Foundations for the Theory of Least Squares

By B. M. HILL

University of Michigan

[Received November 1967. Revised April 1968]

SUMMARY

The Bayesian theory of least squares is founded upon a weaker and more tangible form of prior knowledge than the conventional assumption of normality. The underlying assumption is a form of conditional uniformity on spheres for the "actual errors" in the experiment. This provides a unified theory appropriate for randomization models in the analysis of variance as well as for classical least-squares analysis.

1. INTRODUCTION

THE purpose of this paper is to found the Bayesian theory of least squares on the basis of weaker and more tangible forms of prior knowledge than the conventional assumption of normality. This is done in terms of an assumption of conditional uniformity on spheres for the "actual" errors in the experiment. Section 2 illustrates the approach in some of its most pertinent applications, namely to randomization models in the analysis of variance. Section 3 presents a coordinate-free formulation in a unitary vector space, and explores the relationships between conditional uniformity and conventional normality assumptions. Section 4 motivates the assumption of conditional uniformity on spheres and discusses some general issues.

Results agree with but go beyond those of Jeffreys (1961, p. 147), Savage *et al.* (1963) and Lindley (1965, Ch. 8), who do assume normality. There is also some correspondence with more classical results. See, for example, Scheffé (1959, Chs. 1, 4, 9) and Fisher (1960, Ch. IV).

2. RANDOMIZED BLOCKS

In this section a method of analysis is proposed for randomized block designs, with or without technical errors, but with no treatment-unit interactions. The analysis is presented in a series of four cases, culminating in randomized blocks with blocks random, which illustrate different aspects. Case A sheds light on the general problem of inference when the number of "cells" is equal to the number of observations and the variance is "unknown".

Case A. Completely randomized design, I treatments, one observation per treatment

Let η_{iv} be the conceptual "true" response when treatment i is applied to unit v, and define $\eta = \eta_{..}$, $\alpha_i = \eta_{i.} - \eta$, $\epsilon_{iv} = \eta_{iv} - \eta_{i.}$, so that $\eta_{iv} = \eta + \alpha_i + \epsilon_{iv}$, with $\alpha_. = \epsilon_{i.} = 0$, $i, v = 1, \ldots, I$, where the dot notation indicates the simple average over all possible values of the subscript replaced by a dot. Suppose that treatment-unit interactions are zero, so that in fact $\epsilon_{iv} = \eta_{.v} - \eta \equiv \xi_v$, say, and define $U^2 = \sum \xi_v^2$. Suppose also for the time being that there are no technical errors, so that the observations can be written $y_i = \eta + \alpha_i + \bar{\epsilon}_i$, with $\bar{\epsilon}_i = \sum_v d_{iv} \epsilon_{iv}$, and $d_{iv} = 1$ if treatment i falls on unit v

and otherwise 0, $Id_{i.} = 1$. The $d_{i\nu}$ are assumed known and determined in such a way that they are completely uninformative about $\{\alpha_i\}$, U^2, in the sense that the personal probability distribution for these quantities, given $\{d_{i\nu}\}$, does not depend upon $\{d_{i\nu}\}$. This would be the case, for example, if the $d_{i\nu}$ were determined by a physical randomization. Then $y_{.} = \eta$ and the marginal posterior distribution of $\{\alpha_i\}$ is

$$\begin{aligned}\Pr\{\alpha_i | y_i, d_{i\nu}\} &\propto \Pr\{y_i | \alpha_i, d_{i\nu}\} \Pr\{\alpha_i | d_{i\nu}\} \\ &\propto \Pr\{\tilde{\epsilon}_i = y_i - y_{.} - \alpha_i | \alpha_i, d_{i\nu}\} \Pr\{\alpha_i\} \\ &\propto \Pr\{\tilde{\epsilon}_i = y_i - y_{.} - \alpha_i | \alpha_i, d_{i\nu}, U^2 = \Sigma(y_i - y_{.} - \alpha_i)^2\} \\ &\quad \times \Pr\{U^2 = \Sigma(y_i - y_{.} - \alpha_i)^2 | \alpha_i\} \Pr\{\alpha_i\}. \qquad (1)\end{aligned}$$

This determines $\Pr\{U^2 | y_i, d_{i\nu}\}$ by virtue of $U^2 = \Sigma(y_i - y_{.} - \alpha_i)^2$ for the given data $\{y_i\}$. Here the indices i, ν run through the appropriate sets, and with an abuse of notation I have used the same symbol to denote the random quantity and its value. Also, although the above notation is ordinarily used only when the probability distributions are discrete, so long as care is taken, particularly in regard to degeneracies such as $\alpha_{.} = 0$, $y_{.} = \eta$, $\Sigma(y_i - y_{.} - \alpha_i)^2 = U^2$, there is no harm in using this same notation to represent density functions and even more general types of distributions.

I shall now assume that conditional upon U^2, $\{\alpha_i, d_{i\nu}\}$, the vector $\xi = (\xi_1, \ldots, \xi_I)'$ is uniformly distributed on the intersection of the sphere of radius U centred at the origin with the hyperplane $\xi_{.} = 0$ in I-dimensional Euclidean space. This last form of assumption is at the heart of the present approach to least squares, and will be referred to as a conditional uniformity assumption (C.U.A.). It will be discussed in its most general form in Section 4. For the time being note that in the case of densities C.U.A. is substantially weaker than the corresponding more customary assumption of normality. Although discrete analogues are obvious, the remainder of this article will deal explicitly only with density functions.

From (1) and C.U.A.,

$$\Pr\{\alpha_i | y_i, d_{i\nu}\} \propto [\Sigma(y_i - y_{.} - \alpha_i)^2]^{-(I-3)/2} \Pr\{U^2 = \Sigma(y_i - y_{.} - \alpha_i)^2 | \alpha_i\} \Pr\{\alpha_i\}. \qquad (2)$$

[With density functions it is necessary to distinguish between $\Pr\{\xi_i | \Sigma\xi_i^2\}$ and, say, $\Pr\{\xi_i | (\Sigma\xi_i^2)^{1/2}\}$. C.U.A. implies $\Pr\{\xi_i | \Sigma\xi_i^2\} \propto (\Sigma\xi_i^2)^{-(I-3)/2}$, where $\xi_{.} = 0$.]

In particular, if prior knowledge of $\{\alpha_i\}$, U^2, is formally represented by the improper prior density $\Pr\{\alpha_i, U^2\} \propto \rho(U^2)$, then

$$\left.\begin{aligned}&\Pr\{\alpha_i | y_i, d_{i\nu}\} \propto [\Sigma(y_i - y_{.} - \alpha_i)^2]^{-(I-3)/2} \rho[\Sigma(y_i - y_{.} - \alpha_i)^2], \\ &\Pr\{\xi_i | y_i, d_{i\nu}\} \propto (\Sigma\xi_i^2)^{-(I-3)/2} \rho(\Sigma\xi_i^2) \propto \Pr\{\xi_i\}, \\ &\Pr\{U^2 | y_i, d_{i\nu}\} \propto \rho(U^2).\end{aligned}\right\} \qquad (3)$$

In so far as it is possible to draw inferences about $\{\alpha_i\}$ based upon vague prior knowledge, I believe (3) is the appropriate posterior distribution. This is meant in the sense that for a wide class of prior distributions for $\{\alpha_i\}$, U^2, each of which is suitably gentle in the vicinity of $\alpha_i = y_i - y_{.}$, and not too large elsewhere, the posterior distribution can be adequately approximated by (3). This can be made precise after the fashion of Savage *et al.* (1963, pp. 201 f.). Note, however, that

$$[\Sigma(y_i - y_{.} - \alpha_i)^2]^{-(I-3)/2}$$

has an infinite integral with respect to $\prod d\alpha_i$ so that the above posterior distribution of $\{\alpha_i\}$ is not proper unless the factor $\rho[\sum(y_i-y_.-\alpha_i)^2]$ is such as to tame it. Since

$$\int [\sum(y_i-y_.-\alpha_i)^2]^{-(I-3)/2}\,\rho[\sum(y_i-y_.-\alpha_i)^2]\prod d\alpha_i \propto \int_0^\infty \rho(U^2)\,dU^2,$$

all that is required in order that the posterior distribution of $\{\alpha_i\}$ be proper is that the prior density of U^2 be proper. Finally, the fact that the data are totally uninformative about $\{\xi_i\}$ and U^2 is not surprising. That there is information about $\{\alpha_i\}$ stems from the conditional uniformity assumption about the "error" vector $\boldsymbol{\xi}$, and the prior distributions.

Now suppose that a technical error l_i is added, so $y_i = \eta+\alpha_i+\tilde{\epsilon}_i+l_i$. There do exist situations in which it is possible to draw inferences about both U^2 and characteristics of the distribution of the technical errors. In general, however, it is not possible to untangle the $\tilde{\epsilon}_i$ and the $l_i-l_.$, and it seems appropriate to let C.U.A. apply instead to $\tilde{\epsilon}_i^* = \tilde{\epsilon}_i+l_i-l_.$, given $\{\alpha_i, d_{i\nu}\}$ and $(U^*)^2 = (\sum\tilde{\epsilon}_i^*)^2$. But $y_i-y_. = \alpha_i+\tilde{\epsilon}_i^*$, and if inference about $\{\alpha_i\}$ were based solely upon the data $\{y_i-y_.\}$, then the posterior distribution of $\{\alpha_i\}$ could be evaluated just as in the case of no technical errors. Since $y_. = \eta+l_.$ ordinarily carries only slight and indirect information about all parameters other than η (which is of little interest in a comparative experiment), this procedure is intuitively plausible. The general treatment of technical errors is elucidated in Section 3.

Case B. Completely randomized design, J observations on each of I treatments

Let $\eta_{ij\nu}$ be the conceptual "true" response when treatment i is applied to unit $j\nu$, (the νth unit in the jth block), and define $\eta = \eta_{...}$, $\alpha_i = \eta_{i..}-\eta$, $\epsilon_{ij\nu} = \eta_{ij\nu}-\eta_{i..}$, so that $\eta_{ij\nu} = \eta+\alpha_i+\epsilon_{ij\nu}$, with $\alpha_. = \epsilon_{i..} = 0$, $i,\nu = 1,\ldots,I, j = 1,\ldots,J$. Suppose that treatment-unit interactions vanish, so that $\epsilon_{ij\nu} = \eta_{.j\nu}-\eta_{...} \equiv \xi_{j\nu}$, say, and define $U^2 = \sum\xi_{j\nu}^2$. Suppose also for the time being that there are no technical errors, so that the observations can be written $y_{ij} = \eta+\alpha_i+\tilde{\epsilon}_{ij}$, with $\tilde{\epsilon}_{ij} = \sum d_{ij\nu}\epsilon_{ij\nu}$, $d_{ij\nu} = 1$ if treatment i falls on unit $j\nu$, $d_{ij\nu} = 0$ otherwise, and $Id_{ij.} = 1$. The $d_{ij\nu}$ are again assumed to be known and totally uninformative about $\{\alpha_i\}, U^2$. A natural form of C.U.A. is now that conditional upon $U^2, \{\alpha_i, d_{ij\nu}\}$, the vector $\boldsymbol{\xi} = (\xi_{11},\ldots,\xi_{JI})'$ is uniformly distributed on the intersection of the sphere of radius U centred at the origin with the hyperplane $\xi_{..} = 0$, in IJ-dimensional Euclidean space. Then $y_{..} = \eta$, and

$$\begin{aligned}\Pr\{\alpha_i \mid y_{ij}, d_{ij\nu}\} &\propto \Pr\{\tilde{\epsilon}_{ij} = y_{ij}-y_{..}-\alpha_i \mid \alpha_i, d_{ij\nu}\}\Pr\{\alpha_i \mid d_{ij\nu}\}\\ &\propto [\sum(y_{ij}-y_{..}-\alpha_i)^2]^{-(IJ-3)/2}\Pr\{U^2 = \sum(y_{ij}-y_{..}-\alpha_i)^2 \mid \alpha_i\}\Pr\{\alpha_i\}. \quad (4)\end{aligned}$$

In particular, if vague prior knowledge of $\{\alpha_i\}$ is formally represented by the improper prior density $\Pr\{\alpha_i, U^2\} \propto \rho(U^2)$, then

$$\begin{aligned}\Pr\{\alpha_i \mid y_{ij}, d_{ij\nu}\} &\propto [\sum(y_{ij}-y_{..}-\alpha_i)^2]^{-(IJ-3)/2}\,\rho(\sum(y_{ij}-y_{..}-\alpha_i)^2)\\ &\propto [1+[\{J\sum(y_{i.}-y_{..}-\alpha_i)^2\}/\{\sum(y_{ij}-y_{i.})^2\}]]^{-(IJ-3)/2}\,\rho[\sum(y_{ij}-y_{..}-\alpha_i)^2], \quad (5)\end{aligned}$$

$$\begin{aligned}\Pr\{U^2 \mid y_{ij}, d_{ij\nu}\} &\propto \rho(U^2)\,(U^2)^{-(IJ-3)/2}[U^2-\sum(y_{ij}-y_{i.})^2]^{(I-3)/2}\\ &\propto \rho(U^2)\,U^{-I(J-1)}[1-\{\sum(y_{ij}-y_{i.})^2/U^2\}]^{(I-3)/2}\end{aligned}$$

for $U^2 > \sum(y_{ij}-y_{i.})^2$, and is otherwise zero. Note that it is no longer necessary that $\rho(U^2)$ be proper in order that the posterior distribution of $\{\alpha_i\}$ be proper, as it was in Case A, namely $J = 1$. If in fact $\rho(U^2) \propto (U^2)^{-1}$, then the posterior distribution of $\{\alpha_i\}$ is identical with that based upon conventional normality assumptions and Jeffrey's prior distribution, as is shown in substantial generality in Section 3. Similarly, although $(IJ-1)^{-1}U^2$ does not have the same meaning as the variance of the errors under normal theory, the posterior distribution of this quantity is very much like the usual posterior distribution for such a variance (Lindley, 1965, p. 101). This will be discussed further in Sections 3 and 4. Finally, if a technical error l_{ij} is added, and if C.U.A. applies to $\tilde{\epsilon}^*_{ij} = \tilde{\epsilon}_{ij} + l_{ij} - l_{..}$, given $\{\alpha_i, d_{ij\nu}\}$ and $U^{*2} = \sum\tilde{\epsilon}^{*2}_{ij}$, then often inference about $\{\alpha_i\}$ can be based upon the data $\{y_{ij}-y_{..}\}$ alone just as if there were no technical errors. Clearly $y_{..} = \eta + l_{..}$ carries only slight and indirect information about all parameters other than η.

Case C. Randomized blocks, blocks fixed

Let $\eta_{ij\nu}$ be the conceptual "true" response when treatment i is applied to unit $j\nu$ and define $\eta = \eta_{...}$, $\alpha_i = \eta_{i..} - \eta$, $\beta_j = \eta_{.j.} - \eta$, $\gamma_{ij} = \eta_{ij.} - \eta - \alpha_i - \beta_j$, $\epsilon_{ij\nu} = \eta_{ij\nu} - \eta_{ij.}$, so that $\alpha_. = \beta_. = \gamma_{i.} = \gamma_{.j} = \epsilon_{ij.} = 0$, $i, \nu = 1, \ldots, I$, $j = 1, \ldots, J$. Assume that treatment-unit interactions vanish so in fact

$$\epsilon_{ij\nu} = \eta_{.j\nu} - \eta_{.j.} \equiv \xi_{j\nu},$$

say, and define $U_j^2 = \sum\xi_{j\nu}^2$, $U^2 = \sum U_j^2$. Again suppose first that there are no technical errors, so that the observations can be written

$$y_{ij} = \eta + \alpha_i + \beta_j + \gamma_{ij} + \tilde{\epsilon}_{ij},$$

where $\tilde{\epsilon}_{ij} = \sum d_{ij\nu}\epsilon_{ij\nu}$, with $d_{ij\nu} = 1$ if treatment i falls on the νth unit in the jth block, $d_{ij\nu} = 0$ otherwise, and $Id_{ij.} = 1$. The $d_{ij\nu}$ are assumed known and totally uninformative about $\{\eta_{ij.}, U_j^2\}$. Note that $y_{..} = \eta$, $y_{.j} = \eta + \beta_j$.

There are two different versions of C.U.A. which seem of general interest. First, suppose that conditional upon $\{\eta_{ij.}, U_j^2, d_{ij\nu}\}$, the vector $\boldsymbol{\xi}_j = (\xi_{j1}, \ldots, \xi_{jI})'$ of errors in the jth block is uniformly distributed on the intersection of the sphere of radius U_j centred at the origin with the hyperplane $\xi_{j.} = 0$ in I-dimensional Euclidean space, and furthermore that the vectors $\boldsymbol{\xi}_j$ for different blocks are conditionally independent. Then

$$\begin{aligned}\Pr\{\eta_{ij.} \mid y_{ij}, d_{ij\nu}\} &\propto \Pr\{\tilde{\epsilon}_{ij} = y_{ij} - \eta_{ij.} \mid \eta_{ij.}, d_{ij\nu}\} \Pr\{\eta_{ij.}\} \\ &\propto \prod_j [\sum(y_{ij}-\eta_{ij.})^2]^{-(I-3)/2} \Pr\{U_j^2 = \sum(y_{ij}-\eta_{ij.})^2\} \Pr\{\eta_{ij.}\}, \qquad (6)\end{aligned}$$

and $\Pr\{U_j^2 \mid y_{ij}, d_{ij\nu}\}$ is obtained from $U_j^2 = \sum_i(y_{ij}-\eta_{ij.})^2$. This version of C.U.A. allows one to express different opinions about the U_j^2 for different blocks. The remainder of this discussion concerns a more specific version of C.U.A. which builds in symmetry of opinion about the U_j^2.

In this second version of C.U.A. it is assumed that given U^2, the vector $\boldsymbol{\xi} = (\xi_{11}, \ldots, \xi_{JI})'$ is conditionally uniformly distributed on the intersection of the sphere of radius U centred at the origin with the J hyperplanes $\xi_{j.} = 0$, in IJ-dimensional Euclidean space. Then

$$\begin{aligned}\Pr\{\eta_{ij.} \mid y_{ij}, d_{ij\nu}\} &\propto \Pr\{\tilde{\epsilon}_{ij} = y_{ij} - \eta_{ij.} \mid \eta_{ij.}, d_{ij\nu}\} \Pr\{\eta_{ij.}\} \\ &\propto [\sum(y_{ij}-\eta_{ij.})^2]^{-[J(I-1)-2]/2} \Pr\{U^2 = \sum(y_{ij}-\eta_{ij.})^2\} \Pr\{\eta_{ij.}\}. \qquad (7)\end{aligned}$$

In particular, if prior knowledge is formally represented by the improper prior density $\Pr\{\eta_{ij.}, U^2\} \propto \rho(U^2)$, then

$$\left.\begin{aligned}\Pr\{\eta_{ij.}|y_{ij}, d_{ijv}\} &\propto [\Sigma(y_{ij}-\eta_{ij.})^2]^{-[J(I-1)-2]/2}\,\rho[\Sigma(y_{ij}-\eta_{ij.})^2],\\ \Pr\{U^2|y_{ij}, d_{ijv}\} &\propto \rho(U^2).\end{aligned}\right\} \tag{8}$$

This is much like the corresponding result in Case A, with the data supplying no information about U^2, and $[\Sigma(y_{ij}-\eta_{ij.})^2]^{-[J(I-1)-2]/2}$ having an infinite integral with respect to $\eta_{ij.}, \eta_{.j.} = y_{.j}$. Recalling that $\beta_j = y_{.j}-y_{..}$, $\eta_{ij.} = y_{.j}+\alpha_i+\gamma_{ij}$,

$$\Pr\{\alpha_i|y_{ij}, d_{ijv}\} \propto \int\prod d\gamma_{ij}[\Sigma(y_{ij}-y_{.j}-\alpha_i-\gamma_{ij})^2]^{-[J(I-1)-2]/2}\,\rho\{\Sigma(y_{ij}-y_{.j}-\alpha_i-\gamma_{ij})^2\}.$$

Since $\Sigma(y_{ij}-y_{.j}-\alpha_i-\gamma_{ij})^2 = \Sigma(y_{ij}-y_{.j}-y_{i.}+y_{..}-\gamma_{ij})^2+J\Sigma(y_{i.}-y_{..}-\alpha_i)^2$,

$$\begin{aligned}\Pr\{\alpha_i|y_{ij}, d_{ijv}\} &\propto \int_{U_0^2}^{\infty} dU^2\rho(U^2)\,U^{-[J(I-1)-2]}(U^2-U_0^2)^{\{(I-1)(J-1)-2\}/2}\\ &\propto U_0^{-(I-3)}\int_0^1 dz\rho\left(\frac{U_0^2}{1-z}\right)z^{\{(I-1)(J-1)-2\}/2}(1-z)^{(I-5)/2},\end{aligned} \tag{9}$$

where $U_0^2 = J\Sigma(y_{i.}-y_{..}-\alpha_i)^2$. If in fact $\rho(U^2) \propto U^{-2}$ then

$$\Pr\{\alpha_i|y_{ij}, d_{ijv}\} \propto [\Sigma(y_{i.}-y_{..}-\alpha_i)^2]^{-(I-1)/2},$$

agreeing with (3). This posterior distribution is of course improper, whereas (8) and (9) are proper if $\rho(U^2)$ is proper.

The above posterior distributions are appropriate when $\Pr\{\eta_{ij.}, U^2\}$ is a suitably gentle function. Often, however, it is natural to regard the γ_{ij} as small relative to the α_i. If in fact the interactions vanish, and if $\Pr\{\alpha_i, U^2\} \propto \rho(U^2)$, then from (8),

$$\begin{aligned}\Pr\{\alpha_i|y_{ij}, d_{ijv}\} &\propto [\Sigma(y_{ij}-y_{.j}-\alpha_i)^2]^{-[J(I-1)-2]/2}\,\rho[\Sigma(y_{ij}-y_{.j}-\alpha_i)^2]\\ &\propto [1+[\{J\Sigma(y_{i.}-y_{..}-\alpha_i)^2\}/\{\Sigma(y_{ij}-y_{i.}-y_{.j}+y_{..})^2\}]]^{-[J(I-1)-3]/2}\\ &\quad\times\rho(\Sigma(y_{ij}-y_{.j}-\alpha_i)^2),\end{aligned} \tag{10}$$

$$\Pr\{U^2|y_{ij}, d_{ijv}\} \propto \rho(U^2)\,U^{-(I-1)(J-1)}\,[1-[\{\Sigma(y_{ij}-y_{i.}-y_{.j}+y_{..})^2\}/U^2]^{(I-3)/2}$$

if $U^2 > \Sigma(y_{ij}-y_{i.}-y_{.j}+y_{..})^2$, and is otherwise zero.

Technical errors can be dealt with by basing inference upon the $y_{ij}-y_{.j}$ alone, as in Case D below.

Case D. Randomized blocks, blocks random

Let the J blocks in the experiment now be viewed as a sample from some population of blocks. Then the observations can be written

$$y_{ij} = \eta+\alpha_i+b_j+c_{ij}+\tilde{\epsilon}_{ij}+l_{ij},$$

with $\alpha_{\cdot} = c_{\cdot j} = \tilde{\epsilon}_{\cdot j} = 0$ and l_{ij} as technical error. See, for example, Scheffé (1959, p. 266). Now $y_{.j} = \eta+b_j+l_{.j}$ and thus carries only indirect information about $\{\alpha_i\}$. Although a more refined analysis is possible and sometimes necessary, the primary aim here is to evaluate the posterior distribution of $\{\alpha_i\}$, and ordinarily this can be done with only slight loss by basing inference upon $\{y_{ij}-y_{.j}\}$. But

$$y_{ij}-y_{.j} = \alpha_i+c_{ij}+\tilde{\epsilon}_{ij}+l_{ij}-l_{.j},$$

and often, I believe, it will be natural to assume C.U.A. for

$$\tilde{\epsilon}^*_{ij} \equiv \tilde{\epsilon}_{ij} + c_{ij} + l_{ij} - l_{.j},$$

given $(U^*)^2 \equiv (\sum \tilde{\epsilon}^*_{ij})^2$, $\{\alpha_i, d_{ijv}\}$. This leads to a posterior distribution for $\{\alpha_i\}$ much as in Case C.

3. A General Theory of Least Squares

A general theory of least squares which includes most of the earlier examples as special cases (and can easily be extended to include all) will now be presented. The approach is coordinate-free.

Let us suppose that a vector $\mathbf{Y}$ of observations in an n-dimensional real vector space V can be written $\mathbf{Y} = \boldsymbol{\eta} + \boldsymbol{\xi}$, where $\boldsymbol{\eta}$ lies in a known linear manifold S of dimension s, $\boldsymbol{\xi}$ lies in a known linear manifold T of dimension t, and let P_S, P_T denote the orthogonal projection operators on S and T, respectively. We shall suppose also that the purpose of the experiment is to draw inference about $\boldsymbol{\eta}$ and $U^2 = \|\boldsymbol{\xi}\|^2$, where the norm or length of a vector is defined in terms of the inner product, as usual. Since $\boldsymbol{\eta} - P_T(\boldsymbol{\eta}) = \mathbf{Y} - P_T(\mathbf{Y})$ is observable, so that it is only necessary to consider inference about U^2 and $P_T(\boldsymbol{\eta}) = P_T(\mathbf{Y}) - \boldsymbol{\xi} \in T$, there is thus no real loss in generality in supposing $S \subset T$ to begin with, and we shall do so henceforth.

The form of C.U.A. which I adopt is to suppose that given $\boldsymbol{\eta}, U^2$, $\boldsymbol{\xi}$ is conditionally uniformly distributed on the intersection of T with the "sphere" $\|\boldsymbol{\xi}\|^2 = U^2$. Thus $\Pr\{\boldsymbol{\xi} \mid \boldsymbol{\eta}, U^2\} \propto \phi(U^2)$ for some function ϕ of U^2, and $\boldsymbol{\xi}$ in the intersection. The posterior distribution of $\boldsymbol{\eta}$ is therefore given by

$$\begin{aligned} \Pr\{\boldsymbol{\eta} \mid \mathbf{Y}\} &\propto \Pr\{\mathbf{Y} \mid \boldsymbol{\eta}\} \Pr\{\boldsymbol{\eta}\} \\ &\propto \Pr\{\boldsymbol{\xi} = \mathbf{Y} - \boldsymbol{\eta} \mid \boldsymbol{\eta}\} \Pr\{\boldsymbol{\eta}\} \\ &\propto \phi(\|\mathbf{Y} - \boldsymbol{\eta}\|^2) \Pr\{U^2 = \|\mathbf{Y} - \boldsymbol{\eta}\|^2 \mid \boldsymbol{\eta}\} \Pr\{\boldsymbol{\eta}\}, \quad \boldsymbol{\eta} \in S, \end{aligned} \tag{11}$$

where $\Pr\{\boldsymbol{\eta}, U^2\}$ denotes the prior distribution of $\boldsymbol{\eta}, U^2$. Since $U^2 = \|\mathbf{Y} - \boldsymbol{\eta}\|^2$, this determines the posterior distribution of U^2. In particular, if V is R^n (n-dimensional Euclidean space), the inner product is the usual inner product for R^n, and prior knowledge is formally represented by the improper density $\Pr\{\boldsymbol{\eta}, U^2\} \propto \rho(U^2)$, $\boldsymbol{\eta} \in S$, then

$$\left.\begin{aligned} \Pr\{\boldsymbol{\eta} \mid \mathbf{Y}\} &\propto \|\mathbf{Y} - \boldsymbol{\eta}\|^{-(t-2)} \rho(\|\mathbf{Y} - \boldsymbol{\eta}\|^2), \quad \boldsymbol{\eta} \in S, \\ \Pr\{U^2 \mid \mathbf{Y}\} &\propto \rho(U^2)\, U^{-(t-2)} [U^2 - \|\mathbf{Y} - P_S(\mathbf{Y})\|^2]^{(s-2)/2} \\ &\propto \rho(U^2)\, U^{-(t-s)} [1 - \{\|\mathbf{Y} - P_S(\mathbf{Y})\|^2 / U^2\}]^{(s-2)/2} \end{aligned}\right\} \tag{12}$$

for $U^2 > \|\mathbf{Y} - P_S(\mathbf{Y})\|^2$, and otherwise zero.† I believe these posterior distributions will be appropriate in a wide variety of real situations, as is discussed further in Section 4.

Note that (12) agrees with the corresponding posterior distributions based upon conventional normality assumptions. Indeed, suppose that $\boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, where $\boldsymbol{\xi} \in T$ as before, $\mathbf{I}$ is the $t \times t$ identity matrix, and I am using the same notation for the vector $\boldsymbol{\xi}$ and for its coordinates relative to an orthonormal basis for T. Clearly

$$\Pr\{\boldsymbol{\xi} \mid \|\boldsymbol{\xi}\|^2\} \propto \|\boldsymbol{\xi}\|^{-(t-2)},$$

† It is to be understood that densities are densities with respect to Lebesgue measure restricted to the appropriate manifold, and $s \geqslant 1$.

and if

$$\Pr\{\eta, \sigma^2\} \propto h(\sigma^2), \quad \eta \in S,$$

then

$$\rho(U^2) \propto U^{t-2} \int_0^\infty \sigma^{-t} \exp\left[-(U^2/2\sigma^2)\right] h(\sigma^2)\, d\sigma^2.$$

Thus C.U.A. follows from conventional normality assumptions, and with the indicated choice of $\rho(U^2)$ the corresponding posterior distributions of η and U^2 are identical. Under normality, in addition

$$\Pr\{\sigma^2 | \mathbf{Y}\} \propto h(\sigma^2)\, \sigma^{-(t-s)} \exp - [\| \mathbf{Y} - P_S(\mathbf{Y}) \|^2 / 2\sigma^2].$$

If in fact $h(\sigma^2) \propto (\sigma^2)^{-1}$, then also $\rho(U^2) \propto (U^2)^{-1}$, and $\Pr\{\eta | \mathbf{Y}\} \propto \| \eta - \mathbf{Y} \|^{-t}$, the last agreeing with the results of others (Jeffreys, 1961, p. 147; Lindley, 1965, p. 222).

Another aspect of the relationship between C.U.A. and normality is revealed if $T = V = R^n$, and if the components of ξ are regarded as a sample from a population of errors $\{l_1, \ldots, l_N\}$. Let $\sigma^2 = (N-1)^{-1} \sum^N (l_i - l_\cdot)^2$, and suppose that conditional upon $\sigma^2, l_\cdot$, the vector $\mathbf{l} = (l_1, \ldots, l_N)'$ is uniformly distributed on the sphere of radius $[(N-1)\sigma^2]^{1/2}$ centred at $\mathbf{l}_\cdot = (l_\cdot, \ldots, l_\cdot)'$ in N-dimensional Euclidean space. Let $\xi_1, \ldots, \xi_n$, be any $n < N$ coordinates of $\mathbf{l}$. Then for $\sum^n (\xi_i - l_\cdot)^2 \leqslant (N-1)\sigma^2$,

$$\begin{aligned} \Pr\{\xi_1, \ldots, \xi_n | l_\cdot, \sigma^2\} &\propto \sigma^{-(N-2)} [(N-1)\sigma^2 - \textstyle\sum^n (\xi_i - l_\cdot)^2]^{(N-n-2)/2} \\ &\propto \sigma^{-n} [1 - \{\textstyle\sum (\xi_i - l_\cdot)^2 / (N-1)\sigma^2\}]^{(N-n-2)/2}, \end{aligned}$$

or, if $N-n$ is large, approximately

$$\sigma^{-n} \exp - \tfrac{1}{2} [\{(N-n-2)/(N-1)\} \{\textstyle\sum (\xi_i - l_\cdot)^2 / \sigma^2\}].$$

Thus C.U.A. for $\mathbf{l}$ as above implies both approximate normality and a form of C.U.A. for the vector of errors $\xi = (\xi_1, \ldots, \xi_n)'$.

Now return to the general model in R^n and superpose a technical error $\mathbf{f}$, so $\mathbf{Y} = \eta + \xi + \mathbf{f}$. For example, $\mathbf{f}$ may represent a sample from a larger population of errors $\{l_1, \ldots, l_N\}$, as in the last paragraph, and where now $l_\cdot = 0$. Then $P_T(\mathbf{Y}) = \eta + \xi^*$ and $\mathbf{Y} - P_T(\mathbf{Y}) = \mathbf{f} - P_T(\mathbf{f})$, where $\xi^* = \xi + P_T(\mathbf{f})$. But it is easily shown that the sum of independent random quantities each satisfying C.U.A. (for spheres centred at the origin) will also satisfy C.U.A., so it follows that C.U.A. will be appropriate for ξ^* as well as for ξ and $P_T(\mathbf{f})$. This suggests that the posterior distribution of η and $\| \xi^* \|^2$ might be evaluated as before based only upon the data $P_T(\mathbf{Y})$ and the posterior distribution of σ^2 based only upon the data $\mathbf{Y} - P_T(\mathbf{Y})$. Although there exist states of mind (that is, of prior knowledge) under which this would be misleading, I believe that ordinarily the posterior distributions determined in this way will be good approximations to the corresponding posterior distributions given all the data.

4. Conditional Uniformity on Spheres

Some aspects of the mathematical relationship between conventional forms of normality assumption and the assumption of conditional uniformity on spheres have been explored in Section 3. In particular, it has been seen that C.U.A. for a large population of errors implies approximate normality for a small sample from the population. In general, however, normality is a substantially stronger assumption. Thus, in the notation of Section 3, normality implies both C.U.A. and also that given σ^2, the quantity U^2/σ^2 has the χ^2 distribution with t degrees of freedom. Still another

sense in which C.U.A. is weaker than normality is implied by Maxwell's theorem, which states that spherical symmetry, together with independence of the components, implies normality. Here independence is meant in the frequentist sense, which is equivalent to conditional independence given all parameters for the Bayesian, as is clarified by deFinetti (1964, Ch. III).

The fact that C.U.A. is mathematically weaker than normality and yet leads to the same posterior distribution and therefore the same inference for η would itself be sufficient reason for interest in C.U.A. What seems to me to be more important, however, is that in practice C.U.A. will often be a great deal more natural and plausible as a description of the knowledge or opinions of a person than the corresponding assumption of normality. It is implicit here that I would not necessarily regard the components of the error vector as independent, or the quantity U^2/σ^2 as having a χ^2 distribution. Consider, for example, the completely randomized design of Case B, Section 2. Then in the first place there may be no natural larger population from which to imagine that the IJ units in the experiment have been sampled, and so σ^2, the variance of such a population, would be fictitious. But even if such a population did exist one might prefer to assess directly a prior distribution for U^2 based upon the total perception of the actual units in the experiment, and then draw an inference about $\{\alpha_i\}$ using C.U.A., rather than to draw such an inference indirectly and implicitly in terms of a normality assumption and the parameter σ^2. The essential point is that if the purpose of the experiment is to make an inference about treatment comparisons, then there is no need to introduce the parameter σ^2 unless it is useful in evaluating the prior distribution of U^2, or is of interest in its own right. This leads to a more flexible analysis. (On the other hand, it might be objected that such an approach does not provide estimates of the standard error. This is a side-issue so far as the present paper is concerned, since from the Bayesian viewpoint the whole of the inferential problem regarding the $\{\alpha_i\}$ is contained in the evaluation of their posterior distribution. It is true of course, as can be seen immediately from equation (5), that the degree of concentration of this posterior distribution may be largely determined by much the same kind of sample quantities that others would regard as estimates of the standard error. However, for the Bayesian such quantities are relevant to inference and decision making about the $\{\alpha_i\}$ only indirectly through their effect upon the posterior distribution. This is not to say that U^2 and σ^2 may not be of interest in their own right, but only to point out that this would be another matter entirely. I am aware, of course, that most statisticians would disagree with me on this question. However, to the best of my knowledge the only serious attempt to justify such a viewpoint is that of Fisher (1960, p. 64), who argues that only randomization provides a valid estimate of the standard error and a valid significance test. Thus he argues that a reduction of the true errors, unaccompanied by their elimination in the statistical analysis, yields an inflated estimate of the variance (which is certainly true), and is therefore of no value. Is it unreasonable of me to insist that his argument carried to its logical conclusion implies that even if one could entirely eliminate all sources of error by means of judicious allocation of treatments to units, and thus determine with certainty all contrasts between treatments, this would be of no value because there would then be no valid test of significance?)

The foregoing discussion is intended to show only that C.U.A. constitutes something of an improvement over conventional normality assumptions as a basis for inference. Supposing that this be granted, the question still remains as to how C.U.A. can be justified. It must be understood here that from the standpoint of personal

probability, C.U.A. does not imply that ξ is in any literal sense sampled conditionally uniform on the sphere, but is rather meant as an approximate description of the opinions of a person, given all available evidence. As such it is clearly a way of expressing ignorance or vague prior knowledge, and to me it seems to stand virtually alone as the only general way of doing so which commands any real credibility. I cannot hope to prove this, any more than I can hope to prove that any other coherent way of evaluating probabilities is the "right" way to do so. I can hope that others will find C.U.A. as compelling as I do. It is my belief that in fact an underlying attitude of conditional uniformity for errors is at the heart of the willingness of many people to assume the normal law of errors, and as a consequence to adopt least-squares methods of analysis of data. In this sense conditional uniformity provides a foundation for the theory of least squares. Other attempts to justify least squares and the analysis of variance have been made on the basis of randomization, the Gauss–Markoff theorem, and the like, but these do not seem to be entirely satisfactory even from the frequentist viewpoint, and in any case are hardly appropriate for a Bayesian. The enormously appealing normality assumption, on the other hand, remained more or less of a mystery, since it had not been related to any plausible state of prior knowledge. Thus the famous remark to the effect that "everybody believes in the law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact". I am proposing that conditional uniformity for errors is often a plausible state of prior knowledge and lies at the heart of the theory of least squares.

REFERENCES

DEFINETTI, B. (1964). Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability* (ed. by H. E. Kyburg and H. E. Smokler), pp. 93–158. New York: Wiley.

FISHER, R. A. (1960). *The Design of Experiments* (7th ed.). Edinburgh: Oliver & Boyd.

JEFFREYS, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Clarendon Press.

LINDLEY, D. V. (1965). *Introduction to Probability and Statistics.* Part 2. *Inference.* Cambridge: University Press.

SAVAGE, L. J., LINDMAN, H. and EDWARDS, W. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* New York: Wiley.