THE POWER OF THE MATRIX

BU-1563 -M **March 2001**

Shayle R. Searle

Keywords: matrix proofs, linear models.

Abstract:

The first linear models book by Graybill (1961) was one of the early major works to address many of the problems arising out of extending regression algebra to linear models, analysis of variance and analysis of covariance. These problems included such topics as the calculation of sums of squares, the use of the non-central χ^2 and F-distributions, the F-statistics for testing linear hypotheses and even the estimation of variance components.

In that book the style for developing a difficult result is to begin with a simple theorem in scalars, then extend it slightly to a theorem a little more difficult, and do that again and again, finishing up with a very general theorem. This was (and is) a wonderful teaching method for students needing to know details to enable them to answer exam questions concerning specifics of the intermediary theorems. But it perhaps fails to emphasize the value of the general theorem.

This methodology is also somewhat akin to the introductory teaching of analysis of variance. We start with the completely randomized design, then consider randomized complete blocks, followed maybe by balanced incomplete blocks, Latin squares and split plots. Such a progression gives a sense of a different analysis for each different experiment design. For the beginning student there is little sense of continuity - of some underlying principle by which we get all these sums of squares such as $\sum(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2, \sum\sum(\bar{y}_{ij} - \bar{y}_{i\cdot})^2$ and so on.

Shouldn't we be giving more emphasis to the universality and connectedness of many statistical procedures? And for doing so, make earlier use of matrix notation and matrix manipulations, and some of the general theorems which their use provides. Hopefully, this would make the learning of statistics easier than it is now viewed by so many students.

1563-M

The Power of the Matrix

Shayle R. Searle
Departments of Biometrics and Statistics
Cornell University
Ithaca, N. Y., 14853

BU-1563-M March 2001

Key Words: Matrix proofs, Linear models.

Abstract

A few examples are given of matrix procedures providing broad and general results in statistics.

Introduction

The first linear models book by Graybill (1961) was one of the early major works to address many of the problems arising out of extending regression algebra to linear models, analysis of variance and analysis of covariance. These problems included such topics as the calculation of sums of squares, the use of the non-central χ^2 and F-distributions, the F-statistics for testing linear hypotheses and even the estimation of variance components.

In that book the style for developing a difficult result is to begin with a simple theorem in scalars, then extend it slightly to a theorem a little more difficult, and do that again and again, finishing up with a very general theorem. This was (and is) a wonderful teaching method for students needing to know details to enable them to answer exam questions concerning specifics of the intermediary theorems. But it perhaps fails to emphasize the value of the general theorem.

This methodology is also somewhat akin to the introductory teaching of analysis of variance. We start with the completely randomized design, then consider randomized complete blocks, followed maybe by balanced incomplete blocks, Latin squares and split plots. Such a progression gives a sense of a different analysis for each different experiment design. For

the beginning student there is little sense of continuity — of some underlying principle by which we get all these sums of squares such as $\sum(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$, $\sum\sum(\bar{y}_{ij} - \bar{y}_{i\cdot})^2$ and so on.

Shouldn't we be giving more emphasis to the universality and connectedness of many statistical procedures? And for doing so, make earlier use of matrix notation and matrix manipulations, and some of the general theorems which their use provides. Hopefully, this would make the learning of statistics easier than it is now viewed by so many students.

Compact notation

The generality of matrix notation arises from being able to use the same symbols whether we have five data or 5,000. For any n data we define

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \mathbf{y} \text{ with mean } \mathrm{E}(\mathbf{y}) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_n \end{bmatrix} = \boldsymbol{\mu}$$

with variances and covariances

$$\mathrm{var}(\mathbf{y}) = \begin{bmatrix} \mathrm{var}(y_1) & \mathrm{cov}(y_1, y_2) & \cdots & \mathrm{cov}(y_1, y_i) & \cdots & \mathrm{cov}(y_1, y_n) \\ \mathrm{cov}(y_2, y_1) & \mathrm{var}(y_2) & \cdots & \mathrm{cov}(y_2, y_i) & \cdots & \mathrm{cov}(y_2, y_n) \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathrm{cov}(y_i, y_1) & \mathrm{cov}(y_i, y_2) & \cdots & \mathrm{var}(y_i) & \cdots & \mathrm{cov}(y_i, y_n) \\ \mathrm{cov}(y_n, y_1) & \mathrm{cov}(y_n, y_2) & \cdots & \mathrm{cov}(y_n, y_i) & \cdots & \mathrm{var}(y_i) \end{bmatrix} = \mathbf{V}.$$

This provides endless opportunity for succinctness. For example, if the data in $\mathbf{y}$ come from a multi-normal distribution we simply write

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$$

no matter what size n is. The advantage of these notations is that they provide the means for having results which apply to any number of data, and which can be adapted to all manner of special cases. We give some examples.

Sums of squares and χ^2

The definition of χ^2 is that for

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \mathbf{I}_{n\times n}), \quad \sum_{i=1}^{n} y_i^2 \sim \chi_n^2. \tag{1}$$

Although that is the definition, the more widely used result is Theorem 1.

$$\textbf{Theorem 1} \qquad \mathbf{y} \sim \mathcal{N}(\mu \mathbf{1}_{n\times 1}, \sigma^2 \mathbf{I}_n) \;\Rightarrow\; \sum (y_i - \bar{y}.)^2/\sigma^2 \sim \chi^2_{n-1}. \tag{2}$$

However, the conceptual jump from (1) to (2) is not at all obvious. We need to prove (2), based on (1), and we do this with two different proofs.

The first proof of (2) begins with a simple extension of (1):

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \sigma^2 \mathbf{I}_n) \quad \text{implies} \quad \sum_{i=1}^{n} y_i^2/\sigma^2 \sim \chi^2_n. \tag{3}$$

A second extension uses linear combinations of the y_is which, only with benefit of hindsight, provide the jump from (1) to (2). These functions are

$$z_i = \left(\sum_{r=1}^{i} y_r - i y_{i+1}\right) \Big/ \sqrt{i(i+1)} \text{ for } 1, 2, \cdots, n-1. \tag{4}$$

Examples are $z_1 = (y_1 - y_2)/\sqrt{2}$ and $z_2 = (y_1 + y_2 - 2y_3)/\sqrt{6}$. Important properties of the z_is of (4) based on $\mathbf{y}$ of (2) are that

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}_{(n-1)\times 1}, \sigma^2 \mathbf{I}_{n-1}) \quad \text{and} \quad \sum_{i=1}^{n-1} z_i^2 = \sum_{i=1}^{n} (y_i - \bar{y}.)^2. \tag{5}$$

Thus by (3) and (5) we have (2).

The preceding proof of (2) relies on (4), the introduction of which seems to have no connection to (2) until we observe (5). And there is no obvious way of extending this proof to where $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ rather than $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ of (5), or to where we are interested in some general second-degree form $\mathbf{y}'\mathbf{A}\mathbf{y}$ rather than $\sum(y_i - \bar{y})^2$. But all this can be encompassed in a single, broad theorem, Theorem 2. It requires only a few simple properties of matrices: idempotency, rank, trace; and also acknowledgment of the non-central χ^2 distribution and appreciation of a simple condition which reduces it to the regular (non-central) distribution. And (2) is just one of many special cases of Theorem 2.

> **Theorem 2.** For $\mathbf{y}_{n\times 1} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ $\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2\left[r(\mathbf{AV}), \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\right]$ if and only if $\mathbf{AV}$ is idempotent; and when $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ then $\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2_{r(\mathbf{AV})}$, $r(\mathbf{AV})$ being the rank of $\mathbf{AV}$.

Theorem 1 is now proven by using Theorem 2 with $\mathbf{A} = (\mathbf{I} - \bar{\mathbf{J}})/\sigma^2$, $\boldsymbol{\mu} = \mu\mathbf{1}$, and $\mathbf{V} = \sigma^2\mathbf{I}$, where $\bar{\mathbf{J}}$ is $n \times n$ with every element $1/n$, and $\mathbf{1}$ has 1.0 for every element. It is

then easily established that $\mathbf{AV} = \mathbf{I} - \bar{\mathbf{J}}$, which is idempotent with rank $n-1$, and $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$. Hence $\mathbf{y}'\mathbf{A}\mathbf{y} = \sum(y_i - \bar{y})^2/\sigma^2$ has a χ^2_{n-1} distribution; and so (2) is proven.

Surely Theorem 2 exemplifies what we need more of in today's teaching of statistics — bringing together those parts of the many different analyses which are common to them all. True, the theorem is more advanced than the step-by-step scalar proof of Theorem 1. But no more advanced than requiring a few features of matrices — and acknowledging the non-central χ^2, even though it is dispensable in most real-life situations.

The suggestion is that we increase the teaching of matrix algebra before teaching statistics, so that we can then state, accept and use Theorem 2 and others like it. By no means is this to suggest showing the proof of that theorem, which is quite complicated — especially the necessity part, which is of little practical use anyway. Accepting the theorem without proof takes no more courage than accepting Boyle's Law, or accepting the arithmetic correctness of computer software — and we do that all the time. And today's teaching is so much more connected to computer technology than to algebraic details, which we are coming to accept (with some danger perhaps) as being taken care of in the computing software.

And think of the advantages of having a quite general result like Theorem 2, to be able to apply it to a myriad of special cases. Given a second-degree expression of the ys (usually a sum of squares), written as $\mathbf{y}'\mathbf{A}\mathbf{y}$ so that we know $\mathbf{A}$, then for $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ we know what the χ^2 properties of $\mathbf{y}'\mathbf{A}\mathbf{y}$ are, if $\mathbf{AV}$ is idempotent. Contemplate the importance of this in light of computer software spewing out all manner of sums of squares. Upon ascertaining $\mathbf{A}$ and knowing $\mathbf{V}$, we can quickly decide whether or not $\mathbf{y}'\mathbf{A}\mathbf{y}$ has a χ^2 distribution: is $\mathbf{AV}$ idempotent? And thence whether $\mathbf{y}'\mathbf{A}\mathbf{y}$ will be available for testing a linear hypothesis.

It is true that suggesting the use of a general theorem applicable to many special cases flies in the face of Moore (2001), who states that "Few people learn from basic principles down to special cases." That is a fairly strong statement — in spite of which I think we could gain some real benefit from the "general to special" approach.

Linear models

Of course, one of the largest sections of statistics which has gained from matrices is the

linear model, typified by

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \qquad \mathbf{y} - \mathrm{E}(\mathbf{y}) = \mathbf{e} \tag{6}$$

and hence

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{7}$$

Estimation of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} \qquad \text{with} \qquad \mathbf{X}\boldsymbol{\beta}^0 = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}, \tag{8}$$

invariant to $(\mathbf{X}'\mathbf{X})^-$; and $(\mathbf{X}'\mathbf{X})^- = (\mathbf{X}'\mathbf{X})^{-1}$ if $\mathbf{X}$ has full column rank.

This set-up, along with its myriad of consequences applies, as we all know, to a wide variety of different situations, all of them governed by the form of $\mathbf{X}$ and $\boldsymbol{\beta}$. For simple regression $\boldsymbol{\beta}$ has order 2, and $\mathbf{X}$ has two columns: one is $\mathbf{1}$ and the other is values of the single regressor variable. For multiple regression of n regressor variables $\mathbf{X}$ still has $\mathbf{1}$ as a column with the other n columns each containing the values of one of the regressor variables (without repetition). And so it goes on. For analyzing data from an experiment $\mathbf{X}$ is usually a matrix of just zeros and ones, and is often called an incidence, or design, matrix. For a well-designed and executed experiment $\mathbf{X}$ can have easily-seen patterns of zeros and ones which lead to well-known estimators and to analyses of variance: for example, randomized complete blocks and Latin squares. Even for survey-style data $\mathbf{X}$ will be a series of zeros and ones, but with little or no evidence of easily-seen patterns. Yet the set-up of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ can still be used in any of the preceding situations when, in addition to the zeros and ones, $\mathbf{X}$ has a column or columns of regressor variables, often in that case called covariables, so leading to analysis of covariance. And $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ can even be extended to $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ for multivariate analysis.

So here is an example of how powerful matrices and their algebra are in covering a wide array of statistical analyses, for which the same algebra is applicable to a variety of different situations.

Partitioned linear models

A particular example of this is in the calculation of the reduction in sum of squares for partitioned linear models. The basic result, for $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, is

$$R(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}; \tag{9}$$

and $\boldsymbol{R(\beta)}$ can be used to test $\boldsymbol{H : X\beta = 0}$ or, for $\mathbf{X}$ of full column rank, $\boldsymbol{H : \beta = 0}$.

When $\boldsymbol{\beta}$ is partitioned into two parts, as $\boldsymbol{\beta} = [\boldsymbol{\beta}_1' \quad \boldsymbol{\beta}_2']$, then

$$\begin{aligned} \boldsymbol{R(\beta_2|\beta_1)} &= \boldsymbol{R(\beta_1, \beta_2) - R(\beta_1)} && (10) \\ &= \mathbf{y'M_1X_2(X_2'M_1X_2)^-X_2'M_1y} && (11) \end{aligned}$$

where, for $\mathbf{X}^+$ being the Moore-Penrose inverse of $\mathbf{X}$,

$$\mathbf{M_1 = I - X_1(X_1'X_1)^-X_1' = I - X_1X_1^+} \tag{12}$$

and $\boldsymbol{R(\beta_0|\beta_1)}$ is available for testing

$$\boldsymbol{H : M_1X_2\beta = 0.}$$

But in this hypothesis note that

$$\mathbf{M_1X_2\beta_2 = X_2\beta_2 - X_1X_1^+X_2\beta_2} \tag{13}$$

and $\mathbf{X_1X_1^+}$ can be an identity matrix if $\mathbf{X_1}$ has full row rank, which will seldom be the case. But (13) is $\mathbf{0}$ if $\mathbf{X_2\beta_2 = 0}$, so we can take the hypothesis as

$$\boldsymbol{H : X_2\beta_2 = 0,} \tag{14}$$

which is $\boldsymbol{H : \beta_2 = 0}$ if $\mathbf{X_2}$ has full column rank.

Finally, if $\boldsymbol{\beta}$ is partitioned into three parts so that

$$\mathbf{y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + e}$$

it can be shown (Searle, 1987, Sec. 8.6) that $\boldsymbol{R(\beta_2|\beta_1)}$ can be used to test the hypothesis

$$\boldsymbol{H : M_1X_2\beta_2 + M_1X_2(M_1X_2)^+X_3\beta_3 = 0} \tag{15}$$

where, by the same argument as used in deriving (14), we have (15) being satisfied by

$$\boldsymbol{H : X_2\beta_2 + X_2(M_1X_2)^+X_3\beta_3 = 0.} \tag{16}$$

This hypothesis has complete generality — see Searle (1987), p. 279.

Other occurrences of matrices

The preceding examples concern only one of the major parts of statistics where matrices play an important and broadening role. There are numerous others, of which multivariate analysis is probably the largest, as in Anderson (1958), for example. After all, what could be more powerful and more general than $\left(|\hat{\mathbf{V}}|/|\hat{\mathbf{V}}_{\boldsymbol{H}}|\right)^{\frac{1}{2}}$ as the likelihood ratio criterion for a linear hypothesis H in multivariate linear regression, where $\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}_{\boldsymbol{H}}$ are the maximum likelihood estimators under normality of the variance-covariance matrix of each vector of data, without and with (respectively) assuming H to be true (Anderson, 1958, p. 188). In this connection Rao (1952) has written "The problems of multivariate analysis resolve themselves into the analysis of the dispersion matrix and reduction of determinants."

Other aspects of statistics in which reliance on matrices is so valuable are design of experiments, Markov chains, Leslie matrices for population growth, and in the statistics of pedigree improvement based on selection of phenotype records such as weight gain in beef cattle, egg production in poultry and milk yield in dairy cattle.

The entry of matrices into statistics

A precursory feature of today's widespread use of matrices in statistics is that it began relatively recently. Although matrices seem to have started with Cayley (1858), they did not begin to catch hold in statistics until the 1930s. Even as late as 1951 multiple regression was being taught at Cambridge University without benefit of matrices! As Grattan-Guiness and Lederman (1994) remark, "The rise of matrix theory to staple diet has occurred only since the 1950s." And many early staple statistics books made little or no use of matrices. Rao (1952) and Kempthorne (1952) were early users; Snedecor (1937), Kendall (1943-52) and Mood (1950) were not; neither was Aitken (1939). And even though he published both a matrix book (1939a) and a statistics book (1939b), neither book has anything substantive about the topic of the other! Searle (2000) provides considerable detail on the infusion of matrices into statistics.

References

Aitken, A.C. (1939a) **Determinants and Matrices**. Oliver & Boyd, Edinburgh.

Aitken, A.C. (1939b) **Statistical Mathematics**. Oliver & Boyd, Edinburgh.

Anderson, T.W. (1958) **An Introduction to Multivariate Statistical Analysis**. Wiley, New York.

Cayley, A. (1858) A memoir on the theory of matrices. *Philosophical Transactions of the Royal Society of London*, **148**, 17-37.

Grattan-Guiness, I. and Lederman, W. (1994) Matrix Theory. In **Companion Encyclopedia to the History and Philosophy of the Mathematical Sciences** (I. Grattan-Guiness, ed.) Routledge, London, 775-786.

Graybill, F.A. (1961) **An Introduction to Linear Statistical Models**, Vol I. McGraw-Hill, New York.

Kempthorne, O. (1952) **The Design and Analysis of Experiments**. Wiley, New York.

Kendall, M.G. (1943-52) **The Advanced Theory of Statistics**. Griffin, London.

Moore, D.S. (2001) Undergraduate programs and the future of academic statistics. *The American Statistician* **55**, 1-6.

Rao, C.R. (1952, reprint 1970) **Advanced Statistical Methods in Biometric Research**. Hafner, Darien, Connecticut.

Searle, S.R. (1987) **Matrix Algebra Useful for Statistics**. Wiley, New York.

Searle, S.R. (2000) The infusion of matrices into statistics, *Image, Journal of the International Algebraic Society* **24**, 25-32.

Snedecor, G.W. (1937-56) **Statistical Methods**, editions 1-5. Iowa University Press, Ames, Iowa.

Snedecor, G.W. and Cochran, W.G. (1967-89) **Statistical Methods**, editions 6-8. Iowa University Press, Ames, Iowa.