

# Doing Least Squares: Perspectives from Gauss and Yule

**John Aldrich**

*Department of Economics, University of Southampton, Southampton SO17 1BJ, UK  
e-mail: jca1@soton.ac.uk*

## Summary

**Gauss introduced a procedure for calculating least squares estimates and their precisions. Yule introduced a new system of notation adapted to correlation analysis. This paper describes these formalisms and compares them with the matrix and vector space formalisms used in modern regression analysis.**

*Key words:* Least squares; Gauss; Yule; Fisher; Aitken; Elimination; Partial Correlation; Orthogonalisation; Projection.

## 1 Introduction

Since C.F. Gauss (1777–1855) first used the method in 1795, many wonderful schemes have been devised for doing least squares—for theorising and for computing. I will examine a few from the perspective of two early ones.

My first scheme was introduced in 1809/10 by Gauss. The process survives as Gaussian elimination but the associated notation has gone. That never entered the Pearson–Fisher mainstream of early twentieth century statistics; elsewhere it lasted and lasted—into the 1950s in textbooks of astronomy and geodesy.

My second scheme was introduced nearly a century later by G.U. Yule (1871–1951). Yule’s interest was in correlation which he (1909, p. 722) saw as ‘an application to the purposes of statistical investigation’ of least squares. There were new quantities to calculate and a need for a better formalism. His notation survives—in multivariate analysis—but the process, a way of reasoning based on ‘product-sums of deviations’, has gone.

Sections 2–4 treat Gauss’s scheme and 6–8 Yule’s. Section 5 treats Laplace’s elimination method which has features of both schemes. From a modern viewpoint the schemes are founded on triangularisation and orthogonality. Sections 9–13 sketch the development of these notions and indicate some connections between the schemes and the modern matrix and vector space treatments of regression.

This paper considers how different techniques have been used on the same least squares—or multiple regression—problem. It complements accounts like Stigler’s (1986, Parts 1 & 3), that emphasise the conceptual distance between the least squares work of Gauss and Yule, or Seal’s (1967) that present a growing body of results in a uniform modern notation. Longley (1984, pp. 1–11) has covered some of my ground but in less detail. These works—and Farebrother’s (1988) historically informed survey of computational methods—have influenced my choice of material but the anthology is ultimately a personal one, reflecting the accidents of my reading and background.

## 2 The Gaussian Algorithm

I begin with the version of Gaussian elimination in nineteenth and early twentieth century textbooks of astronomy and geodesy; this is simpler than Gauss's own. The textbooks in English—e.g. Chauvenet (1867, p. 530), Brunt (1917, p. 90), Smart (1958, p. 126) and Rainsford (1957, p. 48)—followed the exposition of Gauss's pupil Encke (1835, pp. 264–72). These texts' eclectic, even self-contradictory, way with inference is discussed by Aldrich (1997).

To illustrate the textbook method, take the case of three unknowns. If there were no measurement error, the observation equations would hold exactly—with the equation for the  $i$ -th observation written

$$a_i p + b_i q + c_i r - m_i = 0.$$

Here  $a_i, b_i, c_i$  and  $m_i$  are measurements and so *known*;  $p, q$  and  $r$  are *unknowns*. However associated with each observation is an error of measurement  $v_i$ ,

$$a_i p + b_i q + c_i r - m_i = v_i.$$

The least squares problem is to choose  $p, q$  and  $r$  to minimise

$$W = \sum v_i^2 = \sum \left( a_i p + b_i q + c_i r - m_i \right)^2. \quad (2.1)$$

The first order conditions for a minimum are three linear equations:

$$\begin{aligned} (aa)p + (ab)q + (ac)r &= (am) \\ (ba)p + (bb)q + (bc)r &= (bm) \\ (ca)p + (cb)q + (cc)r &= (cm). \end{aligned} \quad (2.2)$$

This notation—the use of  $(ab)$  for  $\sum a_i b_i$ , etc.—was introduced by Gauss in 1811. He (1801, p. 18) had previously used the symbol  $(ab)$  less economically:  $\sum (ab)$  as an abbreviation for  $ab + a'b' + a''b'' + \text{etc.}$  In his presentation of least squares Legendre (1805) used the notation  $\int ab$  for  $\sum a_i b_i$ .

The solution process begins by solving the first equation for  $p$ ,

$$p = -\frac{(ab)}{(aa)}q - \frac{(ac)}{(aa)}r + \frac{(am)}{(aa)} \quad (2.3)$$

and then substituting this value into the second and third equations to give

$$\begin{aligned} \left( (bb) - \frac{(ba)(ab)}{(aa)} \right) q + \left( (bc) - \frac{(ba)(ac)}{(aa)} \right) r &= \left( (bm) - \frac{(ba)(am)}{(aa)} \right) \\ \left( (cb) - \frac{(ba)(ac)}{(aa)} \right) q + \left( (cc) - \frac{(ca)(ac)}{(aa)} \right) r &= \left( (cm) - \frac{(ca)(am)}{(aa)} \right) \end{aligned}$$

The coefficient of  $q$  in the first equation is abbreviated to  $(bb, 1)$  where the 1 indicates that the first elimination has been completed. Using such “auxiliaries”, the two equations can be written

$$(bb, 1)q + (bc, 1)r = (bm, 1)$$

$$(cb, 1)q + (cc, 1)r = (cm, 1).$$

The process continues by using the first equation to eliminate  $\check{q}$ , giving

$$\left( (cc, 1) - \frac{(cb, 1)(bc, 1)}{(bb, 1)} \right) r = \left( (cm, 1) - \frac{(cb, 1)(bm, 1)}{(bb, 1)} \right)$$

which, with the formation of further auxiliaries, can be written

$$(cc, 2)r = (cm, 2).$$

Solving for  $r$  gives

$$r = \frac{(cm, 2)}{(cc, 2)}. \tag{2.4}$$

Given  $r$ , the solutions for  $q$  and  $p$  can be obtained working backwards through the  $(\dots, 1)$  and  $(\dots)$  equations. Thus the original system (2.2) has been replaced by the triangular system

$$\begin{aligned} (aa)p + (ab)q + (ac)r &= (am) \\ (bb, 1)q + (bc, 1)r &= (bm, 1) \\ (cc, 2)r &= (cm, 2) \end{aligned} \tag{2.5}$$

which is solved from the bottom up by backwards substitution.

Variants of this procedure proliferated: by 1933 Frisch & Waugh (p. 396n) were grumbling, “Most of the ‘new’ schemes for solving the normal equations that are developed from time to time are nothing but unessential modifications of the Gaussian algorithm”. Farebrother (1988) reviews many of the schemes, both pre- and post-1933. Legendre who first published the method of least squares (in 1805) saw no need for special methods. We now examine Gauss’s own scheme(s); see Stewart (1995) for a treatment from a closely related point of view. Plackett (1972) and Stigler (1986, Part 1) provide background for the discovery of least squares.

### 3 Gauss’s Algorithms: 1809 and 1810

The ancestor of the Gaussian algorithm—the details are given in Section 4—reduces  $W$ , the sum of squared deviations, to the sum of squares

$$W = \frac{p'^2}{\alpha} + \frac{q'^2}{\beta'} + \frac{r'^2}{\gamma''} + \text{constant}, \tag{3.1}$$

where  $p'$  depends on  $p, q$  and  $r$ ,  $q'$  depends on  $q$  and  $r$ , and  $r'$  depends on  $r$ . His followers reduced (2.2) but Gauss reduced (2.1) as Stewart (1995) notes.

In his Bayesian treatment of 1809/10 Gauss specified a uniform prior for  $p, q$  and  $r$ ; he made no inferences about the observation precision,  $h$ —he treated this problem in 1816. With independent normal errors of constant precision  $h$ , the posterior density is proportional to

$$e^{-h^2W}.$$

Maximising this density is equivalent to minimising  $W$ . Gauss used (3.1) to show that the posterior density of each coefficient is normal. Naturally the precisions of  $p, q$  and  $r$  are relative to the unknown value,  $h$ . In the case of  $r$  this precision is the reciprocal of  $\gamma''$ .

The representation (3.1) also underpinned the computational scheme of 1809 (pp. 266–8). The scheme—in matrix notation—involves introducing  $\mathbf{d}$ :

$$\mathbf{d} = \mathbf{X}'\mathbf{X}\beta - \mathbf{X}'\mathbf{y} \quad (3.2)$$

and then solving for  $\beta$  to obtain, say

$$\beta = \mathbf{b} + \mathbf{C}\mathbf{d}. \quad (3.3)$$

Consider the right hand side terms. The intercept gives the least squares vector—corresponding to  $\mathbf{d} = 0$ . The coefficient of  $d_i$  in the equation for  $\beta_i$ ,  $c_{ii}$ , is the reciprocal of the precision of  $\beta_i$ . Gauss used (3.1) to show this.

In 1810 Gauss founded a second computational scheme on (3.1). He had paid no special attention to the task of transforming (3.2) to (3.3). He now took the task of obtaining the least squares values,  $\mathbf{b}$ , more seriously—his work on the planetoid Pallas involved six unknowns. He (1810, p.154) presented a new procedure, claiming “The process of solution, very tedious when the number of unknowns is considerable, can be simplified notably”. Referring to (3.1), he had already shown that  $\alpha$ ,  $\beta'$  and  $\gamma''$  are all positive. So  $W$  is a minimum when

$$p' = q' = r' = 0. \quad (3.4)$$

Gauss proposed solving these equations for  $p$ ,  $q$  and  $r$ . But (3.4) turns out to be the system (2.5). It is time we considered the route from (2.1) to (3.1).

#### 4 The Reduction of $W$

Gauss had already used the reduction in his *Disquisitiones Arithmeticae* (1801, Art. 270) when treating the definiteness of quadratic forms. Lagrange (1759, p. 7) used it even earlier when treating the second order conditions in multivariate calculus; this technique for reducing a quadratic appears in some textbooks (e.g. Bôcher 1907, p. 131) as “Lagrange’s method”.

In 1809 Gauss described the process by which  $\alpha$ ,  $\beta'$ ,  $\gamma''$ ,  $p'$ ,  $q'$  and  $r'$  are formed but, as he did not calculate with them, he did not give explicit expressions. I will reproduce the 1809 (pp. 264–5) argument, inserting the explicit formulae Gauss gave in 1810. The formulae involve auxiliaries.

Begin by writing out  $W$ , as defined in (2.1), in the notation of 1810:

$$W = (aa)p^2 + 2(ab)pq + 2(ac)pr + (bb)q^2 + 2(bc)qr + (cc)r^2 \\ - 2(am)p - 2(bm)q - 2(cm)r + (mm).$$

Gauss (p. 264) introduced  $p'$ :

$$p' = \frac{1}{2} \frac{dW}{dp} = (aa)p + (ab)q + (ac)r - (am). \quad (4.1)$$

In 1809 he wrote only that the expression on the right is linear in  $p$ ,  $q$  and  $r$ . The terms in  $p$  can be eliminated from  $W$  by forming  $W'$ , where

$$W' = W - \frac{p'^2}{\alpha} = (bb, 1)q^2 + 2(bc, 1)qr + (cc, 1)r^2 \\ - 2(bm, 1)q - 2(cm, 1)r + (mm, 1)$$

and  $\alpha = (aa)$ ; the auxiliaries were defined above. Gauss next introduced  $q'$

$$q' = \frac{1}{2} \frac{dW'}{dq} = (bb, 1)q + (bc, 1)r - (bm, 1). \quad (4.2)$$

All terms in  $q$  can be eliminated from  $W'$  by forming  $W''$ , where

$$W'' = W' - \frac{q'^2}{\beta'} = (cc, 2)r^2 - 2(cm, 2)r + (mm, 2)$$

and  $\beta' = (bb, 1)$ . Finally Gauss defined  $r'$  and  $W'''$  by

$$r' = \frac{1}{2} \frac{dW''}{dr}, = (cc, 2)r - (cm, 2) \quad (4.3)$$

and

$$W''' = W'' - \frac{r'^2}{\gamma''} \text{ with } \gamma'' = (cc, 2).$$

Equation (3.1) follows.  $W'''$  is the constant term in (3.1) and the posterior precision of  $r$  is  $\gamma''$ . We now consider how Gauss used the reduction of  $W$  to justify the interpretation of  $\mathbf{C}$  in (3.3).  $\mathbf{d}$  has components  $P$ ,  $Q$  and  $R$  where

$$P = \frac{1}{2} \frac{dW}{dp}, Q = \frac{1}{2} \frac{dW}{dq} \text{ and } R = \frac{1}{2} \frac{dW}{dr}.$$

It can be seen from the definition of  $r'$  that it is a linear function of  $P$ ,  $Q$  and  $R$  with the coefficient of  $R$  equal to unity: so write say,

$$r' = AP + BQ + R.$$

Combining this with (4.3) yields

$$r = \frac{(cm, 2)}{(cc, 2)} + \frac{A}{(cc, 2)}P + \frac{B}{(cc, 2)}Q + \frac{1}{(cc, 2)}R. \quad (4.4)$$

In (4.4) the constant term is the least squares value of  $r$  and the coefficient of  $R$  is the reciprocal of the posterior precision of  $r$ . (4.4) and the analogous equations when  $p$  and  $q$  are the final variables make up (3.3).

In Section I the  $n$ -th order auxiliaries are written in terms of  $(n - 1)$ th order auxiliaries; they can be written in terms of auxiliaries of *all* lower orders. Gauss used both forms: e.g. the auxiliary  $(cc, 2)$  can be written in two ways:

$$(cc, 2) = (cc, 1) - \frac{(cb, 1)(bc, 1)}{(bb, 1)} \quad (4.5)$$

$$(cc, 2) = (cc) - \frac{(ca)(ac)}{(aa)} - \frac{(cb, 1)(bc, 1)}{(bb, 1)}. \quad (4.6)$$

When Gauss (1821, pp. 39–43) switched to a Gauss–Markov justification of least squares the reduction of  $W$  was no longer essential to his inference theory. Yet it remained the basis of his computational scheme—see e.g. (1823, p. 67). Although auxiliaries originated in a theoretical argument,

their fate came to be bound up with a particular computing procedure.

## 5 Laplace's Elimination Method

Gauss devised special notation for the entities created by reduction but the reduction itself was not specialised to least squares. The triangularisation method in Laplace's *Théorie Analytique* (first supplement, 1816, pp. 542–558) is so specialised for it exploits the properties of least squares residuals. It is based on cross-products of transformed variables instead of transformed cross-products. Farebrother (pp. 66–8) discusses the method while Stigler (1986, pp. 143–57) gives a general account of Laplace's work on least squares.

Laplace has unknowns  $z, z', z'', z''', z^{IV}, z^V$  and, indexing the observations by  $i$ , knowns  $p^{(i)}, q^{(i)}, r^{(i)}, t^{(i)}, \gamma^{(i)}, \lambda^{(i)}, \alpha^{(i)}$  and residual  $\varepsilon^{(i)}$ . To facilitate comparisons, I re-express his argument—and some later ones—in a neo-Gaussian vector notation with  $v, a, b, c$  and  $m$  as vectors. The elimination starts from the equation defining the least squares residual vector  $v$ :

$$v = ap + bq + cr - m. \quad (5.1)$$

Multiply by  $a$ , use the normal equation property that  $(av) = 0$  and rearrange:

$$p = -\frac{(ab)}{(aa)}q - \frac{(ac)}{(aa)}r + \frac{(am)}{(aa)}. \quad (5.2)$$

This was obtained as (2.3) above. Now define the transformed vectors

$$b_1 = b - a\frac{(ba)}{(aa)}; c_1 = c - a\frac{(ca)}{(aa)}; m_1 = m - a\frac{(ma)}{(aa)}. \quad (5.3)$$

Substitute from (5.2) into (5.1) and use these new vectors to obtain

$$v = b_1q + c_1r - m_1. \quad (5.4)$$

Multiplying by  $b_1$  and using  $(b_1v) = 0$  allows  $q$  to be eliminated. Following the same process of creating new variables leads to

$$v = c_2r - m_2. \quad (5.5)$$

Laplace multiplied (5.1) through by  $a$ , (5.4) through by  $b_1$  and (5.5) through by  $c_2$  and uses the properties of residuals to obtain the triangular system

$$\begin{aligned} (aa)p + (ab)q + (ac)r &= (am) \\ (b_1b_1)q + (b_1c_1)r &= (b_1m_1) \\ (c_2c_2)r &= (c_2m_2). \end{aligned} \quad (5.6)$$

Laplace used the vinculum not brackets: my  $(c_2m_2)$  is his  $\overline{r_2\alpha_2}$ ; it is formed so

$$(c_2m_2) = (c_1m_1) - \frac{(c_1b_1)}{(b_1, b_1)}(b_1m_1).$$

The Gaussian processes of Sections 2–4 can be expressed in terms of  $b_1$  and  $c_2$  instead of the auxiliaries since

$$(bb_1) = \left( b \left( b - a \frac{(ba)}{(aa)} \right) \right) = (bb) - \frac{(ba)(ab)}{(aa)} = (bb, 1)$$

etc. On making such changes the triangular system (2.5) becomes

$$\begin{aligned} (aa)p + (ab)q + (ac)r &= (am) \\ (bb_1)q + (bc_1)r &= (bm_1) \\ (cc_2)r &= (cm_2). \end{aligned} \tag{5.7}$$

Such a development appears in Bienaymé’s (1853) comparison of Cauchy’s (1836) interpolation method with least squares except that the initial linear equations must cover both methods and Bienaymé follows Cauchy and has  $\Delta b, \Delta^2 c$ , etc. instead of  $b_1, c_2$ , etc. Bienaymé (p. 12) recognised the relationship between his and the Gauss and Laplace schemes. Farebrother (pp. 59–66) discusses the method while Heyde & Seneta (1977, chapter 4) treat the complicated story of Bienaymé’s relations with Cauchy (and Chebyshev).

Much later Frisch & Waugh (1933, p. 396) interpreted the transformed variables—conflating Gauss and Laplace—when they remarked that the Gaussian algorithm is founded on successive bivariate regressions with residuals as variables (for a modern textbook account see Draper & Smith (1966, pp. 107–115 & 180)). The econometricians were showing that there is no conflict between regression with de-trended data (least squares residuals) and regression with unadjusted data and time as an explicit regressor:  $q$  and  $r$  can be obtained from (5.4) or from (5.1). Not that Frisch and Waugh *said* so—they used determinants and wrote in Yule’s notation. The trend-adjustment debate is reviewed by Morgan (1990, Section 5.3).

## 6 Correlation: Yule’s Old System

Yule’s “new system” of 1907 was a completely specialised system of least squares analysis. Like Laplace’s method, it exploits residuals but on a grander scale. Yule, however, was not interested in elimination and his notation was not designed to represent iterated bivariate regressions. It was designed to improve on Karl Pearson’s correlation notation which Yule used in his first work on the theory of partial and multiple correlation. For inference theory Yule relied on Pearson (1896) and Pearson & Filon (1898). This corruption of Bayesian analysis is discussed by Aldrich (1997).

Stigler (1986, Part 3) describes Yule’s view of the relationship between least squares and correlation and Aldrich (1995) his interpretation of correlation. Yule (1897, pp. 832–3) based the development of partial correlation on a pair of lines fitted by least squares, written using deviations from the mean

$$\begin{aligned} x_1 &= b_{12}x_2 + b_{13}x_3 \\ x_2 &= b_{21}x_1 + b_{23}x_3. \end{aligned}$$

Yule *defined*  $\rho_{12}$ , the net or partial correlation between  $x_1$  and  $x_2$  with  $x_3$  held constant, as

$$\rho_{12} = \sqrt{(b_{12}b_{21})} \tag{6.1}$$

by analogy with the relationship between total correlation and the total regressions. The notation does not indicate *which* variable is held constant.

Yule defined partial correlation in terms of the least squares values  $b_{12}$  and  $b_{13}$  but he did not use the computational procedures from the least squares literature. He took the normal equations,

$$\begin{aligned} S(x_1x_2) &= b_{12}S(x_2^2) + b_{13}S(x_2x_3) \\ S(x_1x_3) &= b_{13}S(x_2x_3) + b_{13}S(x_3^2), \end{aligned}$$

and saw that the quantities involved were closely related to those in Pearson's (1896) correlation theory. Yule rewrote the equations in terms of the (sample) correlation coefficients,  $r_{12}$  etc., and (sample) standard deviations,  $\sigma_1$  etc.:

$$\begin{aligned} r_{12}\sigma_1 &= b_{12}\sigma_2 + b_{13}r_{23}\sigma_3 \\ r_{13}\sigma_1 &= b_{12}r_{23}\sigma_2 + b_{13}\sigma_3. \end{aligned}$$

These can be solved for, say,  $b_{12}$

$$b_{12} = \frac{r_{12} - r_{13}r_{23}}{(1 - r_{23}^2)} \cdot \frac{\sigma_1}{\sigma_2}. \quad (6.2)$$

So  $b_{12}$  is generated from the correlations and standard deviations on the right hand side. The normal equations are not *solved* in any numerical sense; they are theoretical relationships underlying such constructions as (6.2).

Using formula (6.2) for  $b_{12}$ , the analogous formula for  $b_{21}$  and (6.1), the definition of the partial correlation correlation, Yule derived the expression

$$\rho_{12} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}. \quad (6.3)$$

Yule knew that similar formulae could be produced for any number of variables but he (p. 836) recognised the "very rapid growth in the complexity of the formulae and arithmetic as the number of variables increases". He discussed the case of four variables but did not give the formula relating  $\rho_{12}$  (with  $x_3$  and  $x_4$  as the variables held constant) to the underlying simple correlations.

## 7 The Calculus of Subscripts

By 1907 Yule was satisfied he had found a way round the increasing complexity of the formulae as the number of variables increases. The new notation was "simple, definite and quite general, thus greatly facilitating the treatment of the subject" (1907, p. 182). It had genuine heuristic value: "The majority of the results . . . were . . . first suggested by the notation itself."

Consider the impact of the new notation on (6.3):  $\rho_{12}$  is replaced by the wholly "definite"  $r_{12.3}$  and an extraordinary extension is "suggested", viz.

$$r_{12.34\dots n} = \frac{r_{12.4\dots n} - r_{13.4\dots n}r_{23.4\dots n}}{\sqrt{(1 - r_{13.4\dots n}^2)(1 - r_{23.4\dots n}^2)}} \quad (7.1)$$

where  $r_{12.34\dots n}$  is the partial correlation of  $x_1$  and  $x_2$  given  $x_3, \dots, x_n$ , and so on. Previously such results had been inaccessible; in the old notation  $r_{12.34\dots n}$  could not even be distinguished from  $r_{12.4\dots n}$ .

Yule introduced notation for the quantities arising from regressions with  $n$  variables  $x_1, x_2, \dots, x_n$ : e.g.  $b_{13.24\dots n}$  denotes the coefficient of  $x_3$  in the regression of  $x_1$  on  $x_2, x_3, \dots, x_n$ ,  $x_{1.23\dots n}$  the (typical) residual from this regression and  $\sigma_{1.23\dots n}$  the standard deviation of the residuals. These are sample quantities though at one point Yule seems to have confused them with population quantities. Refining the notation to distinguish the two was a slow process only completed by Bartlett (1933). Yule did not index observations; it can be convenient to let his  $x_{1.23\dots n}$  represent a *vector*.

Yule stressed *notation* but the accompanying theorems actually did the work. His approach was different from Gauss's. Gauss found an existing sophisticated mathematical apparatus well-suited to least squares. Yule constructed a sophisticated specialised structure using only elementary mathematical apparatus. His calculus of subscripts exploits the properties of product-sums of deviations (residuals), which are close relatives of Laplace's  $\overline{r_2\alpha_2}$ .

The first theorem involves rewriting the normal equations as

$$\sum x_2 x_{1.23\dots n} = \sum x_3 x_{1.23\dots n} = \dots = \sum x_n x_{1.23\dots n} = 0.$$

The subscripts before the stop are "primary" subscripts and those after are "secondary" subscripts. The "order" of the residual/deviation is the number of secondary subscripts; a quantity such as  $x_2$  is a "deviation of order zero". The *first theorem* states: "The product-sum of any deviation of order zero with any deviation of higher order is zero, provided the subscript of the former occur amongst the secondary subscripts of the latter."

The *second theorem* is, "the product-sum of any two deviations of the same order, with the same secondary suffixes, is unaltered by omitting any or all of the secondary subscripts of either and, conversely, the product-sum of any deviation of order  $p$  with a deviation of order  $p + q$ , the  $p$  subscripts being the same in each case, is unaltered by adding to the secondary subscripts of the former any or all of the  $q$  additional subscripts of the latter".

$$\begin{aligned} \sum x_{1.3\dots n} x_{2.3\dots n} &= \sum x_{1.3\dots n} (x_2 - b_{23.4\dots n} x_3 - \dots - b_{2n.3\dots n-1} x_n) \\ &= \sum x_{1.3\dots n} x_2. \end{aligned}$$

Similarly

$$\sum x_{1.3\dots n} x_{2.3\dots n-1} = \sum x_{1.3\dots n} x_2$$

and so on. Therefore, quite generally,

$$\sum x_{1.3\dots n} x_{2.3\dots n} = \sum x_{1.3\dots n} x_{2.3\dots n-1} = \dots = \sum x_{1.3\dots n} x_2.$$

The *third theorem*, that "the product-sum of any two deviations is zero if all the subscripts of the one are contained among the secondary subscripts of the other", comes from combining the first and second theorems.

The system is rigorous, and apart from Theorem 1, self-contained. It is also very formal; Yule had no real interest in deviations and the theorems simply register patterns in subscripts. The reader of the *Introduction* (1911, p. 235) is told "The theorems . . . are of fundamental importance, and should be carefully remembered". The modern reader can try to understand them as well.

## 8 Yule's New Results

Yule considered his first result of theoretical significance only; it characterises the multiple regression coefficient  $b_{12.3\dots n}$  as the simple regression coefficient of  $x_{1.3\dots n}$  on  $x_{2.3\dots n}$ :

$$b_{12.3\dots n} = \frac{\sum x_{1.3\dots n} x_{2.3\dots n}}{\sum x_{2.3\dots n}^2}. \tag{8.1}$$

The proof is a typical application of the second and third theorems:

$$\begin{aligned}
0 &= \sum x_{2.3\dots n}x_{1.23\dots n} = \sum x_{2.3\dots n}(x_1 - b_{12.3\dots n}x_2 - \text{terms in } x_3 \text{ to } x_n) \\
&= \sum x_1x_{2.3\dots n} - b_{12.3\dots n} \sum x_2x_{2.3\dots n} \\
&= \sum x_{1.3\dots n}x_{2.3\dots n} - b_{12.3\dots n} \sum x_{2.3\dots n}^2.
\end{aligned}$$

Yule (p. 185) remarked on (8.1), “Such a relation would not, of course, afford a practical method of calculating the partial coefficients, as the arithmetic would be extremely lengthy”. Yet it is a cousin of (2.4): for a full triangle, (cf. (2.5)) repeat the argument with  $x_{3.4\dots n}, \dots$  and  $x_n$  instead of  $x_{2.3\dots n}$ .

For Yule, the result (8.1) was significant for the interpretation of partial correlation. He combined the new characterisation of  $b_{12.3\dots n}$  and the corresponding one for  $b_{21.3\dots n}$  with the old definition of partial correlation

$$r_{12.3\dots n} = \sqrt{(b_{12.3\dots n}b_{21.3\dots n})},$$

to produce

$$r_{12.3\dots n} = \frac{\sum x_{1.3\dots n}x_{2.3\dots n}}{\sqrt{(\sum x_{1.3\dots n}^2 \sum x_{2.3\dots n}^2)}}. \quad (8.2)$$

The partial correlation of  $x_1$  and  $x_2$  holding  $x_3, \dots, x_n$  constant is the total correlation of the residuals  $x_{1.3\dots n}$  and  $x_{2.3\dots n}$ . By characterising partial correlation as an “actual correlation between determinate variables” Yule could extend properties of total correlations to partial correlations.

Curiously Yule did not use (8.2) to redefine partial correlation, though it is a better match to the idea of partial correlation as correlation after allowing for other factors. (8.2) was also more fruitful: it was the basis for Fisher’s work on the distribution of the partial correlation coefficient—see Section 10 below. Today (8.2) is standard, although (6.3) had a good long run.

Yule established numerous formulae relating statistics—standard deviations, correlations and regression coefficients—of different orders modelled on relation (7.1). These recursive formulae were a vent for Yule’s post-discovery euphoria but they also had a role in computation. Yule went from the total coefficients (zero order coefficients) for *all* pairs of variables to increasingly complicated partial coefficients. Yule’s scheme, unlike elimination, was not directed at a single multiple regression: the preliminary regressions and correlations were of interest too. However his computational scheme did not survive; later correlationists used elimination.

Yule’s formulae appear in other formalisms, as devices to save re-doing all the calculations when regressors are added or removed. Goedseels (1902) and Wright & Hayford (1906, p. 117) treat such problems using Gaussian elimination while Fisher (1938) and Cochran (1938) directly manipulate the normal equations. More recently matrix methods have been used: Lovell (1963) provided a synthesis for the age of stepwise regression. The different formalisms never became totally alien—there were always some bilingualists—yet it was easier to work in the home formalism than to go into another to search for what—if anything—had been done there and translate back.

Yule did not call his paper “Some Properties of Least Squares Residuals with Applications to the Theory of Correlation.” If he had, it might have had some impact on least squares theory. However developments in correlation seem to have had little reaction back on the least squares formalism—even in the work of the same person. Fisher worked in both fields and used geometric methods in correlation and algebraic methods in least squares—see Sections 10 and 11 below. However correlation refreshed the topic of least squares by increasing enormously its range of applications; correlation also influenced the way least squares was expounded—e.g. by starting with simple regression instead of with the mean and by using scatter diagrams for visualising the procedure.

### 9 Orthogonalisation

Yule started with the properties of product-sums of residuals. Modern formalisms start further back: with the residuals or with the operations that convert variables into residuals. In this and later sections we consider the relationship between the modern formalisms and the schemes described above.

Two of Yule’s theorems can be read as expressing the orthogonality properties of certain residual vectors. Laplace’s method also exploits the orthogonality of *the* residual vector to the vectors created in the process of elimination. Moreover those vectors  $a, b_1, c_2$  etc. given by

$$a; b_1 = b - \frac{(ab)}{(aa)}a; c_2 = c_1 - \frac{(c_1b_1)}{(b_1b_1)}b_1; \text{ etc.} \tag{9.1}$$

are mutually orthogonal.

The construction (9.1) is usually called the “modified Gram–Schmidt orthogonalisation procedure”. Farebrother (pp. 66–8) prefers the name *Laplace orthogonalisation* procedure. Yet, though Laplace described a process by which orthogonal vectors are generated, he did not recognise their orthogonality. A triangular system can be obtained by noting this property and the simple form the normal equations take when the regressors are orthogonal. Rewrite (6.1) expressing  $a, b$  and  $c$  in terms of the orthogonal components  $a, b_1$  and  $c_2$  to obtain orthogonal regressors:

$$\begin{aligned} v &= ap + bq + cr - m \\ &= ap + \left( a \frac{(ab)}{(aa)} + b_1 \right)q + \left( a \frac{(ca)}{(aa)} + b_1 \frac{(c_1b_1)}{(b_1b_1)} + c_2 \right)r - m \\ &= a \left( p + \frac{(ab)}{(aa)}q + \frac{(ca)}{(aa)}r \right) + b_1 \left( q + \frac{(c_1b_1)}{(b_1b_1)}r \right) + c_2r - m. \end{aligned}$$

This way of writing  $v$  generates a triangular system of normal equations. The system becomes (5.6) on noting that  $(b_1m) = (b_1m_1)$  and  $(c_2m) = (c_2m_2)$ .

Systematic attention to orthogonal regressors seems to have begun with Chebyshev’s work on polynomial regressors. Expressed in my neo-Gaussian notation, Chebyshev (1858) developed  $m$  in terms of orthogonal regressors as

$$m = ak_0 + b_1k_1 + c_2k_2 + \dots$$

and obtained  $k_2$  as  $(mc_2)/(c_2c_2)$  etc. However in the later paper on least squares computations he (1859, pp. 482–98) found the  $k$ ’s from the equations

$$\begin{aligned} (am) &= k_0(aa) \\ (bm) &= k_0(ba) + k_1(bb_1) \\ (cm) &= k_0(ca) + k_1(cb_1) + k_2(cc_2) \end{aligned}$$

using the orthogonality of  $a$  to  $b_1, c_2, \dots$  and that of  $b$  to  $c_2, \dots$  etc. Chebyshev’s method of constructing the orthogonal polynomials based on a second order recursion—see Seal (p. 9)—does not seem to have had much general impact. Aitken (1933) and Romanovsky (1925) replaced it by the Schmidt method.

Like procedure (9.1), the Schmidt (1908, p. 61) method is based on bivariate regression but each vector is expressed not in terms of deviations of the next lowest order but of deviations of *all* lower orders. Thus the vector  $c_2$  is given by

$$c_2 = c - \frac{(cb_1)}{(b_1b_1)}b_1 - \frac{(ca)}{(aa)}a.$$

The two ways of expressing higher order deviations correspond to the two ways of writing higher order auxiliaries given in (4.5) and (4.6) above. Schmidt did not refer to Laplace or Gauss and of course the notation used is mine.

There is an orthogonalisation procedure implicit in Yule, the formation of the deviations  $x_1$ ,  $x_{2.1}$  and  $x_{3.12}$ . These multiple regression constructions can be related to bivariate processes using Yule's single result on deviations, as distinct from product-sums of deviations. This result conformed to the pattern that any equation between correlations, regressions and standard deviations "holds good for all secondary subscripts" (cf. (6.3) & (7.1)). Yule (p. 190) showed that the expression defining the deviation

$$x_{1.2\dots n} = x_1 - b_{12.3\dots n}x_2 - \dots - b_{1n.23\dots n-1}x_n \quad (9.2)$$

can be extended, by adding secondary subscripts, to

$$x_{1.2\dots kn} = x_{1.k} - b_{12.3\dots kn}x_{2.k} - \dots - b_{1n.23\dots k(n-1)}x_{n.k} \quad (9.3)$$

where  $k$  is any subscript or collection of subscripts.

Yule gave no use for this result but it can be used to justify the sequence,

$$x_1; x_{2.1} = x_2 - b_{2.1}x_1; x_{3.12} = x_{3.1} - b_{32.1}x_{2.1}; \text{ etc.} \quad (9.4)$$

For by (9.3) the expression for  $x_{3.12}$  can be established from the definition of the first-order deviation

$$x_{3.2} = x_3 - b_{32}x_2$$

by adding the secondary subscript, 1. The vectors can be put in modified Gram–Schmidt form by noting that by (8.2) the multiple regression coefficient  $b_{32.1}$  can be obtained by regressing  $x_{3.1}$  on  $x_{2.1}$ .

Yule's notation was not designed around the repeated bivariate regressions of the Laplace scheme but around all possible multiple regressions involving a set of variables. However Yule's multiple regression residuals in (9.4) equal the residuals in (9.1) constructed by repeated bivariate regressions.

## 10 Geometry: Trigonometry and Hilbert Space Theory

Orthogonality is an essential ingredient of the modern geometric conception of regression theory. Although there was earlier work, e.g. Bose (1944), that conception only became fixed in the 1960s with the importation of a geometric system of linear algebra. Before then the interpretation of observations on variables as *points* had been used but in a different way. Herr (1980) discusses some of the material treated here and in Section 13 below.

R.A. Fisher (1890–1962) used geometry throughout his work on distribution theory but his treatment of correlation is of most interest here. In his paper on the exact distribution of the correlation coefficient he (1915, p. 509) wrote, "The five quantities [associated with the bivariate normal] have ... an exceedingly beautiful interpretation in generalised space." The  $n$  observations on a pair of variables can be represented by two points  $P$  and  $Q$  in  $n$ -dimensional space and the correlation coefficient interpreted as the cosine of the angle between  $OP$  and  $OQ$ . Later Fisher (1924, p. 330) observed that when there is a third vector corresponding to point  $R$ , the partial correlation of the

original vectors is the “cosine of the angle between the projections of  $OP$  and  $OQ$  upon the region perpendicular to  $OR$ ”. This is a geometric interpretation of residuals and of (8.2) above. It also realises the abstract projection analysis from Hilbert space theory of Section 13 below.

Using these interpretations Fisher showed that the partial correlation based on  $n$  pairs of observations has the same distribution as the total correlation based on  $(n - 1)$  pairs. He first noted that in the case of total correlation the effect of deleting the first pair of observations is to project everything on to a “region at right angles to one of the axes of coordinates”. He then argued that by choosing new coordinates  $OR$  can be made to lie along an axis. This orthogonal transformation leaves angles unaffected and so the distribution of the angle defining the partial correlation must be the same as that defining total correlation in the case of  $(n - 1)$  observations. The insight reduced the distribution of the partial correlation to a trivial variation on that of the total correlation—something of a self-annihilating insight.

The treatise on the trigonometry of correlation that Pearson (1916, p. 237) thought “greatly to be desired” never materialised. Kendall’s (1961) *Course in the Geometry of  $n$  Dimensions* is a partial offering but it appeared just as a new approach was taking off. This drew on the Hilbert space theory developed in the early part of the century and assembled by Stone (1932): Birkhoff & Kreysig (1984) and Dieudonné (1981) discuss the development. The “projections” that invaded least squares in the 1960s (see Section 13) owed more to Stone (pp. 70–5) and his idempotent operators than to Fisher.

Halmos’s influential *Finite Dimensional Vector Spaces* (1942, p. i) starts from the observation that the “seemingly separate” Hilbert space theory and elementary matrix theory are “intimately associated”, are actually different ways of doing the same thing. The implications of this situation for regression theory were only worked out on any scale in the 1960s, when the matrix formalism was dominant. Kruskal (1961, p. 435) argued that the vector space approach “permits a simpler, more general, more elegant, and more direct treatment of the general theory of linear estimation than do its notational competitors”.

Long before Kruskal, Wold (1938, p. 76) had observed that “the multiple regression theory founded and developed by the English statistical school . . . can be interpreted as a particular branch of theory of approximations in general linear spaces”. Wold’s work led on to Kolmogorov’s (1939) treatment of least squares extrapolation of stationary sequences in Hilbert space. However, though this work was known to some statisticians, it was not the vehicle by which the Hilbert space paradigm entered statistical regression analysis.

An even earlier “non-vehicle” was Schmidt’s paper referred to above, which was one of the main sources of the new geometry. It treats the space of square-summable sequences, the prototype Hilbert space  $l^2$ , using geometric ideas on orthogonality, Pythagoras’ theorem, etc. Schmidt’s paper was part of the Hilbert school’s project on the theory of integral equations—see Kline (pp. 1052–75). Nevertheless Schmidt (1908) referred to Gram’s work on least squares (Section 11 below) and gave one least squares result: he (pp. 64–5) used Pythagoras’s theorem to show that the combination

$$\sum_v (D; B_v) B_v(x)$$

of the orthonormal vectors  $\{B_v(x)\}$  gives a better approximation to an arbitrary point  $D(x)$  than any other linear combination of the same vectors. The notation,  $(A; B)$  for  $\sum_x A(x)B(x)$ , which became standard in Hilbert space theory for expressing ‘such’ quantities seems to have descended from Gauss but Gauss did not conceive of  $A$  and  $B$  as entities in their own right.

The geometric formulation of least squares and the use of Pythagoras’s theorem for obtaining the normal equations—or proving Yule’s first theorem—only became common in statistics much later. However it seems to be implicit in Bartlett’s (1934, p. 329) derivation of the normal equations.

Statisticians took up Schmidt’s orthogonalisation process more readily than the accompanying geometry. When Wold put Yule and Schmidt together in the context of the linear approximation

of random variables he drew on Kowalewski (1909) for orthogonalisation and on Kendall's (1937) updating of Yule's *Introduction*: determinants provided the common framework.

**11 Algebra: Determinants and Matrices**

In the late nineteenth and early twentieth centuries determinants had a central role in treating the linear equations and quadratic forms of applied mathematics, including least squares. By the mid-twentieth century matrices had taken over. Matrices were introduced by Cayley in the 1850s, when the long ascendancy of determinants in least squares work was only beginning: Cayley himself contributed a method of evaluating the constant of proportionality of the multivariate normal distribution to Todhunter's (1865) determinant based analysis. According to Hawkins (1977, p. 83), the theories of determinants and matrices were both "outgrowths" of Gauss's work—his work on number theory. Kline (1972, ch. 33) outlines the relevant history.

Gram (1883) used determinants in developing a procedure for constructing orthogonal regressors. His (pp. 43–5) procedure for constructing orthogonal regressors  $\Phi_1, \Phi_2, \dots, \Phi_n$  from the regressors  $x_1, x_2, \dots, x_n$  is based on the construction of least squares 'predictors' with  $y^{(k)}$  based on regressors  $x_1, x_2, \dots, x_k$ . Gram then exploits the identity

$$y^{(n)} = y^{(1)} + (y^{(2)} - y^{(1)}) + \dots + (y^{(n)} - y^{(n-1)}). \tag{11.1}$$

Gram used elementary determinant theory—Cramer's rule and expansion by alien co-factors—to show that the terms on the right are orthogonal. They are proportional to the orthogonal functions  $\Phi_1, \Phi_2, \dots, \Phi_n$  given by

$$\Phi_k = \begin{vmatrix} p_{11} p_{12} & \dots & p_{1,k-1} x_1 \\ p_{21} p_{22} & \dots & p_{2,k-1} x_2 \\ \dots & \dots & \dots \\ p_{k1} p_{k2} & \dots & p_{k,k-1} x_k \end{vmatrix} \text{ where } p_{ij} = \sum x_i x_j.$$

The  $\Phi$ 's are not formed recursively—they are *multiple* regression quantities—but Schmidt saw that they are proportional to the vectors generated by his own process. Gram referred to Chebyshev for the simple form of the regression coefficients associated with the  $\Phi$ 's. Dieudonné (1981, p. 60) places Gram's work in the history of functional analysis.

Determinants were used in other important work on least squares and correlation—including Pearson (1896), Fisher (1922), Frisch & Waugh (1933) and David & Neyman (1938). Yule did not use them but Pearson (1916) went on to obtain Yule's results by "direct determinantal analysis", a task that required new results on determinants. Determinants were frequently brought to the assistance of the least squares argument but they never carried it, unlike Yule's product-sums or—later—projections.

The use of matrices could unify many of the disparate themes in least squares analysis. Matrices could support determinants and accommodate such processes as *elimination and projection*. Arguments based on manipulating sets of linear equations seem easier to grasp when reduced to rules of matrix algebra. So arguments from earlier writers—from Gauss especially—were given a new life.

The worth of matrices was realised by several writers independently. Bartlett was one of them but the most influential was A.C. Aitken (1895–1967), an algebraist who worked on orthogonal polynomials and the numerical solution of linear equations; he also wrote the widely-used textbook *Determinants and Matrices* (1939). Whittaker & Bartlett (1968) review his life and work.

The treatise by Turnbull & Aitken (1932) used least squares to illustrate the utility of matrices. This was a new departure; their only reference was to Frisch's (1929) study of the correlation matrix. By vector differentiation of the sum of squared deviations, they obtained the normal equations and

then

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{11.2}$$

I have modernised their notation; they considered least squares as a method of solving the inconsistent equations,  $\mathbf{Ax} = \mathbf{h}$ . The familiar use of  $\mathbf{X}$  and  $\mathbf{y}$  only became established around 1950—see Durbin & Watson (1950) and Kempthorne (1952) writing from different corners of the regression universe—and could be described as Fisher in matrices.

More than anybody, Fisher was responsible for the creation of modern regression analysis. His innovations beginning in the 1920s—conditioning on covariates,  $t$ - and  $F$ -tests, analysis of variance and theory of experimental design—are beyond the scope of this paper. I will only consider his influence on the *formalism*. In *Statistical Methods* Fisher (1925, pp. 130–7) set up multiple regression with the expected value of  $y$  (the dependent variable), denoted by  $Y$ , related to the values of the  $x$ 's (the independent variables):

$$Y = b_1x_1 + b_2x_2 + b_3x_3 \tag{11.3}$$

where the variables are measured from their means. The notation is even simpler than early Yule (Section 6 above) for Fisher, like Gauss and Laplace, was primarily interested in a single multiple regression equation. (11.3) is something of a notational lapse for Fisher as it was under his influence that the use of different symbols for the population regression coefficients and the estimates became mandatory; earlier sections of this paper have followed the looser custom of earlier ages. For the emphasis on the need to distinguish between “statistics” and “parameters”—Fisher’s terms—see Aldrich (1997).

Sometimes Fisher used arguments that almost beg to be put into matrices. After writing the normal equations, he (1925, p. 131) added “It frequently happens that, for the same set of values of the independent variates, it is desired to examine the regressions for more than one set of values of the dependent variates.” So it is best to solve once and for all the sets of equations:

$$\begin{aligned} b_1S(x_1^2) &+ b_2S(x_1x_2) &+ b_3S(x_1x_3) &= 1, 0, 0 \\ b_1S(x_2x_1) &+ b_2S(x_2^2) &+ b_3S(x_2x_3) &= 0, 1, 0 \\ b_1S(x_3x_1) &+ b_2S(x_3x_2) &+ b_3S(x_3^2) &= 0, 0, 1. \end{aligned}$$

Fisher denoted the solution of these equations by  $c_{11}, c_{12}, \dots, c_{33}$ . He then described how the regression coefficients are calculated when the  $y$  values are specified. Fisher described the calculations quite clearly but he did not explain why they work. Put into matrices the method of solving the normal equations is less mysterious: it is based on combining

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} \quad \text{and} \quad \mathbf{b} = \mathbf{CX}'\mathbf{y}$$

to form (11.2). In later editions of the *Methods* (e.g. 1938, p.168) Fisher referred to the “ $c$ -matrix” or covariance matrix; it had been known since Gauss (1821) that the sampling variance of  $b_i$  is proportional to  $c_{ii}$ . Some Fisherians—e.g. Yates & Hale (1939)—referred to the “reciprocal matrix” but they did not manipulate matrices. Fisher’s computational procedure was criticised by Deming (1943, p. 218) on grounds of numerical instability, “The use of the reciprocal matrix as a multiplier is in theory very fascinating, but as a practical matter in curve fitting we should not wax too enthusiastic about it.”

Returning to Aitken, his best-known piece (1935) treated a version of—what came to be called—the “Gauss–Markov theorem”. He was concerned with extending a result from the actuarial literature on fitting time polynomials. The theorem became the focus of a lively contest between formalists.

Lidstone (1933, p. 155) was puzzled that two minimisation processes with “completely different logical bases” led to identical results—something Sheppard (1912) had shown in the course of a “very heavy algebraical investigation”. So Lidstone gave another proof, a “rather simple application” of orthogonal polynomials. Aitken (1935, p. 42) pursued the problem, declaring that “to attain full generality with adequate compactness we shall employ the notation of matrices and vectors”. Aitken assumes only that  $\mathbf{X}$  has full rank.

Aitken proved his theorem by an easy application of a “simple but very useful” matrix lemma on constrained minimisation, proved using Lagrange multipliers. David & Neyman (1938, p. 105) judged Aitken’s paper “remarkably clear and elegant” but found it worthwhile to present a result using reasoning of a “more elementary character”—reasoning based on determinants. They saw their work as an “extension” of a neglected Markov theorem on least squares.

Plackett (1949) pointed out that all the results were variations on a neglected result of Gauss (1821); Aitken and Bose had gone beyond Gauss by considering an arbitrary error covariance matrix and less than full rank  $\mathbf{X}$  respectively. Gauss’s proof was clear and elegant *and* elementary: it did not even use multipliers. Plackett put Gauss into matrices and modern textbooks usually give ‘their’ argument. Durbin & Kendall (1951), like Lidstone, were puzzled that two apparently unrelated problems should have the same solution; they gave a geometric explanation based on the duality of lines and points.

## 12 Triangular Matrices: Gauss and Yule

Aitken’s paper demonstrated the power of matrices in obtaining new—or what seemed to be new—results. Matrices were also used to re-present old arguments. Since the 1940s Gaussian elimination has been presented in terms of matrices. Dwyer (1944) presented the process of Section 2 in terms of factorising a positive definite matrix; he (1951, p. 109) reported later that Banachiewicz had seen the relevance of matrix factorisation somewhat earlier. Dwyer (1944, p. 88) showed that the matrix,  $\mathbf{A}$  say, can be written as

$$\mathbf{A} = \mathbf{S}'\mathbf{D}^{-1}\mathbf{S} \quad (12.1)$$

where  $\mathbf{D}$  is diagonal and  $\mathbf{S}$  triangular. Dwyer used (12.1) to motivate a new least squares algorithm. However Cholesky—see Benôit (1924) had already proposed the algorithm, but without any matrix argument. The factorisation is often called after him—see Horn & Johnson (1985, pp. 114 & 407)—though it is just as implicit in Gauss’s reduction of  $W$ . To make the translation from (12.1) to (3.1), write  $W$  in terms of a matrix  $\mathbf{A}$

$$W = [-p - q - r] \begin{bmatrix} (aa) & (ab) & (ac) & (am) \\ (ab) & (bb) & (bc) & (bm) \\ (ac) & (bc) & (cc) & (cm) \\ (am) & (bm) & (cm) & (mm) \end{bmatrix} \begin{bmatrix} -p \\ -q \\ -r \\ 1 \end{bmatrix}$$

Then  $\mathbf{D}$  has elements  $(aa)$ ,  $(bb, 1)$ ,  $(cc, 2)$  and  $(mm, 3)$  and  $\mathbf{S}$  is given by

$$\mathbf{S} = \begin{bmatrix} (aa) & (ab) & (ac) & (am) \\ 0 & (bb, 1) & (bc, 1) & (bm, 1) \\ 0 & 0 & (cc, 2) & (cm, 2) \\ 0 & 0 & 0 & (mm, 3) \end{bmatrix}$$

Dwyer did not use the Gaussian symbols but a system of subscripts resembling Yule’s with  $a_{33}$ ,  $a_{33.1}$  and  $a_{33.12}$  etc., though here the order of the secondary subscripts reflects the order in which variables are eliminated. It seems that Dwyer rediscovered Yule’s *calculus* for he (1941, p. 458)

wrote “the notation . . . is very useful in providing an easy development of various theorems in multiple and partial correlation studies”.

Turnbull and Aitken showed how the Schmidt process can be presented as a matrix theorem showing the interplay of orthogonality and triangularity. The Yulean version of their matrix factorisation theorem is:

$$(x_1, x_{2.1}, x_{3.12}) = (x_1, x_2, x_3) \begin{bmatrix} 1 & -b_{21} & -b_{31.2} \\ 0 & 1 & -b_{32.1} \\ 0 & 0 & 1 \end{bmatrix} \quad (12.2)$$

where  $x_1, x_{2.1}, x_{3.12}$  are orthogonal vectors. (12.2) combines the definition of the residuals  $x_1, x_{2.1}$  and  $x_{3.12}$  and the implication of Yule’s Theorem 2 that they are orthogonal. The orthogonal decomposition of  $(x_1, x_2, x_3)$  can be obtained by inverting the triangular matrix. From this (12.1) may be obtained.

### 13 Projections and Subscripts

Matrix factorisation is the modern way of presenting Gaussian elimination. There is no modern way of presenting Yule’s calculus of subscripts for it has not survived. However some features of the calculus can be re-created using the idempotency properties of the residual-creating transformations.

Aitken (1935) introduced these idempotent matrices when he digressed to give a “simple proof of a classical result”. Gauss (1821, pp. 77–9) had shown—in modern terms—that the residual sum of squares divided by the number of degrees of freedom is an unbiased estimator of  $\sigma^2$ . Aitken’s (pp. 43–4) proof starts by showing that the matrices, written in the usual notation,

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{ and } \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

are symmetric and idempotent. The first converts crude data into “graduated” values, the second into residuals. Aitken’s proof turns on showing that the trace of the second matrix equals the number of degrees of freedom.

Aitken (1945) continued to study such linear transformations. One new result (p. 143)—I have altered his notation again—was that for any matrix  $\mathbf{X}$  with linearly independent columns and submatrix  $\mathbf{X}_1$  made up of the first  $k_1$  of these columns

$$\mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1 \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1.$$

The proof takes off from the identity (11.1) and puts Gram into reverse: the incremental contributions are orthogonal because they can be expressed in terms of orthogonalised regressors.

Aitken’s illustration is that the estimates are the same whether we fit a cubic to crude data or fit a cubic to graduated values obtained from first fitting a quintic. A more vital concern with the components of (11.1) and in idempotent matrices generally came from the analysis of variance. Quadratic forms in normal variables and their independence were established as a major topic by Cochran (1934). Craig (1943) brought idempotent matrices into play.

In the terms of the ‘finitised’ Hilbert space scheme, Aitken’s results concern projections—see e.g. the textbooks by Seber (1977, p. 396) or Davidson & MacKinnon (1993, pp. 9–11). The notion that least squares is all about projections was the key to the construction of a least squares theory that tied together estimation and test theory. However this theory, unlike Yule’s collection of mysterious rules, belonged to mainstream mathematics.

The projection counterparts of Aitken’s results are (roughly) as follows. Let  $\mathbf{P}$  be the operator that projects on to the  $\mathbf{X}$  space and  $\mathbf{P}_1$  that which projects on to the  $\mathbf{X}_1$  space, a subspace of the  $\mathbf{X}$  space; the complementary projections are  $\mathbf{M}$  and  $\mathbf{M}_1$ . The relevant results (see Seber (1977, p. 396) are

$$\begin{aligned} \mathbf{P}^2 &= \mathbf{P} & \text{and} & & \mathbf{M}^2 &= \mathbf{M} \\ \mathbf{P}_1\mathbf{P} &= \mathbf{P}_1 & \text{and} & & \mathbf{M}\mathbf{M}_1 &= \mathbf{M}. \end{aligned}$$

The proofs given are much more direct than those based on manipulating the corresponding matrix expressions—Aitken's 'graduating' matrices.

Yule's theorems from Section 7 can be re-stated and proved using idempotent matrices/projections. His first theorem states that the residual vector  $\mathbf{M}\mathbf{y}$  is orthogonal to the regressor vectors, i.e.

$$\mathbf{y}'\mathbf{M}\mathbf{X} = 0.$$

The second and third theorems follow immediately from the results on iterated projections applied to the projections,  $\mathbf{M}$  and  $\mathbf{M}_1$ . The second theorem states

$$\mathbf{y}'\mathbf{M}\mathbf{M}_1\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{M}\mathbf{y}$$

and the third theorem is a special case of the result

$$\mathbf{X}'_1\mathbf{M}_2\mathbf{M}\mathbf{y} = \mathbf{X}'_1\mathbf{M}\mathbf{y} = 0$$

where  $\mathbf{M}_2$  is associated with  $\mathbf{X}_2$ , a submatrix of  $\mathbf{X}$ .

Yule's results in Section 8 can be given the same treatment. Thus Davidson & MacKinnon (pp. 19–24) present a generalisation of (8.1) with (9.3) as the "Frisch–Waugh–Lovell theorem"—viz. that in the regression of  $\mathbf{y}$  on  $\mathbf{X}_1$  and  $\mathbf{X}_2$ ,

$$\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e},$$

$\mathbf{b}_2$  can be calculated using residuals from the regressions of  $\mathbf{y}$  and  $\mathbf{X}_2$  on  $\mathbf{X}_1$ :

$$\mathbf{b}_2 = (\mathbf{X}'_2\mathbf{M}_1\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{M}_1\mathbf{M}_1\mathbf{y}.$$

The name recognises Lovell's (1963) extensions to the work of Frisch & Waugh (1933)—above Section 5—but Laplace and Yule could also be recognised.

The modern technique does much more than Yule's. It even *shows* what is going on, including the creation of patterns in Yule's subscripts. Yet there is a loss. Yule's elaborate notation makes visible the kinship between

$$b_{12.3\dots n} = \frac{\sum x_{1.3\dots n}x_{2.3\dots n}}{\sum x_{2.3\dots n}^2} \text{ and } b_{12} = \frac{\sum x_1x_2}{\sum x_2^2}$$

$\mathbf{b}_2$  is similarly related to the estimate of the regression of  $\mathbf{y}$  on  $\mathbf{X}_2$  viz.,

$$(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{y},$$

but the simplified modern notation does nothing to declare the kinship.

## 14 Discussion

I have described how some least squares schemes came to be developed. Rather than summarise the account, I will close with some general points about ways of doing least squares. I emphasise again the limited coverage of my review. The caveat applies especially to work on the relationships between the schemes. Every one is surely well understood 'somewhere'—some appear as exercises in Dempster (1969). However to trace the growth of that understanding would be quite a task.

The relationship between different methods of least squares has a logical and a historical dimension. It is clear that—to paraphrase Halmos—“seemingly separate” approaches, based on such devices as auxiliaries, product-sums of deviations and projections *must* be “intimately associated”. This is not to say that there was any simple historical connection. The broadening and deepening apparent in the sequence of devices was *not* the result of a desire to improve on the methods of an earlier ‘stage’. The relationship between Yule and Gauss is typical: Yule took Gauss’s normal equations—not the auxiliaries, which I do not think he knew about—and built his system of product-sums around them.

The “seemingly separate” approaches were designed for different jobs but there was duplication of results—some are described in Section 8. This duplication was unplanned and easily missed. There was also deliberate duplication, as in the Gauss–Markov work described in Section 11. Here the association is in history as well as in logic: the innovation was promoted as a better way of doing the old job and as a way of doing jobs the old methods could not do.

Effectiveness was not the only criterion: “simplicity”, “generality”, “suggestiveness”, “beauty”, “elegance”, “compactness”, “elementariness” and “directness” have been mentioned. The application of these criteria was not straightforward and could be disputed. There were disagreements about rigour. Pearson (1916, p. 235) distrusted Yule’s methods: the proof of (6.1) “seemed to be based on some appeal to general analogy”. Fisher’s geometric proofs were distrusted; his results were only fixed in ‘history’ when they had been established analytically. Bartlett, in conversation with Olkin (1989), has some interesting observations on this point—see also Anderson (1996).

Changes in least squares theorising usually followed changes in the branch of pure mathematics in which least squares was ‘embedded’: quadratic forms and linear equations. Yule’s new system is the grand exception. Gauss and Yule represent extremes in ways of doing least squares. Gauss met all the theoretical and computational requirements of least squares (as of 1809/10) by adapting an existing piece of sophisticated mathematics. Yule used elementary mathematical arguments to build a sophisticated system. The system is specialised or ad hoc, according to taste. Other authors developed specialised structures within a sophisticated mathematical environment or made modest additions to the mathematical apparatus to make it better adapted to the least squares argument. However the additions were never on the scale of those stimulated by the needs of, say, mechanics in earlier centuries: they did not constitute whole new branches of mathematics.

## Acknowledgements

This paper is a revision of “Doing Least Squares: Some Scenes from History” which was presented at the Australasian meeting of the Econometric Society in 1996. I am grateful to Raymond O’Brien and Janne Rayner for helpful discussions. In preparing the revision I am grateful to the referees for their suggestions, in particular for directing me to G.W. Stewart’s recent work.

## References

- Aitken, A.C. (1933). On Fitting Polynomials to Data with Weighted and Correlated Errors. *Proceedings of the Royal Society of Edinburgh*, **54**, 12–16.
- Aitken, A.C. (1935). On Least Squares and Linear Combinations of Observations. *Proceedings of the Royal Society of Edinburgh*, **55**, 42–48.
- Aitken, A.C. (1939). *Determinants and Matrices*. Edinburgh: Oliver & Boyd.
- Aitken, A.C. (1945). Studies in Practical Mathematics. IV. On Linear Approximation by Least Squares. *Proceedings of the Royal Society of Edinburgh*, **62**, 138–146.
- Aldrich, J. (1995). Correlations Genuine and Spurious in Pearson and Yule. *Statistical Science*, **10**, 364–376.
- Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912–22. *Statistical Science*, **12**, 162–176.
- Anderson, T.W. (1996). R. A. Fisher and Multivariate Analysis. *Statistical Science*, **11**, 20–34.
- Bartlett, M.S. (1933). On the Theory of Statistical Regression. *Proceedings of the Royal Society of Edinburgh*, **53**, 260–283.
- Bartlett, M.S. (1933/4). The Vector Representation of a Sample. *Proceedings of the Cambridge Philosophical Society*, **30**, 327–340.
- Benôit, E. (1924). Note sur une Étude de Résolution des Équations Normales etc. *Bulletin Géodésique*, 67–77.

- Bienaymé, I.J.(1853). Remarques sur les Différences qui Distinguent L'Interpolation de M. Cauchy de la Méthode des Moindres Carrés etc. *Compte Rendu de l'Académie des Sciences de Paris*, **37**, 5–13.
- Birkhoff, G. & Kreysig, E. (1984). The Establishment of Functional Analysis. *Historia Mathematica*, **11**, 258–321.
- Bôcher, M. (1907). *Introduction to Higher Algebra*. New York: Macmillan.
- Bose, R.C. (1944). The Fundamental Theorem of Linear Estimation. *Proceedings of the 31st Indian Scientific Congress*, (abstract) 2–3.
- Brunt, D. (1917). *The Combination of Observations*. Cambridge: Cambridge University Press.
- Cauchy, A. (1836). On a New Formula for Solving the Problem of Interpolation in a Manner Applicable to Physical Investigations. *Philosophical Magazine*, **49**, 459–468.
- Chauvenet, W. (1867). *A Manual of Spherical and Practical Astronomy*, volume 2, fifth edition (1891). Reprinted by Dover, New York 1960.
- Chebyshev, P.L. (1858). Sur les Fractions Continues. *Journal de Mathématiques Pures et Appliquées*, II, **3**, 289–323. Chapter 11 in *Œuvres* Tome 1.
- Chebyshev, P.L. (1859). Sur l'Interpolation par la Méthode des Moindres Carrés. *Mémoires de l'Académie des Sciences de St.-Petersbourg*, VII, série, **1**, 1–24. Chapter 18 in *Œuvres* Tome 1, reprinted by Chelsea, New York 1961.
- Cochran, W.G. (1934). The Distribution of Quadratic Forms in a Normal System, with Applications to the Analysis of Covariance. *Proceedings of the Cambridge Philosophical Society*, **30**, 178–191.
- Cochran, W.G. (1938). The Omission or Addition of an Independent Variable in Multiple Linear Regression. Supplement to *Journal of the Royal Statistical Society*, **5**, 171–176.
- Craig, A.T. (1943). Note on the Independence of Certain Quadratic Forms. *Annals of Mathematical Statistics*, **18**, 195–197.
- David, F.N. & Neyman, J. (1938). Extension of the Markoff Theorem on Least Squares. *Statistical Research Memoirs*, University College London, **2**, 105–116.
- Davidson, R. & MacKinnon, J.G.(1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
- Deming, W.E. (1943). *Statistical Adjustment of Data*. Reprinted by Dover, New York 1964.
- Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis*. Reading, Massachusetts: Addison-Wesley.
- Dieudonné, J.(1981). *History of Functional Analysis*. Amsterdam: North-Holland.
- Draper, N.R. & Smith, H. (1966). *Applied Regression Analysis*. New York: Wiley.
- Durbin, J. & Kendall, M.G. (1951). The Geometry of Estimation. *Biometrika*, **38**, 150–158.
- Durbin, J. & Watson, G.S. (1950). Testing for Serial Correlation in Least Squares Regression. *Biometrika*, **37**, 409–428.
- Dwyer, P.S. (1941). The Doolittle Technique. *Annals of Mathematical Statistics*, **14**, 449–458.
- Dwyer, P.S. (1944). A Matrix Presentation of Least Squares and Correlation Theory with Matrix Justification of Improved Methods of Solution. *Annals of Mathematical Statistics*, **17**, 82–89.
- Dwyer, P.S. (1951). *Linear Computations*. New York: Wiley.
- Encke, J.F. (1835). Über die Methode der kleinsten Quadrate. *Berliner Astronomisches Jahrbuch*, 253–320.
- Farebrother, R.W. (1988). *Linear Least Squares Computations*. New York: Dekker.
- Fisher, R.A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, **10**, 507–521.
- Fisher, R.A. (1922). The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients. *Journal of the Royal Statistical Society*, **85**, 597–612.
- Fisher, R.A. (1924). The Distribution of the Partial Correlation Coefficient. *Metron*, **3**, 329–332.
- Fisher, R.A. (1925/38). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd. Seventh edition in 1938.
- Fisher, R.A. (1925a). Applications of 'Student's' Distribution. *Metron*, **5**, 90–104.
- Frisch, R. (1929). Correlation and Scatter in Statistical Variables. *Nordic Statistical Journal*, **8**, 36–102.
- Frisch, R. & Waugh, F.V. (1933). Partial Time Regressions as Compared with Individual Trends. *Econometrica*, **1**, 387–401.
- Gauss, C.F. (1801). *Disquisitiones Arithmeticae*. English translation by A.A. Clarke (1965), New Haven: Yale University Press.
- Gauss, C.F. (1809). *Theoria Motus Corporum Coelestium*. English translation by C.H. Davis, reprinted 1963 Dover, New York.
- Gauss, C.F. (1811). Disquisitio de Elementis Ellipticis Palladis. English translation of extract in pp. 148–155 of Trotter, H. F. (1957). Gauss's Work (1803–26) on the Theory of Least Squares, Technical Report 5, Statistical Techniques Research Group, Princeton University. A translation of *Méthodes des Moindres Carrés*, the authorised French translation of Gauss's writings on least squares by J. Bertrand (1855), Paris: Mallet-Bachelier.
- Gauss, C.F. (1821/-3/-6). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*, in two parts with a supplement. Reprinted with an English translation and notes by G.W. Stewart, (1995), Philadelphia: SIAM.
- Goedseels, E. (1902). Sur l'Application de la Méthode de Cauchy aux Moindres Carrés. *Annales de la Société Scientifique de Bruxelles*, **26**, 148–156.
- Gram, J.P.(1883). Ueber die Entwicklung reeller Functionen in Reihen mittelst der Methode der kleinsten Quadrate. *Journal für die reine und angewandte Mathematik*, **94**, 41–73.
- Halmos, P.R. (1942). *Finite Dimensional Vector Spaces*. Princeton: Princeton University Press.
- Hawkins, T. (1977). Another Look at Cayley and the Theory of Matrices. *Archives Internationales d'Histoire des Sciences*, **26**, 82–112.
- Herr, D.G. (1980). On the History of the Use of Geometry in the General Linear Model. *The American Statistician*, **34**, 43–47.
- Heyde, C.C. & Seneta, E. (1977). *I.J. Bienaymé: Statistical Theory Anticipated*. New York: Springer-Verlag.
- Horn, R.A. & Johnson, C.R. (1985). *Matrix Analysis*. Cambridge: Cambridge University Press.
- Kempthorne, O. (1952). *The Design and Analysis of Experiments*. New York: Wiley.
- Kendall, M.G. (1961). *A Course in the Geometry of n Dimensions*. London: Griffin.

- Kline, M. (1972). *Mathematical Thought from Ancient to Modern Times*. New York: Oxford University Press.
- Kolmogorov, A.N. (1939). Sur l'Interpolation et Extrapolation des Suites Stationnaires. *Compte Rendu de l'Academie des Sciences de Paris*, **208**, 2043–2045.
- Kowalewski, G. (1909). *Einführung in die Determinantentheorie*. Leipzig.
- Kruskal, W. (1961). The Coordinate-Free Approach to Gauss-Markov Estimation, and its Application to Missing and Extra Observations. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 435–451.
- Lagrange, J.L. (1759). Recherches sur la Méthode de Maximis et Minimis, reprinted in *Œuvres*, vol. 1, pp. 3–20. Gauthiers-Villars, Paris 1867.
- Laplace, P.S. (1812). *Théorie Analytiques des Probabilités*, Third edition (1820) with introduction and three supplements reprinted in *Œuvres*, vol. 7, Imprimerie Royale, Paris 1847.
- Legendre, A.M. (1805). *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Paris: Courcier.
- Lidstone, G.J. (1933). Notes on Orthogonal Polynomials and their Application to Least-Square Methods etc. *Journal of the Institute of Actuaries*, **64**, 128–64.
- Longley, J.W. (1984). *Least Squares Computations using Orthogonalization Methods*. New York: Dekker.
- Lovell, M.S. (1963). Seasonal Adjustment of Economic Time Series. *Journal of the American Statistical Association*, **58**, 993–1010.
- Morgan, M.S. (1990). *The History of Econometric Ideas*. New York: Cambridge University Press.
- Olkin, I. (1989). A Conversation with Maurice Bartlett. *Statistical Science*, **4**, 151–163.
- Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society A*, **187**, 253–318.
- Pearson, K. (1916). On Some Novel Properties of Partial and Multiple Correlation in a Universe of Manifold Characteristics. *Biometrika*, **11**, 231–238.
- Pearson, K. & Filon, L.N.G. (1898). Mathematical Contributions to the Theory of Evolution IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation. *Philosophical Transactions of the Royal Society A*, **191**, 229–311.
- Plackett, R.L. (1949). A Historical Note on the Method of Least Squares. *Biometrika*, **36**, 458–460.
- Plackett, R.L. (1972). The Discovery of the Method of Least Squares. *Biometrika*, **59**, 239–251.
- Rainsford, H.F. (1957). *Survey Adjustments and Least Squares*. London: Constable.
- Romanovsky, V. (1925). Sur une Méthode d'Interpolation de Tchebycheff. *Compte Rendu de l'Academie des Sciences de Paris*, **181**, 595–597.
- Schmidt, E. (1908). Über die Auflösung linearer Gleichungen-mit unendlich vielen Unbekannten. *Rendiconti del Circolo Matematico di Palermo*, **25**, 53–77.
- Seal, H.L. (1967). The Historical Development of the Gauss Linear Model. *Biometrika*, **54**, 1–24.
- Seber, G.A.F. (1977). *Linear Regression Analysis*. New York: Wiley.
- Sheppard, W.F. (1912). Reduction of Errors by Means of Negligible Differences. *Proceedings of the Fifth International Congress of Mathematicians (Cambridge)*, **2**, 348–384.
- Smart, W.M. (1958). *The Combination of Observations*. Cambridge: Cambridge University Press.
- Stewart, G.W. (1995). Gauss, Statistics, and Gaussian Elimination. *Journal of Computational and Graphical Statistics*, **4**, 1–11.
- Stigler, S.M. (1986). *The History of Statistics*. Cambridge: Harvard University Press.
- Stone, M.H. (1932). *Linear Transformations in Hilbert Space*. New York: American Mathematical Society.
- Todhunter, I. (1865). On the Method of Least Squares. *Proceedings of the Cambridge Philosophical Society*, **11**, 219–238.
- Turnbull, H.W. & Aitken, A.C. (1932). *An Introduction to the Theory of Canonical Matrices*. London: Blackie.
- Whittaker, J.M. & Bartlett, M.S. (1968). Alexander Craig Aitken, 1895–1967. *Obituary Notices of Fellows of the Royal Society*, **24**, 1–14.
- Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. Uppsala: Almqvist & Wiksells.
- Wright, T.W. & Hayford, J.F. (1906). *The Adjustment of Observations*. New York: Van Nostrand.
- Yates, F. & Hale, R.W. (1939). The Analysis of Latin Squares when Two or More Rows, Columns or Treatments are Missing. *Journal of the Royal Statistical Society B*, **6**, 67–79.
- Yule, G.U. (1897). On the Theory of Correlation. *Journal of the Royal Statistical Society*, **60**, 812–854.
- Yule, G.U. (1907). On the Theory of Correlation for any Number of Variables, treated by a New System of Notation. *Philosophical Transactions of the Royal Society A*, **79**, 182–193.
- Yule, G.U. (1909). The Applications of the Method of Correlation to Social and Economic Statistics. *Journal of the Royal Statistical Society*, **72**, 721–730.
- Yule, G.U. (1911). *An Introduction to the Theory of Statistics*. London: Griffin.
- Yule, G.U. & Kendall, M.S. (1937). *An Introduction to the Theory of Statistics*. London: Griffin.

## Résumé

Gauss a présenté une méthode pour obtenir les estimations de la méthode des moindres carrés et leurs précisions. Yule a présenté un nouveau système de la notation adapté à l'analyse de la corrélation. Cette étude décrit ces notations et les compare avec les notations du calcul matricien et de l'espace vectoriel que on emploie en l'analyse moderne de la régression.

[Received October 1996, accepted June 1997]