Cartesian Statistics: Data Analysis by Linear Algebra and Analytic Geometry

STEPHEN WHITNEY*

God ever geometrizes.
Plato

Abstract. *A linear algebraic approach can be helpful in the computation, interpretation, teaching and understanding of statistics at all levels. Although the reference space has a dimension equal to the sample size, it is not necessary to consider more than 2 or 3 dimensions at a time, for most applications.*

A sample becomes a vector which has an orthogonal decomposition into mean and standard deviation projections. The correlation between two samples is essentially measured by the angle between the vectors. Even linear regression has an interpretation «dual» to the usual scatter diagram: the regression coefficients are scalar projections between the sample vectors. Orthogonal regression, somewhat neglected in the literature, is easily treated here.

This «cartesian» approach can be extended to other branches of statistics, separating, in true object-oriented fashion, the basic deterministic notions from the probabilistic and computational aspects. However, a brief examination of some references indicates only sporadic use of the cartesian approach in the learning and use of statistics.

A mathematics professor who teaches a statistical course for the first time in years (or ever), soon remembers (or realizes) that many areas of statistics can be taught from a «pure» (analytic and/or algebraic) standpoint, rather than an «applied» (probabilistic, computational, or even applications-oriented) one. The author was in such a situation recently, and he became curious about how much statistics could (should?) be presented by «pure» linear algebraic concepts.

The «extreme» linear algebraic approach presented here is not claimed to be the only, or even the best, way of looking at statistics. There are many statistical subjects (hypothesis testing, sampling theory, the definition of the common distributions, etc.) that are not easily seen through linear algebraic spectacles. However, a linear algebraic standpoint in statistics has its advantages (sometimes in common with other standpoints, such as a «pure» mathematically analytic one):

1) Visualization. As we shall see, many concepts of statistics are modelled by linear algebra, which in turn is essentially «seen» as analytic geometry. The vectors of linear algebra are points or vectors in n-space. Linear algebra does not have human visual and conceptual limitations concerning dimensions, and sometimes an n-dimensional «dual space» is easier to work in than is the plane or ordinary 3-space. In any case, a relationship between two or three vectors at a time, in n-space, only involves two or three dimensions, which can be easily visualized on a plane or in ordinary space. Linear transformations, represented numerically by matrices, are transformations of space (like expansions, contractions, projections, etc.) that preserve vector addition triangles and scalar multiplication. Inner products and the Cauchy-Schwarz inequality are related to the angle and the projections between two vectors, and norms are essentially lengths of vectors.

2) Understanding. Many, probably(!) most, of the *deterministic* concepts of statistics can be modelled by linear algebra. This helps to separate the deterministic concepts from the probabilistic ones, enabling a student or user of statistics to concentrate on one aspect at a time.

* *Département d'informatique et de mathématiques, Université du Québec, Chicoutimi, Qué., Canada G7H 2B1. Tel. 418 545 5011, ext. 5060. Fax: -5012, E-mail: swhitney@uqac.uquebec.ca.*

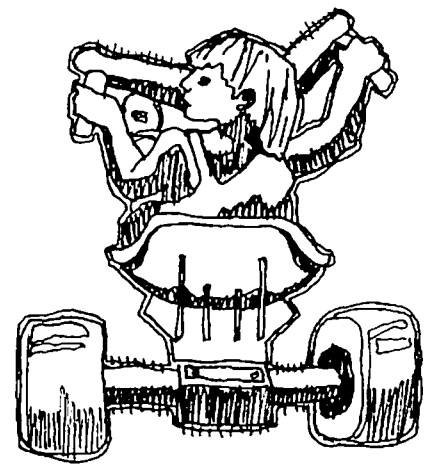
3) Computation. A linear algebraic approach is an object-oriented approach, separating the global parallel aspects of numerical processing from the sequential details of loops and component-by-component operations. Many efficient algorithms and robust programs (EISPACK, IMSL, LINPACK) have been developed for «industrial» use in the solution of systems of linear equations and the search for eigenvalues and eigenvectors, among other things (see, *e.g.*, Johnson). The implementation of a matrix product **AB**, where **A** and **B** are 100 by 100 matrices, may involve more than one «Mflop» (*million* floating point operations). Now that's power! It is interesting to note that manuals devoted to computational aspects of statistics (*e.g.* Jambu and also Thisted) as well as statistical packages (*e.g.*, SAS, SPSS, STATPACK) use linear algebra extensively.

4) Teaching. A similar object-oriented approach could be used in the teaching of statistics. Students or users who have the linear algebraic prerequisites (or possibly corequisites) could master large areas of statistics very quickly. It is hoped that this paper will help identify the linear algebra necessary for a «fast track» acquisition of statistics (especially regression, correlation, analysis of variance, and more advanced topics like principal component and factorial analysis; see Christensen and also Bouroche).

5) Notation. Linear algebraic notation is very compact, as we saw in (3) above. The student and user can concentrate on the notions rather than on the details of dimension and vector components (It's easier for word processing also).

The intermingling of statistical and linear algebraic notions and notations is advantageous for linear algebra as well, by showing which notions of linear algebra are useful, because statistics is a very important application.

Of course, the disadvantage of a linear algebraic approach to statistics, is that it is necessary for a statistics student or user to know some linear algebra. Is that a disadvantage? Some might consider it necessary to have a certain «mathematical maturity» to learn or use this approach. In general, a vectorial point of view would help many people, not only statisticians, in approaching problems from a functional or even object-oriented standpoint. The functional approach seems difficult in both mathematics and computing science, but there also seems to be little formal research on this difficulty.



This paper begins mathematically in Section 1 below, by casting the basic statistical ideas of sample and mean in a linear algebraic light, eliminating the need to mention the dimension and individual sample observations. Then, in Section 2, the light is shone upon variance and its generalization to covariance and correlation, presenting (among other things) the orthogonal decomposition of the sample vector into its mean and standard deviation. In Section 3, regression comes under the light, before the conclusion in Section 4 analyzes several statistical textbooks and manuals for «linear» content.

Statistical terms (which are in **bold face**) are defined and studied with respect to linear algebra, whose own terms (in *italics*) and theorems are given informally. Several references are given, because most of the results are not original, in the sense that they are just applications of well-known linear algebraic ideas (or perhaps just part of the «folklore» of statistics). However, the «extreme» linear algebraic viewpoint appears to be unusual in the literature.

It is hoped that this paper will be of some interest to users as well as teachers of statistics, irrespective of applications. Some knowledge of descriptive statistics, at least at the «user» level (see *e.g.* the first four chapters of Spiegel, 1961), and of linear algebra, at least at the analytic geometry level (see, *e.g.*, the first five chapters of Lipschutz), is necessary on the part of the reader (The previ-

ous two references are from the well-known and well-used «Schaum's Outline Series of Theory and Problems». A concise summary of «linear algebra for statisticians» can be found in Christensen, App. A, B.)

I. Sample and Mean

1.1 An ordered **sample** of n numerical observations $x_1,x_2,...x_n$ can be modelled by a *vector* **x** in *real n-space* R^n, where R, as is usual in mathematics, denotes the set of real numbers (*scalars*, numbers like 2 or $\sqrt{4.2}$ or $-\pi$ or ¾ that can be represented decimally to any precision desired). It is not necessary to visualize n dimensions, because we only study two or three vectors at a time; two (linearly independent) vectors, even in n-space, determine a two-dimensional plane, and three such vectors determine ordinary three-dimensional space.

The statistical distinction between sample and population will not be treated here; there seems to be no distinction from a linear algebraic standpoint. It is possible to «vectorialize» any frequency distribution, discrete (function from R to the natural numbers N) or continuous (R to R), enabling us to study concepts like the median and the interquartile range, but this will not be done here; we shall concentrate on samples of the same size n in this paper.

1.2 The usual **operations** of vector spaces are of interest to statistics. It may be necessary to add or subtract two vectors (because there may be two measurements made in each of n trials, giving two samples in the same space), or even making a *scalar multiplication* of a sample/vector (because of scaling, precisely). These vector operations are immediately visualizable. We can also multiply vectors component by component to form a *vector product* (not often referred to directly in linear algebra, and different from the cross product seen in 3-space) defined by $\mathbf{xy} = (x_1y_1,x_2y_2,....x_ny_n)$. In particular, $\mathbf{x}^2 = \mathbf{xx}$, $\mathbf{x}^3 = \mathbf{xxx}$ and so on. Any vector can be summed: $\Sigma\mathbf{x} = \sum_{i=1}^{n} x_i = x_1 + x_2 + ... x_n$. Finally, another useful multiplication of vectors is the scalar-valued *dot product*: $\mathbf{x}\cdot\mathbf{y} = \Sigma\mathbf{xy}$.

1.3 The **constant unit vector u** = $(1,1,...1) \in R^n$ will be used extensively hereafter; it is more important in statistics than the canonical unit vectors of euclidean n-space. Notice, for example, that $\mathbf{xu} = \mathbf{x}$ (**u** is a unit for the vector product), $\mathbf{u}^2 = \mathbf{u}$, $\Sigma\mathbf{u} = \mathbf{u}\cdot\mathbf{u} = n$, $\mathbf{x}\cdot\mathbf{u} = \Sigma\mathbf{x}$, and $\mathbf{xy}\cdot\mathbf{u} = \mathbf{x}\cdot\mathbf{y}$. We can even define $\mathbf{x}^0 = \mathbf{u}$. It will no longer be necessary to refer to vector components or to use the full «sigma» notation in this paper.

1.4 The **mean** of a sample **x** is just $\bar{x} = \mathbf{x}\cdot\mathbf{u}/n$ ($= \Sigma\mathbf{x}/n$; this last equality is not needed, but it is closer to the original definition). The linearity of the dot product ($(\mathbf{x} + c\mathbf{y})\cdot\mathbf{u} = \mathbf{x}\cdot\mathbf{u} + c(\mathbf{y}\cdot\mathbf{u})$) guarantees that of the mean ($\overline{x+cy} = \bar{x} + c\bar{y}$), and the mean of **u** is 1. We also need the *mean vector* $\mathbf{Mx} = \bar{x}\,\mathbf{u}$, which is the (orthogonal) *projection* of **x** on **u** (Think of **x** as a straight pole stuck in the ground and leaning toward the sun which is on the horizon, without the effects of refraction; **Mx** is the shadow of **x** on a vertical screen behind the pole. Only two dimensions are involved).

1.5 In fact, **M** is a *linear transformation* from R^n onto the one-dimensional *subspace* U of scalar multiples of **u**, and the *kernel* of **M** (subspace of vectors **y** such that $\mathbf{My} = \mathbf{0}$) is the **zero sum hyperplane** Z, *i.e.* the subspace of dimension $n-1$ (of vectors **y** that are) *orthogonal* to **u** ($\mathbf{y}\cdot\mathbf{u} = 0$) (Think of **u** as a vertical pole, and think of Z as the ground (the flat surface perpendicular to the pole); a point in Z (other than the origin) corresponds to a sample with positive and negative numbers, whose sum (and mean) is 0). Other examples of linear transformations include the *identity transformation* **I** and the *sum* $\Sigma = n\mathbf{M}$: $\mathbf{Ix} = \mathbf{x}$ and $\Sigma\mathbf{x} = n\mathbf{Mx} = (\mathbf{x}\cdot\mathbf{u})\mathbf{u}$.

Linear transformations from R^n into R^n are uniquely represented as n by n *matrices*. For example, **I** is represented by the *identity matrix* **I** (1 on the main diagonal, 0 elsewhere), **M** by $\mathbf{uu}'/n$ and Σ by $\mathbf{uu}'$, where **u** is seen as a *column vector* (n by 1 matrix) with *transpose* $\mathbf{u}'$, a *row vector* (1 by n matrix). Note that $\mathbf{MM} = \mathbf{M} = \mathbf{M}'$ (matrix *transpose*); these equalities imply that $\mathbf{M} = \mathbf{M}'\mathbf{M}$ and characterize a projection (Christensen: 335). Thus, **I** is also a projection, but Σ is not.

1.6 Normalization (in the sense of independence from n) is achieved by the mean; we no longer *need* to mention n in this paper, although we do so from time to time to recall the «classical» approach, or if there are several samples, necessitating matrix multiplication; see, *e.g.*,

paragraph 2.8 below. We note that $\overline{xy}$, just like **x**·**y**, defines an *inner product, i.e.* a *symmetric* ($\overline{xy} = \overline{yx}$) *positive* ($\overline{xx} > 0$ except when **x** = **0**) *bilinear* ($\overline{(x+cy)z} = \overline{xz} + c\overline{yz}$ and $\overline{x(y+cz)} = \overline{xy} + c\overline{xz}$) *form*. The *norm* defined by this inner product is $\|x\| = \sqrt{\overline{xx}}$ (= $|\mathbf{x}|/\sqrt{n}$, where $|\mathbf{x}|^2 = \mathbf{x}\cdot\mathbf{x}$ defines the ordinary *length* of **x**).

Geometrically, the dot product of two vectors **x** and **y** is related to the *angle* θ between them, and so is $\overline{xy}$, in exactly the same way: $cos\theta = \mathbf{x}\cdot\mathbf{y}/(|\mathbf{x}|\,|\mathbf{y}|) = \overline{xy}/(\|x\|\,\|y\|)$. The *Cauchy-Schwarz inequality* is still valid, as it is for any inner product: $|\overline{xy}| \le \|x\|\,\|y\|$ (*i.e.* $|cos\theta| \le 1$), and there is equality precisely when **x, y** are *linearly dependent, i.e.* when **x** is a scalar multiple of **y** and/or *vice versa* (one of **x, y** might be **0**).

1.7 The word «**norm**» and its linguistic derivatives are overworked in both linear algebra and statistics. In linear algebra, the «norm» is a length defined (as in the above paragraph) by an inner product in a real vector space, but a vector might be «normal» in the sense of «perpendicular, orthogonal» to a (hyper) surface, at a given point, in a real vector space. In statistics, the «normal» distribution is the gaussian one, represented by the «bell curve», and «normalize» might mean «remove the dependence on *n*» (as above) or «centralize» (as below). In this paper we do not use the word «normal», although we do use «norm» and «normalize».

II. (Co)variance and Correlation

2.1 The **variance** of a sample **x** is defined as $s_x^2 = \overline{(x-\bar{x})^2}$ and the **standard deviation** s_x is the square root of the variance. Note that since we do not distinguish between sample and population, there is no question here of (un)biased estimators; we don't need Greek letters here. We can define the *standard deviation vector* **Sx** = **x** - **Mx** (see Fig. 1). (If **x** is a straight pole stuck in the ground (paragraph 1.4), then **Sx** is the shadow of **x** on the ground, with the sun directly overhead.)

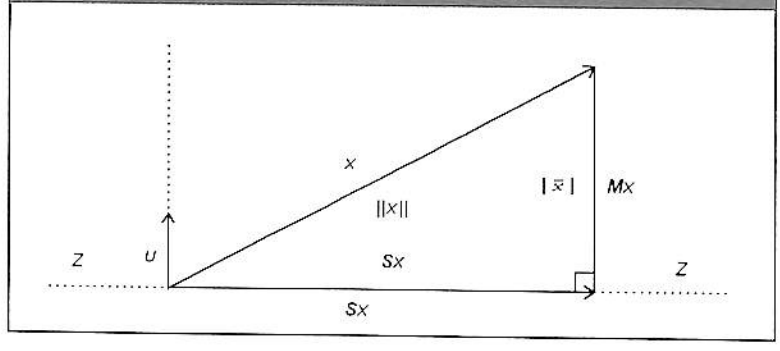
In fact, **S** = **I** - **M** is a vector projection *orthogonal* to **u** (a linear transformation onto Z whose kernel is U, such that SS = S = S' = S'S), represented by the matrix **I-uu'**/*n*. This projection preserves neither distances nor angles (orthogonality). Note, however, that two samples having the same projection differ by a constant vector: if **Sx** = **Sy** then **y-x** = $(\overline{y-x})$**u**. Also, $t = \bar{x}$ gives the constant vector «closest» to **x** in the sense that $\|x - tu\|$ is minimized (at s_x) (Morlat).

2.2 Note that $s_x = \|Sx\|$. Higher-order **centralized moments** are often defined by $m_i = \overline{(Sx)^i}$ (see *e.g.* Spiegel, 1980: 85; note that $m_2 = s_x^2$); thus m_3/s_x^3 measures *skewness* (**x** has a *symmetric* distribution if $m_3 = 0$) and m_4/s_x^4 measures *kurtosis* (a gaussian distribution has a kurtosis of 3).

2.3 It is easy to prove the computational formula $s_x^2 = \overline{x^2} - \bar{x}^2$. The consequences are not only computational. Note that $\|Sx\|^2 + \|Mx\|^2 = \|x\|^2$. Visually, this formula shows that the standard deviation of **x** is the scalar projection of **x** on Z, orthogonal to **u**. The illustration of the right angle triangle in Fig. 1 is two-dimensional, irrespective of *n*. This **central decomposition** of **x** into its mean and standard deviation is important for what follows, but is seldom directly mentioned in the literature. When the mean is calculated, we can *centralize* a sample and continue our analysis (algebra?) on the zero sum hyperplane Z, thus eliminating a dimension (that of **u**). Even though Z has dimension *n-1*, we need only two dimensions for two *non-parallel* (*i.e.* linearly independent) vectors.

2.4 The **covariance** between two samples **x** and **y** is defined by [**x**, **y**] = $\overline{(x-\bar{x})(y-\bar{y})} = \overline{SxSy}$ and can be evaluated by $\overline{xy} - \overline{x}\,\overline{y}$. We cannot use the notation s_{xy} because that already means the

FIGURE 1. CENTRAL ORTHOGONAL DECOMPOSITION OF A SAMPLE *x* INTO ITS MEAN AND STANDAR DEVIATION VECTORS. THE NAMES OF THE VECTORS ARE OUTSIDE THE TRIANGLE, AND THEIR LENGTHS ARE INSIDE.



standard deviation of **xy**, but it is obvious that the variance is a special case of the covariance: $s_x^2 = s_{xx} = [\mathbf{x}, \mathbf{x}]$. In fact, the covariance is a bilinear symmetric non-negative form in R^n (and even a scalar product in Z whose norm is the standard deviation s !). We can say that **x, y** are *centrally orthogonal* if **[x, y]** = 0; this means that there is a right angle in Z between **Sx** and **Sy** (but not necessarily between **x** and **y**; orthogonality is not preserved by **S**, in general). Remember also that s and even s^2 are not linear: $s_{x-y}^2 = s_x^2 - 2[\mathbf{x},\mathbf{y}] + s_y^2$ (cosine law in Z) and $s_{x+y} \leq s_x + s_y$ (triangular inequality).

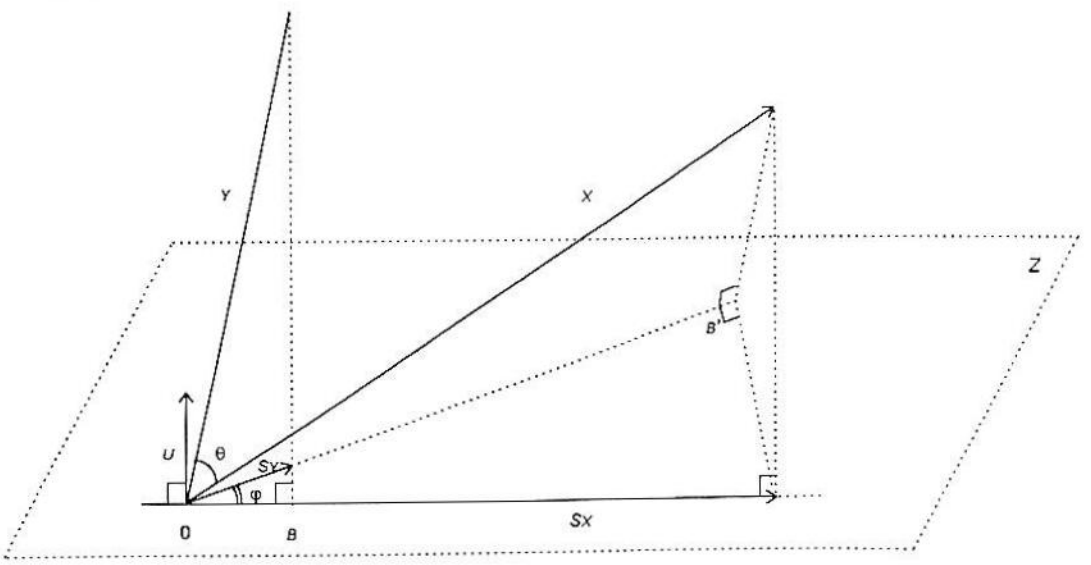
2.5 The Cauchy-Schwarz inequality (valid in general for non-negative forms) gives us $|[\mathbf{x}, \mathbf{y}]| < s_x s_y$ except when **x, y, u** are *linearly dependent* (*i.e.* there exist scalars *a, b, c* such that $a\mathbf{x}+b\mathbf{y}+c\mathbf{u} = \mathbf{0}$; this is true because it is equivalent to the linear dependence of **Sx, Sy**, *i.e.* the existence of *a, b* such that $a\mathbf{Sx}+b\mathbf{Sy} = \mathbf{S}(a\mathbf{x}+b\mathbf{y}) = \mathbf{0}$). This linear algebraic dependence puts **x, y, u** in the same (two-dimensional) plane and is equivalent to a perfect **linear statistical dependence** between **x** and **y**.

2.6 From a linear algebraic standpoint, the covariance is an example of «**deflation**»: if <, > is an inner product represented by the positive definite matrix **A**, then $\langle \mathbf{x}, \mathbf{y}\rangle - \langle \mathbf{x}, \mathbf{v}\rangle\langle \mathbf{y}, \mathbf{v}\rangle$ is a bilinear symmetric non-negative form represented by the positive semidefinite (non-negative definite) matrix **A - vv'** and an inner product on the subspace orthogonal to **v**, for any unit vector **v** ($\langle \mathbf{v}, \mathbf{v}\rangle = 1$). This «deflation» is not often seen in elementary manuals of linear algebra, but it is associated with *Householder transformations* in numerical analysis (see, *e.g.* Johnson: 118), and even *Gram-Schmidt orthogonalization* (see, *e.g.* Auer: 387), which is carried out, in fact, by projections).

2.7 The **correlation** (coefficient) between **x** and **y** is defined by $r = [\mathbf{x}, \mathbf{y}]/(s_x s_y)$, which is geometrically the cosine of the angle φ between **Sx** and **Sy**, the projections on Z (If **x** and **y** are straight poles stuck into the ground at the same place, with an angle θ between them, then φ is the shadow of θ on the ground, with the sun directly overhead. See Fig. 2 below). We see that $|r| \leq 1$ by Cauchy-Schwarz, that $|r| = 1$ when **x, y, u** are linearly dependent, and that $r = 0$ when **x, y** are centrally orthogonal.

2.8 More generally, if $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k$ are samples of the same size n, then they can be viewed as the *columns* of an n by k matrix **X**. The (row) vector of means is $\overline{X} = \mathbf{u}'\mathbf{X}/n$ and the **covariance matrix** is $[\mathbf{X}, \mathbf{X}] = (\mathbf{SX})'\mathbf{SX}/n = \mathbf{X}'\mathbf{S}'\mathbf{SX}/n = \mathbf{X}'\mathbf{SX}/n = \mathbf{X}'(\mathbf{I} - \mathbf{uu}'/n)\mathbf{X}/n = \mathbf{X}'\mathbf{X}/n - \mathbf{X}'\mathbf{uu}'\mathbf{X}/n^2 = \overline{X'X} - \overline{X}'\overline{X}$, of dimension k by k (To eliminate n from the discussion, it would be necessary to redefine matrix multiplication by dividing by n when n is the «internal» dimension of the multiplication).

FIGURE 2. PROJECTIONS ON Z OF SAMPLE VECTORS X, Y AND REGRESSION BETWEEN THEM. THE VECTORS SX AND SY ARE IN THE ZERO SUM HYPERPLANE Z. THE DISTANCE 0B IS BS$_y$ AND 0B' IS B'S$_y$.



Note that **[X, X]** is a *non-negative definite* matrix representing a symmetric non-negative bilinear form: **b'[X, X]b = (SXb)' (SXb)**/n = $\|SXb\|^2 \geq 0$ for any k-dimensional vector **b**. Also, **b** = $\overline{X}$ is the point in k-space «closest» to the points in k-space that are *rows* of **X**, in the sense that $\|X - ub'\|$ is minimized. The covariance matrix gives an idea of the orthogonality (in Z) of the column vectors of **SX**: if the main diagonal is *dominant* (variances much larger than covariances), then the vectors are more orthogonal linearly and more independent statistically.

III. Regression

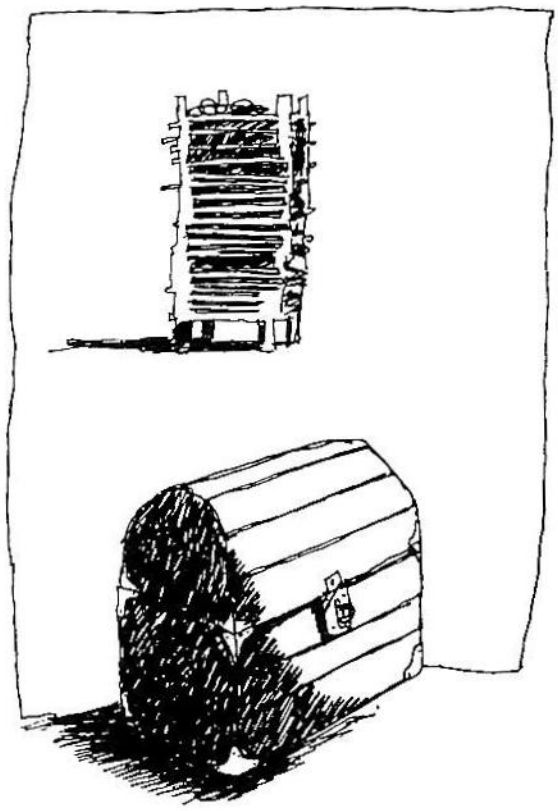
3.1 The two-dimensional visualization of regression is well known: find a curve of a certain type (we usually begin with a straight line) that best fits a scatter diagram of points in a plane, in the sense of minimizing the sum of the squares of distances from the points to the curve. This visualization can be extended to three dimensions and even beyond (best (hyper) plane or best (hyper) surface). In fact, the linear model of regression is the best-known application of linear algebra in statistics. We suggest a «dual» or «transposed» version below.

3.2 The remarks in paragraphs 2.5 and 2.7 above can be collected in order to state the following equivalences: **x, y, u** are linearly dependent, iff (if and only if) **x, y, u** are in the same two-dimensional plane, iff $|[\mathbf{x}, \mathbf{y}]| = s_x s_y$, iff $|r| = 1$, iff **Sx, Sy** are linearly dependent, iff **Sx, Sy** are in the same one-dimensional line on Z. In **linear regression**, we find the coefficients for **x, y, u** that come «closest» (in the least-squares sense) to linear dependence. (In classical linear algebra, a set of vectors is linearly dependent or it is not; however, in numerical linear algebra, we are aware of *ill-conditioned* situations that are «close» to linear dependence, in the sense that a determinant is «close» to 0, or that some matrix norm or condition index is «far» from 1 (see, *e.g.* Johnson: 45-49). Linear regression gives us another way of measuring the tendency toward linear dependence).

3.3 To «**regress**» **y** on **x**, we calculate scalars a, b such that $\|au + bx - y\|$ is minimized; this happens when b = **[x, y]/[x, x]** = rs_y/s_x and $a = \overline{y} - b\overline{x}$. In fact, bs_x is the scalar projection of **y** on **Sx**, and, not surprisingly for an analytic geometer, the minimal (normalized) distance, often called the **standard error of estimate** $s_{y.x}$ (Spiegel: 262), is the orthogonal distance from the point **y** to the plane determined by **u** and **Sx**, *i.e.* the orthogonal distance from the point **Sy** to the line determined by **Sx**, in Z. (See Fig. 2). This distance is the square root of the **error sum of squares** *SSE* (Neter: 602), which is equal to **[y, y]**-b**[x, y]** = **[y-**b**x, y]**. The regression of **x** on **y** is the same, *mutatis mutandis*, if $b' = rs_x/s_y$ is the corresponding coefficient of **y**, then $r^2 = bb'$ (Kenney: 259; Spiegel, 1980: 261).

3.4. One can also perform **linear orthogonal regression**: find α, c such that the (orthogonal) distance $\|u\cos\alpha + x\sin\alpha - y\cos\alpha\|$ from «**x, y**» to the «straight line» **y** *cos* α = **x** *sin* α + c**u** is minimized. The solution is given by *tan* 2α = 2**[x, y]**/([**x, x**]-[**y, y**]) and $c = \overline{y}\cos\alpha - \overline{x}\sin\alpha$; all three regression lines pass through $(\overline{x}, \overline{y})$, and $|b| \leq |\tan\alpha| \leq s_y/s_x \leq |1/b'|$. This solution is not easy to find in the statistical literature, because it is claimed to be of little use for experimental reasons (Kenney: 279), but **x** and **y** are in the same space, and orthogonal regression is «legitimate» from the linear algebraic point of view.

3.5 In **general linear regression**, the «best» least-squares linear fit $c_0\mathbf{u} + c_1\mathbf{x}_1 + \ldots c_k\mathbf{x}_k$ for **y** is the (hyper) plane whose coefficients are given by **X' Xc = X' y**, where **X** is an n by $k+1$ matrix whose first column is $\mathbf{x}_0 = \mathbf{u}$. We shall prove this, in order to illustrate that derivatives can be easily handled vectorially. To minimize $\|y - Xc\|$ we minimize $n\|y - Xc\|^2$ = **y' y - 2y'Xc + c' X' Xc** by (partially) differentiating it by **c**, remembering that **X** and **y** are constant; this gives **0 = -2y 'X + 2X' Xc** , whence $\mathbf{c} = (X'X)^{-1}X'y$. More over, if $\mathbf{P} = X(X'X)^{-1}X'$, then **PP = P = P'** and **PX = X**, making **P** a projection onto the subspace generated by (the columns of) **X**; hence the standard error of estimate, *i.e.* the minimal distance $\|y - Xc\| = \|y - Py\|$ is the (orthogonal) distance from **y** to this subspace, *i.e.* the (orthogonal) distance (in Z) from **Sy** to the subspace generated by (the columns of) **SX**.



In the above analysis, it is necessary that $(X'X)^{-1}$ exists; this existence is equivalent to the linear independence of the columns of **X** (including **u**), or the linear independence of the columns of **SX** (excluding **Su = 0**) (see Christensen: 84). Another way of calculating the coefficient vector **c** is to orthogonalize the columns of **SX** in *Z*, use the covariances with **y**, then «de-orthogonalize». The **least-squares polynomials** are special cases of general linear regression, using $x_o = \mathbf{u}$, $x_1 = x$, ... $x_k = x^k$. It is well known that in this case **X'X** is invertible if **x** is not constant.

3.6 One can also do **general orthogonal regression** by minimizing the sum of the squares of the (orthogonal) distances between the *rows* of **X** and a (hyper) plane $a\mathbf{u} = \mathbf{Yb}$ where $\mathbf{b}\cdot\mathbf{b} = 1$. To minimize $\|au - Xb\|^2 = a^2 - 2a\overline{X}b + b'\overline{X'X}b$ using *Lagrange multipliers*, we add the term $\lambda(1-\mathbf{b}\cdot\mathbf{b})$ before differentiating, to get $0 = 2a - 2\overline{X}b$ (and therefore $a = \overline{X}b$, not surprisingly: the hyperplane passes through $\overline{X}$) and $\mathbf{0} = -2a\overline{X}' + 2\overline{X'X}b - 2\lambda b = -2\overline{X}'\overline{X}b + 2\overline{X'X}b - 2\lambda b = 2[\mathbf{X},\mathbf{X}]b - 2\lambda b = 2([\mathbf{X}, \mathbf{X}] - \lambda\mathbf{I})b$, making λ an eigenvalue of the covariance matrix and **b** its corresponding unit-length eigenvector. In fact, $\lambda = \lambda\mathbf{b}.\mathbf{b} = \mathbf{b}'\lambda\mathbf{b} = \mathbf{b}'[\mathbf{X}, \mathbf{X}]\mathbf{b} = [\mathbf{Xb}, \mathbf{Xb}] = s^2_{Xb}$. If the covariance matrix **[X, X]** is diagonally dominant, the eigenvalues will be reasonably close to the diagonal values, *i.e.* the variances.

3.7 More **specialized topics** such as analysis of variance and covariance, multicollinearity, factor analysis, principal component analysis, can be treated «linearly». See, *e.g.*, Bouroche and Christensen.

(Temporary) Conclusion

1. There is no agreement concerning the linear algebraic level to be used to present statistics in teaching and in actual professional use. This is illustrated by the great difference in the linear algebraic levels used in statistical textbooks and articles. Part of the explanation comes from the varied mathematical background of students and users of statistics. There have been changes in mathematical education. It took a century for linear algebra to be generally understood as the theoretical foundation of «elementary» analytic geometry. During the last half-century, geometry in general (and analytic geometry in particular) has undergone neglect in many curricula, although a recent rise in importance of computer graphics and robotics may turn the tide.

2. The variety of approaches may be part of a wider educational tradition. It appears generally that a European (*e.g.* French; see Morlat) presentation of statistics has more linear algebra than an American (U.S.A.; see, *e.g.*, Neter or Spiegel 1980) approach, as part of a wider trend in Europe to present «applied» science with a solid «pure» foundation. European students in engineering and even social sciences have had more exposure to pure mathematics than their American counterparts. (The author received most of his mathematical education in the American tradition, in English Canada, but has taught for almost twenty years, *en français*, in Québec, where (as in many parts of Latin America) both traditions are influential).

3. Among the references below, that which has the most «extreme» linear algebraic approach is by **Christensen**, who extensively uses orthogonality, projections and even generalized matrices in the analysis of linear models of estimation, hypothesis testing, regression, variance, covariance, and so on. Even Christensen draws back from a «coordinate free approach» because «too many people never make the jump from coordinate free theory back to practical applications» (p. vii). His book therefore has sigma notation and references to *n* and vector components –but no diagrams.

4. The «**French School**» has an approach almost as algebraic and more geometric than Christensen. The encyclopedia article by Morlat refers to *n*-space, *k*-space and some orthogonal distances. This is extended by Bouroche (in the popular series «Que sais-je?» to direct (illustrated) geometrical (pp. 19, 26 *et passim*) and linear algebraic (pp. 23, 25 *et passim*) references. But these authors still use sigma notation and a coordinate approach.

5. Other books do not use much analytic geometry beyond that necessary to visually present regression. This includes applications-oriented books from France (*e.g.* Mialaret) and most statistical books from North America (*e.g.* Brase, Neter, Spiegel), including Québec (*e.g.* Bertaud). More advanced books develop matrices to present multiple regression (*e.g.* Gunst), multivariate normal distributions (*e.g.*Hogg, Cramér), and advanced computation techniques (Thisted). Kenney gives a geometrical interpretation of the correlation coefficient (p. 260). This «survey» is far from complete, but this author has found little evidence of any approach more «extreme» than that of Christensen.

6. It may be noted that the creators (discoverers?) of **modern statistical theory** seem to have been well aware of the geometrical aspects at least. Witness the title of a seminal article in 1901 on regression: «On lines and planes of closest fit to systems of points in space» (Pearson, cited in Bouroche: 3). It should be added that least-squares analysis (the technique used in regression) goes back to work by Legendre and especially Gauss at the beginning of the 19th century, before the development of linear algebra by Cayley, Sylvester and others (see, *e.g.* Bell: 259, 379 *et passim*).

7. The author would appreciate hearing about teaching or professional experiences, or other references, using a **cartesian approach** (from a linear algebraic and/or analytic geometric standpoint) to statistics. This paper is part of a larger project to «algebrize» probability and statistics. In the quadricentennial year (1996) of the birth of Rene Descartes (1596-1648), the author makes this humble addition to the cartesian contribution to mathematics (analytic geometry) and philosophy («*Cogito* (*Mathematizo*?) *ergo sum*») (Bell: 38). ◆

BIBLIOGRAPHY

Auel, J. (1991). *Linear Algebra with Applications.* Prentice-Hall. Scarborough, Ont.

Bell, E. (1937). *Men of Mathematics.* Simon and Schuster, New York.

Bertaud, M. & Bernard, Ch. (1989). *Initiation à la Statistique et aux Probabilités.* Presses de l'Univ. de Montréal.

Bouroche, J-M. & Saporta, C. (1980). *L'analyse des Données.* Presses Universitaires de France, Paris.

Brase, Ch. & Brase, C.

_____ (1991). *Understanding Statistics.* Heath. Lexington, Mass.

_____ (1994). *Pour Comprendre la Statistique.* Guérin. Montréal.

Christensen, R. (1987). *Plane Answers to Complex Questions.* Springer-Verlag, New York.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

Gunst, R. & Mason, R. (1980). *Regression Analysis and its Application.* Marcel Dekker, New York.

Hogg, R. & Craig, A. (1965). *Introduction to Mathematical Statistics.* Macmillan, New York.

Jambu, M. (1989) *Exploration Informatique et Statistique des Données.* Dunod, Paris.

Johnson, L. & Riess, R. *Numerical Analysis.* Addison-Wesley. Reading, Mass.

Kenney, J. & Keeping, E. (1954). *Mathematics of Statistics.* Van Nostrand. Princeton.

Lipschutz, S. (1980). *Linear Algebra.* McGraw-Hill, New York (Schaum Outline Series).

Mialaret, G. & Pham, D. (1967). *Statistique à l'usage des Éducateurs.* Presses Universitaires de France, Paris.

Morlat, G. (1990). «Statistique», *in Encyclopaedia Universalis*, v. 21, pp. 553-562. Paris.

Mosteller, Fr., *et al.* (1974). «Statistics», in *Encyclopaedia Britannica*, v. 28 pp. 239-239. Chicago.

Neter, J. *et al.* (1988). *Applied Statistics.* Allyn and Bacon (Simon & Schuster). Newton, Mass.

Spiegel, M.

_____ (1961). *Statistics.* Schaum. New York (Schaum's Outline Series of Theory and Problems).

_____ (1980). *Probability and Statistics.* McGraw-Hill. Singapore (Schaum's Outline Series).

Thisted, R. (1988). *Elements of Statistical Computing: Numerical Computation.* Chapman and Hall, London.