1590

STATISTICAL ANALYSIS OF DATA FROM DESIGNED EXPERIMENTS

Review by

Shayle R. Searle
Department of Biometrics
College of Agriculture and Life Sciences
Cornell University
Ithaca, NY 14853
BU-1590-M
December 17, 2001

of

"A First Course in the Design of Experiments:
A Linear Models Approach"

by

Donald C. Weber and John H. Skillings

Boca Raton, CRC Press, 2000, 680 pp.
ISBN 0-8493-9671-9
Price: not shown on book.

General Comments

This is not a psychology book. It is a straightout statistics book, written by two professors at Miami University, Oxford, Ohio. It deals with the statistical analysis of data from well-known, well-designed, and well-executed experiments. Its first five chapters deal largely with linear model theory (supplemented by a matrix algebra appendix) and its last thirteen chapters deal with analyzing data, mostly from experiments. A multitude of well-known standard results is presented, along with quite extensive numerical details for many good examples.

Assessing intended readership is not easy because the book contains so much information, from elementary to advanced: matrix algebra, linear model theory (but with no statistical appendix), many numerical examples, and SAS GLM computing, both commands and data output, the latter in excessively many-decimal numbers. These are no easier to follow, for the beginning student, than the algebra they are illustrating. At least a few thoughtfully planned hypothetical examples with easy-to-read and easy-to-calculate numbers would be more useful.

In the literature of mathematical statistics the convention is to use capital letters for random variables and lower case for realizations thereof. This clashes with the custom in matrix algebra of capitals for matrices and lower case for vectors (usually both in bold face). To its credit, this book tries to use both these notations. For example, equation (2.1.4), on page 10 is $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a vector (of random variables), but $\mathbf{X}$ is a matrix (of known constants), and $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are vectors. Then on page 19 there is $\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{e}$, described as the sample form of (2.1.4), with $\mathbf{y}$ being a data vector and $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in (2.3.5). This is fine, but page 66 has $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ and exercise 4-14 on page 97 has $\mathrm{Var}(\hat{\boldsymbol{\beta}})$. But, and it is a big "but", a "hat" over a symbol representing an estimator (or estimate) is so firmly entrenched in the literature that its

use in $\hat{\boldsymbol{\beta}}$ where the **Y** therein is not data is hard to accept. It is, of course, very correct, because then $\hat{\boldsymbol{\beta}}$ is the estimator and **b** is the estimate, as is made clear on page 66. However, when the notation gets extended on page 115 to $\text{Var}(\ell'\hat{\boldsymbol{\beta}}) = \sigma^2\, \ell'(\mathbf{X}'\mathbf{X})^{-}\ell$ and to (5.3.4) which has $\text{Var}(\ell'\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\ell'(\mathbf{X}'\mathbf{X})^{-}\ell$ it is harder to keep track. Then atop page 116, comes "the point estimate of the random variable $\text{Var}(\ell'\hat{\boldsymbol{\beta}})$ is $s^2\ell'(\mathbf{X}'\mathbf{X})^{-}\ell$ where s^2 is the observed value of the random variable $\hat{\sigma}^2$ ". This is too much! The authors' attempt at rigorously distinguishing between random variables and their realizations is highly commendable, but it becomes cumbersome. Most writings of linear models avoid this maze by having **y** do double duty, both for data, the realization of a random vector, and for that random vector itself. Similarly $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$ and $\hat{\sigma}^2$ each represent either a random variable or a realization thereof, as determined by context. This clears the air notationally, and seldom leads to confusion. *This notation is used here, in this review.*

A starting point for discussing a linear statistical model is usually an equation such as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ of (2.1.4) on page 10. It is referred to there as a linear model. That is not so. It is a model equation. A model is a model equation plus description of its elements. Describing **y** as a random variable, as on page 9, and $\boldsymbol{\varepsilon}$ as a random error term as on page 11 tells us nothing about what $\boldsymbol{\varepsilon}$ represents. What sort of error? A satisfying answer is to use $\text{E}(\cdot)$ for expectation over repeated sampling and to assume

$$\text{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \qquad [1]$$

and define

$$\boldsymbol{\varepsilon} \equiv \mathbf{y} - \mathrm{E}(\mathbf{y}) \qquad [2]$$

Then $\boldsymbol{\varepsilon}$ is the amount by which $\mathbf{y}$ differs from its expected value, its mean. [1] and [2] have three important consequences. First, they lead to model equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ of (2.1.4). Second, [2] shows what we are defining $\boldsymbol{\varepsilon}$ to represent. Third, [2] provides a reason for $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, without it being an unexplained assumption as in (a) of (4.1.1) on page 63. The basic assumption is [1]; and [2] is a definition.

Actually, it is probably more correct to "attribute" to E(y) the right-hand side of [1] than to "assume" it. A dictionary says "assuming" is "taking for granted", whereas to "attribute" is to "consider as belonging" which seems appropriate here. Likewise in the equation (4.1.1) on page 63 we attribute (rather than assume) σ^2 and 0 to be the forms of $\mathrm{Var}(\varepsilon_i)$ and $\mathrm{cov}(\varepsilon_i, \varepsilon_j)$ respectively.

Another advantage of [2] is for describing least squares estimation of $\boldsymbol{\beta}$. We want to find $\hat{\boldsymbol{\beta}}$ as the value of $\boldsymbol{\beta}$ which minimizes the sum of squares of the elements of $\boldsymbol{\varepsilon}$, namely $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. Since $\boldsymbol{\beta}$ is an unknown constant we can minimize $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ with respect to $\boldsymbol{\beta}$ only by temporarily treating $\boldsymbol{\beta}$ as a mathematical variable. Doing this gives the minimizing value as $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$ which we call $\hat{\boldsymbol{\beta}}$, but denoted by $\mathbf{b}$ in (3.2.2) on page 44.

Technical Matters

Least squares estimation of $\boldsymbol{\beta}$ for $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ leads to normal equations with a solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$ for $(\mathbf{X}'\mathbf{X})^{-}$ being a generalized inverse of $\mathbf{X}'\mathbf{X}$. Because there are many values

of $(\mathbf{X}'\mathbf{X})^-$ when, as is usual, $\mathbf{X}$ has less than full column rank, there are advantages to using β^o for $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$ to distinguish it from denoting an estimate of β as $\hat{\beta}$, which β^o is not.

Although for many of the book's examples (involving as they do, balanced data), $(\mathbf{X}'\mathbf{X})^-$ is not difficult to derive, an easier procedure is to apply side conditions directly to the normal equations. Example 3.4.1 on page 52 illustrates this. But what is not said is that this simplification does not necessarily apply easily for unbalanced data. The minimal discussion of such data (pages 441-7) summarizes their difficulties quite well, except for some errors of omission such as the following. The expected value three lines above (13.7.1) should include $\bar{\beta}_{\cdot} + (\overline{\alpha\beta})_{i\cdot}$, providing $n_{ij\cdot} > 0 \; \forall j$. And the second line below the table on page 447 requires "the specific" to precede "interaction". And at the bottom of that page the necessity for the data to be connected needs to be said.

The book's endpoint is well-known analyses of variance for designed experiments. Derivations are based on linear models procedures with little or no use of the basic identities that, for example

$$(y_{ij} - \bar{y}_{\cdot\cdot}) \equiv (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) + (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}) \qquad [3]$$

and, for balanced data, the same identity holds for sums of squares of each parenthesized term in [3]. That identity is behind many an analysis of variance table. Instead of that, the authors use what they call (page 127) the "principle of conditional error" (a phrase new to me, found in neither the table of contents nor the index). It uses a reduced model, which is the full model with $y = \mathbf{X}\beta + \varepsilon$ amended by assuming a hypothesis $\mathbf{L}\beta = \mathbf{0}$ is true. With SSe and SSe* being error

(residual) sums of squares for fitting the full and reduced models, respectively, the book repeatedly uses (what in this review is called) Δ:

$$\Delta = \text{SSe}^* - \text{SSe} \qquad [4]$$

For many experiments a familiar summation formula is given (or derived) for Δ, e.g. pages 196-7, 323 and 393. But sometimes (e.g., pages 365-6) using computer software to calculate SSe* and SSe is suggested, without giving any specific formulae.This occurs with incomplete block designs (IBDs) of chapter 11. As described there (page 363) these are indistinguishable from 2-factor unbalanced data with 0 or 1 observation per cell. Hence no summation formulae *are* available, as indicated at pages 446-7. But the page 363 illustration of an IBD is, without saying so, a special case, a balanced incomplete block design (BIBD), the most usual form of IBD, and it does have familiar summation formulae (page 378).

The book professes to use linear models (and hence matrix) procedures. Unfortunately they are not used to the fullest extent available. For example, Δ of [4] has a general matrix expression on page 129. With some straightforward matrix manipulation one can readily show, when the reduced model is simply the full model with some factors removed, that Δ is a sum of squares, with a quadratic form in $\mathbf{y}$ having an idempotent matrix [e.g., (82) of Searle (1987; p. 264)]. These are very important properties of Δ, yet they are barely alluded to, on page 213, along with suggesting that their derivation can be found in the succeeding exercises: but I found no such exercise. And theorem 5.1.5 provides another opportunity for the advantage of a succinct matrix treatment over the tedious scalar proof of page 109.

For computing needs, this book relies entirely on SAS GLM software. It mentions no other. Yet it also gives no information whatever as to where SAS comes from. All that SAS gets is nine words in the preface; it merits more than that.

In illustrating SAS GLM (to an excess in my opinion) only the sums of squares labeled Types I and III are explained (pages 152-162) in terms of regression models. But describing Type III as "principle of conditional error sums of squares" (page 162) is not correct for unbalanced data for linear models having interactions, especially for some-cells-empty data. And the ignored Type II is very useful for unbalanced data, even for Latin squares.

A problem with a book having extensive use of any computing package is that even long-lasting software are constantly being up-graded. Hence a book's descriptions can often be out of date before they are published.

Finally, for computing purposes, it is quite unnecessary to present a reparameterization of models solely for using regression software (e.g., sections 7.6, 10.6, 11.3, 12.4, and 13.5, comprising some 30 pages, nearly 5% of the book). This is not needed in today's computing environment: whoever would have regression software without also having GLM software of some sort?

Final Remarks

The early part of the book deals primarily with theoretical underpinnings. The resulting methodology is then applied in pretty much a uniform manner to designs such as the completely random design, randomized complete blocks, balanced incomplete blocks, Latin squares, and extensions thereof, complete with the standard forms of analyses of variance, F-tests, contrasts, standard errors, confidence intervals, multiple comparisons and so on. The technical parts of this are all good. And the many exercises, both numeric and theoretic, are excellent. But I do find there to be too many words, unnecessary verboseness. For example, on page 282, the first paragraph, of 53 words, is easily written in 20: "These procedures utilize the t-distribution; Bonferroni's (section 8.4) controls overall Type I error, whereas Fisher's controls per

comparison error". And chapter 7, on the one-way classification, has 68 pages! This is 10% of the book, which is excessive for what is quite the easiest model to understand. Overall, there is considerable lack of the succinctness that so often brings clarity to, and which is a hallmark of, good mathematical writing.

A distinguished mathematician allegedly once said "a mathematics book without at least one typographical error on every page is a sign of wasted effort!" For its size (680 pages), this book has remarkably few typos, most of which, being obvious, will already be known to the authors. A couple of less obvious ones are: (i) on page 166 the second denominator in p_2 should be $p_1'p_1$: it is currently $p_0'p_1$, which equals zero. *(ii)* In the last line of exercise 6-14, on page 181, the α_1 should be α_2.

On two occasions (hopefully no more) technical terms occur which have not earlier been defined. They are (I) "Type I error" (on page 221) which is not even in the index; and (ii) "power", on page 219, which is defined on page 251, but with the serious omission of "only when H_0 is true" which needs to come after "probability of rejection".

Some errors of omission are unfortunate: theorem 3.2.3 requires $\mathbf{X}$ to be real. For page 104, proof or a reference is needed that the sum of *n* independent χ_1^2 variables is a χ_n^2 variable (and chi-square is not in the index!). Errors in variables need mentioning on page 64. The $\mathbf{X}$ that is called a design matrix on page 190 is now more generally called a model matrix. And on page 523, in the middle of item 3 "there is little interest in estimating the β_js [no apostrophe needed, β_j does not possess anything] since these are random . . . " is plain wrong. There is considerable

literature on predicting random effects. See, for example, Searle, Casella and McCulloch (1992, chapter 7), and McCulloch and Searle (2001, chapter 9).

Publicity on the book's back cover suggests its availability as a reference book. It certainly qualifies as such in terms of its wealth of good, standard material. But for success as a reference it has serious deficiencies. First, the running heads contain only chapter titles; and without section or sub-section titles and numbers, the running heads are useless. Sub-sections are also not in the table of contents. Second, the index lacks many important entries; e.g., hypothesis, tests of hypotheses, Types I and II errors, SAS, and on and on. All of this makes it difficult to locate a topic of interest.

Finally, the book's list, and use, of references is woefully poor, with many supporting books and papers notably absent. Moreover, references in the text seldom have page numbers (e.g., page 221). Are readers expected to read a whole book or paper to find what they want? And in some cases (e.g., Fisher and Bonferroni, at pages 282-3) not even dates are given; and these two are not in the reference list (pages 661-2). And what is de Morgan's law (page 270)? Also, the headings of Tables B.3, B.5, and B.7 (are 8 pages of this really needed?) are incomplete.

In summary then, I could not use the book as a reference despite its containing, for analyzing experimental data, a wealth of useful information, little of which is new and most of which is readily available in a raft of established texts; e.g., Kempthorne (1952), Federer (1955) and Winer (1971). And I would not try to teach from the book because, for that purpose, it is too verbose and repetitive.

References

Federer, W.T. (1955) *Experimental Design.* Macmillan, New York.

Kempthorne, O. (1952) *The Design and Analysis of Experiments*. John Wiley & Sons, NewYork.

McCulloch C.E. and Searle S.R. (2001) *Generalized, Linear, and Mixed Models.* John Wiley & Sons, New York.

Searle, S.R. (1987) *Linear Models for Unbalanced Data.* John Wiley & Sons, New York.

Searle, S.R., Casella G, and McCulloch C.E.(1992) *Variance Components*. John Wiley & Sons, New York.

Winer B.J.. (1971) *Statistical Principles in Experimental Design* (2nd Ed.) McGraw-Hill, N.Y.