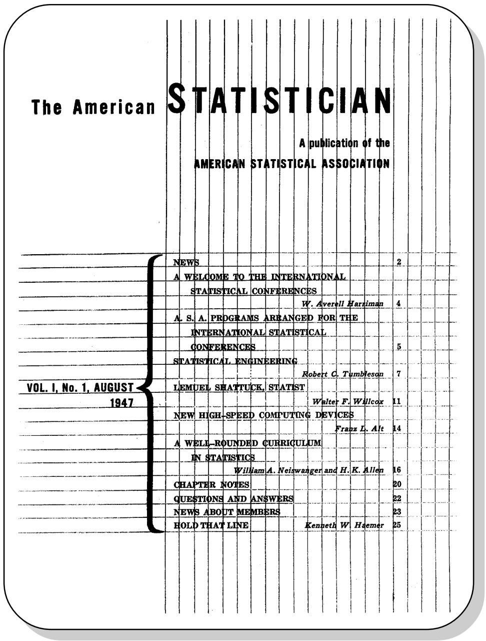
History Corner



The American STATISTICIAN

A publication of the
AMERICAN STATISTICAL ASSOCIATION

VOL. I, No. 1, AUGUST
1947

NEWS 2
A WELCOME TO THE INTERNATIONAL
STATISTICAL CONFERENCES
W. Averell Harriman 4
A. S. A. PROGRAMS ARRANGED FOR THE
INTERNATIONAL STATISTICAL
CONFERENCES 5
STATISTICAL ENGINEERING
Robert C. Tumbleson 7
LEMUEL SHATTUCK, STATIST
Walter F. Willcox 11
NEW HIGH-SPEED COMPUTING DEVICES
Franz L. Alt 14
A WELL-ROUNDED CURRICULUM
IN STATISTICS
William A. Neiswanger and H. K. Allen 16
CHAPTER NOTES 20
QUESTIONS AND ANSWERS 22
NEWS ABOUT MEMBERS 23
HOLD THAT LINE *Kenneth W. Haemer* 25

The first issue of TAS *appeared in August 1947.*

The Victory of Least Squares and Orthogonality in Statistics

Ralph DELAUBENFELS

This article gives a short history of the origin of least squares from a geometric perspective. It describes techniques used to deal with contradictory data in the second half of the eighteenth century, and their applications to problems in astronomy and geodesy. It is interesting that analogues of least squares—with maximum deviation and sums of absolute deviations, respectively, replacing sums of squared deviations—preceded least squares, which first appeared publicly in 1805, by 50 years. Geometry—specifically, an inner product being used to produce angles and orthogonality—is offered as the reason for least squares becoming preferable. More generally, we briefly outline how definitions and fundamental results in the general linear model, analysis of variance, conditional probability, independence, sufficiency, and time series can be unified and clarified as deriving from the inner product.

KEY WORDS: Conditional probability; History of statistics; Inner product; Least squares; Sufficiency; Time series.

1. INTRODUCTION

Here is a simple example of a statistical decision. I jump on my cheap scales and get 195 pounds as my alleged weight. This sounds wrong to me, so I weigh myself again and get 190 pounds. Now I have a promising trend, so I again weigh myself, this time getting 194 pounds. Stopping at this point, my decision is what to give as an estimate of my weight to an insistent authority figure.

To the modern perspective, it is almost a reflex to deal with this contradictory data, that is, the inconsistent system of equations

$$\begin{aligned} 195 &= \omega \\ 190 &= \omega \\ 194 &= \omega \end{aligned} \qquad (1)$$

by taking the average $\overline{y} \equiv \frac{1}{3}(195 + 190 + 194) = 193$ as my estimate of ω. This is a good starting point for a history of least squares, because for the system (1) $\overline{y}$ is simultaneously

(a) the least-squares solution of (1) (hence the "least-squares estimator of ω"); and

(b) the solution of the equation obtained from (1) by adding the three equations together.

Assertion (b) is a special case of the *method of averages*, which deals with an inconsistent linear system by solving a new linear system obtained by taking linear combinations of the original equations.

Because this article intends to be both historical and expository, we will briefly explain, on the eve of our history, the assertion (a), first using the mathematics available when our history

Ralph deLaubenfels, 1841 Drew Avenue, Columbus, OH 43235 (E-mail: *exmathprof@yahoo.com*). Much of this material is in the author's master's thesis in statistics, at Ohio State University. The author is indebted to his adviser, Steve MacEachern, and to Douglas Critchlow and Mario Peruggia, for useful advice and information

©2006 American Statistical Association DOI: 10.1198/000313006X151109

began (eighteenth century). Let

$$L(\omega) \equiv (\omega - 195)^2 + (\omega - 190)^2 + (\omega - 194)^2,$$

the sum of the squares of the errors. Taking derivatives, we find that $L''(\omega) = 6$ for all ω and $L'(\omega) = 0$ only when $\omega = 193 = \overline{y}$, thus the miracles of calculus tell us that $L(\omega)$ is minimized when (and only when) $\omega = \overline{y}$. See the exposition near the end of Section 3 for the more modern (nineteenth century) geometric approach.

2. ORIGINS OF LEAST SQUARES

The method of averages was first used succesfully by the cartographer and astronomer Tobias Mayer, in 1750, when he studied the libration of the moon (Stigler 1986, p. 16; Hald 1998, p. 94). This refers to the fact that the moon does not show precisely the same face to the earth at all times; about 60% of the moon's face is visible to us at some time. His observations led to 27 equations in three variables. He broke his 27 equations into three groups of 9 equations each and, within each group, added the equations together. He then solved the resulting system of three equations in three variables.

The mathematician Leonhard Euler, in 1749, studied the effects of Jupiter and Saturn on each other's motions; more precisely, Jupiter being larger, he studied the effect of Jupiter on Saturn's motion (Stigler 1986, p. 25). This "three-body problem" (Jupiter, Saturn, and the Sun) is immensely more complex than the "two-body problem" of calculating Saturn's motion due only to the influence of the Sun. Euler came up with 75 equations in eight variables. He was willing to combine some equations, to solve for two of the variables, but was unwilling to combine data that was too dissimilar (the observations were spread out over the period 1582–1745), hence did not solve for all the variables.

The mathematician Laplace, also studying Jupiter and Saturn in 1787, had a system of 24 equations with four variables (Stigler 1986, p. 31; Hald 1998, p. 107). He obtained four equations in the same variables with a much more elaborate combination of observations. One equation was the sum of all 24 equations; another

$$(1\text{st} + 2\text{nd} + \cdots + 12\text{th}) - (13\text{th} + 14\text{th} + \cdots + 24\text{th});$$

another

$$(3\text{rd} + 4\text{th} + 10\text{th} + 11\text{th} + 17\text{th} + 18\text{th} + 23\text{rd} + 24\text{th}) \\ -(1\text{st} + 7\text{th} + 14\text{th} + 20\text{th});$$

and, finally,

$$(2\text{nd} + 8\text{th} + 9\text{th} + 15\text{th} + 16\text{th} + 21\text{st} + 22\text{nd}) \\ -(5\text{th} + 6\text{th} + 12\text{th} + 13\text{th} + 19\text{th}).$$

See Stigler (1986) and Hald (1998) for much more detailed descriptions of this work in astronomy by Mayer, Euler, and Laplace.

The method of averages was the preferred method for dealing with inconsistent linear systems during the second half of the eighteenth century [until least squares appeared, very early in the nineteenth century; see Hald (1998, pp. 107–108) and Farebrother (1988)]. The method of averages is simple, and is guaranteed to produce a solution, for *some* choice of linear combination of the original equations; certainly the linear system consisting solely of a single homogeneous equation equal to a linear combination of the original equations would be consistent. Two disadvantages are soon apparent. First, there was no prescribed method for choosing *which* linear combinations of equations one took; yet different linear combinations would yield different linear systems with different solutions. Second, these linear combinations involved a loss of information. For example, if we replace the two equations

$$x = y, \quad y = z$$

with their sum

$$x + y = y + z, \quad \text{equivalent to} \quad x = z,$$

then we have lost the information $x = y$.

It is interesting that techniques for dealing with inconsistent linear systems, that are actually very close to least squares, also appeared in the second half of the eighteenth century. By "close" I mean their goal was to minimize some sort of measurement of *error*, that is, the distance between the model and the data.

This was first done by the Jesuit Roger Boscovich in 1757 (Stigler 1986, p. 39; Hald 1998, p. 97). Like so many of us, he was concerned about what shape the world was in. Specifically, he was studying the ellipticity of the earth by measuring a one-degree arc of longitude at two locations, one near the equator and one near the north pole; of particular interest was whether the earth was an oblate ellipsoid, meaning flattened at the poles.

Boscovich's data can be written as what we now call simple regression,

$$y_i = a + bx_i, \quad 1 \leq i \leq n, \tag{2}$$

where (x_i, y_i) are data, a and b are parameters to be estimated, and (for his data) $n = 5$. He sought a, b such that, writing e_i for error,

$$e_i \equiv (y_i - (a + bx_i)),$$

$$\text{(i)} \ \sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \text{(ii)} \ \sum_{i=1}^{n} |e_i| \ \text{is minimized} .$$

Boscovich used Newton's "geometric" style of argument, which means drawing pictures in lieu of an analytic argument. See Eisenhart (1961) for a description. Laplace in 1789 gave an analytic presentation of Boscovich's technique, calling it the "method of situation" (Stigler 1986, pp. 50–55; Hald 1998, p. 112).

Euler in 1749 and Lambert in 1765 introduced an L^∞ analogue of (ii) above; that is, they sought a and b that minimize

$$\max_{1 \leq i \leq n} |e_i|;$$

a popular shorthand for this is "minimax" (see Sheynin 1966). Laplace gave an explicit technique for this in 1783 (Plackett 1972; Hald 1998, p. 108).

The similarity of these techniques to least squares is transparent if we use vector terminology. Denote by $\vec{e}$ the vector of errors

$$\vec{e} \equiv (e_1, e_2, \ldots, e_n).$$

Then both techniques above consist of minimizing $\|\vec{e}\|$, the norm (intuitively, the size or magnitude) of $\vec{e}$. The question that remains is which norm to use. The method of situation uses the L^1, or *absolute* norm

$$\|\vec{a}\|_1 \equiv \sum_{i=1}^{n} |a_i|, \tag{3}$$

while the Euler-Lambert-Laplace technique uses the L^∞, or *maximum* norm

$$\|\vec{a}\|_\infty \equiv \max_{1 \le i \le n} |a_i|. \tag{4}$$

Least-squares means minimizing the L^2 norm $\|\vec{e}\|_2$, where

$$\|\vec{a}\|_2^2 \equiv \sum_{i=1}^{n} |a_i|^2. \tag{5}$$

This did not appear until 1805, introduced by the mathematician Legendre at the end of a paper studying comets; he then applied it to Boscovich's problem (Stigler 1986, p. 55; Hald 1998, p. 118).

More generally, Legendre applied least squares to the general linear system

$$y_i = \sum_{j=1}^{m} x_{ij}\beta_j \quad (1 \le i \le n), \tag{6}$$

where the x_{ij}'s are given and the β_j's are to be estimated; in matrix language this is

$$\vec{y} = X\vec{\beta} \quad (\vec{y} \in \mathbf{R}^n,\ \vec{\beta} \in \mathbf{R}^m,\ X \text{ an } n \times m \text{ matrix}), \tag{7}$$

and derived the normal equations, whose solution, call it $\hat{\beta}$, minimizes

$$\|\vec{y} - X\vec{\beta}\|_2^2 \equiv \sum_{i=1}^{n} \left(\vec{y} - X\vec{\beta}\right)_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j \right)^2$$

by taking derivatives $\frac{\partial}{\partial \beta_j}$ and setting them equal to zero.

The choice of norm did not appear to be influenced by the application, except to the extent that the calculations were feasible. L^1 and L^∞ were simply considered the only natural choice for about 50 years, then Legendre introduced L^2, as being "…more general, more exact, and more easy to apply…" (Hald 1998, p. 119).

At this point, our history takes an unfortunate, but all too common, turn. Gauss, possibly the greatest mathematician of all time, kept what you might call a "secret journal": a written record of results he had obtained but not published. After Legendre published his 1805 least-squares results, Gauss announced that he had created this technique in 1795. Legendre naturally objected, noting that anyone can *say* they previously discovered a published result. But Gauss then produced witnesses who had heard Gauss discuss these results before 1805; see Plackett (1972) for a discussion, including interesting correspondence from the time.

It should be mentioned here, where it is historically apropos (it will be discussed in more detail in Section 4, when we survey analysis of variance), that Gauss's 1809 publication of the least-squares technique also included much more: a probability distribution on (in matrix notation, as in (7)) $(\vec{y} - X\vec{\beta})$, along with an interesting argument for why this distribution should be normal. He then showed that, under normality, the least-squares estimator is the maximum likelihood estimator, and found the posterior distribution of $\hat{\beta}$, given independent uniform prior distributions on β_i (Hald 1998, chaps. 18.1 and 19). See also Farebrother (1985) and Seal (1967).

The general idea of placing a probability distribution on the error $(\vec{y} - \vec{\theta})$, when $\vec{y}$ is data measuring $\vec{\theta}$, began with the self-taught British mathematician Thomas Simpson in 1755 (Eisenhart 1961; Plackett 1972; Stigler 1986, p. 90). This was a line of development simultaneous with that of least-squares and its predecessors, with Gauss merging the two lines.

It seems only fair to mention here that an American mathematician, R. Adrain, in solving a surveying problem, independently developed least squares from a probabilistic point of view in 1808 (see Hald 1998, pp. 368–373; Stigler 1980).

See Kendall and Plackett (1977), Pearson and Kendall (1970), Kendall (1960), and deLaubenfels (2004) for more history.

3. MAXIMUM AND ABSOLUTE ERROR VERSUS SQUARED ERROR

It is not surprising that maximum error and absolute error appeared first. These are more natural and simple to describe. Maximum error is like the strength of a chain: only as strong as the weakest link. The difference between absolute error and squared error is like the difference between mean deviation ("average distance from the mean") and standard deviation ("square root of the average of the squared distance from the mean"). The latter definition sounds suspiciously awkward. The fact that it takes so many more words to define standard deviation should make one suspicious of its use. More generally, the squaring in squared error looks artificial; why are we exaggerating the effect of larger errors by squaring? A reflection of this exaggeration is the fact that the mean, the minimizer of squared error with one parameter, is more sensitive to extreme values than the median, the minimizer of absolute error with one parameter. In other words, the median is more robust with respect to extreme values than the mean. See Hampel, Ronchetti, Rousseeuw, and Stahel (1986) for a very thorough discussion of robustness, including many comparisons between the mean and the median for robustness (Hampel, Ronchetti, Rousseeuw, and Stahel 1986, p. 22, pp. 88–90 and 96–99) and an interesting method for quantifying (see in particular the definition of *influence function* on p. 84) robustness.

When we think of a set of data as a vector, then, at least in two or three dimensions where we may represent vectors as arrows, the L^2 norm, as the Euclidean length of a vector (or its "arrow"), becomes natural. But this linear algebra geometric outlook did not appear until the nineteenth century: the purely algebraic representation of linear equations in terms of vectors and matrices in the middle of the nineteenth century, and the more geometric analysis of vectors, what we now call "vector analysis," in the late nineteenth century. Thus, in our eighteenth century research environment, the L^1 and L^∞ norms were more natural, as a way of measuring the difference between two sets of data; this is why 50 years elapsed between minimizing absolute error or maximum error, as in the work by Boscovich, Euler, Lambert, and Laplace, and minimizing squared error, as done by Legendre and Gauss.

Unlike its absolute (3) and maximum (4) analogues mentioned above, least squares, after its belated appearance, quickly replaced the method of averages as the most popular method for dealing with inconsistent linear systems. Why did the L^2 least squares "win" over the similar L^1 and L^∞ techniques?

Historically, squared error came to be used instead of maximum or absolute error, because the calculations are much easier; compare Laplace's work in Stigler (1986, pp. 51–55) and Hald (1998, pp. 112–115), for absolute, and in Plackett (1972) and Hald (1998, p. 108), for maximum, to the least-squares technique. See also Eisenhart (1961, pp. 209–210).

But this begs the question: What is it about squared error that makes it easier to use than maximum or absolute error? The answer is that it is a norm that comes from an inner product:

$$\sum_{k=1}^{n} |e_k|^2 = \langle \vec{e}, \vec{e} \rangle \quad \text{where} \quad \langle \vec{x}, \vec{y} \rangle \equiv \sum_{k=1}^{n} x_k y_k. \tag{8}$$

For such norms, closed, convex sets are not only guaranteed to have a unique point of minimum norm (not true for maximum and absolute error: consider the line $x + y = 1$, in the plane $\mathbf{R}^2$, with absolute norm $\|(x, y)\| \equiv |x| + |y|$, where every point in $\{(x, y) \,|\, x + y = 1, 0 \leq x \leq 1\}$ is a point of minimum norm, or, similarly, the line $y = 1$ with the maximum norm $\|(x, y)\| \equiv \max\{|x|, |y|\}$), but also the concept of orthogonality (by definition, when the inner product equals zero) gives us an intuitive and straightforward way to obtain that point. The best approximation of a vector $\vec{x}$ from a subspace W is that point in W, call it $P_W(\vec{x})$, the *orthogonal projection of* $\vec{x}$ *onto* W, such that $(\vec{x} - P_W(\vec{x}))$ is orthogonal to W.

In particular, the least squares estimator $\hat{\beta}$ of $\vec{\beta}$ in (7), choosing $W = \{X\vec{\beta} \in \mathbf{R}^n \,|\, \vec{\beta} \in \mathbf{R}^m\}$, satisfies, for all vectors $\vec{\beta}$, by definition of orthogonality,

$$0 = \left\langle (\vec{y} - X\hat{\beta}), X\vec{\beta} \right\rangle = \left\langle X^T(\vec{y} - X\hat{\beta}), \vec{\beta} \right\rangle, \qquad \text{hence} \quad X^T(\vec{y} - X\hat{\beta}) = 0;$$

we immediately obtain the normal equations

$$X^T\vec{y} = X^T X\hat{\beta}.$$

For the weight problem (1) with which our article began, we have (7), with $\vec{y} = [195\ 190\ 194]^T$, $X = [1\ 1\ 1]^T$, $\vec{\beta} = \omega$, so that the normal equations become

$$(195 + 190 + 194) = [1\ 1\ 1]\,\vec{y} = X^T\vec{y} = X^T X\omega = 3\omega,$$

forcing ω to be $\overline{y}$.

Returning briefly to a special case of squared error, conventional wisdom sometimes sees variance as preferable to, for example, mean deviation (absolute error), because the variance of the sum of independent random variables is the sum of the variances. But this is merely an immediate special case of the inner-product geometric perspective and the fact that the squared-error norm comes from an inner product. For any orthogonal set of vectors $\{x_k\}_{k=1}^n$, with respect to an inner product $\langle \cdot, \cdot \rangle$,

$$\|\sum_{k=1}^{n} x_k\|^2 \equiv \left\langle \sum_{k=1}^{n} x_k, \sum_{j=1}^{n} x_j \right\rangle = \sum_{k,j=1}^{n} \langle x_k, x_j \rangle$$
$$= (\text{ by definition of orthogonality}) \sum_{k=1}^{n} \langle x_k, x_k \rangle \equiv \sum_{k=1}^{n} \|x_k\|^2;$$

note that this is precisely the Pythagorean theorem. For variance, the inner product is covariance, thus the Pythagorean theorem says much more than the conventional wisdom just alluded to, that the variance of the sum of uncorrelated random variables equals the sum of the variances (recall that independence implies zero covariance).

4. A SURVEY OF GEOMETRY UNDERLYING SOME STATISTICS

The inner product unifies a good deal of statistical theory. We have, up to now, addressed the oldest and simplest of such theory, the general linear model (6), in some detail, in the hope of actively engaging the reader. In the interest of space and accessibility, we will now indulge in a lighter survey of some other statistical theory that may be seen as similar manifestations of the inner product, referring the reader interested in details to deLaubenfels (2004, 2006), where histories of these topics, including a more detailed history of least squares, may also be found. Statements made without proof should not be assumed obvious.

The use of (finite-dimensional) orthogonal projections in the general linear model and analysis of variance is well known. Conditional expectation as an orthogonal projection may be found, for example, in Dudley (2002, theorem 10.2.9). The characterization of sufficiency in terms of inner products we believe to be new.

4.1 General Linear Model and Analysis of Variance

For inference on the general linear model (7), give it the statistical setting introduced by Gauss

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon}, \tag{9}$$

where $\vec{Y}$ and $\vec{\epsilon}$ are n-vectors of random variables, $\vec{\beta} \in \mathbf{R}^m$, X is an $n \times m$ matrix. We have already seen how orthogonality with respect to the inner product (8) produces the least-squares estimator

$$\hat{\beta} = (X^T X)^{-1} X^T \vec{Y}.$$

Further inference—in particular, confidence intervals and hypothesis tests—on $\vec{\beta}$ similarly are based on orthogonality. Under the popular hypotheses of ϵ_k, $1 \leq k \leq n$, being independent and normally distributed, with zero means and equal standard deviations, the fundamental fact in getting the desired behavior (t, χ^2, or F distributions) of test statistics is the behavior of $\|P\vec{Y}\|^2$, when P is an orthogonal projection on $\mathbf{R}^n$.

Analysis of variance may be set up as a special case of the general linear model (9), with X a matrix consisting entirely of ones and zeroes. For example, suppose I am a very lazy researcher speculating differences in average weights, among Americans, Canadians, and Mexicans (chair design at NAFTA conferences comes to mind). I weigh three Americans, obtaining 191, 197, and 194 pounds, two Canadians, obtaining 191 and 187 pounds, and a single Mexican at 186 pounds. Writing μ_{A} for the average weight, in pounds, of all Americans, and similarly for Canadians and Mexicans, our model is

$$\begin{array}{ll} 191 = y_{\mathrm{A},1} = \mu_{\mathrm{A}} + \epsilon_{\mathrm{A},1}, & 191 = y_{\mathrm{C},1} = \mu_{\mathrm{C}} + \epsilon_{\mathrm{C},1}, \\ 197 = y_{\mathrm{A},2} = \mu_{\mathrm{A}} + \epsilon_{\mathrm{A},2}, & 187 = y_{\mathrm{C},2} = \mu_{\mathrm{C}} + \epsilon_{\mathrm{C},2} \\ 194 = y_{\mathrm{A},3} = \mu_{\mathrm{A}} + \epsilon_{\mathrm{A},3}, & 186 = y_{\mathrm{M},1} = \mu_{\mathrm{M}} + \epsilon_{\mathrm{M},1}. \end{array}$$

This may be written as (9) (except for the custom of replacing the random vector $\vec{Y}$ with the data $\vec{y}$), with

$$\begin{aligned} \vec{y} &\equiv \begin{bmatrix} 191 \\ 197 \\ 194 \\ 191 \\ 187 \\ 186 \end{bmatrix}, \\ \vec{\beta} &\equiv \begin{bmatrix} \mu_{\mathrm{A}} \\ \mu_{\mathrm{C}} \\ \mu_{\mathrm{M}} \end{bmatrix}, \\ \vec{\epsilon} &\equiv \begin{bmatrix} \epsilon_{\mathrm{A},1} \\ \epsilon_{\mathrm{A},2} \\ \epsilon_{\mathrm{A},3} \\ \epsilon_{\mathrm{C},1} \\ \epsilon_{\mathrm{C},2} \\ \epsilon_{\mathrm{M},1} \end{bmatrix}, \\ X &\equiv \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

The least-squares estimator comes as no surprise here:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T \vec{y} \\ &= \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} (191+197+194) \\ (191+187) \\ 186 \end{bmatrix} \\ &= \begin{bmatrix} 194 \\ 189 \\ 186 \end{bmatrix}; \end{aligned}$$

that is, $\hat{\mu}_{\mathrm{A}} = 194$, the average of the three Americans weighed, and so on.

Perhaps of more interest or subtlety is to test my weight speculation:

$$H_0 : \mu_{\mathrm{A}} = \mu_{\mathrm{C}} = \mu_{\mathrm{M}} \quad \text{versus} \quad H_a : \text{at least one mean different.}$$

Writing again W for $\{X\vec{\beta} \mid \vec{\beta} \in \mathbf{R}^3\}$, equal, in this case, to $\{\vec{y} \in \mathbf{R}^6 \mid y_1 = y_2 = y_3,\ y_4 = y_5\}$, so that $P_W(\vec{y}) = X\hat{\beta} =$ $[194\ 194\ 194\ 189\ 189\ 186]^T$, and $\vec{1} \equiv [1\ 1\ 1\ 1\ 1\ 1]^T$, the usual alphabet soup of "ANOVA tables" becomes, after calculation,

$$\begin{aligned} \mathrm{SSE} &= \|\vec{y} - P_W(\vec{y})\|^2 = 26, \\ \mathrm{SSR} &= \|P_W(\vec{y}) - (\overline{y})\vec{1}\|^2 = 60, \\ \mathrm{SST} &= \|\vec{y} - (\overline{y})\vec{1}\|^2 = 86, \end{aligned}$$

and may be most easily visualized as the right triangle with hypotenuse SST, legs SSE and SSR.

Note in particular that the sum of squares formula SST = SSE + SSR is the Pythagorean theorem. Under the hypotheses of $\vec{\epsilon}$ consisting of normal, independent random variables of equal variance and zero mean, the orthogonality of appropriate matrices implies that, under H_0, SSE and SSR are values of independent χ^2 random variables, so that, up to three decimal places,

$$f = \frac{\mathrm{SSR}/2}{\mathrm{SSE}/3} = 3.462$$

is the value of an F distribution, with P value $P(F_{2,3} > 3.462) = .166$. Note that f is a constant times the square of the tangent of the angle between SST and SSR.

4.2 Conditional Expectation and Variance Shrinking

Under either of the inner products

$$\langle X, Y\rangle_1 \equiv E(XY) \quad \text{or} \quad \langle X, Y\rangle_2 \equiv \operatorname{cov}(X, Y),$$

on the vector space of random variables X and Y with finite variance, the conditional expectation $E(X|Y)$ is the orthogonal projection of X onto (equivalence classes of){$g(Y) \mid g$ is Borel measurable } (equal to the set of all random variables of finite variance measurable with respect to the sigma algebra generated by Y—see Chung (1974, p. 209); this explains why conditioning appears so commonly when minimizing variance, as when considering unbiased estimators. The Rao-Blackwell theorem, for example,

$$\operatorname{var}(E(W|T)) \leq \operatorname{var}(W)$$

is an immediate consequence of the Pythagorean theorem, with $\langle\cdot,\cdot\rangle_2$, applied to W, $E(W|T)$ and $(W - E(W|T))$; in fact, the Pythagorean theorem tells us precisely how much variance is lost by conditioning:

$$\begin{aligned} &\operatorname{var}(W) - \operatorname{var}(E(W|T)) \\ &\qquad = \operatorname{var}(W - E(W|T)) = E\left((W - E(W|T))^2\right), \end{aligned}$$

with the latter equality following from the fact that $E(W) = E(E(W|T))$.

The Cramer-Rao inequality (see any graduate-level mathematical statistics textbook for a precise statement and proof), providing a lower bound on the variance of an estimator, is another famous application of inner products, proven with a clever application of the Cauchy-Schwarz inequality

$$|\langle a, b\rangle| \leq \|a\|\|b\|,$$

with, again, the inner product $\langle\cdot,\cdot\rangle_2$ above. This was actually first proven by the functional analyst (mathematician studying normed spaces, including inner product spaces) Frechet (1943); see Le Cam and Yang (2000, p. 236).

It may be of interest to show here how switching between $\langle\cdot,\cdot\rangle_1$ and $\langle\cdot,\cdot\rangle_2$, in the characterization of $E(X|Y)$ as an orthogonal projection, yields, with a single picture (the Pythagorean theorem), the well-known formula

$$\text{var}\,(X) = \text{var}\,(E(X|Y)) + E\,(\,\text{var}\,(X|Y))\,.$$

Apply the Pythagorean theorem to $E(X|Y)$, $(X - E(X|Y))$, and

$$X = E(X|Y) + (X - E(X|Y)),$$

first with $\langle\cdot,\cdot\rangle_1$, then with $\langle\cdot,\cdot\rangle_2$, to obtain

$$E\left((X - E(X|Y))^2\right) = E(X^2) - E\left((E(X|Y))^2\right), \qquad (10)$$

followed by

$$\text{var}\,(X - E(X|Y)) = \text{var}\,(X) - \text{var}\,(E(X|Y)). \qquad (11)$$

Because $E(X) = E\,(E(X|Y))$, the left-hand sides, hence the right-hand sides, of (10) and (11), are equal; that is,

$$\begin{aligned}\text{var}\,(X) - \text{var}\,(E(X|Y)) &= E(X^2) - E\left((E(X|Y))^2\right)\\ &= E\left(E(X^2|Y)\right) - E\left((E(X|Y))^2\right)\\ &= E\left[E(X^2|Y) - (E(X|Y))^2\right]\\ &\equiv E\,(\,\text{var}\,(X|Y))\,.\end{aligned}$$

4.3 Sufficiency

From the characterization of conditional expectation by orthogonality it follows that sufficiency can be characterized in terms of orthogonal projections: the statistic $T = t(\vec{X})$ is sufficient with respect to the parameter θ if and only if the orthogonal projection, $E_\theta(k(\vec{X})|T)$, of $k(\vec{X})$ onto (equivalence classes of) $\{g(T)\}$, is constant with respect to θ, for Borel $k : \mathbf{R}^n \to \mathbf{R}$ such that $k(\vec{X})$ has finite variance. Since we are dealing with equivalence classes of random variables, to make sense of the word "constant," define two equivalence classes in two different measure spaces to be equal if their intersection is nonempty. Note that this is technically not much different from the traditional definition of sufficiency, that $P(\vec{X} \in A|T = t)$ be constant with respect to θ, for all Borel sets A, t in the range of T; we are replacing probability with expectation. The potential advantage of defining sufficiency in terms of conditional expectation is the geometry. See deLaubenfels (2004, Appendix D) for simpified proofs of basic results such as the "factorization theorem" characterization of sufficiency.

4.4 Correlation and Independence

Being uncorrelated of course is orthogonality with respect to the covariance inner product $\langle\cdot,\cdot\rangle_2$ above. Since independence of X and Y is equivalent to $h(X)$ and $g(Y)$ being uncorrelated, for all L^2 h, g, it follows that independence is characterized as orthogonality of the subspaces $\{h(X)\}$ and $\{g(Y)\}$.

4.5 Time Series

The study of time series $\{X_t\}_{t\in T}$, a family of random variables indexed by time with index set T, is perhaps the most obvious example of the inner product in statistics: a stationary time series is precisely a family of random variables of finite variance and equal means that preserves the covariance inner product $\langle\cdot,\cdot\rangle_2$ above, meaning that

$$\begin{aligned}\langle X_r, X_s\rangle_2 &\equiv \text{cov}\,(X_r, X_s)\\ &= \text{cov}\,(X_{r+t}, X_{s+t}) \equiv \langle X_{r+t}, X_{s+t}\rangle_2\,,\end{aligned}$$

for all r, s, t in T. (For simplicity, let us take T to be the integers or the real line.) Consider, for example, X_t defined to be the air pressure at time t, at a fixed point in a house next to an elementary school playground. Note that X_t here is a function primarily of released children's inexorable sound waves. To make this a stationary time series, with statistical uncertainty highly relevant, restrict t to be chosen only during recesses. Being stationary here means that the relationship between the sounds at time r and time s is a function purely of $(s - r)$, the time elapsed.

See Brockwell and Davis (1991) for basic material on time series, including extensive use of inner-product space techniques in treating stationary time series. It must be mentioned, however, that the full power and consequences of the inner product perspective are not used in books such as this. Consider in particular the fundamental analysis of stationary time series via its spectral representation known sometimes as the "frequency domain." In the school playground example just given, the frequency domain pulls out the different pitches in the chaotic collection of sounds creating our time series.

Any time series may be thought of as a family of vectors in the inner-product space of random variables of finite variance. For linear operators (linear functions from a vector space to itself), there is a powerful theorem known as the *spectral theorem* that gives a simple canonical representation of invertible linear operators that preserve the inner product; in finite dimensions, the representation is as a diagonal matrix with diagonal entries of absolute value one, but an equally useful reduction exists in infinite dimensions; see, for example, Rudin (1973, theorem 12.23). If one writes inner-product preserving sequences $\{X_k\}_{k\in\mathbf{Z}}$ as the orbit $X_k = A^k X_0 (k \in \mathbf{Z})$ of an invertible inner-product preserving linear operator A, then the spectral theorem, characterizing A as an integral of an operator-valued measure, may be immediately translated into a representation of an inner-product preserving sequence as an integral of a vector-valued measure. See deLaubenfels (2006) for details; in particular see example 2.5 for applications to time series. See also, for example, Brockwell and Davis (1991, chap. 2), for the constructions that are thereby bypassed.

An orthogonal time series, with respect to the covariance inner product above, means a family of uncorrelated random variables, such as, but not limited to, an independent family.

5. CONCLUSION

Finding a single concept underlying many seemingly different ideas or techniques is a reduction analogous to the application of a sufficient statistic to data. The subsequent unification of formerly disparate ideas creates a more useful and real understanding.

Least squares historically did not come naturally because the representation of data as a vector, and the algebraic characterization of length and angle, that generalizes not only to $\mathbf{R}^n$, $n \geq 3$, but to infinite dimensions, had not appeared. In hindsight, this representation of data leads to the L^2 norm as the Euclidean length of a vector; in particular, least-squares error is the length of the vector representing the difference between observed data and expected data. Realizing that L^2 norms come from inner products leads to a geometric outlook that is simple and intuitive; especially desirable is explicit construction of unique best approximations as orthogonal projections. This same geometric outlook may be applied similarly to many other areas of statistics, including infinite-dimensional settings where random variables are thought of as vectors, as we have surveyed in Section 4.

[Received August 2005. Revised July 2006.]

REFERENCES

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods* (2nd ed.), New York: Springer Series in Statistics.

Chung, K. L. (1974), *A Course in Probability Theory* (2nd ed.), New York: Academic Press.

deLaubenfels, R. (2004), "Early History of the Inner Product in Statistics," master's thesis, Ohio State University.

——— (2006), "Sequences of Vectors that are Orbits of Operators," *Journal of Mathematical Analysis and Applications*, 318, 459–466.

Dudley, RM(2002), *Real Analysis and Probability*, Cambridge: Cambridge University Press.

Eisenhart, C. (1961), "Boscovich and the Combination of Observations," in *Roger Joseph Boscovich,* ed. L. L. Whyte, London: Allen and Unwin, pp. 200–212; see also Kendall and Plackett (1977, pp. 88–100).

Farebrother, R. W. (1985), "The Statistical Estimation of the Standard Linear Model, 1756–1853," in *Proceedings of the First International Tampere Seminar on Linear Statistical Models and Their Applications,* eds. T. Pukkila and S. Puntanen, Dept. of Mathematical Sciences, Univ. of Tampere, pp. 77–99.

——— (1988), "The Historical Development of the Method of Averages, 1750–1987," unpublished manuscript, Dept. of Econometrics, University of Manchester, UK.

Frechet, M. (1943), "Sur l'extension de Certaines Evaluations Statistiques de petis Echantillons," *International Statistical Review*, 11, 182–205.

Hald, A. (1998), *History of Mathematical Statistics, From 1750–1930,* New York: Wiley.

Hampel, F. R., Ronchetti, EM, Rousseeuw, PJ, and Stahel, WA(1986), *Robust Statistics: The Approach Based On Influence Functions,* New York: Wiley.

Kendall, M. G. (1960), "Where Shall the History of Statistics Begin?" *Biometrika*, 47, 447–449; see also Pearson and Kendall (1970, pp. 45–46).

Kendall, M. G., and Plackett, R. L. (eds.) (1977), *Studies in the History of Statistics and Probability*, (Vol. 2) London: Griffin.

Le Cam, L. M., and Yang, G. L. (2000), *Asymptotics in Statistics: Some Basic Concepts* (2nd ed.), New York: Springer.

Pearson, E. S., and Kendall, M. G. (eds.) (1970), *Studies in the History of Statistics and Probability* (Vol. 1), London: Charles Griffin.

Plackett, R. L. (1958), "The Principle of the Arithmetic Mean," *Biometrika*, 45, 130–135; see also Pearson and Kendall (1970, pp. 21–126).

——— (1972), "The Discovery of the Method of Least Squares," *Biometrika* 59, 239–251; see also Kendall and Plackett (1977, pp. 279–291).

Rudin, W. (1973), *Functional Analysis*, New York: McGraw-Hill.

Seal, H. L. (1967), "The Historical Development of the Gauss Linear Model," *Biometrika*, 54, 1–24; see also Pearson and Kendall (1970, pp. 207–230).

Sheynin, O. B. (1966), "Origin of the Theory of Errors," *Nature (London)*, 211, 1003–1004.

Stigler, S. M. (1980), *American Contributions to Mathematical Statistics in the Nineteenth Century* (2 vols.), New York: Arno Press.

——— (1986), *The History of Statistics: The Measurement of Uncertainty before 1900,* Cambridge, MA: Belknap Press of Harvard University Press.