Bayes Estimates for the Linear Model

By D. V. LINDLEY AND A. F. M. SMITH

University College, London

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1971, Mr M. J. R. HEALY in the Chair]

SUMMARY

The usual linear statistical model is reanalyzed using Bayesian methods and the concept of exchangeability. The general method is illustrated by applications to two-factor experimental designs and multiple regression.

Keywords: LINEAR MODEL; LEAST SQUARES; BAYES ESTIMATES; EXCHANGEABILITY; ADMISSIBILITY; TWO-FACTOR EXPERIMENTAL DESIGN; MULTIPLE REGRESSION; RIDGE REGRESSION; MATRIX INVERSION.

INTRODUCTION

ATTENTION is confined in this paper to the linear model, $E(\mathbf{y}) = \mathbf{A}\boldsymbol{\theta}$, where $\mathbf{y}$ is a vector of observations, $\mathbf{A}$ a known design matrix and $\boldsymbol{\theta}$ a vector of unknown parameters. The usual estimate of $\boldsymbol{\theta}$ employed in this situation is that derived by the method of least squares. We argue that it is typically true that there is available prior information about the parameters and that this may be exploited to find improved, and sometimes substantially improved, estimates. In this paper we explore a particular form of prior information based on de Finetti's (1964) important idea of exchangeability.

The argument is entirely within the Bayesian framework. Recently there has been much discussion of the respective merits of Bayesian and non-Bayesian approaches to statistics: we cite, for example, the paper by Cornfield (1969) and its ensuing discussion. We do not feel that it is necessary or desirable to add to this type of literature, and since we know of no reasoned argument against the Bayesian position we have adopted it here. Nevertheless the reader not committed to this approach may like to be reminded that many techniques of the sampling-theory school are basically unsound: see the review by Lindley (1971b). In particular the least-squares estimates are typically unsatisfactory: or, in the language of that school, are inadmissible in dimensions greater than two. This follows since, by a well-known device in least-squares theory (see, for example, Plackett, 1960, p. 59), we may write the linear model after transformation in the form $E(z_i) = \xi_i$ for $i \leqslant p$ and $E(z_i) = 0$ for $i > p$. Here the z's are transforms of the data, and the ξ's of the parameters. Adding the assumption of normality, we can appeal to the results of Brown (1966), generalizing those of Stein (1956), which show that for a very wide class of loss functions the estimate of ξ_i by z_i, for $i \leqslant p$ is inadmissible. In Section 1 of this paper we do comment on the admissibility of the Bayesian estimates and try to show, in a way that might appeal to an adherent of orthodox ideas, that they are likely to be superior, at least in some situations, to the least-squares estimates.

1. EXCHANGEABILITY

We begin with a simple example. Suppose, in the general linear model, that the design matrix is the unit matrix so that $E(y_i) = \theta_i$ for $i = 1, 2, \ldots, n$, and that $y_1, y_2, \ldots, y_n$ are independent, normally distributed with known variance σ^2. Such a simple model might arise if y_i was the observation on the ith variety in a field trial, of average yield θ_i. In considering the prior knowledge of the θ_i it may often be reasonable to assume their distribution *exchangeable*. That is, that it would be unaltered by any permutation of the suffixes: so that, in particular, the prior opinion of θ_7 is the same as that of θ_4, or any other θ_i; and similarly for pairs, triplets and so on. Now one way of obtaining an exchangeable distribution $p(\boldsymbol{\theta})$ is to suppose

$$p(\boldsymbol{\theta}) = \int \prod_{i=1}^{n} p(\theta_i|\mu)\, dQ(\mu),$$

where $p(\theta_i|\mu)$, for each μ, and $Q(\mu)$ describe arbitrary probability distributions. In other words, $p(\boldsymbol{\theta})$ is a *mixture*, by $Q(\mu)$, of independent and identical distributions, given μ. Indeed, Hewitt and Savage (1955), in generalization of de Finetti's original result, have shown that if exchangeability is assumed for *every* n, then a mixture is the *only* way to generate an exchangeable distribution.

In the present paper we study situations where we have exchangeable prior knowledge and assume this exchangeability described by a mixture. In the example this implies $E(\theta_i) = \mu$, say, a common value for each i. In other words there is a linear structure to the *parameters* analogous to the linear structure supposed for the observations $\mathbf{y}$. If we add the premise that the distribution from which the θ_i appear as a random sample is normal, the parallelism between the two stages, for $\mathbf{y}$ and $\boldsymbol{\theta}$, becomes closer. In this paper we study the situation in which the parameters of the general linear model themselves have a general linear structure in terms of other quantities which we call *hyperparameters*.† In this simple example there is just one hyperparameter, μ.

Indeed, we shall find it necessary to go further and let the hyperparameters also have a linear structure. This will be termed a *three-stage model* and is analysed in detail in the next section. There are straightforward extensions to any number of stages.

Returning to the simple example with $E(y_i) = \theta_i$, $E(\theta_i) = \mu$ and respective variances σ^2 and τ^2, say, the situation will be completely specified once a prior distribution has been given for μ. (Effectively this is the third stage just mentioned.) Supposing μ to have a uniform distribution over the real line—a situation usually described by saying there is vague prior knowledge of μ—Lindley (1971a) has obtained the posterior distribution of θ_i and found its mean to be

$$E(\theta_i|\mathbf{y}) = \frac{y_i/\sigma^2 + y_.\/\tau^2}{1/\sigma^2 + 1/\tau^2}, \tag{1}$$

where $y_. = \sum y_i/n$. The detailed analysis has been given in the reference just cited, so we content ourselves with a brief discussion to serve as an introduction to the general theory in the next section.

The estimates, (1), will be referred to as *Bayes* estimates, and it is these that we propose as substitutes for the usual least-squares estimates. We denote them by θ_i^*,

† We believe we have borrowed this terminology from I. J. Good but are unable to trace the reference.

and reserve the usual notation, $\hat{\theta}_i$, for the ordinary estimates. Notice that θ_i^* is a weighted average of $y_i = \hat{\theta}_i$ and the overall mean, $y_.$, with weights inversely proportional to the variances of y_i and θ_i. Hence the natural estimates are pulled towards a central value $y_.$, the extreme values experiencing most shift. We shall find the weighted average phenomenon will persist even within the general model. Of course the estimate (1) depends on σ^2 and τ^2, which will typically be unknown, but their estimation presents no serious difficulties. If, for each i, there is replication of the y_i then σ^2 may be estimated as the usual within variance. Since we have replication (from the distribution $N(\mu, \tau^2)$ underlying the exchangeability assumption) for the θ_i, τ^2 may be estimated. For example $\sum(\theta_i^* - \theta_.^*)^2/(n-1)$ might be a reasonable estimate of τ^2, although in fact the reference just cited shows this can be improved upon. These estimates of σ^2 and τ^2 can be used in place of the known values used in (1) and the cycle repeated.

Let us now digress from the Bayesian viewpoint and try to persuade an orthodox statistician that (1) is a sensible estimate for him to consider, and indeed is better than the least-squares estimate. Of course, θ_i^* is a biased estimate of θ_i, so its merit cannot be judged by its variance. We use instead the mean-square error $E(\theta_i^* - \theta_i)^2$. This is just a criterion for judging the merit of one of the n estimates, so let us look at the average mean-square error over the n values. Simple, but tedious, calculations enable this to be found and compared with the corresponding quantity for $\hat{\theta}_i$, namely σ^2. The condition for the average m.s.e. for θ_i^* to be less than that for $\hat{\theta}_i$ is that

$$\sum(\theta_i - \theta_.)^2/(n-1) < 2\tau^2 + \sigma^2. \tag{2}$$

The m.s.e. for θ_i^* depends on θ_i and hence this condition does also. Consequently the Bayes estimates are not always superior to least-squares. But consider when (2) obtains. The θ_i are, by supposition, given μ, τ^2, a random sample from $N(\mu, \tau^2)$ so that the left-hand side of (2) is the usual estimate of τ^2, had the θ_i been known. Hence the condition is that the estimate of τ^2 be less than $2\tau^2 + \sigma^2$. The distribution of the estimate is a multiple of χ^2 and simple calculations show that the chance—according to the $N(\mu, \tau^2)$ distribution—of (2) being satisfied is high for n as low as 4 and rapidly tends to 1 as n increases. But τ^2, as we have seen, can itself be estimated, so with this in (1) we are almost certain to have a smaller m.s.e. for θ_i^* than for $\hat{\theta}_i$. In particular the expectation (over the θ-distribution) is always in favour of the Bayes estimate.

That argument is heuristic. Our estimates are similar to those proposed by Stein (1956), which he rigorously showed to be superior (in the average m.s.e. sense) to the least-squares estimates. It has been pointed out to us by L. Brown (personal communication) that (1), with known σ^2, τ^2, is an admissible estimate. Essentially this is because the impropriety in our prior distribution is confined to one dimension—in μ. We digress to amplify this statement.

If a *proper* prior distribution (that is, one whose integral over the whole space is unity) and a *bounded* utility function are used, then the estimate obtained by using as an estimate that value which maximizes the expected (over the parameter distribution) utility is always admissible. This is easy to demonstrate since, under the two conditions stated, all the usual mathematical operations, such as reversals of order of integration, are valid. Difficulties arise if either of the italicized conditions above are violated. Quadratic loss, leading to m.s.e. is unbounded, but can conveniently be replaced by

$$1 - \exp\{-(\boldsymbol{\theta} - \mathbf{e})^{\mathrm{T}} \boldsymbol{\Lambda} (\boldsymbol{\theta} - \mathbf{e})\} \tag{3}$$

for estimate $\mathbf{e}$, where $\mathbf{\Lambda}$ is positive semi-definite and, in particular, a unit matrix. The use of vague prior knowledge, with a uniform, and therefore improper, prior distribution does cause difficulties and it is this feature, at least in dimensions higher than two, that gives rise to inadmissible estimates, as Stein was the first to show. In the general theory of the next section all our estimates will be admissible in terms of the bounded loss function (3) provided the prior distribution is proper; we conjecture admissibility if the impropriety is confined to at most two dimensions.

Returning, then, to the inequality (2), we see that there is good reason within the orthodox framework for preferring the new estimates to the old. Further justification may be found in papers by Hoerl and Kennard (1970a, b) who discuss a special case of the estimates that we shall develop in Section 5.3. We do not take these justifications very seriously, feeling that the Bayesian viewpoint is supported by so many general considerations in which criteria, like mean-square error, play little or no part, that the additional validation they provide is of small consequence.

Before proceeding to the general discussion one point must be emphasized. In the example we have assumed an exchangeable prior distribution. The estimates (1) are therefore only suggested when this assumption is practically realistic. It is the greatest strength of the Bayesian argument that it provides a formal system within which any inference or decision problem can be described. In passing from the real-world problem to its mathematical formulation it becomes necessary to make, and to expose, the assumptions. (This applies to any formalism, Euclidean geometry, for example, and not just to Bayesian statistics.) Here exchangeability is one such assumption, and its practical relevance must be assessed before the estimates based on it are used. For example, if, as suggested above, our model described the observed yields of n varieties in an agricultural field trial, the exchangeability assumption would be inappropriate if one or more varieties were controls and the remainder were experimental. However, the assumption might be modified to one of exchangeability within controls and separately within experimental varieties. Similarly with a two-way classification into rows and columns, it might be reasonable to assume separately that the rows and the columns were exchangeable. In any application the particular form of the prior distribution has to be carefully considered.

It should be noted that in assigning a prior distribution to the θ_i of the above form, whilst we are effectively regarding them as a random sample from $N(\mu, \tau^2)$, we are not thereby passing to a Model II, random effects, situation such as has been discussed by Fisk (1967) and Nelder (1968). We are interested in the estimation of the *fixed* effects. One of us (A. F. M. S.) has studied the genuine Model II situation and obtained estimates for μ ($\boldsymbol{\theta}_2$ in the general model below) but this will be reported separately.

We now turn to the general theory. The mathematics is not difficult for someone familiar with matrix algebra, and the main result is stated as a theorem with corollaries. The results in Section 2 all assume *known* variances. The extensions to unknown variances will be described later.

2. General Bayesian Linear Model

The notation $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{D})$ means that the column vector $\mathbf{y}$ has a (multivariate) normal distribution with mean $\boldsymbol{\mu}$, a column vector, and dispersion $\mathbf{D}$, a positive semi-definite matrix.

Lemma. Suppose, given $\boldsymbol{\theta}_1$, a vector of p_1 parameters,

$$\mathbf{y} \sim N(\mathbf{A}_1 \boldsymbol{\theta}_1, \mathbf{C}_1) \tag{4}$$

and that, given $\boldsymbol{\theta}_2$, a vector of p_2 hyperparameters,

$$\boldsymbol{\theta}_1 \sim N(\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_2). \tag{5}$$

Then (a) the marginal distribution of $\mathbf{y}$ is

$$N(\mathbf{A}_1\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}}), \tag{6}$$

and (b) the distribution of $\boldsymbol{\theta}_1$, given $\mathbf{y}$, is $N(\mathbf{Bb}, \mathbf{B})$ with

$$\mathbf{B}^{-1} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1+\mathbf{C}_2^{-1} \tag{7}$$

and

$$\mathbf{b} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}+\mathbf{C}_2^{-1}\mathbf{A}_2\boldsymbol{\theta}_2. \tag{8}$$

(Here $\mathbf{y}$ is a vector of n elements and $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{C}_1$ and $\mathbf{C}_2$ are known positive-definite matrices of obvious dimensions.)

The lemma is well known but we prove it here, both for completeness and because the proof has an unexpected byproduct.

To prove (a) we write (4) in the form $\mathbf{y} = \mathbf{A}_1\boldsymbol{\theta}_1+\mathbf{u}$, where $\mathbf{u}\sim N(\mathbf{0}, \mathbf{C}_1)$ and (5) as $\boldsymbol{\theta}_1 = \mathbf{A}_2\boldsymbol{\theta}_2+\mathbf{v}$ where $\mathbf{v}\sim N(\mathbf{0}, \mathbf{C}_2)$. Hence, putting these two equalities together, we have $\mathbf{y} = \mathbf{A}_1\mathbf{A}_2\boldsymbol{\theta}_2+\mathbf{A}_1\mathbf{v}+\mathbf{u}$. But, by the standard properties of normal distributions, $\mathbf{A}_1\mathbf{v}+\mathbf{u}$, a linear function of independent normal random variables, is $N(\mathbf{0}, \mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}})$ and the result follows.

To prove (b) we use Bayes's theorem,

$$p(\boldsymbol{\theta}_1|\,\mathbf{y}) \propto p(\mathbf{y}|\,\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_1).$$

The product on the right-hand side is $e^{-\frac{1}{2}Q}$ where Q is given by

$$\begin{aligned}(\mathbf{y}-\mathbf{A}_1\boldsymbol{\theta}_1)^{\mathrm{T}}\mathbf{C}_1^{-1}(\mathbf{y}-\mathbf{A}_1\boldsymbol{\theta}_1)+(\boldsymbol{\theta}_1-\mathbf{A}_2\boldsymbol{\theta}_2)^{\mathrm{T}}\mathbf{C}_2^{-1}(\boldsymbol{\theta}_1-\mathbf{A}_2\boldsymbol{\theta}_2)\\ = \boldsymbol{\theta}_1^{\mathrm{T}}\mathbf{B}^{-1}\boldsymbol{\theta}_1-2\mathbf{b}^{\mathrm{T}}\boldsymbol{\theta}_1+\{\mathbf{y}^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}+\boldsymbol{\theta}_2^{\mathrm{T}}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2\boldsymbol{\theta}_2\}\end{aligned}$$

on collecting the quadratic and linear terms in $\boldsymbol{\theta}_1$ together, and using the expressions (7) and (8) for $\mathbf{b}$ and $\mathbf{B}$. Completing the square in $\boldsymbol{\theta}_1$, Q may finally be written

$$(\boldsymbol{\theta}_1-\mathbf{Bb})^{\mathrm{T}}\mathbf{B}^{-1}(\boldsymbol{\theta}_1-\mathbf{Bb})+\{\mathbf{y}^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}+\boldsymbol{\theta}_2^{\mathrm{T}}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2\boldsymbol{\theta}_2-\mathbf{b}^{\mathrm{T}}\mathbf{Bb}\}. \tag{9}$$

The term in braces is a constant as far as the distribution of $\boldsymbol{\theta}_1$ is concerned, and the remainder of the expression demonstrates the truth of (b).

The proof of the lemma is complete, but by combining the separate proofs of (a) and (b) an interesting result can be obtained. On integrating $e^{-\frac{1}{2}Q}$, with Q given by (9), with respect to $\boldsymbol{\theta}_1$, the result is proportional to the density of $\mathbf{y}$, already obtained in (a). The integration does not affect the term in braces in (9) so that, in particular, the quadratic term in $\mathbf{y}$ in (9)—remembering that $\mathbf{b}$ contains $\mathbf{y}$—may be equated to the quadratic term obtained directly from (6), with the result that

$$\mathbf{C}_1^{-1}-\mathbf{C}_1^{-1}\mathbf{A}_1\mathbf{B}\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1} = \{\mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}}\}^{-1}.$$

We therefore have the

Matrix lemma. For any matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{C}_1$ and $\mathbf{C}_2$ of appropriate dimensions and for which the inverses stated in the result exist, we have

$$\mathbf{C}_1^{-1}-\mathbf{C}_1^{-1}\mathbf{A}_1(\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1+\mathbf{C}_2^{-1})^{-1}\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1} = \{\mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}}\}^{-1}. \tag{10}$$

The result follows from the last equation on inserting the form for $\mathbf{B}$, equation (7). It is, of course, easy to prove the result (10) directly once its truth has been conjectured: furthermore $\mathbf{C}_1$ and $\mathbf{C}_2$ do not have to be positive definite. It suffices to multiply the left-hand side of (10) by $\mathbf{C}_1+\mathbf{A}_1\mathbf{C}_2\mathbf{A}_1^{\mathrm{T}}$ and verify that the result is a unit matrix. The above proof is interesting because it does not require an initial conjecture and because it uses a probabilistic argument to derive a purely algebraic result. The matrix lemma is important to us since it provides simpler forms than would otherwise be available for our estimates. This result has been given by Rao (1965, Exercise 2.9, p. 29).

We next proceed to the main result. As explained in Section 1, we are dealing with the linear model, which is now written in the form $E(\mathbf{y}) = \mathbf{A}_1\boldsymbol{\theta}_1$, the suffixes indicating that this is the first stage in the model. We generalize to an arbitrary dispersion matrix, $\mathbf{C}_1$, for $\mathbf{y}$. The prior distribution of $\boldsymbol{\theta}_1$ is expressed in terms of hyperparameters $\boldsymbol{\theta}_2$ as another linear model, $E(\boldsymbol{\theta}_1) = \mathbf{A}_2\boldsymbol{\theta}_2$ with dispersion matrix $\mathbf{C}_2$. This can proceed for as many stages as one finds convenient: it will be enough for us to go to three, supposing the mean, as well as the dispersion, known at the final stage. For our inferences, and in particular for estimation, we require the posterior distribution of $\boldsymbol{\theta}_1$. This is provided by the following result.

Theorem. Suppose that, given $\boldsymbol{\theta}_1$,

$$\mathbf{y} \sim N(\mathbf{A}_1\boldsymbol{\theta}_1, \mathbf{C}_1), \tag{11.1}$$

given $\boldsymbol{\theta}_2$,

$$\boldsymbol{\theta}_1 \sim N(\mathbf{A}_2\boldsymbol{\theta}_2, \mathbf{C}_2) \tag{11.2}$$

and given $\boldsymbol{\theta}_3$,

$$\boldsymbol{\theta}_2 \sim N(\mathbf{A}_3\boldsymbol{\theta}_3, \mathbf{C}_3). \tag{11.3}$$

Then the posterior distribution of $\boldsymbol{\theta}_1$, given $\{\mathbf{A}_i\}$, $\{\mathbf{C}_i\}$, $\boldsymbol{\theta}_3$ and $\mathbf{y}$ is $N(\mathbf{Dd}, \mathbf{D})$ with

$$\mathbf{D}^{-1} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1 + \{\mathbf{C}_2+\mathbf{A}_2\mathbf{C}_3\mathbf{A}_2^{\mathrm{T}}\}^{-1} \tag{12}$$

and

$$\mathbf{d} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y} + \{\mathbf{C}_2+\mathbf{A}_2\mathbf{C}_3\mathbf{A}_2^{\mathrm{T}}\}^{-1}\mathbf{A}_2\mathbf{A}_3\boldsymbol{\theta}_3. \tag{13}$$

(Here $\boldsymbol{\theta}_i$ is a vector of p_i elements and the dispersion matrices, $\mathbf{C}_i$, are all supposed non-singular.)

The joint distribution of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is described in (11.2) and (11.3). The use of part (a) of the lemma enables the marginal distribution of $\boldsymbol{\theta}_1$ to be written down as

$$\boldsymbol{\theta}_1 \sim N(\mathbf{A}_2\mathbf{A}_3\boldsymbol{\theta}_3, \mathbf{C}_2+\mathbf{A}_2\mathbf{C}_3\mathbf{A}_2^{\mathrm{T}}). \tag{14}$$

(Notice that this is the prior distribution of $\boldsymbol{\theta}_1$ free of the hyperparameters $\boldsymbol{\theta}_2$. We could have expressed the prior in this way but in applications we find the hierarchical form more convenient.)

Then, with (14) as prior, (11.1) as likelihood, part (b) of the lemma shows that the posterior distribution of $\boldsymbol{\theta}_1$ is as stated.

In particular the mean of the posterior distribution may be regarded as a point estimate of $\boldsymbol{\theta}_1$ to replace the usual least-squares estimate. The form of this estimate is a generalization of the form noted in the example of Section 1; namely, it is a weighted average of the least-squares estimate $(\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}$ and the prior mean $\mathbf{A}_2\mathbf{A}_3\boldsymbol{\theta}_3$ (equation (14)) with weights equal to the inverses of the corresponding dispersion matrices, $\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1$ for the least-squares values, $\mathbf{C}_2+\mathbf{A}_2\mathbf{C}_3\mathbf{A}_2^{\mathrm{T}}$ for the

prior distribution (14). For the simple example considered in Section 1 we produced an heuristic argument to show that, with respect to our prior distribution, we were confident of satisfying inequality (2) and thus achieving smaller mean square error than with the least-squares estimate. This result can be shown to hold generally for Bayes's estimates derived from hierarchical prior structures, as in (11.1)–(11.3), and will be presented in a future paper.

The matrix lemma enables us to obtain several alternative forms for the term in braces in (12), and hence for the posterior mean and variance, both of which involve this expression. These alternatives look more complicated than those already stated but are often useful in applications. Notice that a computational advantage of the matrix lemma is that its use reduces the order of the matrices to be inverted. The matrix on the right-hand side of (10) is of order n, whereas on the left-hand side, apart from $\mathbf{C}_1$ which is usually of a simple structure (often $\mathbf{C}_1 = \sigma^2\mathbf{I}$), the matrix to be inverted is of order p_1, typically much less than n.

Corollary 1. An alternative expression for $\mathbf{D}^{-1}$ (equation (12)) is

$$\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1+\mathbf{C}_2^{-1}-\mathbf{C}_2^{-1}\mathbf{A}_2(\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2+\mathbf{C}_3^{-1})^{-1}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}. \tag{15}$$

This is immediate on applying (10), with the suffixes all increased by one, to the second term in (12).

In most applications of these results the design of the experiment rather naturally suggests the second stage, (11.2), in the hierarchy but at the third stage we find ourselves in a position where the prior knowledge is weak. (Least-squares results apply when the second-stage prior knowledge is weak.) It is natural to express this by supposing the third-stage dispersion matrix $\mathbf{C}_3$ to be large, or to let its inverse, the precision matrix, be zero. In the original form of (12) and (13) it is not easy to see what happens when $\mathbf{C}_3^{-1} = \mathbf{0}$, but (15) enables the form to be seen easily.

Corollary 2. If $\mathbf{C}_3^{-1} = \mathbf{0}$, the posterior distribution of $\boldsymbol{\theta}_1$ is $N(\mathbf{D}_0\mathbf{d}_0, \mathbf{D}_0)$ with

$$\mathbf{D}_0^{-1} = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1+\mathbf{C}_2^{-1}-\mathbf{C}_2^{-1}\mathbf{A}_2(\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1} \tag{16}$$

and

$$\mathbf{d}_0 = \mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{y}. \tag{17}$$

The form for $\mathbf{D}_0^{-1}$ follows by direct substitution of $\mathbf{C}_3^{-1} = \mathbf{0}$ in (15). That for $\mathbf{d}_0$ follows by remarking that if the second and third terms in (15) are postmultiplied by $\mathbf{A}_2$ the result is zero, but such postmultiplication takes place in the original expression for $\mathbf{d}$, equation (13).

This corollary is the form we shall most often use in applications.

It is possible to extend the theorem to cases where some or all of the dispersion matrices $\mathbf{C}_i$ are singular. This can be accomplished using generalized inverses and will be the subject of a separate paper. Notice that we have not assumed, as in the usual least-squares theory, that $\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1$ is non-singular. (The case $\mathbf{C}_1 = \sigma^2\mathbf{I}$ will be more familiar.) In the standard exposition it is usual to constrain the individual parameters in the vector $\boldsymbol{\theta}_1$ to preserve identifiability in the likelihood function. Identifiability problems do not arise in the Bayesian formulation since, provided the prior distribution is proper, so is the posterior, whether or not the parameters referred to in these two distributions are identifiable or not in the likelihood function. An example below will help to make this clear.

The situation described in Section 1 has already been discussed in detail by Lindley (1971a), though not within the general framework which was briefly described

in Lindley (1969). The interested reader can easily fit the example into the argument of this section. Corollary 2 is relevant and it is an easy matter to perform the necessary matrix calculations. We proceed to the discussion of other examples.

3. EXAMPLES

3.1. *Two-factor Experimental Designs*

Consider t "treatments" assigned to n experimental units arranged in b "blocks". If the ith treatment is applied within the jth block and yields an observation y_{ij}, the usual model is

$$E(y_{ij}) = \mu + \alpha_i + \beta_j \quad (1 \leqslant i \leqslant t, 1 \leqslant j \leqslant b)$$

with the errors independent $N(0, \sigma^2)$. In the general notation of (11.1)

$$\boldsymbol{\theta}_1^{\mathrm{T}} = (\mu, \alpha_1, \alpha_2, \ldots, \alpha_t, \beta_1, \beta_2, \ldots, \beta_b)$$

and $\mathbf{A}_1$ describes the design used.

For the second stage we argue as follows. It might be reasonable to assume that our prior knowledge of the treatment constants $\{\alpha_i\}$ was exchangeable, and similarly that of the block constants $\{\beta_j\}$, but that these were independent. We emphasize the word "might" in the last sentence. In repetition of the point made in Section 1, we remind the reader that this *assumption* is not always appropriate and our recipes below are not necessarily sensible when this form of exchangeability is unreasonable. For example, it may be known that the treatments are ordered, say $\alpha_1 \leqslant \alpha_2 \leqslant \ldots \leqslant \alpha_t$. In this case other forms of prior information are available and alternative estimates are sensible: these will be reported on in a separate paper.

Adding the assumptions of normality we therefore describe the second stage (11.2) by

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad \beta_j \sim N(0, \sigma_\beta^2), \quad \mu \sim N(\omega, \sigma_\mu^2),$$

these distributions being independent. The means of α_i and β_j have been chosen to be zero. Any other value would do since the likelihood provides no information about them, but the choice of zero mean is convenient, since it leads to straightforward comparisons of the Bayes and (constrained) least-squares estimates as deviations from an average level. We shall consider the case where the prior knowledge of μ is vague, so that $\sigma_\mu^2 \to \infty$; ω will then be irrelevant. A third stage is not necessary. We proceed to calculate expressions (12) and (13) for the posterior distribution of $\boldsymbol{\theta}_1$.

The matrix $\mathbf{C}_2$ is diagonal, so the same is true of $\mathbf{C}_2^{-1}$ and its leading diagonal is easily seen to be

$$(\sigma_\mu^{-2}, \sigma_\alpha^{-2}, \ldots, \sigma_\alpha^{-2}, \sigma_\beta^{-2}, \ldots, \sigma_\beta^{-2})$$

and as $\sigma_\mu^2 \to \infty$, the first element tends to zero. $\mathbf{C}_1$ is the unit matrix times σ^2. We can therefore substitute these values into (12) and (13), remembering that $\mathbf{C}_3 = \mathbf{0}$ and $(\mathbf{A}_3 \boldsymbol{\mu})^{\mathrm{T}} = (\omega, 0, \ldots, 0)$ and easily obtain

$$\mathbf{D}^{-1} = \sigma^{-2} \mathbf{A}_1^{\mathrm{T}} \mathbf{A}_1 + \mathbf{C}_2^{-1}$$

and

$$\mathbf{d} = \sigma^{-2} \mathbf{A}_1^{\mathrm{T}} \mathbf{y}.$$

Hence $\boldsymbol{\theta}_1^*$, the Bayes estimate $\mathbf{Dd}$, satisfies the equations

$$(\mathbf{A}_1^{\mathrm{T}} \mathbf{A}_1 + \sigma^2 \mathbf{C}_2^{-1}) \boldsymbol{\theta}_1^* = \mathbf{A}_1^{\mathrm{T}} \mathbf{y}. \tag{18}$$

These differ from the least-squares equations only in the inclusion of the extra term $\sigma^2 \mathbf{C}_2^{-1}$.

In the case of a complete randomized-block design where each treatment occurs exactly once in each block we have, on arranging the elements of $\mathbf{y}$ in lexicographical order,

$$(\mathbf{A}_1^{\mathrm{T}} \mathbf{A}_1 + \sigma^2 \mathbf{C}_2^{-1}) = \begin{pmatrix} bt & b\mathbf{1}_t^{\mathrm{T}} & t\mathbf{1}_b^{\mathrm{T}} \\ b\mathbf{1}_t & (b+\sigma^2/\sigma_\alpha^2)\mathbf{I}_t & \mathbf{J}_{t,b} \\ t\mathbf{1}_b & \mathbf{J}_{b,t} & (t+\sigma^2/\sigma_\beta^2)\mathbf{I}_b \end{pmatrix}, \tag{19}$$

where $\mathbf{1}_m$ is a vector of m 1's, $\mathbf{I}_m$ is the unit matrix of order m and $\mathbf{J}_{m,n}$ is a matrix of order $m \times n$ all of whose elements are 1. As usual

$$(\mathbf{A}_1^{\mathrm{T}} \mathbf{y})^{\mathrm{T}} = (bty_{..}, by_{1.}, \ldots, by_{t.}, ty_{.1}, \ldots, ty_{.b}).$$

Notice that the matrix (19) is non-singular and the solution to (18) is easily seen to be

$$\mu^* = y_{..}, \quad \alpha_i^* = (b\sigma_\alpha^2 + \sigma^2)^{-1} b\sigma_\alpha^2 (y_{i.} - y_{..}), \quad \beta_j^* = (t\sigma_\beta^2 + \sigma^2)^{-1} t\sigma_\beta^2 (y_{.j} - y_{..}). \tag{20}$$

Consequently the estimators of the treatment and block effects (on being measured from the overall mean) are shrunk towards zero by a factor depending on the ratio of σ^2 to σ_α^2 or σ_β^2 respectively. This is in agreement with the result, equation (1), quoted above. Because this is an orthogonal design the magnitude of the "shrinkage" of the treatment effect does not depend on the exchangeability for the blocks, and vice versa. With a non-orthogonal design, such as balanced incomplete blocks, the same remark is not true.

3.2. *Exchangeability Between Multiple Regression Equations*

The following practical example stimulated our extension from the example of Section 1 to the general model, and we shall report on its use in Section 5.2. The context was educational measurement where variables x and y were related with the usual linear regression structure. However the values of the regression parameters depended on the school the student had attended. Novick (personal communication) suggested to us that improved estimates might be obtained for any one school by combining the data for all schools. This is just what the Bayes estimates do, and would seem to be appropriate whenever exchangeability *between* regressions (schools) is a sensible assumption. The mathematics for p regressor variables goes as follows.

Suppose

$$\mathbf{y}_j \sim N(\mathbf{X}_j \boldsymbol{\beta}_j, \mathbf{I}_{n_j} \sigma_j^2) \tag{21}$$

for $j = 1, 2, \ldots, m$ and $\boldsymbol{\beta}_j$ a vector of p parameters: that is m linear, multiple regressions on p variables. In the notation of the Theorem, $\mathbf{A}_1$, expressed in terms of submatrices, is diagonal with $\mathbf{X}_j$ as the jth diagonal submatrix; $\boldsymbol{\theta}_1^{\mathrm{T}}$ is $(\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_m^{\mathrm{T}})$ of mp elements. The exchangeability of the individual $\boldsymbol{\beta}_j$ added to normality gives us the second stage as

$$\boldsymbol{\beta}_j \sim N(\boldsymbol{\xi}, \boldsymbol{\Sigma}) \tag{22}$$

say. Here $\mathbf{A}_2$ is a matrix of order $mp \times p$, all of whose $p \times p$ submatrices are unit matrices, and $\boldsymbol{\theta}_2 = \boldsymbol{\xi}$. We shall suppose vague prior knowledge of $\boldsymbol{\xi}$ and use the special form of Corollary 2.

Simple calculations show that $(\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2)^{-1} = m^{-1}\boldsymbol{\Sigma}$ and then that

$$\mathbf{C}_2^{-1}\mathbf{A}_2(\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_2^{\mathrm{T}}\mathbf{C}_2^{-1}$$

is a matrix of order *mp all* of whose $p \times p$ submatrices are $m^{-1}\boldsymbol{\Sigma}^{-1}$. In the usual way $\mathbf{A}_1^{\mathrm{T}}\mathbf{C}_1^{-1}\mathbf{A}_1$, expressed in terms of submatrices, is diagonal with $\sigma_j^{-2}\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j$ as the jth diagonal submatrix. The equations for the Bayes estimates $\boldsymbol{\beta}_j^*$ are then found to be

$$\begin{pmatrix} \sigma_1^{-2}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1+\boldsymbol{\Sigma}^{-1} & \cdots & \mathbf{0} \\ \cdots & \sigma_2^{-2}\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2+\boldsymbol{\Sigma}^{-1} & \cdots \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \sigma_m^{-2}\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m+\boldsymbol{\Sigma}^{-1} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \\ \vdots \\ \boldsymbol{\beta}_m^* \end{pmatrix} - \boldsymbol{\Sigma}^{-1}\begin{pmatrix} \boldsymbol{\beta}_\cdot^* \\ \boldsymbol{\beta}_\cdot^* \\ \vdots \\ \boldsymbol{\beta}_\cdot^* \end{pmatrix} = \begin{pmatrix} \sigma_1^{-2}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} \\ \sigma_2^{-2}\mathbf{X}_2^{\mathrm{T}}\mathbf{y} \\ \vdots \\ \sigma_m^{-2}\mathbf{X}_m^{\mathrm{T}}\mathbf{y} \end{pmatrix}, \tag{23}$$

where $\boldsymbol{\beta}_\cdot^* = \sum \boldsymbol{\beta}_i^*/m$. These equations are easily solved for $\boldsymbol{\beta}_\cdot^*$ and then, in terms of $\boldsymbol{\beta}_\cdot^*$, the solution is

$$\boldsymbol{\beta}_j^* = (\sigma_j^{-2}\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j+\boldsymbol{\Sigma}^{-1})^{-1}(\sigma_j^{-2}\mathbf{X}_j^{\mathrm{T}}\mathbf{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_\cdot^*), \tag{24}$$

a compromise between the least-squares estimate and an average of the various estimates. The example of Section 1 is a special case with $p = 1$.

Noting that $\mathbf{D}_0^{-1}$, given in Corollary 2 (16), may, for this application, be written in the form,

$$\begin{pmatrix} \sigma_1^{-2}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1+\boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \vdots & \vdots \\ \mathbf{0} & \sigma_m^{-2}\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m+\boldsymbol{\Sigma}^{-1} \end{pmatrix} - m^{-1}\begin{pmatrix} \boldsymbol{\Sigma}^{-1} \\ \vdots \\ \boldsymbol{\Sigma}^{-1} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{pmatrix} (\boldsymbol{\Sigma}^{-1} \ldots \boldsymbol{\Sigma}^{-1})$$

and thus may be inverted by the matrix Lemma (10), we can obtain an explicit form for $\boldsymbol{\beta}_j^*$. After some algebra we obtain the weighted form of (24) with $\boldsymbol{\beta}_\cdot^*$ replaced by $\sum \mathbf{W}_i\hat{\boldsymbol{\beta}}_i$ where,

$$\mathbf{W}_i = \left[\sum_{j=1}^{m}(\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j\sigma_j^{-2}+\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j\sigma_j^{-2}\right]^{-1}(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\sigma_i^{-2}+\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\sigma_i^{-2}.$$

This shows explicitly how the information from the ith regression equation is combined with the information from all equations.

3.3. *Exchangeability Within Multiple Regression Equations*

In contrast to the last section suppose that we have a single multiple regression situation

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n\sigma^2). \tag{25}$$

In the educational context, the p regressor variables might be the results of p tests applied to students and the dependent variable, y, a measure of the students' performance after training. We are interested in the case where the individual regression coefficients in $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \beta_2, \ldots, \beta_p)$ are exchangeable. To achieve this it may be necessary to rescale the regressor variables: for example, to write (25) in correlation form in which the diagonal elements of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ are unity and the off-diagonals are the sample correlations. (Again we emphasize the point that this is an assumption and may not be appropriate). If the assumption is sensible then we may fit it into our general model by supposing

$$\beta_j \sim N(\xi, \sigma_\beta^2). \tag{26}$$

There are at least two useful possibilities: (i) to suppose vague prior knowledge for ξ (Corollary 2), (ii) to put $\xi = 0$, reflecting a feeling that the β_i are small.

In (i) simple but tedious calculations analogous to those of Section 3.2 show that

$$\boldsymbol{\beta}^* = \{\mathbf{I}_p + k(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{I}_p - p^{-1}\mathbf{J}_p)\}^{-1}\hat{\boldsymbol{\beta}}, \tag{27}$$

where $k = \sigma^2/\sigma_\beta^2$. Similar calculations in (ii), using only a two-stage model, give

$$\boldsymbol{\beta}^* = \{\mathbf{I}_p + k(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\}^{-1}\hat{\boldsymbol{\beta}}. \tag{28}$$

The estimates (27) and (28) are very similar to those proposed by Hoerl and Kennard (1970a). The main difference is that k in their argument is a constant introduced for various intuitively sensible reasons, whereas here it is a variance ratio. Also the derivation is different: Hoerl and Kennard argue within the orthodox, sampling theory framework, whereas we use the formal theory. We do not attempt to reproduce their most convincing argument against the least-squares estimates and in favour of (27) and (28), merely referring the sampling-theorist to it and saying that we agree with its conclusions with the reservation that we feel that the estimates may not be so sensible if the exchangeability within the regression equation is inappropriate. We return to this example in Section 5.3 where the estimation of k is discussed.

Examples 3.2 and 3.3 may be combined when there is exchangeability between *and* within regressions. We omit the details of this and many other extensions and instead consider how we might remove the major impediment to the application of the general theory, namely the assumption that all the variances are known. In the next section we show that the simple device of replacing the known variances by estimated values in the Bayes estimates is satisfactory.

4. Estimation with Unknown Covariance Structure

For the purpose of the immediate exposition denote by θ the parameters of interest in the general model and by ϕ the nuisance parameters. The latter will include the dispersion matrices $\mathbf{C}_i$ when these are unknown. Consider how the Bayesian treatment proceeds. We first assign a joint prior distribution to θ and ϕ—instead of just to θ—and combine this with the likelihood function to provide the joint posterior distribution $p(\theta, \phi \mid \mathbf{y})$. This distribution then has to be integrated with respect to ϕ, thus removing the nuisance parameters and leaving the posterior for θ. Finally, if we are using quadratic loss or generally one of the forms given by (3), we shall require the mean of this distribution, necessitating another integration. The calculation of the mean will also require the constant of proportionality in Bayes's formula to be evaluated, involving yet another integration. Any reasonable

prior distributions for ϕ that we have considered lead to integrals which cannot all be expressed in closed form and, as a result, the above argument is technically most complex to execute. We therefore consider an approximation to it which is technically much simpler and yet yields the bulk, though not unfortunately all, of the information required for the estimation.

The first approximation consists in using the *mode* of the posterior distribution in place of the *mean*. Secondly, we mostly use the mode of the *joint* distribution rather than that of the *θ-margin*. The modal values satisfy the equations

$$\frac{\partial}{\partial\theta}p(\theta,\phi|\mathbf{y}) = \frac{\partial}{\partial\phi}p(\theta,\phi|\mathbf{y}) = 0.$$

These equations may be re-written in terms of conditional and marginal distributions. In particular that for θ may be expressed as

$$\frac{\partial}{\partial\theta}p(\theta|\phi,\mathbf{y})p(\phi|\mathbf{y}) = 0$$

or, assuming $p(\phi|\mathbf{y})\neq 0$, as

$$\frac{\partial}{\partial\theta}p(\theta|\phi,\mathbf{y}) = 0. \tag{29}$$

But the conditional density $p(\theta|\phi,\mathbf{y})$ in (29) is exactly what has been found in the general theory of Section 2, where it was shown to be normal, with mode consequently equal to the mean. Hence we have the result that the θ-value of the posterior mode of the joint distribution of θ and ϕ is equal to the mode of the conditional distribution of θ evaluated at the modal value of ϕ. Consequently all we have to do is to take the estimates derived in Section 2 and replace the unknown values of the nuisance parameters by their modal estimates. For example, the simple estimate (1) is replaced by

$$\frac{y_i/s^2+y_./t^2}{1/s^2+1/t^2},$$

where s^2 and t^2 are respectively modal estimates of σ^2 and τ^2. This approach avoids the integrations referred to above. The modal estimates of ϕ may, analogous to (29), be found by supposing θ known, and then replacing θ in the result by their modes.

It is reasonably clear that the approximations are only likely to be good if the samples are fairly large and the resulting posterior distributions approximately normal. Also the approach does not provide information about the precision of the estimates, such as a standard error (of the posterior, not the sampling-theoretic distribution!) would provide. But as a first step on the way to a satisfactory description of the posterior distribution, it seems to go a long way and has the added merit of being intuitively sensible. In practice we shall find it convenient to proceed as follows. For an assumed ϕ calculate the mode $\theta^{(1)}$, say. Treating $\theta^{(1)}$ as known we can find the mode for ϕ, $\phi^{(1)}$ say. This may be used to find $\theta^{(2)}$, and so on. This sequence of iterations typically converges and only involves equations for the modes of one parameter, knowing the value of the other.

We now proceed to apply these ideas to the situations discussed in Section 3. At the moment we have no general theory to parallel that of Section 2. The reason for this is essentially that we do not have an entirely satisfactory procedure for

estimating the dispersion matrix of a multivariate normal distribution. This might appear an odd statement to make when there are numerous texts on multivariate analysis available that discuss this problem. But just as the usual estimates of the means are inadmissible, so are those of the variances and covariances (Brown, 1968), and are, in any case, obtained from unrealistic priors. We hope to report separately on this problem and defer discussion of a general theory.

5. Examples with Unknown Covariance Structure

5.1. *Two-factor Experimental Designs*

We saw in Section 3.1 that there were three variances in this situation: σ^2 the usual residual variance contributing to the likelihood function, and $\sigma_\alpha^2, \sigma_\beta^2$ being respectively the variances of the treatment and block effects. ($\sigma_\mu^2 \to \infty$ so does not enter.) It is first necessary to specify prior distributions for these and this we do through the appropriate conjugate family, which is here inverse-χ^2, assuming the three variances independent. This conjugate family involves two parameters and is sufficiently flexible for most applications. Specifically we suppose

$$\frac{\nu\lambda}{\sigma^2} \sim \chi_\nu^2, \quad \frac{\nu_\alpha \lambda_\alpha}{\sigma_\alpha^2} \sim \chi_{\nu_\alpha}^2 \quad \text{and} \quad \frac{\nu_\beta \lambda_\beta}{\sigma_\beta^2} \sim \chi_{\nu_\beta}^2. \tag{30}$$

The joint distribution of all quantities involved can then be written down as proportional to

$$\begin{aligned}
&(\sigma^2)^{-\frac{1}{2}(n+\nu+2)} \exp\left[-\frac{1}{2\sigma^2}\{\nu\lambda + S^2(\mu,\alpha,\beta)\}\right] \\
&\times (\sigma_\alpha^2)^{-\frac{1}{2}(t+\nu_\alpha+2)} \exp\left[-\frac{1}{2\sigma_\alpha^2}\{\nu_\alpha\lambda_\alpha + \textstyle\sum \alpha_i^2\}\right] \\
&\times (\sigma_\beta^2)^{-\frac{1}{2}(b+\nu_\beta+2)} \exp\left[-\frac{1}{2\sigma_\beta^2}\{\nu_\beta\lambda_\beta + \textstyle\sum \beta_j^2\}\right],
\end{aligned} \tag{31}$$

where $S^2(\mu,\alpha,\beta)$ is the sum of squares $\sum(y_{ij}-\mu-\alpha_i-\beta_j)^2$.

If σ^2, σ_α^2 and σ_β^2 are known, the mode of this distribution has been found—equation (18), or in the balanced case, equation (20). We have only to substitute the modal estimates of the three variances into these expressions. To find these modal estimates we can, reversing the roles of θ and ϕ in the general argument of the previous paragraph, suppose μ, α and β known. Using the corresponding Roman letters for these modes, we easily see them to be, from (31),

$$\left.\begin{aligned}
s^2 &= \{\nu\lambda + S^2(\mu^*,\alpha^*,\beta^*)\}/(n+\nu+2), \\
s_\alpha^2 &= \{\nu_\alpha\lambda_\alpha + \textstyle\sum \alpha_i^{*2}\}/(t+\nu_\alpha+2), \\
s_\beta^2 &= \{\nu_\beta\lambda_\beta + \textstyle\sum \beta_j^{*2}\}/(b+\nu_\beta+2).
\end{aligned}\right\} \tag{32}$$

These equations, together with (18) (or (20)), can now be solved iteratively. With trial values of σ^2, σ_α^2 and σ_β^2, (18) can be solved for μ^*, α^* and β^*. These values can be inserted into (32) to give revised values for s^2, s_α^2 and s_β^2, which can again be used in (18). The cycle can be repeated until the values converge.

A few points about these solutions are worth noting. Firstly, the value of S^2 that occurs is not the usual residual sum of squares, which is evaluated about

the least-squares value, but the sum about the Bayes estimates. Since the former minimizes the sum of squares, our S^2 is necessarily greater than the residual: s^2 could therefore be larger than the usual estimate. Secondly, whilst it would be perfectly possible to put $\nu = 0$ (referring to σ^2), so avoiding the specification of a value for λ and thereby taking the usual vague prior for a variance, one cannot put ν_α and ν_β zero. If this is done the modal estimates for the treatment and block effects are all zero. The point is discussed in detail in connection with the example of Section 1 in Lindley (1971a). Essentially the estimation of σ_α^2 and σ_β^2 is difficult, in the sense that the data contain little information about them, when they are small in comparison with σ^2: the residual "noise" is too loud. In the contrary case where σ_α^2 and σ_β^2 are large in comparison with σ^2, the actual values of ν_α, λ_α, ν_β and λ_β do not matter much provided the ν's are both small.

5.2. *Exchangeability Between Multiple Regression Equations*

We continue the discussion of Section 3.2 but mainly confine our attention to the homoscedastic case where $\sigma_j^2 = \sigma^2$, say, for all j. It is only necessary to specify prior distributions for σ^2 and $\boldsymbol{\Sigma}$, the dispersion matrix of the regression coefficients (equation (22)). As in the last example we suppose $\nu\lambda/\sigma^2 \sim \chi_\nu^2$. The conjugate distribution for $\boldsymbol{\Sigma}$ is to suppose $\boldsymbol{\Sigma}^{-1}$ has a Wishart distribution with ρ, say, degrees of freedom and matrix $\mathbf{R}$. We are not too happy with this assumption but at least it provides a large-sample solution (see the remarks at the end of Section 4). $\boldsymbol{\Sigma}$ and σ^2 are supposed independent.

The joint distribution of all the quantities is now

$$\begin{aligned}
&(\sigma^2)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)\right\} \\
&\times |\boldsymbol{\Sigma}|^{-\frac{1}{2}m} \exp\left\{-\frac{1}{2}\sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\xi})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\xi})\right\} \\
&\times |\boldsymbol{\Sigma}|^{-\frac{1}{2}(\rho-p-1)} \exp\{-\tfrac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\mathbf{R}\} \\
&\times (\sigma^2)^{-\frac{1}{2}(\nu+2)} \exp\{-\nu\lambda/2\sigma^2\}, \qquad (33)
\end{aligned}$$

assuming $\boldsymbol{\xi}$ to have a uniform distribution over p-space. (The four lines of (33) come respectively from the likelihood, the distribution of $\boldsymbol{\beta}$, (22), the Wishart distribution for $\boldsymbol{\Sigma}^{-1}$ and the inverse-χ^2 for σ^2.) The integration with respect to $\boldsymbol{\xi}$ is straightforward and effectively results in the usual loss of one degree of freedom. The joint posterior density for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Sigma}^{-1}$ is then proportional to

$$\begin{aligned}
&(\sigma^2)^{-\frac{1}{2}(n+\nu+2)} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{m}\{m^{-1}\nu\lambda + (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)\}\right] \\
&\times |\boldsymbol{\Sigma}|^{-\frac{1}{2}(m+\rho-p-2)} \exp\left[-\tfrac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\left\{\mathbf{R} + \sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\cdot)(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\cdot)^{\mathrm{T}}\right\}\right], \qquad (34)
\end{aligned}$$

where

$$\boldsymbol{\beta}_\cdot = m^{-1}\sum_{j=1}^{m}\boldsymbol{\beta}_j.$$

The modal estimates are then easily obtained. Those for $\boldsymbol{\beta}_j$ are as before, equation (24), with $\boldsymbol{\Sigma}$ and σ^2 replaced by modal values. The latter are seen to satisfy

$$s^2 = \sum_{j=1}^{m} \{m^{-1}\nu\lambda + (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_1^*)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j^*)\}/(n+\nu+2),$$

and

$$\boldsymbol{\Sigma}^* = \left\{\mathbf{R} + \sum_{j=1}^{m}(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_\cdot^*)(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_\cdot^*)^{\mathrm{T}}\right\} \Big/ (m+\rho-p-2). \tag{35}$$

It is possible in this case, as in Section 5.1 and 5.3, to proceed a little differently and obtain the posterior distribution of the $\boldsymbol{\beta}_j$'s, free of σ^2 and $\boldsymbol{\Sigma}$ and consider the modes of this. This is because the integration of (34) with respect to $\boldsymbol{\Sigma}^{-1}$ and σ^2 is possible in closed form. The result is

$$\left[\sum_{j=1}^{n}\{m^{-1}\nu\lambda + (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)\}\right]^{-\frac{1}{2}(n+\nu)}$$

$$\times \left|\mathbf{R} + \sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\cdot)(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\cdot)^{\mathrm{T}}\right|^{-\frac{1}{2}(m+\rho-1)}. \tag{36}$$

The mode of this distribution can be used in place of the modal values for the wider distribution. The differentiation is facilitated by using the result that, with $\mathbf{V}$ equal to the matrix whose determinant appears in (36),

$$\frac{\partial}{\partial\boldsymbol{\beta}_i}\log|\mathbf{V}| = 2\mathbf{V}^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_\cdot).$$

It is then possible to verify that the modes for $\boldsymbol{\beta}_j$ satisfy the same equations as before, (24), with $\boldsymbol{\Sigma}$ and σ^2 replaced by values given by (35) except that the divisors on the right-hand sides are $(n+\nu)$ and $(m+\rho-1)$ rather than $(n+\nu+2)$ and $(m+\rho-p-2)$.

It is possible to extend this model significantly by reverting to the heteroscedastic case as originally considered, (21). Here we have to specify a joint distribution for the σ_j^2. A possible device is to suppose that, like the means, the σ_j^2 are exchangeable. A convenient distribution to generate the exchangeability is to suppose $\nu\lambda/\sigma_j^2 \sim \chi_\nu^2$. In the context of several means (Section 1) Lindley (1971a) has shown how the estimates of the variances get pulled towards a central value. The details are so similar here that we do not repeat them.

As explained in Section 3.2, it was Novick's suggestion to consider this problem in an educational context, and we conclude this section by briefly reporting on an application that he, in conjunction with Jackson, Thayer and Cole (1972), have made of these results. We are most grateful to them for permission to include the details here. Their analysis used data from the American College Testing Program on the prediction of grade-point average at 22 colleges from the results of 4 tests; namely, English, Mathematics, Social Studies and Natural Sciences. We therefore have the situation studied in this section with $p = 5$ (one variable corresponding to the mean), $m = 22$, and n_j varying from 105 to 739. They used the heterosecdastic model but the basic equations (24) and (35) are essentially as here described. With the substantial amounts of data available the prior constants, ν, λ, ρ and $\mathbf{R}$ scarcely affect the analysis: the first three were taken to be small and changes of origin of the regressor variables effected to make the prior judgment that $\mathbf{R}$ was diagonal. With ρ small the diagonal elements again play little role. We omit details of how the calculations were performed and refer the interested reader to their paper.

Data were available for 1968 and 1969. The approach was to use the 1968 data to estimate the regressions, to use these estimated equations on the 1969 x-data to estimate the y's, the grade-point averages, and then to compare these predictions with the actual 1969 y-values, using as a criterion of prediction the mean of the squares of the differences. This operation was done twice; once with the full 1968 data, and once with a random 25 per cent sample from each College. The results are summarized in Table 1.

TABLE 1

Comparison of predictive efficiency

	Average mean-square error	
	Least-squares	*Bayes*
All data	0·5596	0·5502
25% sample	0·6208	0·5603

The first row refers to the analysis of the whole data and shows that the Bayesian method only reduces the error by under 2 per cent. With such large samples there is little room for improvement. With the quarter sample, however, in the second row of the table, the reduction is up to 9 per cent and most strikingly the error is almost down to the value reached with the least-squares estimates for all the data. In other words, 25 per cent of the data and Bayes are as good as all the data and least squares: or the Bayesian method provides a possible 75 per cent saving in sample size. They also provide details of the comparisons between the two estimates of the regression coefficients. These tend to be "shrunk" towards a common value (for each regressor variable) and in some cases with the quarter sample the shrinkage is substantial.

It would be dangerous to draw strong conclusions from one numerical study but the analysis should do something to answer the criticism of those who have said that Bayesian methods are not "testable". We favour the method because of its coherence, but the pragmatists may like to extend the method of Novick *et al.* to other data sets, remembering, of course, that we have made an assumption of exchangeability, and the method cannot be expected to work when this is unreasonable.

5.3. *Exchangeability Within Multiple Regression Equations*

In this section we briefly indicate how the analysis of Section 3.3 proceeds when σ^2, the residual regression variance, and σ_β^2, the variance of the regression coefficients, are both unknown. As before, we assume that independently

$$\nu\lambda/\sigma^2 \sim \chi^2_\nu, \quad \nu_\beta \lambda_\beta/\sigma_\beta^2 \sim \chi^2_{\nu_\beta}.$$

As in Section 5.2 the integration with respect to ξ, the mean of the β_j's, may be performed and the result is that the posterior distribution of $\boldsymbol{\beta}$, σ^2 and σ_β^2 is proportional to

$$(\sigma^2)^{-\frac{1}{2}(n+\nu+2)} \exp\left[-\frac{1}{2\sigma^2}\{\nu\lambda+(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}\right]$$

$$\times (\sigma_\beta^2)^{-\frac{1}{2}(p+\nu_\beta+1)} \exp\left[-\frac{1}{2\sigma_\beta^2}\left\{\nu_\beta\lambda_\beta+\sum_{j=1}^{p}(\beta_j-\beta_\cdot)^2\right\}\right], \qquad (37)$$

where

$$\beta_{\cdot} = p^{-1}\sum_{j=1}^{p}\beta_j.$$

The modal equations are then easily seen to be (the first coming from (23))

$$\left.\begin{aligned}\boldsymbol{\beta}^* &= \{\mathbf{I}_p + k^*(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{I}_p - p^{-1}\mathbf{J}_p)\}^{-1}\hat{\boldsymbol{\beta}},\\ s^2 &= \{\nu\lambda + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*)^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*)\}/(n+\nu+2),\\ s_\beta^2 &= \left\{\nu_\beta\lambda_\beta + \sum_{j=1}^{p}(\beta_j^*-\beta_{\cdot}^*)^2\right\}\Big/(p+\nu_\beta+1).\end{aligned}\right\} \tag{38}$$

The value of k^* is of course s^2/s_β^2. The marginal posterior distribution of $\boldsymbol{\beta}$ can be obtained in a manner similar to that described in the last section.

We are now in a position to compare our method with that of Hoerl and Kennard (1970b). We have taken the example of a 10-factor, non-orthogonal multiple regression summarized in Gorman and Toman (1966) and re-analysed by Hoerl and Kennard using their ridge regression method. The results are summarized in Table 2.

TABLE 2

10-factor multiple regression example

Estimate	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	k
Least-squares	−0·185	−0·221	−0·359	−0·105	−0·469	0·813	0·285	0·383	0·092	0·094	0·000
Bayes	−0·256	−0·178	−0·326	−0·086	−0·289	0·592	0·195	0·349	0·117	0·116	0·039
Ridge	−0·295	−0·110	−0·245	−0·050	−0·040	0·325	0·050	0·240	0·125	0·125	0·250

As already explained, the main difference between ridge regression and the Bayes approach lies in the choice of $k(=\sigma^2/\sigma_\beta^2)$ in equation (23). This has the value zero for least-squares, is chosen subjectively in the ridge method by selecting it so large that the regression estimates stabilize, and is estimated from the data in the Bayes method. In applying the Bayes method we started with $k^* = 0$ in (38), obtained estimates $\boldsymbol{\beta}^*$, which were then used in the other equations in (38) to obtain s^2 and s_β^2. It was found that 10 iterations were needed until the cycle converged. The solution is fairly insensitive to changes in the small, positive values of ν and ν_β and these were set to zero.

In the case of non-orthogonal data, the least-squares procedure has a tendency to produce regression estimates which are too large in absolute value, of incorrect sign and unstable with respect to small changes in the data. The ridge method attempts to avoid some of these undesirable features. The Bayesian method reaches the same conclusion but has the added advantage of dispensing with the rather arbitrary choice of k and allows the data to estimate it. It will be seen from Table 2 that except for β_1, β_9 and β_{10}, all the estimates are pulled towards zero, the effect being greater with the ridge method than with Bayes, the latter choosing a considerably larger value of k than the data suggest.

ACKNOWLEDGEMENTS

We are very grateful to Melvin R. Novick who first stimulated our interest in these problems, has continually made fruitful suggestions and, with his colleagues, has allowed us to include the example in Section 5.2. We are indebted to the referees for constructive comment on a first draft of this paper. The second author would like to thank the Science Research Council and the Central Electricity Generating Board for financial support during the course of this research.

REFERENCES

BROWN, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.*, **37**, 1087–1136.

—— (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Ann. Math. Statist.*, **39**, 29–48.

CORNFIELD, J. (1969). The Bayesian outlook and its application. *Biometrics*, **25**, 617–657.

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

FISK, P. R. (1967). Models of the second kind in regression analysis. *J. R. Statist. Soc.*, B, **29**, 266–281.

GORMAN, J. W. and TOMAN, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics*, **8**, 27–51.

HEWITT, E. and SAVAGE, L. J. (1955). Symmetric measures on cartesian products. *Trans. Amer. Math. Soc.*, **80**, 470–501.

HOERL, A. E. and KENNARD, R. W. (1970a). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

—— (1970b). Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.

LINDLEY, D. V. (1969). Bayesian least squares. *Bull. Inst. Internat. Statist.*, **43**(2), 152–153.

—— (1971a). The estimation of many parameters. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 435–455. Toronto: Holt, Rinehart and Winston.

—— (1971b). *Bayesian Statistics, A Review*. Philadelphia: SIAM.

NELDER, J. A. (1968). Regression, model-building and invariance. *J. R. Statist. Soc.*, A, **131**, 303–315.

NOVICK, M. R., JACKSON, P. H., THAYER, D. T. and COLE, N. S. (1972). Estimating multiple regressions in *m*-groups; a cross-validation study. *Brit. J. Math. Statist. Psychology*, (to appear).

PLACKETT, R. L. (1960). *Principles of Regression Analysis*. Oxford: Clarendon Press.

RAO, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: Wiley.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Sympos.*, **1**, 197–206. Berkeley: University of California Press.

DISCUSSION ON THE PAPER BY PROFESSOR LINDLEY AND DR SMITH

Dr J. A. NELDER (Rothamsted Experimental Station): I welcome this paper as shedding interesting new light on an old topic, the linear model with normal errors. The authors develop a thoroughly Bayesian argument, but this does not mean that we have to be Bayesians in order to make use of their ideas, as I shall try to show.

The authors have laid great stress on their estimates being *Bayesian* estimates, and have compared their estimates with least-squares estimates which they interpret only within a sampling-theoretic framework. Essentially what they are doing is incorporating in their model extra information about a set of (say) population means to the effect that they can be taken as a random sample from a "hyper-population". Such situations undoubtedly occur: e.g. the authors' own example about a random subset of the many lines produced in a breeding program. The same notion underlies designed experiments in incomplete blocks, where the blocks of the design are allocated at random to the finite population in the field. Here the Lindley-Smith estimates of the treatment effects when the block effects are assumed exchangeable but nothing is assumed about the treatments are the familiar