

$$\begin{aligned}
S(\beta_1, \beta_2) = &[6 - (\beta_1 + 1\beta_2)]^2 + [5 - (\beta_1 + 2\beta_2)]^2 \\
&+ [7 - (\beta_1 + 3\beta_2)]^2 + [10 - (\beta_1 + 4\beta_2)]^2 \\
&= 4\beta_1^2 + 30\beta_2^2 + 20\beta_1\beta_2 - 56\beta_1 - 154\beta_2 + 210.
\end{aligned}$$

The minimum is determined by calculating the partial derivatives of $S(\beta_1, \beta_2)$ with respect to β_1 and β_2 and setting them to zero

$$\begin{aligned}
\frac{\partial S}{\partial \beta_1} &= 0 = 8\beta_1 + 20\beta_2 - 56 \\
\frac{\partial S}{\partial \beta_2} &= 0 = 20\beta_1 + 60\beta_2 - 154.
\end{aligned}$$

This results in a system of two equations in two unknowns, called the normal equations, which when solved give

$$\begin{aligned}
\beta_1 &= 3.5 \\
\beta_2 &= 1.4
\end{aligned}$$

and the equation $y = 3.5 + 1.4x$ of the line of best fit. The residuals, that is, the discrepancies between the y values from the experiment and the y values calculated using the line of best fit, are then found to be 1.1, -1.3, -0.7, and 0.9 (see the diagram on the right). The minimum value of the sum of squares of the residuals is $S(3.5, 1.4) = 1.1^2 + (-1.3)^2 + (-0.7)^2 + 0.9^2 = 4.2$.

More generally, one can have n regressors x_j, and a linear model

$$y = \beta_1 + \sum_{j=2}^{n+1} \beta_j x_{j-1}.$$

Using a quadratic model

Importantly, in "linear least squares", we are not restricted to using a line as the model as in the above example. For instance, we could have chosen the restricted quadratic model $y = \beta_1 x^2$. This model is still linear in the β_1 parameter, so we can still perform the same analysis, constructing a system of equations from the data points:

$$\begin{aligned}
6 &= \beta_1 (1)^2 \\
5 &= \beta_1 (2)^2 \\
7 &= \beta_1 (3)^2 \\
10 &= \beta_1 (4)^2
\end{aligned}$$

The partial derivatives with respect to the parameters (this time there is only one) are again computed and set to 0:

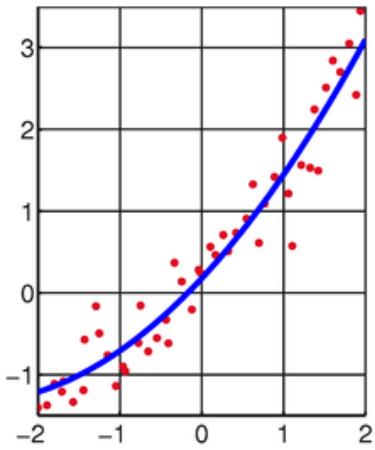
$$\frac{\partial S}{\partial \beta_1} = 0 = 708\beta_1 - 498$$

and solved

$$\beta_1 = 0.703$$

leading to the resulting best fit model $y = 0.703x^2$.

The general problem



The result of fitting a quadratic function $y = \beta_1 + \beta_2 x + \beta_3 x^2$ (in blue) through a set of data points (x_i, y_i) (in red). In linear least squares the function need not be linear in the argument x, but only in the parameters β_j that are determined to give the best fit.

Consider an overdetermined system

$$\sum_{j=1}^{n} X_{ij}\beta_j = y_i, \ (i = 1, 2, \ldots, m),$$

of *m* linear equations in *n* unknown coefficients, $\beta_1, \beta_2, \ldots, \beta_n$, with $m > n$. (Note: for a linear model as above, not all of $\boldsymbol{X}$ contains information on the data points. The first column is populated with ones, $\boldsymbol{X_{i1}} = \mathbf{1}$, only the other columns contain actual data, and *n* = number of regressors + 1.) This can be written in matrix form as

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y},$$

where

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Such a system usually has no solution, so the goal is instead to find the coefficients $\boldsymbol{\beta}$ which fit the equations "best",in the sense of solving thequadratic minimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}),$$

where the objective function*S* is given by

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} X_{ij}\beta_j \right|^2 = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2.$$

A justification for choosing this criterion is given in properties below. This minimization problem has a unique solution, provided that the *n* columns of the matrix $\mathbf{X}$ are linearly independent, given by solving thenormal equations

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

The matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is known as the Gramian matrix of $\mathbf{X}$, which possesses several nice properties such as being a positive semi-definite matrix, and the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is known as the moment matrix of regressand by regressors.[1] Finally, $\hat{\boldsymbol{\beta}}$ is the coefficient vector of the least-squareshyperplane, expressed as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Example implementation

MATLAB

The following MATLAB code shows implementation of this approach on the data used in the first example above.

```
% MATLAB code for finding the best fit line using least squares method
input = [...                % input in the form of matrix
    1, 6;...                % rows contain points
    2, 5;...
    3, 7;...
    4, 10];
m = length(input);              % number of points
X = [ones(m,1), input(:,1)];    % forming X of X beta = y
y = input(:,2);                 % forming y of X beta = y
betaHat = (X' * X) \ (X' * y);   % computing projection of matrix X on y, giving beta
% display best fit parameters
disp(betaHat);
% plot the best fit line
xx = linspace(0, 5, 2);
yy = betaHat(1) + betaHat(2)*xx;
plot(xx, yy)
% plot the points (data) for which we found the best fit
hold on
plot(input(:,1), input(:,2), 'or')
hold off
```

Python

Python 3.6 code using essentially the same variable naming as the MATLAB code above:

```
import numpy as np
import matplotlib.pyplot  as plt
data = np.array([
    [1, 6],
    [2, 5],
    [3, 7],
    [4, 10]
])
m = len(data)
X = np.array([np.ones(m), data[:, 0]]).T
y = np.array(data[:, 1]).reshape(-1, 1)
betaHat = np.linalg.solve(X.T.dot(X), X.T.dot(y))
print(betaHat)
plt.figure(1)
xx = np.linspace(0, 5, 2)
yy = np.array(betaHat[0] + betaHat[1] * xx)
plt.plot(xx, yy.T, color='b')
plt.scatter(data[:, 0], data[:, 1], color='r')
plt.show()
```

Julia (programming language)

```
using Plots
pyplot() #choose plotting backend
input = [
    1 6
    2 5
    3 7
    4 10]
m = size(input)[1]
X = [ones(m) input[:,1]]
y = input[:,2]
betaHat = X \ y #backslash computes LS-solution (X'X)\X'y (as in MATLAB)
print(betaHat)
plot(x->betaHat[2]*x + betaHat[1],0,5,label="curve fit")
scatter!(input[:,1],input[:,2],label="data")
```

R (programming language)

```
m <- 4
n <- 2
input <- matrix(c(1, 6, 2, 5, 3, 7, 4, 10), ncol = n, byrow = T)
k <- rep(1,m)
X <- cbind(k, input[,1])
y <- input[,2]
X.T <- t(X)
betaHat <- solve(X.T%*%X) %*% X.T %*%y
print(betaHat)
plot(input)
abline(betaHat[1], betaHat[2])
```

Derivation of the normal equations

Define the ith **residual** to be

$$r_i = y_i - \sum_{j=1}^{n} X_{ij}\beta_j.$$

Then S can be rewritten

$$S = \sum_{i=1}^{m} r_i^2.$$

Given that S is convex, it is minimized when its gradient vector is zero (This follows by definition: if the gradient vector is not zero, there is a direction in which we can move to minimize it further – see maxima and minima.) The elements of the gradient vector are the partial derivatives of S with respect to the parameters:

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{m} r_i \frac{\partial r_i}{\partial \beta_j} \qquad (j = 1, 2, \ldots, n).$$

The derivatives are

$$\frac{\partial r_i}{\partial \beta_j} = -X_{ij}.$$

Substitution of the expressions for the residuals and the derivatives into the gradient equations gives

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{m}\left(y_i - \sum_{k=1}^{n} X_{ik}\beta_k\right)(-X_{ij}) \qquad (j = 1, 2, \ldots, n).$$

Thus if $\hat{\boldsymbol{\beta}}$ minimizes S, we have

$$2\sum_{i=1}^{m}\left(y_i - \sum_{k=1}^{n} X_{ik}\hat{\beta}_k\right)(-X_{ij}) = 0 \qquad (j = 1, 2, \ldots, n).$$

Upon rearrangement, we obtain the **normal equations**

$$\sum_{i=1}^{m}\sum_{k=1}^{n} X_{ij}X_{ik}\hat{\beta}_k = \sum_{i=1}^{m} X_{ij}y_i \qquad (j = 1, 2, \ldots, n).$$

The normal equations are written in matrix notation as

$$(\mathbf{X}^\mathbf{T}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\mathbf{T}\mathbf{y}$$ (where X^T is the matrix transpose of X).

The solution of the normal equations yields the vector $\hat{\boldsymbol{\beta}}$ of the optimal parameter values.

Derivation directly in terms of matrices

The normal equations can be derived directly from a matrix representation of the problem as follows. The objective is to minimize

$$S(\boldsymbol{\beta}) = \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathbf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^\mathbf{T}\mathbf{y} - \boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y} - \mathbf{y}^\mathbf{T}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}\boldsymbol{\beta}.$$

Here $(\boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y})^\mathbf{T} = \mathbf{y}^\mathbf{T}\mathbf{X}\boldsymbol{\beta}$ has the dimension 1x1 (the number of columns of $\mathbf{y}$), so it is a scalar and equal to its own transpose, hence $\boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y} = \mathbf{y}^\mathbf{T}\mathbf{X}\boldsymbol{\beta}$ and the quantity to minimize becomes

$$S(\boldsymbol{\beta}) = \mathbf{y}^\mathbf{T}\mathbf{y} - 2\boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y} + \boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}\boldsymbol{\beta}.$$

Differentiating this with respect to $\boldsymbol{\beta}$ and equating to zero to satisfy the first-order conditions gives

$$-\mathbf{X}^\mathbf{T}\mathbf{y} + (\mathbf{X}^\mathbf{T}\mathbf{X})\boldsymbol{\beta} = 0,$$

which is equivalent to the above-given normal equations. A sufficient condition for satisfaction of the second-order conditions for a minimum is that $\mathbf{X}$ have full column rank, in which case $\mathbf{X}^\mathbf{T}\mathbf{X}$ is positive definite.

Derivation without calculus

When $\mathbf{X}^\mathbf{T}\mathbf{X}$ is positive definite, the formula for the minimizing value of $\boldsymbol{\beta}$ can be derived without the use of derivatives. The quantity

$$S(\boldsymbol{\beta}) = \mathbf{y}^\mathbf{T}\mathbf{y} - 2\boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{y} + \boldsymbol{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}\boldsymbol{\beta}$$

can be written as

$$\langle\boldsymbol{\beta}, \boldsymbol{\beta}\rangle - 2\langle\boldsymbol{\beta}, (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}\rangle + \langle(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}, (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}\rangle + C,$$

where C depends only on $\mathbf{y}$ and $\mathbf{X}$, and $\langle\cdot,\cdot\rangle$ is the inner product defined by

$$\langle x, y\rangle = x^\mathbf{T}(\mathbf{X}^\mathbf{T}\mathbf{X})y.$$

It follows that $S(\boldsymbol{\beta})$ is equal to

$$\langle\boldsymbol{\beta} - (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}, \boldsymbol{\beta} - (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}\rangle + C$$

and therefore minimized exactly when

$$\boldsymbol{\beta} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = 0.$$

Generalization for complex equations

In general, the coefficients of the matrices $\mathbf{X}, \boldsymbol{\beta}$ and $\mathbf{y}$ can be complex. By using a Hermitian transpose instead of a simple transpose, it is possible to find a vector $\hat{\boldsymbol{\beta}}$ which minimizes $S(\boldsymbol{\beta})$, just as for the real matrix case. In order to get the normal equations we follow a similar path as in previous derivations:

$$S(\boldsymbol{\beta}) = \langle \mathbf{y} - \mathbf{X}\boldsymbol{\beta}, \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle - \overline{\langle \mathbf{X}\boldsymbol{\beta}, \mathbf{y} \rangle} - \overline{\langle \mathbf{y}, \mathbf{X}\boldsymbol{\beta} \rangle} + \langle \mathbf{X}\boldsymbol{\beta}, \mathbf{X}\boldsymbol{\beta} \rangle = \mathbf{y}^{\mathrm{T}}\overline{\mathbf{y}} - \boldsymbol{\beta}^{\dagger}\mathbf{X}^{\dagger}\mathbf{y} - \mathbf{y}^{\dagger}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\overline{\mathbf{X}\boldsymbol{\beta}},$$

where † stands for Hermitian transpose.

We should now take derivatives of $S(\boldsymbol{\beta})$ with respect to each of the coefficients β_j, but first we separate real and imaginary parts to deal with the conjugate factors in above expression. For the β_j we have

$$\beta_j = \beta_j^R + i\beta_j^I$$

and the derivatives change into

$$\frac{\partial S}{\partial \beta_j} = \frac{\partial S}{\partial \beta_j^R}\frac{\partial \beta_j^R}{\partial \beta_j} + \frac{\partial S}{\partial \beta_j^I}\frac{\partial \beta_j^I}{\partial \beta_j} = \frac{\partial S}{\partial \beta_j^R} - i\frac{\partial S}{\partial \beta_j^I} \quad (j = 1, 2, 3, \ldots, n).$$

After rewriting $S(\boldsymbol{\beta})$ in the summation form and writing β_j explicitly, we can calculate both partial derivatives with result:

$$\frac{\partial S}{\partial \beta_j^R} = -\sum_{i=1}^{m}\left(\overline{X}_{ij}y_i + \overline{y}_i X_{ij}\right) + 2\sum_{i=1}^{m} X_{ij}\overline{X}_{ij}\beta_j^R + \sum_{i=1}^{m}\sum_{k\neq j}^{n}\left(X_{ij}\overline{X}_{ik}\overline{\beta}_k + \beta_k X_{ik}\overline{X}_{ij}\right),$$

$$-i\frac{\partial S}{\partial \beta_j^I} = \sum_{i=1}^{m}\left(\overline{X}_{ij}y_i - \overline{y}_i X_{ij}\right) - 2i\sum_{i=1}^{m} X_{ij}\overline{X}_{ij}\beta_j^I + \sum_{i=1}^{m}\sum_{k\neq j}^{n}\left(X_{ij}\overline{X}_{ik}\overline{\beta}_k - \beta_k X_{ik}\overline{X}_{ij}\right),$$

which, after adding it together and comparing to zero (minimization condition for $\hat{\boldsymbol{\beta}}$) yields

$$\sum_{i=1}^{m} X_{ij}\overline{y}_i = \sum_{i=1}^{m}\sum_{k=1}^{n} X_{ij}\overline{X}_{ik}\overline{\hat{\beta}}_k \quad (j = 1, 2, 3, \ldots, n).$$

In matrix form:

$$\mathbf{X}^{\mathrm{T}}\overline{\mathbf{y}} = \mathbf{X}^{\mathrm{T}}\overline{\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right)} \quad \text{or} \quad \left(\mathbf{X}^{\dagger}\mathbf{X}\right)\hat{\boldsymbol{\beta}} = \mathbf{X}^{\dagger}\mathbf{y}.$$

Computation

A general approach to the least squares problem $\min \|\mathbf{y} - X\boldsymbol{\beta}\|^2$ can be described as follows. Suppose that we can find an *n* by *m* matrix **S** such that **XS** is an orthogonal projection onto the image of **X**. Then a solution to our minimization problem is given by

$$\boldsymbol{\beta} = S\mathbf{y}$$

simply because

$$X\boldsymbol{\beta} = X(S\mathbf{y}) = (XS)\mathbf{y}$$

is exactly a sought for orthogonal projection of $\mathbf{y}$ onto an image of **X** (see the picture below and note that as explained in the next section the image of **X** is just a subspace generated by column vectors of **X**). A few popular ways to find such a matrix *S* are described below

Inverting the matrix of the normal equations

The algebraic solution of the normal equations with a full-rank matrix $X^{\mathrm{T}}X$ can be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{+}\mathbf{y}$$

where X^+ is the Moore–Penrose pseudoinverse of X. Although this equation is correct and can work in many applications, it is not computationally efficient to invert the normal-equations matrix (the Gramian matrix). An exception occurs in numerical smoothing and differentiation where an analytical expression is required.

If the matrix $X^{\mathrm{T}}X$ is well-conditioned and positive definite, implying that it has full rank, the normal equations can be solved directly by using the Cholesky decomposition $R^{\mathrm{T}}R$, where R is an upper triangular matrix, giving:

$$R^{\mathrm{T}}R\hat{\boldsymbol{\beta}} = X^{\mathrm{T}}\mathbf{y}.$$

The solution is obtained in two stages, a forward substitution step, solving for $\mathbf{z}$:

$$R^{\mathrm{T}}\mathbf{z} = X^{\mathrm{T}}\mathbf{y},$$

followed by a backward substitution, solving for $\hat{\boldsymbol{\beta}}$:

$$R\hat{\boldsymbol{\beta}} = \mathbf{z}.$$

Both substitutions are facilitated by the triangular nature of R.

Orthogonal decomposition methods

Orthogonal decomposition methods of solving the least squares problem are slower than the normal equations method but are more numerically stable because they avoid forming the product $X^{\mathrm{T}}X$.

The residuals are written in matrix notation as

$$\mathbf{r} = \mathbf{y} - X\hat{\boldsymbol{\beta}}.$$

The matrix X is subjected to an orthogonal decomposition, e.g., the QR decomposition as follows.

$$X = Q\begin{pmatrix} R \\ 0 \end{pmatrix},$$

where Q is an $m\times m$ orthogonal matrix ($Q^{\mathrm{T}}Q=I$) and R is an $n\times n$ upper triangular matrix with $r_{ii} > 0$.

The residual vector is left-multiplied by Q^{T}.

$$Q^{\mathrm{T}}\mathbf{r} = Q^{\mathrm{T}}\mathbf{y} - \left(Q^{\mathrm{T}}Q\right)\begin{pmatrix} R \\ 0 \end{pmatrix}\hat{\boldsymbol{\beta}} = \begin{bmatrix} \left(Q^{\mathrm{T}}\mathbf{y}\right)_n - R\hat{\boldsymbol{\beta}} \\ \left(Q^{\mathrm{T}}\mathbf{y}\right)_{m-n} \end{bmatrix} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$$

Because Q is orthogonal, the sum of squares of the residuals, s, may be written as:

$$s = \|\mathbf{r}\|^2 = \mathbf{r}^{\mathrm{T}}\mathbf{r} = \mathbf{r}^{\mathrm{T}}QQ^{\mathrm{T}}\mathbf{r} = \mathbf{u}^{\mathrm{T}}\mathbf{u} + \mathbf{v}^{\mathrm{T}}\mathbf{v}$$

Since $\mathbf{v}$ doesn't depend on $\boldsymbol{\beta}$, the minimum value of s is attained when the upper block, $\mathbf{u}$, is zero. Therefore, the parameters are found by solving:

$$R\hat{\boldsymbol{\beta}} = \left(Q^{\mathrm{T}}\mathbf{y}\right)_n.$$

These equations are easily solved as R is upper triangular.

An alternative decomposition of X is the singular value decomposition (SVD)[2]

$$X = U\Sigma V^{\mathrm{T}},$$

where U is m by m orthogonal matrix, V is n by n orthogonal matrix and $\boldsymbol{\Sigma}$ is an m by n matrix with all its elements outside of the main diagonal equal to *0*. The pseudoinverse of $\boldsymbol{\Sigma}$ is easily obtained by inverting its non-zero diagonal elements and transposing. Hence,

$$\mathbf{X}\mathbf{X}^{+} = U\Sigma V^{\mathrm{T}}V\Sigma^{+}U^{\mathrm{T}} = UPU^{\mathrm{T}},$$

where P is obtained from $\boldsymbol{\Sigma}$ by replacing its non-zero diagonal elements with ones. Since $\left(\mathbf{X}\mathbf{X}^{+}\right)^{*} = \mathbf{X}\mathbf{X}^{+}$ (the property of pseudoinverse), the matrix UPU^{T} is an orthogonal projection onto the image (column-space) of X. In accordance with a general approach described in the introduction above (find $\mathbf{XS}$ which is an orthogonal projection),

$$S = \mathbf{X}^{+},$$

and thus,

$$\beta = V\Sigma^{+}U^{\mathrm{T}}\mathbf{y}$$

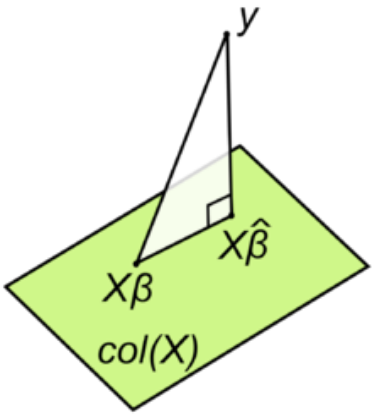
is a solution of a least squares problem. This method is the most computationally intensive, but is particularly useful if the normal equations matrix, $X^{\mathrm{T}}X$, is very ill-conditioned (i.e. if its condition number multiplied by the machine's relative round-off error is appreciably large). In that case, including the smallest singular values in the inversion merely adds numerical noise to the solution. This can be cured with the truncated SVD approach, giving a more stable and exact answer, by explicitly setting to zero all singular values below a certain threshold and so ignoring them, a process closely related to factor analysis.

Properties of the least-squares estimators

The gradient equations at the minimum can be written as

$$(\mathbf{y} - X\hat{\boldsymbol{\beta}})^{\mathrm{T}} X = 0.$$

A geometrical interpretation of these equations is that the vector of residuals, $\mathbf{y} - X\hat{\boldsymbol{\beta}}$ is orthogonal to the column space of X, since the dot product $(\mathbf{y} - X\hat{\boldsymbol{\beta}}) \cdot X\mathbf{v}$ is equal to zero for *any* conformal vector, $\mathbf{v}$. This means that $\mathbf{y} - X\hat{\boldsymbol{\beta}}$ is the shortest of all possible vectors $\mathbf{y} - X\boldsymbol{\beta}$, that is, the variance of the residuals is the minimum possible. This is illustrated at the right.



The residual vector, $y - X\hat{\beta}$, which corresponds to the solution of a least squares system, $y = X\beta + \epsilon$, is orthogonal to the column space of the matrix X.

Introducing $\hat{\boldsymbol{\gamma}}$ and a matrix K with the assumption that a matrix $[X\ K]$ is non-singular and $K^{\mathrm{T}} X = 0$ (cf. Orthogonal projections), the residual vector should satisfy the following equation:

$$\hat{\mathbf{r}} \triangleq \mathbf{y} - X\hat{\boldsymbol{\beta}} = K\hat{\boldsymbol{\gamma}}.$$

The equation and solution of linear least squares are thus described as follows:

$$\mathbf{y} = \begin{bmatrix} X & K \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix},$$

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{bmatrix} X & K \end{bmatrix}^{-1} \mathbf{y} = \begin{bmatrix} (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \\ (K^{\mathrm{T}}K)^{-1}K^{\mathrm{T}} \end{bmatrix} \mathbf{y}.$$

If the experimental errors, $\boldsymbol{\epsilon}$, are uncorrelated, have a mean of zero and a constant variance, $\boldsymbol{\sigma}$, the Gauss–Markov theorem states that the least-squares estimator, $\hat{\boldsymbol{\beta}}$, has the minimum variance of all estimators that are linear combinations of the observations. In this sense it is the best, or optimal, estimator of the parameters. Note particularly that this property is independent of the statistical distribution function of the errors. In other words, *the distribution function of the errors need not be a normal distribution*. However, for some probability distributions, there is no guarantee that the least-squares solution is even possible given the observations; still, in such cases it is the best estimator that is both linear and unbiased.

For example, it is easy to show that the arithmetic mean of a set of measurements of a quantity is the least-squares estimator of the value of that quantity. If the conditions of the Gauss–Markov theorem apply, the arithmetic mean is optimal, whatever the distribution of errors of the measurements might be.

However, in the case that the experimental errors do belong to a normal distribution, the least-squares estimator is also a maximum likelihood estimator.[3]

These properties underpin the use of the method of least squares for all types of data fitting, even when the assumptions are not strictly valid.

Limitations

An assumption underlying the treatment given above is that the independent variable, x, is free of error. In practice, the errors on the measurements of the independent variable are usually much smaller than the errors on the dependent variable and can therefore be ignored. When this is not the case, total least squares or more generally errors-in-variables models, or *rigorous least squares*, should be used. This can be done by adjusting the weighting scheme to take into account errors on both the dependent and independent variables and then following the standard procedure.[4][5]

In some cases the (weighted) normal equations matrix $X^{\mathrm{T}}X$ is ill-conditioned. When fitting polynomials the normal equations matrix is a Vandermonde matrix. Vandermonde matrices become increasingly ill-conditioned as the order of the matrix increases. In these cases, the least squares estimate amplifies the measurement noise and may be grossly inaccurate. Various regularization techniques can be applied in such cases, the most common of which is called ridge regression. If further information about the parameters is known, for example, a range of possible values of $\hat{\boldsymbol{\beta}}$, then various techniques can be used to increase the stability of the solution. For example, see constrained least squares.

Another drawback of the least squares estimator is the fact that the norm of the residuals, $\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|$ is minimized, whereas in some cases one is truly interested in obtaining small error in the parameter $\hat{\boldsymbol{\beta}}$, e.g., a small value of $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|$. However, since the true parameter $\boldsymbol{\beta}$ is necessarily unknown, this quantity cannot be directly minimized. If a prior probability on $\hat{\boldsymbol{\beta}}$ is known, then a Bayes estimator can be used to minimize the mean squared error, $E\left\{\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2\right\}$. The least squares method is often applied when no prior is known. Surprisingly, when several parameters are being estimated jointly, better

estimators can be constructed, an effect known as Stein's phenomenon. For example, if the measurement error is Gaussian, several estimators are known which dominate, or outperform, the least squares technique; the best known of these is the James–Stein estimator. This is an example of more general shrinkage estimators that have been applied to regression problems.

Weighted linear least squares

In some cases the observations may be weighted—for example, they may not be equally reliable. In this case, one can minimize the weighted sum of squares:

$$\operatorname{arg\,min}_{\boldsymbol{\beta}} \sum_{i=1}^{m} w_i \left| y_i - \sum_{j=1}^{n} X_{ij}\beta_j \right|^2 = \operatorname{arg\,min}_{\boldsymbol{\beta}} \left\| W^{1/2}(\mathbf{y} - X\boldsymbol{\beta}) \right\|^2.$$

where $w_i > 0$ is the weight of the ith observation, and W is the diagonal matrix of such weights.

The weights should, ideally, be equal to the reciprocal of the variance of the measurement.[6][7] The normal equations are then:

$$\left(X^{\mathrm{T}} W X\right) \hat{\boldsymbol{\beta}} = X^{\mathrm{T}} W \mathbf{y}.$$

This method is used in iteratively reweighted least squares.

Parameter errors and correlation

The estimated parameter values are linear combinations of the observed values

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W \mathbf{y}.$$

Therefore, an expression for the residuals (i.e., the *estimated* errors in the parameters) can be obtained by error propagation from the errors in the observations. Let the variance-covariance matrix for the observations be denoted by M and that of the parameters by M^β. Then

$$M^\beta = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W M W^{\mathrm{T}} X (X^{\mathrm{T}} W^{\mathrm{T}} X)^{-1}.$$

When $W = M^{-1}$, this simplifies to

$$M^\beta = (X^{\mathrm{T}} W X)^{-1}.$$

When unit weights are used ($W = I$, the identity matrix), it is implied that the experimental errors are uncorrelated and all equal: $M = \sigma^2 I$, where σ^2 is the *a priori* variance of an observation. In any case, σ^2 is approximated by the reduced chi-squared χ^2_ν:

$$M^\beta = \chi^2_\nu (X^{\mathrm{T}} X)^{-1},$$
$$\chi^2_\nu = S/\nu,$$

where S is the minimum value of the (weighted) objective function:

$$S = r^{\mathrm{T}} W r.$$

The denominator, $\nu = m - n$, is the number of degrees of freedom; see effective degrees of freedom for generalizations for the case of correlated observations.

In all cases, the variance of the parameter β_i is given by M^β_{ii} and the covariance between parameters β_i and β_j is given by M^β_{ij}. Standard deviation is the square root of variance, $\sigma_i = \sqrt{M^\beta_{ii}}$, and the correlation coefficient is given by $\rho_{ij} = M^\beta_{ij} / (\sigma_i \sigma_j)$. These error estimates reflect only random errors in the measurements. The true uncertainty in the parameters is larger due to the presence of systematic errors, which, by definition, cannot be quantified. Note that even though the observations may be uncorrelated, the parameters are typically correlated.

Parameter confidence limits

It is often *assumed*, for want of any concrete evidence but often appealing to the central limit theorem—see Normal distribution#Occurrence—that the error on each observation belongs to a normal distribution with a mean of zero and standard deviation σ. Under that assumption the following probabilities can be derived for a single scalar parameter estimate in terms of its estimated standard error se_β (given here):

68% that the interval $\hat{\beta} \pm se_\beta$ encompasses the true coefficient value
95% that the interval $\hat{\beta} \pm 2se_\beta$ encompasses the true coefficient value

99% that the interval $\hat{\beta} \pm 2.5 s e_\beta$ encompasses the true coefficient value

The assumption is not unreasonable when *m* >> *n*. If the experimental errors are normally distributed the parameters will belong to a Student's t-distribution with *m* − *n* degrees of freedom. When *m* >> *n* Student's t-distribution approximates a normal distribution. Note, however, that these confidence limits cannot take systematic error into account. Also, parameter errors should be quoted to one significant figure only, as they are subject to sampling error.[8]

When the number of observations is relatively small, Chebychev's inequality can be used for an upper bound on probabilities, regardless of any assumptions about the distribution of experimental errors: the maximum probabilities that a parameter will be more than 1, 2 or 3 standard deviations away from its expectation value are 100%, 25% and 1% respectively.

Residual values and correlation

The residuals are related to the observations by

$$\hat{\mathbf{r}} = \mathbf{y} - X\hat{\boldsymbol{\beta}} = \mathbf{y} - H\mathbf{y} = (I - H)\mathbf{y}$$

where *H* is the idempotent matrix known as the hat matrix:

$$H = X\left(X^{\mathrm{T}} W X\right)^{-1} X^{\mathrm{T}} W$$

and *I* is the identity matrix. The variance-covariance matrix of the residuals, $\mathbf{M}^{\mathbf{r}}$ is given by

$$M^{\mathbf{r}} = (I - H)\, M (I - H)^{\mathrm{T}}.$$

Thus the residuals are correlated, even if the observations are not.

When $W = M^{-1}$,

$$M^{\mathbf{r}} = (I - H)\, M.$$

The sum of residual values is equal to zero whenever the model function contains a constant term. Left-multiply the expression for the residuals by X^{T}:

$$X^{\mathrm{T}}\hat{\mathbf{r}} = X^{\mathrm{T}}\mathbf{y} - X^{\mathrm{T}}X\hat{\boldsymbol{\beta}} = X^{\mathrm{T}}\mathbf{y} - (X^{\mathrm{T}}X)(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbf{y} = \mathbf{0}$$

Say, for example, that the first term of the model is a constant, so that $X_{i1} = 1$ for all *i*. In that case it follows that

$$\sum_i^m X_{i1}\hat{r}_i = \sum_i^m \hat{r}_i = 0.$$

Thus, in the motivational example, above, the fact that the sum of residual values is equal to zero it is not accidental but is a consequence of the presence of the constant term, α, in the model.

If experimental error follows a normal distribution, then, because of the linear relationship between residuals and observations, so should residuals,[9] but since the observations are only a sample of the population of all possible observations, the residuals should belong to a Student's t-distribution. Studentized residuals are useful in making a statistical test for an outlier when a particular residual appears to be excessively large.

Objective function

The optimal value of the objective function, found by substituting in the optimal expression for the coefficient vector, can be written as (assuming unweighted observations)

$$S = \mathbf{y}^{\mathrm{T}}(I - H)^{\mathrm{T}}(I - H)\mathbf{y} = \mathbf{y}^{\mathrm{T}}(I - H)\mathbf{y},$$

the latter equality holding, since (*I* − *H*) is symmetric and idempotent. It can be shown from this[10] that under an appropriate assignment of weights the expected value of *S* is *m* − *n*. If instead unit weights are assumed, the expected value of *S* is $(m - n)\sigma^2$, where σ^2 is the variance of each observation.

If it is assumed that the residuals belong to a normal distribution, the objective function, being a sum of weighted squared residuals, will belong to a chi-squared (χ^2) distribution with *m* − *n* degrees of freedom. Some illustrative percentile values of χ^2 are given in the following table.[11]

$m-n$	$\chi^2_{0.50}$	$\chi^2_{0.95}$	$\chi^2_{0.99}$
10	9.34	18.3	23.2
25	24.3	37.7	44.3
100	99.3	124	136

These values can be used for a statistical criterion as to the goodness of fit. When unit weights are used, the numbers should be divided by the variance of an observation.

Constrained linear least squares

Often it is of interest to solve a linear least squares problem with an additional constraint on the solution. With constrained linear least squares, the original equation

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

must be fit as closely as possible (in the least squares sense) while ensuring that some other property of $\boldsymbol{\beta}$ is maintained. There are often special-purpose algorithms for solving such problems efficiently. Some examples of constraints are given below:

- Equality constrained least squares: the elements of $\boldsymbol{\beta}$ must exactly satisfy $\mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ (see Ordinary least squares#Constrained estimation).
- Regularized least squares: the elements of $\boldsymbol{\beta}$ must satisfy $\|\mathbf{L}\boldsymbol{\beta} - \mathbf{y}\| \leq \boldsymbol{\alpha}$ (choosing $\boldsymbol{\alpha}$ in proportion to the noise standard deviation of $\boldsymbol{y}$ prevents over-fitting).
- Non-negative least squares (NNLS): The vector $\boldsymbol{\beta}$ must satisfy the vector inequality $\boldsymbol{\beta} \geq \mathbf{0}$ defined componentwise—that is, each component must be either positive or zero.
- Box-constrained least squares: The vector $\boldsymbol{\beta}$ must satisfy the vector inequalities $\boldsymbol{lb} \leq \boldsymbol{\beta} \leq \boldsymbol{ub}$, each of which is defined componentwise.
- Integer-constrained least squares: all elements of $\boldsymbol{\beta}$ must be integers (instead of real numbers).
- Phase-constrained least squares: all elements of $\boldsymbol{\beta}$ must have the same phase (or must be real rather than complex numbers, i.e. phase = 0).

When the constraint only applies to some of the variables, the mixed problem may be solved using **separable least squares** by letting $\mathbf{X} = [\mathbf{X_1}\,\mathbf{X_2}]$ and $\beta^{\mathbf{T}} = [{\beta_1}^{\mathbf{T}}\,{\beta_2}^{\mathbf{T}}]$ represent the unconstrained (1) and constrained (2) components. Then substituting the least-squares solution for $\boldsymbol{\beta_1}$, i.e.

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = {\mathbf{X_1}}^{+}(\mathbf{y} - \mathbf{X_2}\boldsymbol{\beta_2})$$

back into the original expression gives (following some rearrangement) an equation that can be solved as a purely constrained problem in $\boldsymbol{\beta_2}$.

$$\mathbf{P}\mathbf{X_2}\boldsymbol{\beta_2} = \mathbf{P}\mathbf{y},$$

where $\mathbf{P} := \mathbf{I} - \mathbf{X_1}{\mathbf{X_1}}^{+}$ is a projection matrix. Following the constrained estimation of $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$ the vector $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ is obtained from the expression above.

Typical uses and applications

- Polynomial fitting: models are polynomials in an independent variable, *x*:
  - Straight line: $f(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 x$.[12]
  - Quadratic: $f(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 x + \beta_3 x^2$.
  - Cubic, quartic and higher polynomials. For regression with high-order polynomials, the use of orthogonal polynomials is recommended.[13]
- Numerical smoothing and differentiation — this is an application of polynomial fitting.
- Multinomials in more than one independent variable, including surface fitting
- Curve fitting with B-splines [4]
- Chemometrics, Calibration curve, Standard addition, Gran plot, analysis of mixtures

Uses in data fitting

The primary application of linear least squares is in data fitting. Given a set of *m* data points $y_1, y_2, \ldots, y_m$, consisting of experimentally measured values taken at *m* values $x_1, x_2, \ldots, x_m$ of an independent variable (x_i may be scalar or vector quantities), and given a model function $y = f(x, \boldsymbol{\beta})$, with $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$, it is desired to find the parameters β_j such that the model function "best" fits the data. In linear least squares, linearity is meant to be with respect to parameters β_j, so

$$f(x, \boldsymbol{\beta}) = \sum_{j=1}^{n} \beta_j \phi_j(x).$$

Here, the functions ϕ_j may be **nonlinear** with respect to the variable **x**.

Ideally, the model function fits the data exactly, so

$$y_i = f(x_i, \boldsymbol{\beta})$$

for all $i = 1, 2, \ldots, m$. This is usually not possible in practice, as there are more data points than there are parameters to be determined. The approach chosen then is to find the minimal possible value of the sum of squares of the residuals

$$r_i(\boldsymbol{\beta}) = y_i - f(x_i, \boldsymbol{\beta}), \ (i = 1, 2, \ldots, m)$$

so to minimize the function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} r_i^2(\boldsymbol{\beta}).$$

After substituting for r_i and then for f, this minimization problem becomes the quadratic minimization problem above with

$$X_{ij} = \phi_j(x_i),$$

and the best fit can be found by solving the normal equations.

Further discussion

The *numerical methods for linear least squares* are important because linear regression models are among the most important types of model, both as formal statistical models and for exploration of data-sets. The majority of statistical computer packages contain facilities for regression analysis that make use of linear least squares computations. Hence it is appropriate that considerable effort has been devoted to the task of ensuring that these computations are undertaken efficiently and with due regard to round-off error.

Individual statistical analyses are seldom undertaken in isolation, but rather are part of a sequence of investigatory steps. Some of the topics involved in considering numerical methods for linear least squares relate to this point. Thus important topics can be

- Computations where a number of similar, and often nested, models are considered for the same data-set. That is, where models with the same dependent variable but different sets of independent variables are to be considered, for essentially the same set of data-points.
- Computations for analyses that occur in a sequence, as the number of data-points increases.
- Special considerations for very extensive data-sets.

Fitting of linear models by least squares often, but not always, arise in the context of statistical analysis. It can therefore be important that considerations of computation efficiency for such problems extend to all of the auxiliary quantities required for such analyses, and are not restricted to the formal solution of the linear least squares problem.

Rounding errors

Matrix calculations, like any other, are affected by rounding errors. An early summary of these effects, regarding the choice of computation methods for matrix inversion, was provided by Wilkinson.[14]

See also

- Line-line intersection#Nearest point to non-intersecting lines, an application

References

1. Goldberger, Arthur S. (1964). "Classical Linear Regression" (https://books.google.com/books?id=KZq5AAAAIAAJ&pg=PR156). *Econometric Theory*. New York: John Wiley & Sons. pp. 156–212 [p. 158]. ISBN 0-471-31101-4.
2. Lawson, C. L.; Hanson, R. J. (1974). *Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall. ISBN 0-13-822585-0.
3. Margenau, Henry; Murphy, George Moseley (1956). *The Mathematics of Physics and Chemistry*. Princeton: Van Nostrand.
4. Gans, Peter (1992). *Data fitting in the Chemical Sciences*. New York: Wiley. ISBN 0-471-93412-7.
5. Deming, W. E. (1943). *Statistical adjustment of Data*. New York: Wiley.
6. This implies that the observations are uncorrelated. If the observations are correlated, the expression $S = \sum_k \sum_j r_k W_{kj} r_j$ applies. In this case the weight matrix should ideally be equal to the inverse of the variance-covariance matrix of the observations.
7. Strutz, T. (2016). *Data Fitting and Uncertainty (A practical introduction to weighted least squares and beyond)*. Springer Vieweg. ISBN 978-3-658-11455-8, chapter 3
8. Mandel, John (1964). *The Statistical Analysis of Experimental Data*. New York: Interscience.
9. Mardia, K. V.; Kent, J. T.; Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press. ISBN 0-12-471250-9.

10. Hamilton, W. C. (1964). *Statistics in Physical Science* New York: Ronald Press.
11. Spiegel, Murray R. (1975). *Schaum's outline of theory and problems of probability and statistics* New York: McGraw-Hill. ISBN 0-585-26739-1.
12. Acton, F. S. (1959). *Analysis of Straight-Line Data* New York: Wiley.
13. Guest, P. G. (1961). *Numerical Methods of Curve Fitting* Cambridge: Cambridge University Press.
14. Wilkinson, J.H. (1963) "Chapter 3: Matrix Computations", *Rounding Errors in Algebraic Processes*, London: Her Majesty's Stationery Office (National Physical Laboratory Notes in Applied Science, No.32)

Further reading

- Bevington, Philip R.; Robinson, Keith D. (2003) *Data Reduction and Error Analysis for the Physical Sciences* McGraw-Hill. ISBN 0-07-247227-8.
- Barlow, Jesse L. (1993), "Chapter 9: Numerical aspects of Solving Linear Least Squares Problems", in Rao, C. R., *Computational Statistics*, Handbook of Statistics, **9**, North-Holland, ISBN 0-444-88096-8
- Björck, Åke (1996). *Numerical methods for least squares problems* Philadelphia: SIAM. ISBN 0-89871-360-9.
- Goodall, Colin R. (1993), "Chapter 13: Computation using the QR decomposition", in Rao, C. R., *Computational Statistics*, Handbook of Statistics, **9**, North-Holland, ISBN 0-444-88096-8
- National Physical Laboratory (1961), "Chapter 1: Linear Equations and Matrices: Direct Methods", *Modern Computing Methods*, Notes on Applied Science, **16** (2nd ed.), Her Majesty's Stationery Office
- National Physical Laboratory (1961), "Chapter 2: Linear Equations and Matrices: Direct Methods on Automatic Computers", *Modern Computing Methods*, Notes on Applied Science, **16** (2nd ed.), Her Majesty's Stationery Office

External links

- Least Squares Fitting – From MathWorld
- Least Squares Fitting-Polynomial – From MathWorld