

Roundoff Error

A round-off error (rounding error), is the difference between the calculated approximation of a number and its exact mathematical value.

Numerical analysis specifically tries to estimate this error when using approximation equations and/or algorithms, especially when using finite digits to represent infinite digits of real numbers. This is a form of quantization error.

Roundoff Error

Example

Notation	Represent	Approximate	Error
$1/7$	$0.\overline{142857}$	0.142857	0.000000 $\overline{142857}$
$\ln 2$	0.69314718055994530941...	0.693147	0.00000018055994530941...
$\log_{10} 2$	0.30102999566398119521...	0.3010	0.00002999566398119521...
$\sqrt[3]{2}$	1.25992104989487316476...	1.25992	0.00000104989487316476...
$\sqrt{2}$	1.41421356237309504880...	1.41421	0.00000356237309504880...
e	2.71828182845904523536...	2.718281828459045	0.00000000000000023536...
π	3.14159265358979323846...	3.141592653589793	0.00000000000000023846...

There are at least two ways of performing the termination at the limited digit place:

- **truncation**: simply chop off the remaining digits.
 $0.\overline{142857} \approx 0.142$ (dropping all significant digits after 3rd)
- **rounding**: add 5 to the next digit and then chop it. The result may **round up** or **round down**.
 $0.\overline{142857} \approx 0.143$ (rounding the 4th significant digit. This is rounded up because $8 \geq 5$)
 $0.\overline{142857} \approx 0.14$ (rounding the 3rd significant digit. This is rounded down because $2 < 5$)

Linear Algebra Software

- **LINPACK** is a software library for performing numerical linear algebra. It was written in Fortran by Jack Dongarra, Jim Bunch, Cleve Moler, and Pete Stewart, and was intended for use on supercomputers in the 1970s and early 1980s. It has been largely superseded by LAPACK, which will run more efficiently on modern architectures. LINPACK makes use of the BLAS (Basic Linear Algebra Subprograms) libraries for performing basic vector and matrix operations.
- **LAPACK**, the Linear Algebra PACKage, is a software library for numerical computing originally written in FORTRAN 77 and now written in Fortran 90. It provides routines for solving systems of simultaneous linear equations, least-squares solutions of linear systems of equations, eigenvalue problems, Householder transformation to implement QR decomposition on a matrix and singular value problems. **Lapack95** uses features of Fortran 95 to simplify the interface of the routines.
- **ScaLAPACK** (or Scalable LAPACK) is a parallel version of **LAPACK**.
- **Eigen 2** is a C++ template library for linear algebra: vectors, matrices, and related algorithms.

Solution of Linear Equations

Linear Equations

A linear equation is an algebraic equation in which each term is either a constant or the product of a constant and (the first power of) a single variable. Linear equations can have one, two, three or more variables.

Linear equations occur with great regularity in applied mathematics. While they arise quite naturally when modeling many phenomena, they are particularly useful since many non-linear equations may be reduced to linear equations by assuming that quantities of interest vary to only a small extent from some “background” state.

Linear Equations

A common form of a linear equation is:

$$y = mx + b$$

This can easily be rearranged to have the general form:

$$Ax + By + C = 0$$

Linear Equations

A linear equation can involve more than two variables. The general linear equation in n variables is:

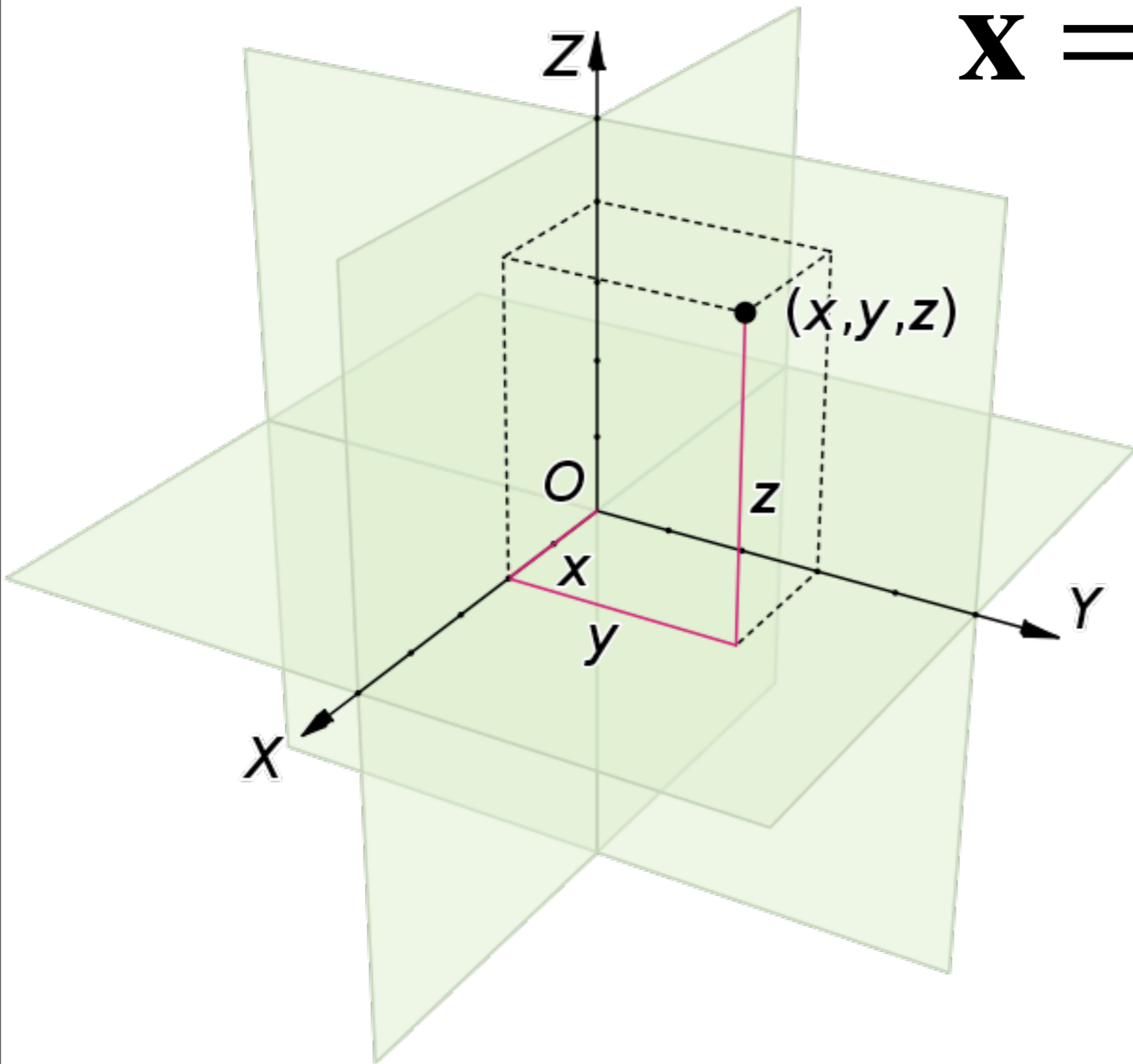
$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b_1$$

In this form, a_1, a_2, \dots, a_n are the coefficients, x_1, x_2, \dots, x_n are the variables, and b_1 is the constant. When dealing with three or fewer variables, it is common to replace x_1 with just x , x_2 with y , and x_3 with z , as appropriate.

Such an equation will represent an $(n-1)$ -dimensional hyperplane in n -dimensional Euclidean space (for example, a plane in 3-space).

Euclidean space

$$\mathbf{X} = (x_1, x_2, \dots, x_n)$$



Every point in three-dimensional Euclidean space is determined by three coordinates.

The term “Euclidean” distinguishes these spaces from the curved spaces of non-Euclidean geometry and Einstein's general theory of relativity, and is named for the Greek mathematician Euclid of Alexandria.

Linear Equations

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

...

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

In general there could be N unknowns x_1 through x_n related by M equations. If $N=M$ we have as many equations as unknowns.

Linear Equations

We can rewrite this as a matrix equation:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_m \end{bmatrix}$$

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$$

$$\mathbf{b} = \mathbf{A} \cdot \mathbf{x} \Leftrightarrow b_i = \sum_j a_{ij} x_j$$

Linear Equations

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$$

can be rewritten as

$$\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$$

Where \mathbf{A}^{-1} is the inverse of \mathbf{A} . However, in the vast majority of cases it is both unnecessary and inadvisable to do so. For example,

$$7x = 21 \Rightarrow x = \frac{21}{7} = 3$$

use of the inverse would have given

$$x = 7^{-1} \times 21 = 0.142857 \times 21 = 2.99997$$

The inverse requires more arithmetic and produces a less accurate answer.

Linear Equations

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_m \end{bmatrix}, \mathbf{A} \cdot \mathbf{x} = \mathbf{b}$$

We note that interchanging any two rows of **A** and the corresponding rows of **b** does not change the solution in any way.

Likewise, if we replace any row of **A** with a linear combination of itself and any other row and do the corresponding replacement for **b**, we do not change the solution.

Tasks of Linear Algebra

When $M=N$

- Solution of $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ for an unknown \mathbf{x} .
- Solution of more than one matrix equation, $\mathbf{A} \cdot \mathbf{x}_j = \mathbf{b}_j$ with \mathbf{A} held constant and different right hand side, \mathbf{b} .
- Calculation of the inverse square matrix \mathbf{A}^{-1} .
- Calculation of the determinant of a square matrix \mathbf{A} .

Tasks of Linear Algebra

When $M < N$ (or $M = N$ but the equations are degenerate), then there are effectively fewer equations than unknowns. In this case there can be either no solution or more than one solution vector \mathbf{x} . If there is more than one solution, the solution space consists of a particular solution \mathbf{x}_p added to any linear combination of vectors which are said to be the nullspace of the matrix \mathbf{A} . Finding the solution space of \mathbf{A} involves

- Singular value decomposition of a matrix \mathbf{A} .

Tasks of Linear Algebra

When $M > N$ (more equations than unknowns), there is in general no solution vector, and the set of equations is said to be *overdetermined*. However, we can seek for the best ‘compromise’ solution, the one that comes closest to simultaneously satisfying all the equations. If closeness is defined in a least squares sense. This is called the

- Linear least squares problem $(\mathbf{A}^T \cdot \mathbf{A}) \cdot \mathbf{x} = (\mathbf{A}^T \cdot \mathbf{b})$

Pivoting

Pivoting is a process performed on a matrix in order to improve numerical stability.

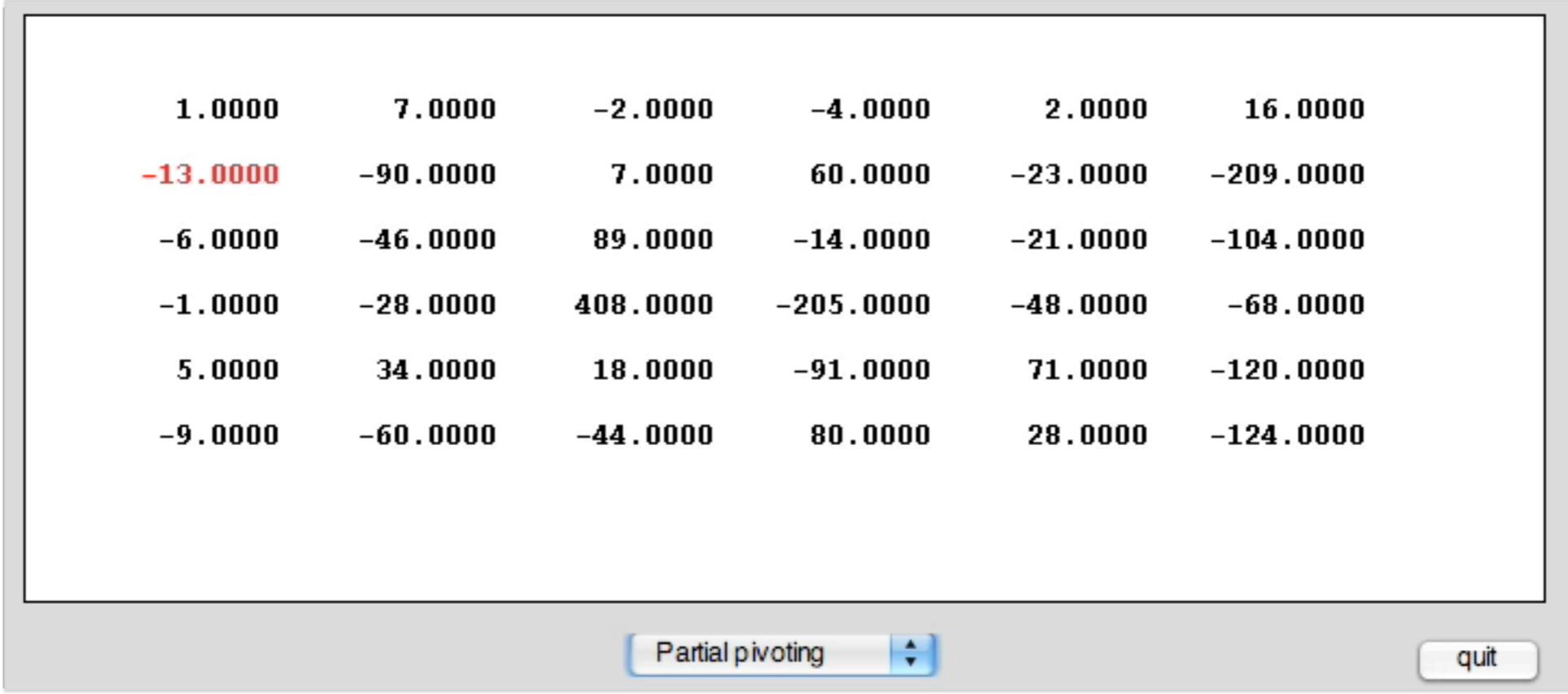
Partial pivoting of an $n \times n$ matrix is the sorting of the rows of the matrix so that row i contains the maximum absolute column value for column i among all rows in i, \dots, n . That is, we begin by swapping row 1 with the row that has the largest absolute value for the first column, then swap row 2 with the row that has the largest magnitude for the second column (among rows 2 and below), and so on.

Complete pivoting is a reordering of both rows and columns, using the same method as above. It is usually not necessary to ensure numerical stability.

Pivoting of a matrix A can be represented as multiplication by a permutation matrix, P , namely, AP .

Partial Pivoting

Interchanging any two equations (rows) is called pivoting.



1.0000	7.0000	-2.0000	-4.0000	2.0000	16.0000
-13.0000	-90.0000	7.0000	60.0000	-23.0000	-209.0000
-6.0000	-46.0000	89.0000	-14.0000	-21.0000	-104.0000
-1.0000	-28.0000	408.0000	-205.0000	-48.0000	-68.0000
5.0000	34.0000	18.0000	-91.0000	71.0000	-120.0000
-9.0000	-60.0000	-44.0000	80.0000	28.0000	-124.0000

Partial pivoting

quit

In partial pivoting, the algorithm considers all entries in the column of the matrix that is currently being considered, picks the entry with **largest absolute value**, and finally swaps rows such that this entry is the pivot in question. This improves the numerical stability. Complete pivoting considers all entries in the whole matrix. Complete pivoting is usually not necessary to ensure numerical stability.

Diagonal Pivoting

1.0000	7.0000	-2.0000	-4.0000	2.0000	16.0000
-13.0000	-90.0000	7.0000	60.0000	-23.0000	-209.0000
-6.0000	-46.0000	89.0000	-14.0000	-21.0000	-104.0000
-1.0000	-28.0000	408.0000	-205.0000	-48.0000	-68.0000
5.0000	34.0000	18.0000	-91.0000	71.0000	-120.0000
-9.0000	-60.0000	-44.0000	80.0000	28.0000	-124.0000

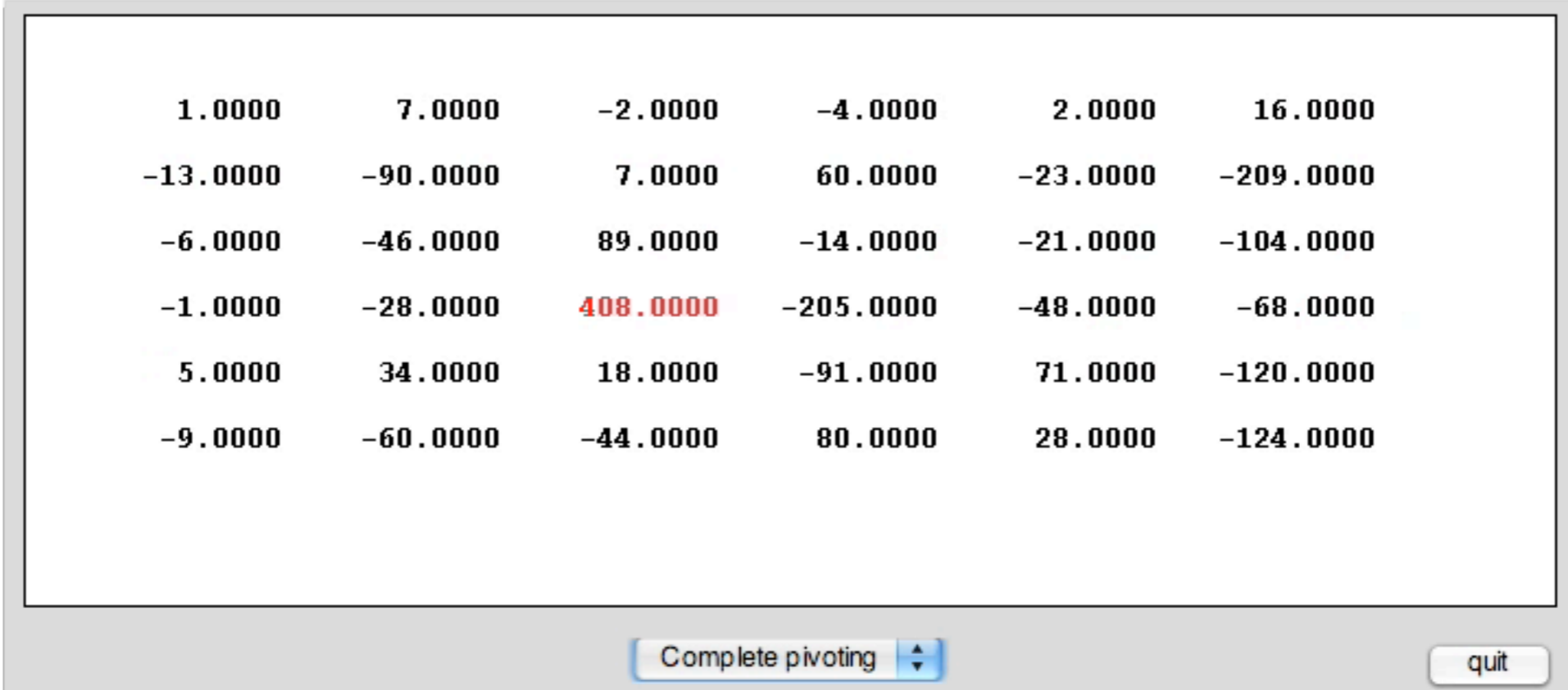
Diagonal pivoting

quit

Complete Pivoting

For certain systems and algorithms, complete pivoting (or maximal pivoting) may be required for acceptable accuracy. Complete pivoting considers all entries in the whole matrix, interchanging rows and columns to achieve the highest accuracy. Complete pivoting is usually not necessary to ensure numerical stability and, due to the additional computations it introduces, it may not always be the most appropriate pivoting strategy.

Complete Pivoting



1.0000	7.0000	-2.0000	-4.0000	2.0000	16.0000
-13.0000	-90.0000	7.0000	60.0000	-23.0000	-209.0000
-6.0000	-46.0000	89.0000	-14.0000	-21.0000	-104.0000
-1.0000	-28.0000	408.0000	-205.0000	-48.0000	-68.0000
5.0000	34.0000	18.0000	-91.0000	71.0000	-120.0000
-9.0000	-60.0000	-44.0000	80.0000	28.0000	-124.0000

For certain systems and algorithms, complete pivoting (or maximal pivoting) may be required for acceptable accuracy. Complete pivoting considers all entries in the whole matrix, interchanging rows and columns to achieve the highest accuracy. Complete pivoting is usually not necessary to ensure numerical stability and, due to the additional computations it introduces, it may not always be the most appropriate pivoting strategy.

LU Decomposition

In linear algebra, the LU decomposition is a matrix decomposition which writes a matrix as the product of a lower triangular matrix (elements only on the diagonal and below) and upper triangular matrix (elements only on the diagonal and above).

$$\mathbf{L} \cdot \mathbf{U} = \mathbf{A}$$

LU Decomposition

For example:

$$\begin{pmatrix} \alpha_{11} & 0 & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} \cdot \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ 0 & \beta_{22} & \beta_{23} \\ 0 & 0 & \beta_{33} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

We can use a decomposition to solve the linear set

$$\mathbf{A} \cdot \mathbf{x} = (\mathbf{L} \cdot \mathbf{U}) \cdot \mathbf{x} = \mathbf{L} \cdot (\mathbf{U} \cdot \mathbf{x}) = \mathbf{b}$$

by first solving the vector \mathbf{y} such that

$$(\mathbf{L} \cdot \mathbf{y}) = \mathbf{b}$$

and then solving

$$(\mathbf{U} \cdot \mathbf{x}) = \mathbf{y}$$

LU Decomposition

What is the advantage of doing this?

The solution of a triangular set of equations is trivial.

$$y_1 = \frac{b_1}{\alpha_{11}}$$

$$y_i = \frac{1}{\alpha_{ii}} \left[b_i - \sum_{j=1}^{i-1} \alpha_{ij} x_j \right], i = 2, 3, \dots, N$$

$$x_N = \frac{y_N}{\beta_{NN}}$$

$$x_i = \frac{1}{\beta_{ii}} \left[y_i - \sum_{j=i+1}^N \beta_{ij} x_j \right], i = N-1, N-3, \dots, 1$$

Once we have the LU decomposition of **A** we can solve for as many right hand sides as we want.

LU Decomposition

Look up the matlab `lu` function and go through the examples you find in `doc lu`

Singular Value Decomposition

There are a set of techniques for dealing with sets of equations that are either singular or close to singular (zero determinant).

The singular values are simply the absolute values of the eigenvalues.

If a matrix **A** has a matrix of eigenvectors that is not invertible (for example, the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has the noninvertible system of eigenvectors $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$), then **A** does not have an eigen decomposition, and so LU decomposition and Gaussian elimination will fail.

Singular Value Decomposition

However, if A is an $M \times N$ real matrix with $M > N$, then A can be written using a so-called singular value decomposition of the form

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

\mathbf{U} and \mathbf{V} have orthogonal columns such that

$$\mathbf{U}\mathbf{U}^T = \mathbf{I}$$

$$\mathbf{V}\mathbf{V}^T = \mathbf{I}$$

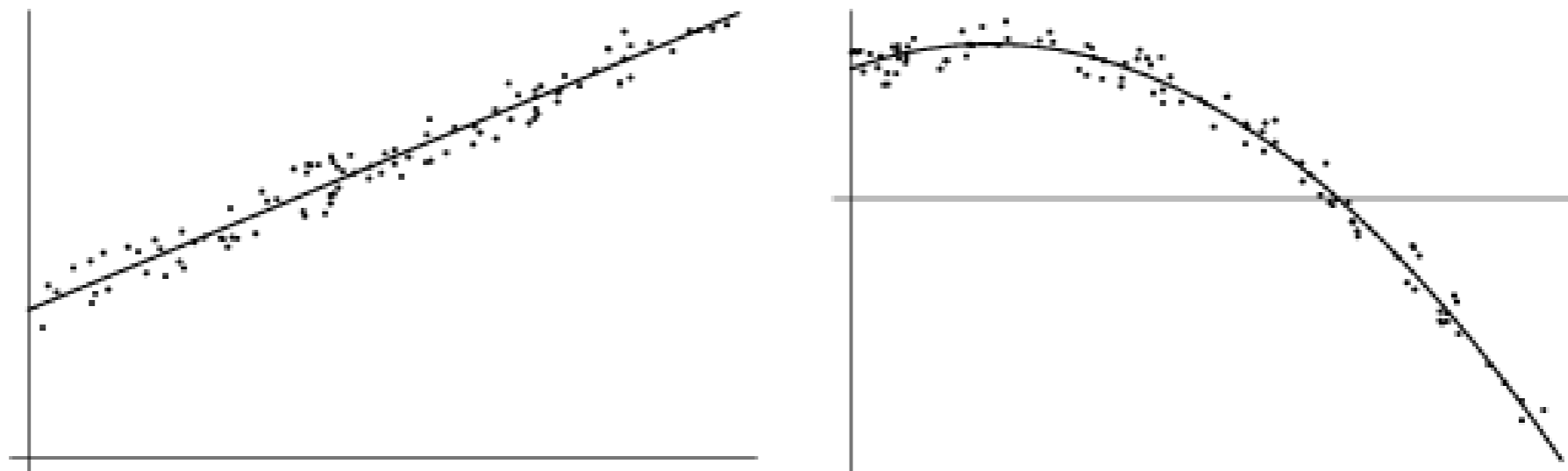
This decomposition can always be done.

Singular Value Decomposition

Singular Value Decomposition is the method of choice for linear least squares problems, with several applications in signal processing and statistics.

doc svd

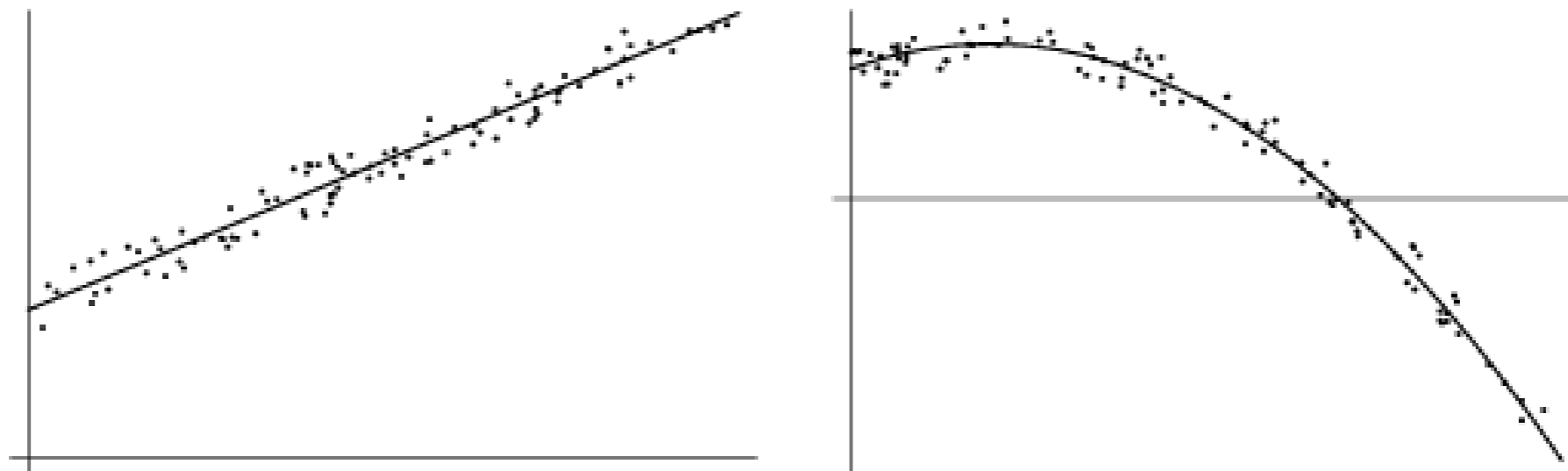
Least Squares Fitting



A mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets (“the residuals”) of the points from the curve.

The sum of the squares of the offsets is used instead of the offset absolute values because this allows the residuals to be treated as a continuous differentiable quantity.

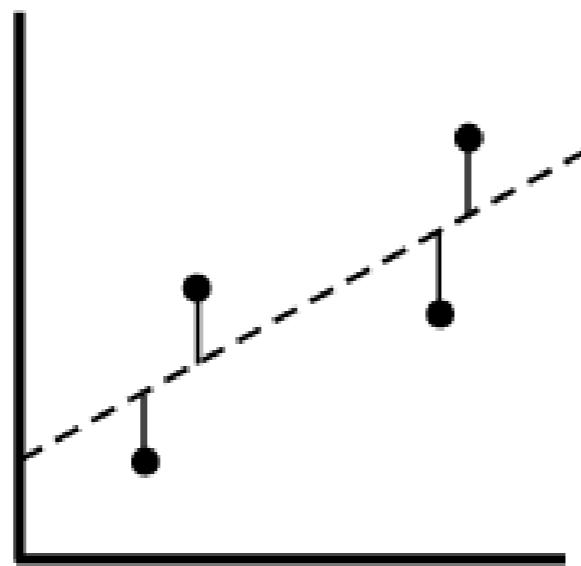
Q. Because squares of the offsets are used what effect will outlying points have?



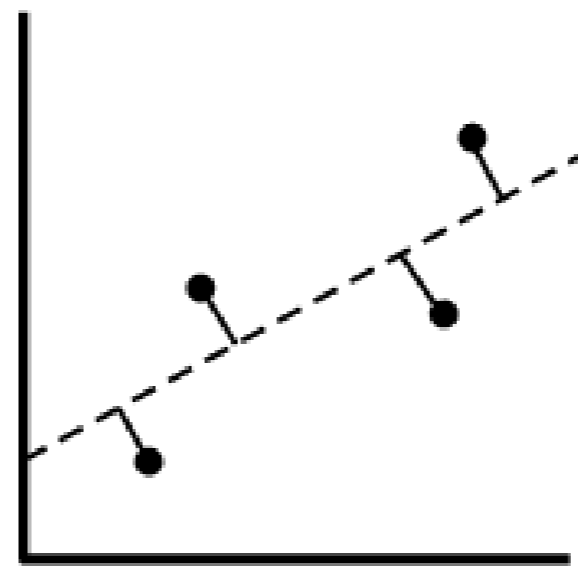
A mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets (“the residuals”) of the points from the curve.

The sum of the squares of the offsets is used instead of the offset absolute values because this allows the residuals to be treated as a continuous differentiable quantity.

Because squares of the offsets are used, outlying points can have a disproportionate effect on the fit, a property which may or may not be desirable depending on the problem at hand.



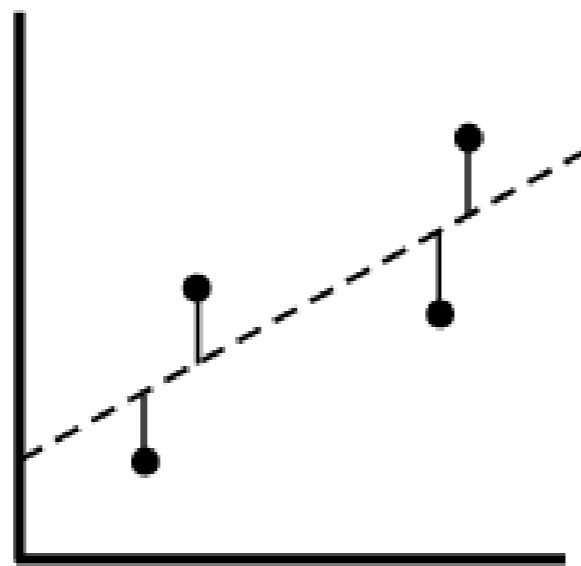
vertical offsets



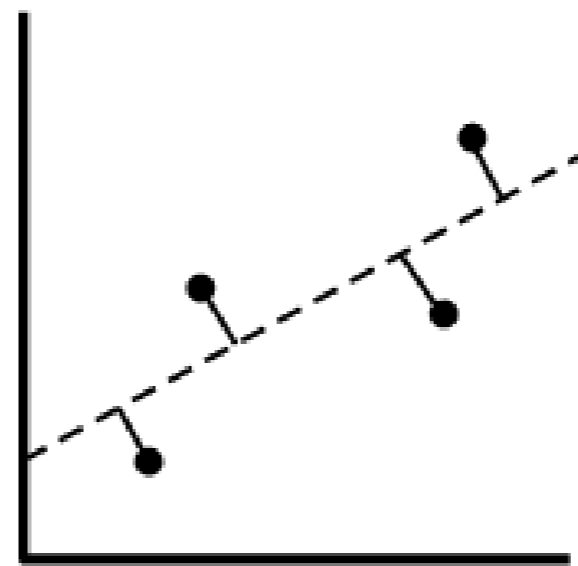
perpendicular offsets

In practice, the vertical offsets from a line (polynomial, surface, hyperplane, etc.) are almost always minimized instead of the perpendicular offsets.

Q. Why do you think this is done?



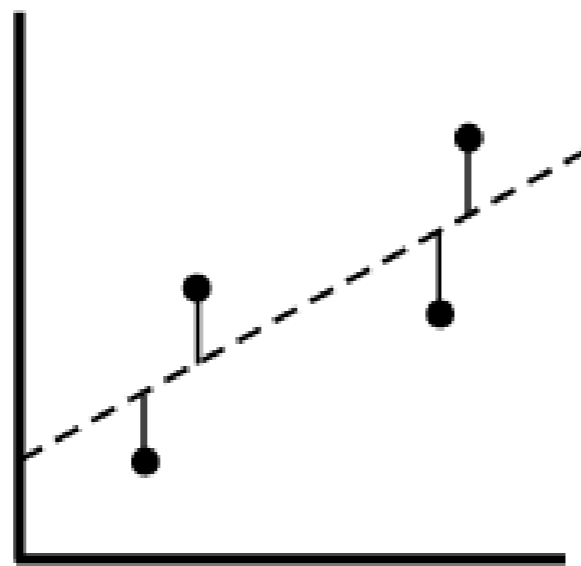
vertical offsets



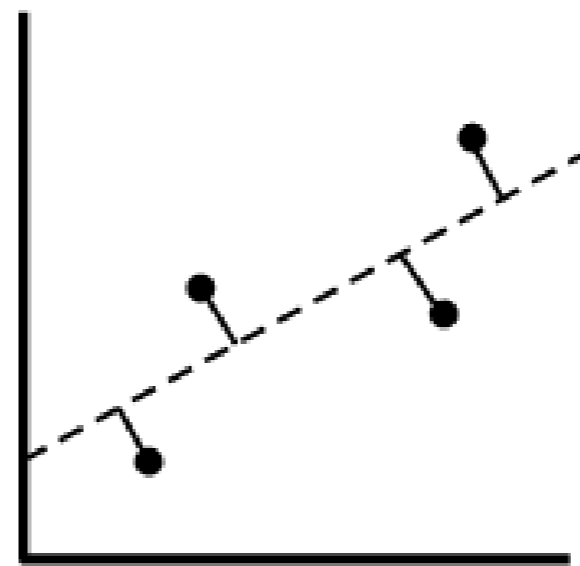
perpendicular offsets

In practice, the vertical offsets from a line (polynomial, surface, hyperplane, etc.) are almost always minimized instead of the perpendicular offsets.

This provides a fitting function for the independent variable X that estimates y for a given x (most often what an experimenter wants), allows uncertainties of the data points along the x - and y -axes to be incorporated simply, and also provides a much simpler analytic form for the fitting parameters than would be obtained using a fit based on perpendicular offsets. In addition, the fitting technique can be easily generalized from a best-fit line to a best-fit polynomial when sums of vertical distances are used. In any case, for a reasonable number of noisy data points, the difference between vertical and perpendicular fits is quite small.



vertical offsets



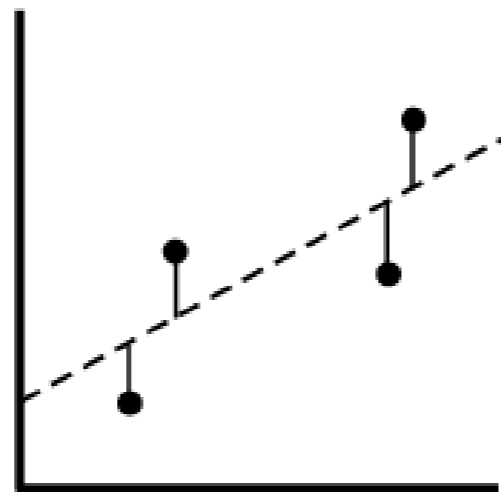
perpendicular offsets

$$R = \sum [y_i - f(x_i, a_1, \dots, a_n)]^2$$

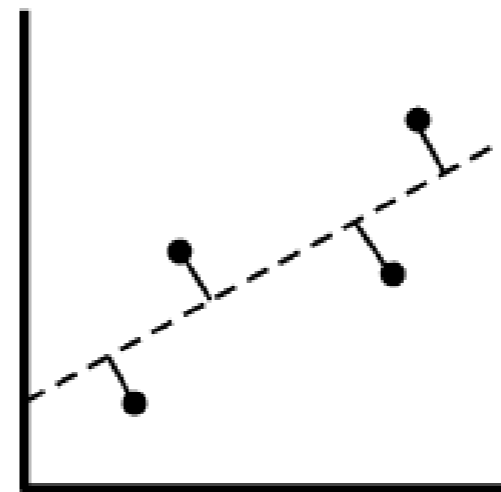
The residual is the sum of deviations from a best-fit curve of arbitrary form.

Error & Residual

- An error is the amount by which an observation differs from its expected value.
- A residual, on the other hand, is an observable estimate of the unobservable error.



vertical offsets



31 *perpendicular offsets*

The linear least squares fitting technique is the simplest and most commonly applied form of linear regression and provides a solution to the problem of finding the best fitting straight line through a set of points.

Regression is a method for *fitting* a curve (not necessarily a straight line) through a set of points using some *goodness-of-fit criterion*.

Linear regression is a regression that is *linear* in the unknown parameters used in the fit.


In fact, if the functional relationship between the two quantities being graphed is known to within additive or multiplicative constants, it is common practice to **transform the data** in such a way that the resulting line is a **straight line**, say by plotting T vs. \sqrt{l} instead of T vs. l in the case of analyzing the period T of a pendulum as a function of its length l .

The formulas for linear least squares fitting were independently derived by Gauss and Legendre.

Johann Carl Friedrich Gauss



Carl Friedrich Gauss, painted by Christian Albrecht Jensen

Born	30 April 1777 Brunswick, Germany
Died	23 February 1855 Göttingen, Hanover, Germany
Residence	 Germany
Nationality	 German
Field	Mathematician and physicist
Institution	Georg-August University
Alma mater	Helmstedt University
Academic advisor	Johann Friedrich Pfaff
Notable students	Friedrich Bessel Christoph Gudermann Christian Ludwig Gerling J. W. Richard Dedekind Johann Encke Johann Listing Bernhard Riemann
Known for	Number theory The Gaussian Magnetism

Adrien-Marie Legendre



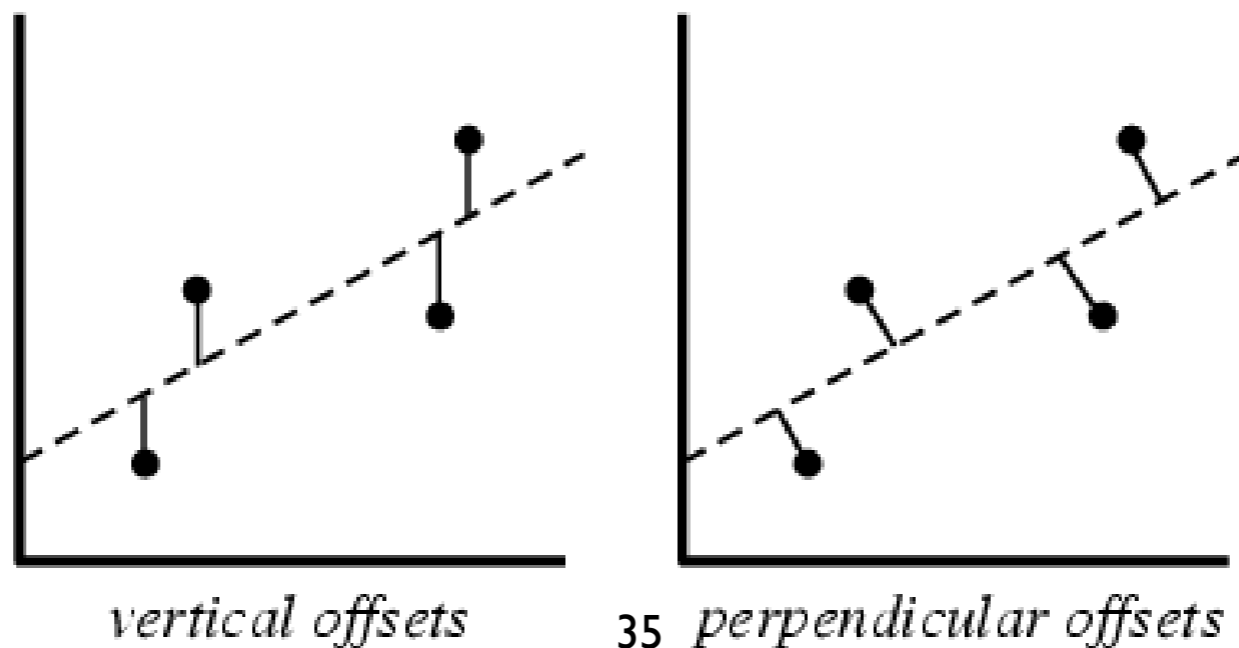
Adrien-Marie Legendre

Born	September 18, 1752 Paris, France
Died	January 10, 1833 Paris, France
Residence	 France
Nationality	 French
Field	Mathematician
Institution	École Militaire
Alma mater	Collège Mazarin
Known for	Lagrangian and elliptic functions

Vertical least squares fitting proceeds by finding the sum of the squares of the vertical deviations R^2 of a set of n data points

$$R = \sum [y_i - f(x_i, a_1, \dots, a_n)]^2$$

from a function f . Note that this procedure does not minimize the actual deviations from the line (which would be measured perpendicular to the given function).



Reading Assignment

Numerical Recipes

Chapter 2