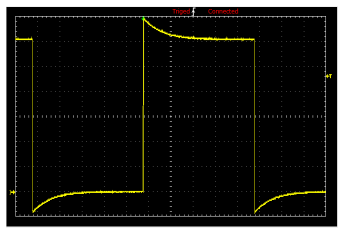


Vladimir Dobrushkin
https://math.uri.edu/~dobrush/

Linear Algebra with Mathematica

Least Squares

The scientific custom of taking multiple observations of the same quantity and then selecting a single estimate that best represents it has its origin the early part of the 16th century. At the beginning of the 19th century, two of the foremost mathematicians of the time, the German C.F. Gauss (1777--1855) and the Frenchman A.M. Legendre (1752--1833), proposed, independently of each other, an optimal criterion that bears the same relation to the average.

The context for Legendre's proposal of the least squares was that of geodesy. At that time, France's scientists had decided to adopt a new measurement system and proposed to use a unit length to be a meter, which would be equal 1/40,000,000 of the circumference of the earth. This necessitated an accurate determination of the said circumference that, in turn, depended on the exact shape of the earth.

The credibility of the method of least squares were greatly enhanced by the Ceres incident. On January 1, 1801 the Italian astronomer Giuseppe Piazzi sighted a heavenly body that he strongly suspected to be a new planet. He announced his discovery and named it Ceres. Unfortunately, 6 weeks later, before enough observations had been taken to make possible accurate determination of its orbit, so as to ascertain that it was indeed a planet, Ceres disappeared behind the sun and was not expected to reemerge for nearly a year. Interest in this possibly new planet was widespread, and the young Gauss proposed that an area of the sky be searched that was quite different from those suggested by other astronomers; he turned out to be right! He became a celebrity upon his discovery, which includes two mathematical methods: the row echelon reduction and the least square method.

In science and business, we often need to predict the relationship between two given variables. In many cases, we begin by performing an appropriate experiments or statistical analysis to obtain the necessary data. However, even if a simple law governs the behavior of the variables, this law may not be easy to find because of errors introduced in measuring or sampling. In practice, therefore, we are often content with a functional equation that provides a close approximation. There are know three general techniques for finding functions which are closest to a given curve:

- linear regression using linear polynomials (matching straight lines);
- general linear regression (polynomials, etc.), and
- transformations to linear regression (for matching exponential or logarithmic functions).

Historically, besides to curve fitting, the least square technique is proved to be very useful in statistical modeling of noisy data, and in geodetic modeling. We discuss three standard ways to solve the least square problem: the normal equations, the QR factorization, and the singular value decomposition.

Consider a typical application of least squares in curve fitting. Suppose that we are given a set of points (x_i, y_i) for $i = 0, 1, 2, \ldots, n$, for which we may not be able (or may not want) to find a function that passes through all points, but rather, we may want to find a function *f(x)* of a particular form that passes as closely as possible to the points. The differences

$$e_i = f(x_i) - y_i \quad \text{for} \quad i = 1, 2, \ldots n,$$

are called the errors or deviations or residuals. There are several norms that can be used with the residuals to measure how the curve $y = f(x)$ lies from data:

- maximum error: $\| e \|_{\infty} = \max_{1 \le i \le n} \{ | f(x_i) - y_i | \}$;
- average error: $\| e \|_1 = \frac{1}{n} \sum_{i=1}^{n} | f(x_i) - y_i |$;
- root-mean square error (or L^2-error): $\| e \|_2 = \left(\frac{1}{n} \sum_{i=1}^{n} | f(x_i) - y_i |^2 \right)^{1/2}$.

Example. Compare the maximum error, average error, and rms error for the linear approximation $y = f(x) = 3.5 - 1.3\,x$ to the data points (-1,5), (0,4), (1,2), (2,1), (3,0), (4,-2), (5,-3), (6,-4).

The errors are summarized in the following table:

x_i	y_i	$f(x_i) = 3.5 - 1.3\,x_i$	$\lvert e_i \rvert$	e_i^2
-1	5	4.8	0.2	0.04
0	4	3.5	0.5	0.25
1	2	2.2	0.2	0.04
2	1	0.9	0.1	0.01
3	0	-0.4	0.4	0.16
4	-2	-1.7	0.3	0.09
5	-3	-1	0	0
6	-4	-4.3	0.3	0.09

We can see that the maximum error $\| e \|_{\infty} = 0.5$ is largest because

$$\| e \|_1 = \frac{1}{8} \sum_{i=1}^{8} | e_i | = 0.25, \qquad \| e \|_2 = \frac{1}{2\sqrt{2}} \left(\sum_{i=1}^{8} | e_i |^2 \right)^{1/2} \approx 0.291548. \qquad \blacksquare$$

Let $\{(x_i, y_i)\}_{i=1}^n$ be set of *n* points, where the abscissas x_i are distinct. The **least-squares line** $y = m\,x + b$ is the line that minimizes the root-mean-square error e_2.

The quantity $\|e\|_2$ will be minimum if and only if the quantity $\sum_{i=1}^n (m\,x_i + b - y_i)^2$ is a minimum. The letter is visualized geometrically by minimizing the sum of the squares of the vertical distances from the points to the line.

Example. Suppose we're given a collection of data points *(x,y):*

$$(1,1), \quad (2,2), \quad (3,2)$$

and we want to find the closest line $y = mx + b$ to that collection. If the line went through all three points, we'd have:

$$\begin{aligned} m + b &= 1, \\ 2m + b &= 2, \\ 3m + b &= 2, \end{aligned}$$

which is equivalent to the vector equation:

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \qquad \text{or, in vector form:} \qquad \mathbf{A}\,\mathbf{x} = \mathbf{b}.$$

In our example the line does not go through all three points, so this equation is not solvable.

In statistical modeling, one often wishes to estimate certain parameters x_j based on some observations, where the observations are contaminated by noise. For instance, one wishes to predict the college grade point average (GPA) of freshman applications, which we denote by *b*, based on their high school GPA, denoted by a_1, and two Scholastic Aptitude Test scores, one verbal (a_2) and one quantitative (a_3), as part of the college admission process. It is a custom to use a linear model in this case: $b = a_1x_1 + a_2x_2 + a_3x_3$. ■

Now we formulate the least square problem. Suppose that we have a set (usually linearly independent, but not necessarily) of vectors $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ and a given vector **b**. We seek coefficients $x_1, x_2, \ldots, x_n$ that produce a minimal error

$$\left\| \mathbf{b} - \sum_{i=1}^{n} x_i\,\mathbf{a}_i \right\|.$$

with respect to the Euclidean inner product on $\mathbb{R}^n$. As long as we are interested only in linear combinations of a finite set $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$, it is possible to transform the problem into one involving finite columns of numbers. In this case, define a matrix **A** with columns given by the linearly independent column vectors $\mathbf{a}_i$, and a vector **x** whose entries are the unknown coefficients x_i. Then matrix **A** has dimensions *m* by *n*, meaning that $\mathbf{a}_i$ are vectors of length *m*. The problem then can be reformulated by choosing **x** that minimizing $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ because $\mathbf{A}\mathbf{x} = \mathbf{b}$ is an inconsistent linear system of *m* equations in *n* unknowns. We call such a vector, if it exists, a **least squares solution** $\mathbf{A}\hat{\mathbf{x}} = \mathbf{b}$, and call $\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}$ the least squares error vector. Statisticians call it *linear regression*.

If column vectors $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ are linearly independent (which will be assumed), matrix **A** becomes full column rank. The vector $\hat{\mathbf{x}} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n$ that is closest to a given vector **b** is its orthogonal projection onto the subspace spanned by the **a**'s. The hat over $\hat{\mathbf{x}}$ indicates the best choice (in the sense that it minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$) that gives the closest vector in the column space. Therefore, to solve the least square problem is equivalent to find the orthogonal projection matrix **P** on the column space such that $\mathbf{P}\mathbf{b} = \mathbf{A}\,\hat{\mathbf{x}}$.

The vector $\hat{\mathbf{x}}$ is a solution to the least squares problem when the error vector $\mathbf{e} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}}$ is perpendicular to the subspace. Therefore, this error vector makes right angle with all the vectors $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$. That gives the *n* equations we need to find $\hat{\mathbf{x}}$:

$$\begin{matrix} \mathbf{a}_1^{\mathrm{T}}\,(\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}) = 0 \\ \vdots \quad \vdots \\ \mathbf{a}_n^{\mathrm{T}}\,(\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}) = 0 \end{matrix} \qquad \text{or} \qquad \begin{bmatrix} - - - \,\mathbf{a}_1^{\mathrm{T}}\, - - - \\ \vdots \\ - - - \,\mathbf{a}_n^{\mathrm{T}}\, - - - \end{bmatrix} \begin{bmatrix} \\ \mathbf{b} - \mathbf{A}\hat{\mathbf{x}} \\ \\ \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The matrix in the right equation is $\mathbf{A}^{\mathrm{T}}$. The *n* equations are exactly

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\,\hat{\mathbf{x}} = \mathbf{A}^{\mathrm{T}}\,\mathbf{b}.$$

This equation is called the **normal equation**.
Now we derive the normal equation using another approach: calculus. To find the best approximation, we look for the **x** where the gradient of $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = (\mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\,(\mathbf{A}\mathbf{x} - \mathbf{b})$ vanishes. So we want

$$\begin{aligned} 0 &= \lim_{\epsilon \mapsto 0} \frac{(\mathbf{A}(\mathbf{x} + \epsilon) - \mathbf{b})^{\mathrm{T}}\,(\mathbf{A}(\mathbf{x} + \epsilon) - \mathbf{b}) - (\mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\,(\mathbf{A}\mathbf{x} - \mathbf{b})}{\|\epsilon\|_2} \\ &= \lim_{\epsilon \mapsto 0} \frac{2\epsilon^{\mathrm{T}}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{A}^{\mathrm{T}}\mathbf{b}\right) + \epsilon^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\epsilon}{\|\epsilon\|_2}. \end{aligned}$$

The second term

$$\frac{\left|\epsilon^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\epsilon\right|}{\|\epsilon\|_2} \leq \frac{\|\mathbf{A}\|_2^2\|\epsilon\|_2^2}{\|\epsilon\|_2} = \|\mathbf{A}\|_2^2\|\epsilon\|_2 \;\mapsto\; 0$$

approaches 0 as ε goes to 0, so the factor $\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{A}^{\mathrm{T}}\mathbf{b}$ in the first term must also be zero, or $\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{A}^{\mathrm{T}}\mathbf{b}$. This is a system of *n* linear equations in *n* unknowns, the normal equation.

Why is $\hat{\mathbf{x}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\,\mathbf{b}$ the minimizer of $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$? We can note that the Hessian $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is positive definite, which means that the function is strictly convex and any critical point is a global minimum. ■

Theorem: Suppose that $\{(x_i, y_i)\}_{i=1}^n$ are *n* points, where the abscissas x_i are distinct. The coefficients of the least squares line

$$y = m\,x + b$$

are the solution to the following linear system, known as the normal equations:

$$\left(\sum_{i=1}^{n} x_i^2\right) m + b\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} x_i y_i,$$
$$\left(\sum_{i=1}^{n} x_i\right) m + n\,b = \sum_{i=1}^{n} y_i.$$

A matrix is **full row rank** when each of the rows of the matrix are linearly independent and **full column rank** when each of the columns of the matrix are linearly independent. For a square matrix these two concepts are equivalent and we say the matrix is **full rank** if all rows and columns are linearly independent. A square matrix is full rank if and only if its determinant is nonzero.

For a non-square matrix with *m* rows and *n* columns, it will always be the case that either the rows or columns (whichever is larger in number) are linearly dependent. Hence when we say that a non-square matrix is full rank, we mean that the row and column rank are as high as possible, given the shape of the matrix. So if there are more rows than columns (*m*>*n*), then the matrix is full rank if the matrix is full column rank.

Suppose that $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ is an inconsistent a linear system of *m* algebraic equations in *n* unknowns. We suspect that inconsistency is caused either by too many equations or by errors in the entries of matrix **A** or input vector **b**. Usually, there are more equations than unknowns (*m* is greater than *n*). The *n*-dimensional column space constitutes a small part of *m*-dimensional space, and **b** is outside the column space. Since no exact solution is possible, we seek a vector $\hat{\mathbf{x}}$ that comes as close as possible.

Least Squares Problem: Given a linear system $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ of *m* equations in *n* unknowns, find a vector **x** that minimizes $\|\mathbf{b} - \mathbf{A}\,\mathbf{x}\|$ with respect to the Euclidean inner product on $\mathbb{R}^m$. Such vector, if it exists, is called a **least squares solution** of $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, which is usually denoted as $\hat{\mathbf{x}}$ The difference $\mathbf{e} = \mathbf{b} - \mathbf{A}\,\mathbf{x}$ is referred to as the **least squares error vector**, and we call $\|\mathbf{b} - \mathbf{A}\,\mathbf{x}\|$ the **least squares error**.

Example: Simple least squares problem: fitting a straight line.

Theorem (Best Approximation Theorem): If *U* is a finite-dimensional subspace of an inner product space *V*, and if **b** is a vector in *V*, then its projection on *U* is the best approximation to **b** from *U* in the sense that

$$\|\mathbf{b} - \mathrm{proj}_U \mathbf{b}\| < \|\mathbf{b} - \mathbf{u}\|$$

for every vector **u** in *U* that is different from proj$_U$**b**.

Proof:

From the above theorem, it follows that if $V = \mathbb{R}^n$ and *U* is the column space of matrix **A**, then the best approximation to **b** from the column space is its projection. But every vector in the column space of **A** is expressible in the form **Ax** for some vector **x**, so there is at least one vector $\hat{\mathbf{x}}$ in the column space of **A** for which $\mathbf{A}\,\hat{\mathbf{x}} = \mathrm{proj}_{column\ space}\mathbf{b}$. Each such vector is a least squares solution of **Ax** = **b**. Although there may be more than one least squares solution of **Ax** = **b**, each such solution $\hat{\mathbf{x}}$ has the same error vector $\mathbf{b} - \mathbf{A}\,\hat{\mathbf{x}}$.

If the kernel of **A** is nontrivial, there can be many vectors $\hat{\mathbf{x}}$ that satisfy the equation $\mathbf{A}\,\hat{\mathbf{x}} = \mathrm{proj}_U\mathbf{b}$, where *U* is the column space of matrix **A**. Fortunately, all these projections are the same. If the matrix **A** is of full column rank, then the square symmetric matrix $\mathbf{A}^{\mathrm{T}}\,\mathbf{A}$ is invertible. Upon application of its inverse, we find the unique least squares solution:

$$\hat{\mathbf{x}} - \left(\mathbf{A}^{\mathrm{T}}\,\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b} \qquad \text{or} \qquad \hat{\mathbf{x}} - \left(\mathbf{A}^{*}\,\mathbf{A}\right)^{-1}\mathbf{A}^{*}\mathbf{b}$$

if matrix **A** has complex entries.

Let **A** be a rectangular m × n matrix. A **pseudoinverse** of **A** is a rectangular n × m matrix, denoted by $\mathbf{A}^{\dagger}$, that satisfies the following four criteria, known as the Moore--Penrose conditions.

1. $\mathbf{A}\,\mathbf{A}^{\dagger}\,\mathbf{A} = \mathbf{A}$ (matrix $\mathbf{A}\,\mathbf{A}^{\dagger}$ maps all column vectors of **A** to themselves)
2. $\mathbf{A}^{\dagger}\mathbf{A}\,\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}$
3. $\left(\mathbf{A}\,\mathbf{A}^{\dagger}\right)^{*} = \mathbf{A}\,\mathbf{A}^{\dagger}$ (matrix $\mathbf{A}\,\mathbf{A}^{\dagger}$ is self-adjoint)
4. $\left(\mathbf{A}^{\dagger}\mathbf{A}\right)^{*} = \mathbf{A}^{\dagger}\mathbf{A}$ (matrix $\mathbf{A}^{\dagger}\mathbf{A}$ is self-adjoint)

Here $\mathbf{A}^{*} = \overline{\mathbf{A}^{\mathrm{T}}}$ is the adjoint of matrix **A**. The pseudoinverse $\mathbf{A}^{\dagger}$ exists for any matrix **A**, but when the latter has full rank, $\mathbf{A}^{\dagger}$ can be expressed as a simple algebraic formula

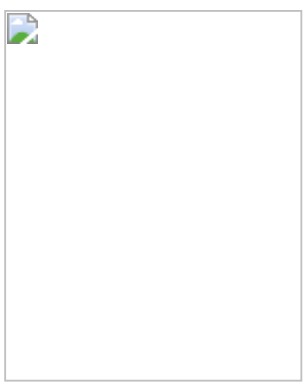
$$\mathbf{A}^{\dagger} = (\mathbf{A}^{*}\,\mathbf{A})^{-1}\mathbf{A}^{*}.$$

This particular pseudoinverse constitutes a **left inverse**, since, in this case, $\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{I}$, the identity m × m matrix.

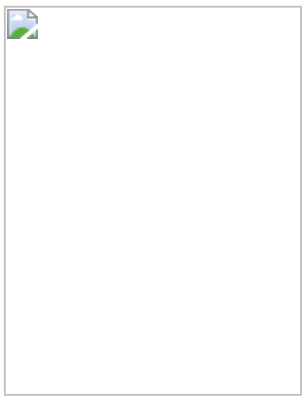
When **A** has linearly independent rows (matrix$\mathbf{A}\,\mathbf{A}^{\dagger}$ is invertible), $\mathbf{A}^{\dagger}$ can be computed as:

$$\mathbf{A}^{\dagger} = \mathbf{A}^{*}(\mathbf{A}\,\mathbf{A}^{*})^{-1}.$$

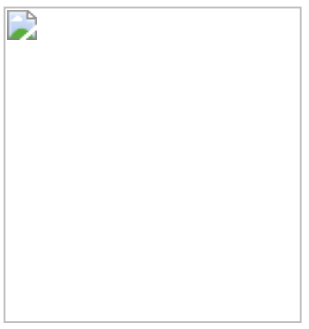
This is a right inverse because $\mathbf{A}\,\mathbf{A}^{\dagger} = \mathbf{I}$. ■



E. H. Moore.



Roger Penrose, 2011.



Erik Ivar Fredholm.

The term pseudoinverse was independently described by E. H. Moore (1862--1932) in 1920, Arne Bjerhammar (1917--2011) in 1951, and Roger Penrose (born in 1931) in 1955. Earlier, Erik Ivar Fredholm (1866--1927) had introduced the concept of a pseudoinverse of integral operators in 1903.

The American mathematician Eliakim Hastings Moore learned mathematics at Yale University. He got the Ph.D. in 1885 with a thesis, supervised by Hubert Anson Newton, on some work of William Kingdon Clifford and Arthur Cayley. Newton encouraged Moore to study in Germany, and thus he spent an academic year at the University of Berlin, attending lectures by Kronecker and Weierstrass. On his return to the United States, Moore taught at Yale and at Northwestern University. When the University of Chicago opened its doors in 1892, Moore was the first head of its mathematics department, a position he retained until his death.

Arne Bjerhammar was a Swedish geodesist. He was professor at Royal Institute of Technology in Stockholm. He was born in Båstad, Scania in the south of Sweden. He developed a method used to determine the geoid in gravimetric data, as well as a system for electro-optical measuring of distances. Seven years after introducing pseudoinverse, fascinated by M.S. Molodensky's new approach to solve the basic problems of physical geodesy, he presented his original idea of analytical downward continuation of the gravity anomaly to an internal sphere.

Sir Roger Penrose is an English mathematical physicist, mathematician and philosopher of science. Penrose told a Russian audience that his grandmother had left St. Petersburg in the late 1880s. His uncle was artist Roland Penrose, whose son with photographer Lee Miller is Antony Penrose. Penrose is the brother of physicist Oliver Penrose and of chess Grandmaster Jonathan Penrose. Penrose attended University College School and University College, London, where he graduated with a first class degree in mathematics. In 1955, while still a student, Penrose reintroduced the generalized matrix inverse. Penrose finished his PhD at Cambridge in 1958, with a thesis on "tensor methods in algebraic geometry" under algebraist and geometer John A. Todd. He devised and popularized the Penrose triangle in the 1950s, describing it as "impossibility in its purest form" and exchanged material with the artist M. C. Escher, whose earlier depictions of impossible objects partly inspired it.

The Swedish mathematician Erik Ivar Fredholm is remembered for his work on integral equations and operator theory foreshadowed the theory of Hilbert spaces. He obtained his PhD at Uppsala University in 1898, under the supervision of Gösta Mittag-Leffler. He was docent at Stockholm University from 1898 to 1906 and professor from 1906 until his death.

Example: Standard basis

Example: Set of rectangular matrices

Example: Set of polynomials of degree up to n

Example: Infinite set of all polynomials.