The Mathematician Educator
2023, Vol 4, No. 1, 55-65

Least Square Estimation and Orthogonality

YAP Von Bing
Department of Statistics and Data Science
National University of Singapore

Two measurement problems give rise to intractable systems of linear equations. By making a heuristic assumption on the measurement errors, least square estimates emerge as approximate solutions. An elegant generalisation to more complicated problems is afforded by the theory of Euclidean space, particularly orthogonal projection. This is an attempt to make the connection explicit. Orthogonality still plays a crucial role in the current, statistical view on least square estimation, which lies in the heart of the ubiquitous technique of linear regression.

Keywords: system of linear equations, least square estimation, Euclidean space, dot product, linear combination, subspace, basis, orthogonal projection, linear regression, random variables

Least square estimation, in linear regression or other kinds of problem, is one of the most widely used numerical techniques in data analysis. In undergraduate education, the topic is treated in greater detail in statistic courses than mathematics courses, but neither are likely to draw out the intimate connection with orthogonal projection in Euclidean spaces. Perhaps the mathematicians think it is too simple, and the statisticians think it is too hard. The statistical view is now predominant, though both deterministic and stochastic versions were invented at the about the same time, around 200 years ago. For a fascinating account of the least square's origin in the context of astronomical investigations, see Stigler (1990).

An outline of this article is as follows. In two simplest measurement problems, the desired quantities are unattainable, for they are solutions of intractable systems of linear equations. By making heuristic assumptions on the measurement errors, it becomes possible to estimate the quantities by minimising a function: the least square estimates. They turn out to be intimately connected with the Euclidean space, which is motivated by coordinate geometry. The orthogonal projection is defined and applied to the two problems, before its connection to a general measurement problem is elicited. Finally, a current statistical formulation of least square estimation is presented, where certain stochastic assumptions are made on the measurement errors, which underpins the application of regression in all kinds of problems.

Case 1: Measuring a Constant

A number of measurements are made of a physical constant, such as the mass of a piece of metal. The same protocol is used throughout, and environmental factors are kept as constant as possible. Nevertheless, if the instrument is sufficiently sensitive, say the weighing scale reads to the nearest microgram, the measurements are unlikely to be all the same. There are errors in the measurements, which may be taken to be additive. Let m be the true mass, and $y_1, \dots, y_n$ be the measurements. Then we can write

$$y_i = m + e_i, \qquad i = 1, \dots, n, \qquad (1)$$

where $e_1, \dots, e_n$ are the measurement errors. The goal is to determine the value of m.

The measurements $(y_1, \dots, y_n)$ are known. $(e_1, \dots, e_n, m)$ are unknown, and are a solution to the system of equations

$$x_i + x_{n+1} = y_i, \ i = 1, \dots, n. \qquad (2)$$

Since the system has only n equations, it has no unique solution. In fact, there are infinitely many solutions. Clearly, one solution is $(y_1, \dots, y_n, 0)$. It cannot be correct, for $m > 0$. Let λ be any real number. Then $(y_1 - \lambda, \dots, y_n - \lambda, \lambda)$ is another solution. We have obtained an infinite set of solutions, which can be written as

$$\{(y_1 - \lambda, \dots, y_n - \lambda, \lambda) : \lambda \in \mathbb{R}\}. \qquad (3)$$

Conversely, any solution must belong to this set. Let $e_1^*, \dots, e_n^*, m^*$ be a solution, i.e.,

$$e_i^* + m^* = y_i, \qquad i = 1, \dots, n.$$

Consequently, for every i, $e_i^* = y_i - m^*$. Hence, it is indeed of the given form, with $\lambda = m^*$. The correct solution, with $\lambda = m$, lies in the set, but we have no way of identifying it from the infinite possibilities.

It seems likely that numerous experiences gave rise to the hunch that the errors tend to cancel each other out, so that their mean $\bar{e}$ should be quite close to 0. This is a key insight, for then the mean of the measurements will be close to m:

$$\bar{y} = m + \bar{e} \approx m.$$

We say $\bar{y}$ is an estimate of m. The deviations of the measurements are

$$d_i = y_i - \bar{y}, \qquad i = 1, \dots, n.$$

Since $e_i = y_i - m \approx y_i - \bar{y}$, d_i is an estimate of e_i. Note that $\sum_{i=1}^n d_i = 0$: the deviations cancel each other like ideal errors, while the actual errors may not.

Let us look at the problem from another angle. Suppose we use z as an estimate of m. The heuristic that errors should cancel suggests the quality of z can be gauged by how close it is to all the measurements. A possible distance is the sum of absolute differences $\sum_{i=1}^n |y_i - z|$, though for ease of manipulation the sum of squared differences is preferable:

$$S(z) = \sum_{i=1}^{n} (y_i - z)^2.$$

It turns out that $z = \bar{y}$ is the unique minimiser of $S(z)$. Hence $\bar{y}$ is called the least square estimate of m. Indeed, since the deviations sum to 0,

$$S(z) = \sum_{i=1}^{n} ([y_i - \bar{y}] + [\bar{y} - z])^2 = \sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n} d_i(\bar{y} - z) + n(\bar{y} - z)^2$$

$$= \sum_{i=1}^{n} d_i^2 + n(\bar{y} - z)^2,$$

which has the minimum value $\sum_{i=1}^n d_i^2$ at $z = \bar{y}$, and at no other value.

Clearly, the further $\bar{e}$ is away from 0, the worse $\bar{y}$ estimates m. It is not possible to verify that $\bar{e} \approx 0$, for the mean of the deviations is 0. Rather, the measurement protocol has to be tested against an external standard, such as an object of known weight. Then the associated errors are known too. This kind of comparison is regularly conducted in many government laboratories, and plays a key role in calibrating measurement systems.

In summary, here is a guide on analysing measurements of an unknown constant m:

$$y_i = m + e_i, \qquad i = 1, \dots, n.$$

Suppose the errors $e_1, \dots, e_n$ roughly cancel each other, i.e., their mean $\bar{e} \approx 0$. Then m can be estimated satisfactorily by the mean of the measurements $\bar{y}$, which is the least square estimate.

Case 2: Measuring a Constant Effect

Suspend a weight from a metal rod, and its length will increase by an amount proportional to the weight, provided the weight does not exceed a certain amount, called the elastic limit. Suppose for $i = 1, \dots, n$, the length y_i is measured when the known weight x_i is suspended. We assume $x_1, \dots, x_n$ are all less than the elastic limit and are not all equal. Then we have

$$y_i = c + bx_i + e_i, \qquad i = 1, \dots, n \qquad (4)$$

where c and b are constants to be estimated. c is the natural length of the rod; b is the extension per unit weight in some unit, which might be called the effect of weight on length[1]. The errors made in the length measurements $e_1, \dots, e_n$ are also unknown.

Like in the previous problem, we assume that

$$\bar{e} = \frac{1}{n}\sum_i^n e_i \approx 0, \qquad \overline{xe} = \frac{1}{n}\sum_i^n x_i\, e_i \approx 0. \qquad (5)$$

The first says the errors roughly cancel each other. The second says that the errors are independent of the weights. We seek the least square estimates of c and b, by minimising

$$S(z_1, z_2) = \sum_{i=1}^n (y_i - [z_1 + z_2 x_i])^2.$$

For an elementary treatment of this minimisation, see the meticulously instructive Freedman & Lane (1981). Here, we use calculus. The two partial derivatives are

$$S_1(z_1, z_2) = -2\sum_{i=1}^n (\, y_i - z_1 - z_2 x_i),$$

$$S_2(z_1, z_2) = -2\sum_{i=1}^n x_i(\, y_i - z_1 - z_2 x_i).$$

Setting both to 0 gives

$$\bar{y} - z_1 - z_2\bar{x} = \overline{xy} - z_1\bar{x} - z_2\overline{x^2} = 0, \qquad (6)$$

where

$$\bar{x} = \frac{1}{n}\sum_i^n x_i\,, \qquad \bar{y} = \frac{1}{n}\sum_i^n y_i, \qquad \overline{x^2} = \frac{1}{n}\sum_i^n x_i^2, \qquad \overline{xy} = \frac{1}{n}\sum_i^n x_i\, y_i\,.$$

Solving the equations gives the unique stationary point:

$$z_2 = \hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \qquad z_1 = \hat{c} = \bar{y} - \hat{b}\bar{x},$$

That this is indeed a minimum can be confirmed by checking that the second derivative matrix, the Hessian, has positive diagonal entries and a positive determinant. As will be seen, the theory of Euclidean space offers a purely algebraic justification.

[1] The constants have units, like metre for c and metre per kilogram for b, though this does not concern us here.

Let us verify that the estimates make sense. Putting (5) into (4) gives $\bar{y} \approx c + b\bar{x}$ and $\overline{xy} \approx c\bar{x} + b\overline{x^2}$. Hence

$$\hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \approx \frac{c\bar{x} + b\overline{x^2} - \bar{x}(c + b\bar{x})}{\overline{x^2} - \bar{x}^2} = b,$$

meaning $\hat{b}$ is quite accurate, which implies the same for $\hat{c}$, since $\hat{c} = \bar{y} - \hat{b}\bar{x} \approx c$. The residuals are

$$d_i = y_i - (\hat{c} + \hat{b}x_i), \qquad i = 1, \dots, n,$$

which are estimates of the errors, like the deviations in the previous case. Substituting $z_1 = \hat{c}$ and $z_2 = \hat{b}$ into (6) gives

$$\sum_i^n d_i = \sum_i^n x_i d_i = 0.$$

Hence the residuals behave like ideal errors. As before, assumptions (5) must be checked by pitting the protocol for measuring length against an external standard, such as a rod of known length and known extension per unit weight.

An Orthogonal Interlude

The two cases are the simplest examples of least square estimation. The associated mathematics is worth studying due to its practical utility, and because it contains the germ for the general theory. This section attempts to present the concepts from Euclidean spaces that are necessary to grasp the relevance of orthogonal projection. Two key theorems are proved using a collection of intuitive facts. For further details on these prerequisite results, the interested reader may consult any of the numerous textbooks, such as Blyth & Robertson (2002) or Lang (1997).

Imagine a straight rod sticking out of flat ground at an angle, i.e., OP is not vertical, where O is the point of contact with the ground and P is the other end of the rod. A vertical light casts a shadow OQ of the rod on the ground, so that PQ is vertical, or perpendicular to the ground. We say the shadow is the orthogonal projection of the rod on the ground. Similarly, in the plane, given a line through the origin O and a point P not on the line, we can construct a point Q on the line, so that OQP is a right angle. OQ is the orthogonal projection of OP on the line. It turns out that orthogonal projection in high-dimensional spaces offers a beautiful algebraic solution of the minimization problem in least square estimation. To get there, we need some working knowledge of the Euclidean space, which will be presented in three stages.

The first stage is an outgrowth of coordinate geometry. For a positive integer n, the n-dimensional Euclidean space $\mathbb{R}^n$ is the set of vectors consisting of n real numbers. Intuitively, the vector $\boldsymbol{x} = (x_1,\dots, x_n)$ specifies a point P in an abstract space. Let $\boldsymbol{y} = (y_1,\dots, y_n)$ specify the point Q. The dot product of $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^n x_i y_i \,.$$

The distance between P and Q is $|\boldsymbol{x} - \boldsymbol{y}| = \sqrt{(\boldsymbol{x} - \boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})}$. In particular, $|\boldsymbol{x}|$ is the distance between P and the origin O, specified by $\boldsymbol{0} = (0,\dots, 0)$. If $\boldsymbol{x} \cdot \boldsymbol{y} = 0$, $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal. Let $\boldsymbol{z} = (z_1,\dots, z_n)$ specify the point R, and suppose P, Q and R are distinct. The line segments PR and QR are perpendicular, written $PR \perp QR$, if $\boldsymbol{z} - \boldsymbol{x}$ and $\boldsymbol{z} - \boldsymbol{y}$ are orthogonal. The Pythagoras Theorem says that if $PR \perp QR$, then $|\boldsymbol{x} - \boldsymbol{y}|^2 = |\boldsymbol{z} - \boldsymbol{x}|^2 + |\boldsymbol{z} - \boldsymbol{y}|^2$. The definitions of distance, orthogonality and perpendicularity, and hence the

Pythagoras Theorem, are in complete accord with the geometry of Euclid. This is because the definitions are chosen such that $\mathbb{R}^3$ and $\mathbb{R}^2$ correspond to space and a plane respectively.

In low dimensions ($n = 2, 3$), the distance formula is a fact[2] about the plane and the space. Similarly, $\boldsymbol{x} \cdot \boldsymbol{y} = |\boldsymbol{x}||\boldsymbol{y}| \cos\theta$, where θ is the angle between OP and OQ. Suppose $\boldsymbol{x}$ is not equal to a constant multiple of $\boldsymbol{y}$, i.e., OPQ is not a straight line, and that θ is acute, as in the figure. Let $\boldsymbol{u} = \boldsymbol{x}/|\boldsymbol{x}|$, so that $|\boldsymbol{u}| = 1$. It follows from trigonometry that OR, the orthogonal projection of OQ on OP, has length $|\boldsymbol{y}| \cos\theta = \boldsymbol{u} \cdot \boldsymbol{y}$. Hence the point R is specified by $(\boldsymbol{u} \cdot \boldsymbol{y})\boldsymbol{u} = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}|^2}\boldsymbol{x}$. Indeed, the fact that $OR \perp RQ$ is readily verified by

$$\boldsymbol{x} \cdot \left(\boldsymbol{y} - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}|^2}\boldsymbol{x}\right) = 0.$$

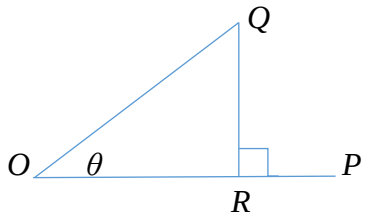

Figure: OR is the orthogonal projection of *OQ* to *OP*

Definition 1. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ be non-zero vectors, such that $\boldsymbol{x}$ is not equal to a constant multiple of $\boldsymbol{y}$. The orthogonal projection of $\boldsymbol{y}$ on the line containing $\boldsymbol{x}$ is defined as $\frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}|^2}\boldsymbol{x}$.

This is the simplest case of orthogonal projection, but is sufficient to apply to Case 1.

Example: Measuring a constant The measurements $y_1, \dots, y_n$ form a vector $\boldsymbol{y} \in \mathbb{R}^n$. Let $\boldsymbol{z} = z\boldsymbol{x}$, where $\boldsymbol{x} = (1, \dots, 1)$, so that $S(z) = |\boldsymbol{y} - \boldsymbol{z}|^2$. Since $S(z)$ is minimised at $z = \bar{y}$, its minimum value is $|\boldsymbol{y} - \bar{y}\boldsymbol{x}|^2$. Now we have $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^n y_i$ and $|\boldsymbol{x}|^2 = n$, so the orthogonal projection of $\boldsymbol{y}$ on the line containing $\boldsymbol{x}$ is $\bar{y}\boldsymbol{x}$. The proximity of the orthogonal projection to the least square estimate is not a coincidence, as we will see later. Notice that (1) can be written

$$\boldsymbol{y} = m\boldsymbol{x} + \boldsymbol{e}$$

where $\boldsymbol{e} = (e_1, \dots, e_n)$ is the vector of measurement errors.

The second stage makes the rough idea of the "line containing $\boldsymbol{x}$" in Definition 1 exact, which is often part of a course on vector spaces. Let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_h \in \mathbb{R}^n$ be non-zero vectors. Given coefficients $\lambda_1, \dots, \lambda_h \in \mathbb{R}$, $\lambda_1\boldsymbol{u}_1 + \cdots + \lambda_h\boldsymbol{u}_h$ is called a linear combination, which is also a vector. Denote the set of all linear combinations of $\boldsymbol{u}_1, \dots, \boldsymbol{u}_h$ by

$$\langle \boldsymbol{u}_1, \dots, \boldsymbol{u}_h \rangle = \{\lambda_1\boldsymbol{u}_1 + \cdots + \lambda_h\boldsymbol{u}_h : \lambda_1, \dots, \lambda_h \in \mathbb{R}\}.$$

This set is called the subspace generated by $\boldsymbol{u}_1, \dots, \boldsymbol{u}_h$. Clearly, every subspace contains $\boldsymbol{0}$. In $\mathbb{R}^2$, lines that pass through the origin are one-dimensional subspaces. In $\mathbb{R}^3$, planes and lines that pass through the origin are respectively two- and one-dimensional subspaces. Here are some general facts and definitions about the subspace $V = \langle \boldsymbol{u}_1, \dots, \boldsymbol{u}_h \rangle$ of $\mathbb{R}^n$. It is always possible to choose a subset $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\} \subseteq \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_h\}$, where $k \leq h$, such that for any $\boldsymbol{v} \in V$,

[2] This is empirical: the formula gives the distance between two points in a real plane or space to a high accuracy.

there is only one set of coefficients $c_1, \dots, c_k$ such that $\boldsymbol{v} = c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k$. In particular, $V = \langle \boldsymbol{v}_1, \dots, \boldsymbol{v}_k \rangle$. The dimension of V is $\dim(V) = k$, and $c_1, \dots, c_k$ are the coordinates of $\boldsymbol{v}$ with respect to the basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$. Among the generating vectors, we might be able to choose different bases, but the number of vectors in any basis is always k. For example, let $\boldsymbol{v}_1 = (1,0,0)$, $\boldsymbol{v}_2 = (0,1,0)$, $\boldsymbol{v}_3 = (1,1,0)$. Since $(0,0,0) = 0\boldsymbol{v}_1 + 0\boldsymbol{v}_2 + 0\boldsymbol{v}_3 = 1\boldsymbol{v}_1 + 1\boldsymbol{v}_2 - 1\boldsymbol{v}_3$, the three vectors are not a basis of $V = \langle (1,0,0), (0,1,0), (1,1,0) \rangle$, the xy-plane in $\mathbb{R}^3$. Clearly, $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$ is a basis, so $\dim(V) = 2$. In fact, any two vectors form a basis. For $k < n$, a k-dimensional subspace of $\mathbb{R}^n$ is like a copy of $\mathbb{R}^k$ in $\mathbb{R}^n$.

The standard basis of $\mathbb{R}^n$ consists of the n vectors $(1,0,0,\dots,0,0)$, $(0,1,0,\dots,0,0)$, ..., $(0,0,0,\dots,0,1)$. The coordinates of any vector with respect to the standard basis are none other than its component entries. A basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ is orthonormal if $i = 1, \dots, k$, $|\boldsymbol{v}_i| = 1$ and $\boldsymbol{v}_i \cdot \boldsymbol{v}_j = 0$ for any $i \neq j$. The standard basis of $\mathbb{R}^n$ is orthonormal. Any basis can be converted into an orthonormal basis for the same subspace, by the Gram-Schmidt process.

We are now ready for the general concept of orthogonal projection.

Theorem 1. Let V be a subspace with an orthonormal basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$. Given $\boldsymbol{y} \in \mathbb{R}^n$, define

$$\boldsymbol{y}_V = (\boldsymbol{y} \cdot \boldsymbol{v}_1)\boldsymbol{v}_1 + \cdots + (\boldsymbol{y} \cdot \boldsymbol{v}_k)\boldsymbol{v}_k \, .$$

(i) $\boldsymbol{y}_V \in V$, and $\boldsymbol{y} - \boldsymbol{y}_V \perp V$, meaning $\boldsymbol{y} - \boldsymbol{y}_V$ is orthogonal to every vector of V.

(ii) The squared distance $|\boldsymbol{y} - \boldsymbol{z}|^2$, with $\boldsymbol{z} \in V$, is uniquely minimized by $\boldsymbol{z} = \boldsymbol{y}_V$.

Proof. Being a linear combination of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$, $\boldsymbol{y}_V \in V$. Since $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ are an orthonormal basis, for each $\boldsymbol{v}_i$, $\boldsymbol{y}_V \cdot \boldsymbol{v}_i = \boldsymbol{y} \cdot \boldsymbol{v}_i$, or $(\boldsymbol{y} - \boldsymbol{y}_V) \cdot \boldsymbol{v}_i = 0$. Hence $\boldsymbol{y} - \boldsymbol{y}_V \perp V$. Next, given $\boldsymbol{z} \in V$, we have $\boldsymbol{y}_V - \boldsymbol{z} \in V$. From (i), $(\boldsymbol{y} - \boldsymbol{y}_V) \cdot (\boldsymbol{y}_V - \boldsymbol{z}) = 0$, implying

$$|\boldsymbol{y} - \boldsymbol{z}|^2 = |(\boldsymbol{y} - \boldsymbol{y}_V) + (\boldsymbol{y}_V - \boldsymbol{z})|^2 = |\boldsymbol{y} - \boldsymbol{y}_V|^2 + |\boldsymbol{y}_V - \boldsymbol{z}|^2.$$

Hence $|\boldsymbol{y} - \boldsymbol{z}|^2$ is minimised by $\boldsymbol{z} = \boldsymbol{y}_V$. It is unique, for if there is another minimiser $\boldsymbol{y}_V^*$, then replacing $\boldsymbol{z}$ by $\boldsymbol{y}_V^*$ in the equation implies $|\boldsymbol{y}_V - \boldsymbol{y}_V^*|^2 = 0$, which contradicts the assumption that $\boldsymbol{y}_V^* \neq \boldsymbol{y}_V$. ■

Definition 2. Let $\boldsymbol{y} \in \mathbb{R}^n$, and V be a subspace. $\boldsymbol{y}_V$ is the orthogonal projection of $\boldsymbol{y}$ on V.

Theorem 1 (ii) implies that $\boldsymbol{y}_V$ is the unique vector in V with property (i). Since $\boldsymbol{y} - \boldsymbol{y}_V \perp \boldsymbol{y}_V$,

$$|\boldsymbol{y}|^2 = |\boldsymbol{y} - \boldsymbol{y}_V|^2 + |\boldsymbol{y}_V|^2,$$

which can be visualised as a right triangle, in any dimension.

The third stage involves elementary facts of matrix algebra, to prepare for Theorem 2, which generalises the formula in Definition 1, with multiplication by a matrix inverse replacing division. A $k \times k$ matrix $\boldsymbol{A}$ is invertible if there is a $k \times k$ matrix $\boldsymbol{B}$ such that $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$. $\boldsymbol{B}$ is called the inverse of $\boldsymbol{A}$, and we write $\boldsymbol{B} = \boldsymbol{A}^{-1}$. Given an $n \times k$ matrix $\boldsymbol{X}$ with entry a_{ij} at row i and column j, the transpose of $\boldsymbol{X}$ is the $k \times n$ matrix with a_{ij} at row j and column i, denoted by $\boldsymbol{X}'$. The rank of an $n \times k$ matrix is the dimension of the subspace of $\mathbb{R}^n$ generated by its columns. Hence its rank cannot be larger than n or k. If the rank of an $n \times k$ matrix $\boldsymbol{X}$ is k, then $\boldsymbol{X}'\boldsymbol{X}$ is invertible.

Theorem 2. Let $\boldsymbol{X}$ be the $n \times k$ matrix with $\boldsymbol{v}_j$ in column j, where the k vectors form a basis of a subspace V of $\mathbb{R}^n$. Let $\boldsymbol{y} \in \mathbb{R}^n$ be written as a column vector. Then the orthogonal projection of $\boldsymbol{y}$ on V is

$$\boldsymbol{y}_V = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Furthermore, the coordinates of $\boldsymbol{y}_V$ is $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Proof. Since $\boldsymbol{X}'\boldsymbol{X}$ is invertible, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is well-defined. $\boldsymbol{v}_i \cdot (\boldsymbol{y} - \boldsymbol{y}_V)$ is the i-th entry of the matrix product:

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{y}_V) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) = \boldsymbol{0}.$$

Hence $\boldsymbol{y} - \boldsymbol{y}_V$ is orthogonal to every vector in V. By Theorem 1, $\boldsymbol{y}_V$ is the orthogonal projection of $\boldsymbol{y}$ on V. Write the $k \times 1$ matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ as $(\lambda_1, \dots, \lambda_k)'$. Then $\boldsymbol{y}_V = \boldsymbol{X}(\lambda_1, \dots, \lambda_k)' = \lambda_1\boldsymbol{v}_1 + \cdots + \lambda_k\boldsymbol{v}_k$, which shows that $\boldsymbol{y}_V \in V$, and its coordinates are $\lambda_1, \dots, \lambda_k$. ■

Example: Measuring a constant effect Let $\boldsymbol{z} = z_1\boldsymbol{1} + z_2\boldsymbol{x}$, where $\boldsymbol{x} = (x_1, \dots, x_n)$ is the vector of weights. Since $S(z_1, z_2) = |\boldsymbol{y} - \boldsymbol{z}|^2$, Theorem 1 says the minimising $\boldsymbol{z}$ is the orthogonal projection $\boldsymbol{y}_V$, where $V = \langle \boldsymbol{1}, \boldsymbol{x} \rangle$. Let $\boldsymbol{X} = [\boldsymbol{1}\ \boldsymbol{x}]$. It is straightforward, if tedious, to check that $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\hat{c}\,, \hat{b})'$, the least square estimates. Since $x_1, \dots, x_n$ are not all equal, $\dim(V) = 2$. By Theorem 2, $\boldsymbol{y}_V = \boldsymbol{X}(\hat{c}\,, \hat{b})' = \hat{c}\boldsymbol{1} + \hat{b}\boldsymbol{x}$. Indeed, (6) holds with $z_1 = \hat{c}, z_2 = \hat{b}$, which says that $\boldsymbol{y} - (\hat{c}\boldsymbol{1} + \hat{b}\boldsymbol{x})$ is orthogonal to both $\boldsymbol{1}$ and $\boldsymbol{x}$, and therefore also orthogonal to every vector in V. Note that (4) can be written as $\boldsymbol{y} = \boldsymbol{X}(c, b)' + \boldsymbol{e}$, where $\boldsymbol{y}$ and $\boldsymbol{e}$ are the column vectors of measurements and errors.

It is marvelous that the orthogonal projection is the global minimiser of S in the two cases, so that there is no need for tedious computations involving the Hessian. The second case offers such a good glimpse into the general case that the subsequent connection should seem rather familiar.

General Case

Here is the general measurement problem. For $i = 1, \dots, n$, we fix the values of p variables, called covariates, at $x_{i1}, \dots, x_{ip}$, and measure the response variable to get y_i. Suppose there are constants $\beta_1, \dots, \beta_p$ such that

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i$$

where e_i is an unknown measurement error. For $j = 1, \dots, p$, β_j is the effect of the j-th covariate on the response variable, i.e., the change in the response caused by increasing the j-th covariate by 1, while keeping all other variables fixed. With $\boldsymbol{y} = (y_1, \dots, y_n)'$, $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)'$, and $\boldsymbol{e} = (e_1, \dots, e_n)'$, the equations can be written

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

It is impossible to determine $\boldsymbol{\beta}$, because there are more unknowns than equations. Suppose the $n \times p$ matrix $\boldsymbol{X} = (x_{ij})$ has rank p. Let $\boldsymbol{z} = (z_1, \dots, z_p)'$. The least square estimate of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, is the unique minimiser of the sum of squares

$$S(\boldsymbol{z}) = \sum_{i=1}^{n} |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{z}|^2.$$

By Theorem 2,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

and $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is the orthogonal projection of $\boldsymbol{y}$ on the column space of $\boldsymbol{X}$.

For $\widehat{\boldsymbol{\beta}}$ to be a satisfactory estimate of $\boldsymbol{\beta}$, some conditions on $\boldsymbol{e}$ are needed. A sufficient condition is that $\boldsymbol{e}$ is roughly orthogonal to V, in the sense that $\boldsymbol{X'e} \approx \boldsymbol{0}$. Indeed, the condition implies

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta} + \boldsymbol{e}) = \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'e} \approx \boldsymbol{\beta}.$$

The residuals

$$\boldsymbol{d} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

are estimates of the errors, and behave like ideal errors:

$$\boldsymbol{X'd} = \boldsymbol{X'y} - \boldsymbol{X'X}\widehat{\boldsymbol{\beta}} = \boldsymbol{0}.$$

The assumption $\boldsymbol{X'e} \approx \boldsymbol{0}$ has to be checked using another set of data where actual errors are known.

Stochastic Error

So far, our treatment of least square estimation is "deterministic", in the sense that the measurement errors are fixed unknown constants, and we are not concerned about how they come about. If they satisfy certain properties approximately, then our estimates are reasonable. It was clear to Gauss, one of the originators of least square estimation, that a useful theory of estimation is obtained if it is assumed that the errors were generated by certain random variables. This probabilistic or statistical view is a significant breakthrough, which has profound impact on the quantitative sciences even to this day.

Let us return to the first case, on measuring a constant. If the errors sum to about 0, then $\bar{y}$ is close to m. This condition is satisfied if $e_1, \ldots, e_n$ has an upward trend, going from negative to positive. However, the trend should concern us, for it suggests something is wrong with the measurement protocol. If the trend persists, $\bar{y}$ becomes increasingly larger than m as n gets larger. A similar issue arises if the errors have a systematic trend, which can be revealed by plotting the deviations $d_1, \ldots, d_n$ against $1, \ldots, n$. Perhaps the conditions were changing systematically despite the precautions. The protocol should be checked and rectified before repeating the measurements. If the errors were generated from a random mechanism, the graph should show no trend. This idea lies behind a statistical model for measurement, known as the Gauss model (Freedman et al., 2007).

Here are the details for case 1. Let $\epsilon_1, \ldots, \epsilon_n$ be independent and identically distributed (IID) random variables with $\mathrm{E}(\epsilon_i) = 0$, and $\mathrm{var}(\epsilon_i) = \sigma^2$. Define the random variables $Y_1, \ldots, Y_n$ by

$$Y_i = m + \epsilon_i, \qquad i = 1, \ldots, n.$$

For $i = 1, \ldots, n$, let e_i be a realisation of ϵ_i, which induces the realisation y_i of Y_i as follows:

$$y_i = m + e_i, \qquad i = 1, \ldots, n.$$

The y's are known, but m and the e's are unknown. These are exactly (1): the Gauss model generates the equations that started the discussion.

The stochastic assumption on the errors, that they come from IID random variables with expectation 0, is clearly an analogue of the assumption that $\bar{e} \approx 0$. $\bar{y}$ is a realization of the random variable

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

which has expectation m. Statisticians say $\bar{Y}$ is an unbiased estimator, and $\bar{y}$ is an unbiased estimate. This means that if the whole measurement protocol were repeated many times, the

respective estimates cluster around m. However, there is no assurance that any single estimate will be close to m. Thus unbiasedness is a property of the estimator $\bar{Y}$, not of an individual estimate. This point must be borne in mind if $\mathrm{E}(\bar{Y}) = m$ gives the impression that Gauss model has solved the basic difficulty of determining m.

Nevertheless, the Gauss model offers new insight on the error in the estimate. It implies that

$$\mathrm{var}(\bar{Y}) = \mathrm{E}(\bar{Y} - m)^2 = \frac{\sigma^2}{n},$$

meaning $\bar{y}$ gets closer to m as $n \to \infty$. Furthermore, the magnitude of the error, $\bar{y} - m$, is roughly the standard deviation of $\bar{Y}$. Thus, a statistician speaks of $\bar{y}$ as having a standard error (SE) of $\frac{\sigma}{\sqrt{n}}$. Since σ is unknown, it has to be estimated from the deviations

$$\sigma \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} d_i^2}.$$

Given that the deviations sum to 0, knowing any $n-1$ of them suffices to determine the last value. They are said to have $n-1$ degrees of freedom, and it turns out that $\frac{1}{n-1}\sum_{i=1}^{n} d_i^2$ is unbiased for σ^2. It is a good idea to plot $d_1, \ldots, d_n$ against $1, \ldots, n$ to check the stochastic assumption. If there is a clear trend, it is a sign that the Gauss model may not work well, i.e., the estimate and the SE may be unreliable.

Now we outline the general statistical model, known as linear regression. Let $\boldsymbol{X}$ be a known $n \times p$ matrix of rank p. Let $\epsilon_1, \ldots, \epsilon_n$ be IID random variables with $\mathrm{E}(\epsilon_i) = 0$, and $\mathrm{var}(\epsilon_i) = \sigma^2$. Define the random variables $Y_1, \ldots, Y_n$ by

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \qquad i = 1, \ldots, n.$$

Thus

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ $\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n)$. Let e_i be a realisation of ϵ_i, which induces the realisation y_i of Y_i, giving the equation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}.$$

The least square estimate of $\boldsymbol{\beta}$, $(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$, is a realisation of the random vector $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$. It is striking that the assumption $\mathrm{E}(\epsilon_i) = 0$ implies $\widehat{\boldsymbol{\beta}}$ is unbiased. The analogous deterministic assumption, that $\boldsymbol{X'e} \approx \boldsymbol{0}$, is more complicated. Since

$$\mathrm{var}\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X'X})^{-1}\sigma^2,$$

the standard error of $\hat{\beta}_j$ is the square root of the (j,j)-entry of $(\boldsymbol{X'X})^{-1}$, multiplied by σ. σ can be estimated as $\sqrt{\frac{1}{n-p}\sum_{i=1}^{n} d_i^2}$, where $\boldsymbol{d} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are the residuals. Like in the special case, it is a good idea to plot $d_1, \ldots, d_n$ against $1, \ldots, n$ to check that the errors seem random. As for the assumption that the errors have expectation 0, and in particular $\widehat{\boldsymbol{\beta}}$ is unbiased, the procedure must be tested against an external standard. There are many textbooks on the linear regression model. Some influential ones are Scheffé (1999) and Rao (2001).

Unlike earlier examples, the linear regression model is not explicitly set up as a measurement protocol. However, it is common to interpret β_j as the effect of the j-th covariate on the

response variable, i.e., it is the expected change in the response upon increasing the j-th covariate by 1, while keeping all other variables unchanged, as if the experimenter has control over the covariates. Clearly, this interpretation is not justified if the study is observational, where the covariate values are already fixed and merely measured by the investigator. The main issue is confounding: the effect of unobserved variables that may be wrongly attributed to the p covariates. For more details on this issue, see the insightful book by Freedman (2005).

Conclusion

The solution of the minimization problem in least square estimation via orthogonal projection is an algebraic *tour de force*, which completely dispenses with the intricacies of the calculus approach. If the initial influence of least square estimation was mainly scientific, it has recently also shone brightly in the underbelly of various machine learning and artificial intelligence algorithms, mainly as part of a stochastic view of measurement error. Going forward, its impact on a lot of computational activity is expected to be substantial. As such, a legitimate case might be made that as far as feasible, the theory should be widely disseminated. Perhaps this article can provide some initial impetus for such an endeavor.

Here is a summary of the path taken to orthogonal projection in a Euclidean space. It starts with dot product, distance between two points and orthogonality (or perpendicularity), which may be familiar in the low dimensions, say from coordinate geometry. Then enter the crucial concept of subspace and associated ideas like basis, coordinates, dimension, and orthonormal basis, which lead to the general construction of orthogonal projection (Theorem 1). The last part deals with requisite facts from matrix algebra for Theorem 2, which relates directly to least square estimation in linear regression. Besides featuring orthogonality early, the course deviates from the usual narratives in vector space or linear algebra in that subspaces are defined as the set of linear combinations generated by some vectors. The concrete definition is a good preparation for the standard equivalent definition: V is a subspace if (i) $\mathbf{0} \in V$, (ii) $\boldsymbol{v}_1, \boldsymbol{v}_2 \in V$, $\lambda \in \mathbb{R}$ imply $\lambda \boldsymbol{v}_1 + \boldsymbol{v}_2 \in V$. Moreover, a basis is defined as a set of generating vectors that assign unique coordinates for each vector in the subspace, without going through the concept of linear independence. This suffices for grasping the content of the theorems, though it is likely necessary to introduce linear independence in order to obtain more streamlined proofs of the supporting facts.

The orthogonal projection is an elegant abstraction of intuitive knowledge from low-dimensional geometry. Whenever a linear regression model is fitted to data, and this happens countless number of times everyday, an orthogonal projection is done, from which least square estimates may be extracted. Thus, the theory of Euclidean space deserves some degree of acquaintance by students of statistics, data science and other quantitative fields. It is also an excellent pitstop for those who venture onto the abstract world of vector spaces.

References

Blyth, T. S., & Robertson, E.F. (2002). *Basic Linear Algebra*. Springer.

Lang, S. (1997). *Introduction to Linear Algebra*. Springer.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Freedman, D. A., & Lane, D. (1981). *Mathematical Methods in Statistics*. Norton.

Freedman, D. A., Pisani, R. & Purves, R. (2007). *Statistics*. Norton.

Rao, C. R. (2001). *Linear Statistical Inference and Its Applications*. Wiley.

Scheffé, H. (1999). *The Analysis of Variance*. Wiley.

Stigler, S. M. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press.

Author:

YAP Von Bing, Associate Professor, Department of Statistics and Data Science, National University of Singapore, Singapore. E-mail: stayapvb@nus.edu.sg