TECHNOMETRICS©, VOL. 17, NO. 1, FEBRUARY 1975

A Bayesian Approach to the Linear Model with Unequal Variances

Tom Leonard

Department of Statistics,
University of Warwick,
Coventry CV4 7AL, England

Estimation procedures are proposed for the means and variances of several normal distributions. They are valid in cases where the parameters of different distributions may be unequal but are thought a priori to be related in certain ways. A log-linear model is assumed for the variances, together with a linear model for the means. Some general Bayesian results are obtained, and some special cases are discussed including the important situation where parameters of different distributions are a priori exchangeable. The posterior estimates then adjust the standard estimates, by shrinking them towards central values based on collateral information. A numerical example concerns the simultaneous estimation of the variances of the observed breaking strengths of six fabrics.

KEY WORDS

Linear models
Heteroscedastic variances
Log-linear models
Multivariate normal prior distributions
Posterior means and modes
Exchangeability
Shrinkage of estimates
Partial pooling processes
Bartlett's test
Regression models
Two-way Analysis of Variance
Autoregressive Processes

1. DISCUSSION OF THE MODEL

Attention is confined to the linear model with possibly unequal variances, where the observations x_{ij} are arranged in m populations and may be unequally replicated. Given θ_i and ϕ_i $(i = 1, \cdots, m)$ we take the x_{ij} to be independent and normally distributed, with

$$E(x_{ij} \mid \theta_i, \phi_i) = \theta_i$$

and

$$\text{var}\,(x_{ij} \mid \theta_i, \phi_i) = \phi_i \; (i = 1, \cdots, m; j = 1, \cdots, n_i)$$

where θ_i and ϕ_i respectively denote the mean and variance for the ith population.

In [7] Lindley provides a method for the simultaneous estimation of the population means θ_i under the assumption that they are a priori *exchangeable* i.e. the joint distribution of any subset of $\{\theta_1, \cdots, \theta_m\}$ is invariant under a permutation of the suffices. The resultant posterior estimates shrink the sample means $x_{1.}, \cdots, x_{m.}$ towards a central value based on collateral information. There is therefore a partial pooling process where information about all the θ_i is used to improve the estimates for each individual θ_i. The estimates bear resemblances to those proposed by James and Stein [2], under a classical approach, in proving *inadmissibility* for $m \geq 3$ of the standard estimates $x_{1.}, \cdots, x_{m.}$ with respect to a quadratic loss function. The results in [7] are generalised in [8] to situations where more general relationships are thought a priori to exist between the θ_i.

In [7] the situation is also discussed where the ϕ_i, as well as the θ_i, are a priori exchangeable. In constructing a suitable exchangeable distribution, a common inverse chi-squared distribution is assumed for the ϕ_i at the first-stage of a two-stage prior model. Owing to technical difficulties, this method does not appear to show promise of capability of generalisation to situations where more complex relationships are thought to exist between the ϕ_i e.g. they may be thought to be related in an ordered fashion, or to depend upon some explanatory variables.

We will provide a solution to this problem under a general formulation which will permit such relationships between the ϕ_i. As a special case we will provide new results in the situation where the ϕ_i are exchangeable. We make no claims of superiority of our own estimates in the simple exchangeable situation, although it is our personal opinion that the results are simpler and easier to interpret.

In [8] the following linear model is assumed for the mean vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$:

$$\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\beta} \qquad (1)$$

Received December, 1973; revised September 1974

where $\mathbf{A}$ is a full rank $n \times p$ matrix of known constants, and β is a $p \times 1$ vector of unknown parameters. The authors mainly restrict themselves to cases where the ϕ_i are either assumed equal or considered separately. They mention that it is possible to use the ideas in [7] to generalise their results to situations where the ϕ_i are unequal, but exchangeable. Under our approach it will be possible to permit the elements of $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)^T$ to be related in a general manner.

In [8] a hierarchical prior structure is proposed for the vector β. This employs multivariate normal distributions at each stage of the hierarchy. If the precision matrices of these distributions are assumed known, then the distribution of β, obtained by combining the various stages in the prior model, is multivariate normal. We assume in general that, given $\boldsymbol{\mu}_\beta$ and $\mathbf{H}_\beta$, the distribution of β is multivariate normal, with mean vector $\boldsymbol{\mu}_\beta$, and precision matrix $\mathbf{H}_\beta$. When $\mathbf{H}_\beta$ is of full rank, the inverse $\mathbf{H}_\beta^{-1}$ provides the prior convariance matrix of β.

On p. 114 of [6], the conditional posterior distribution of β, given $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)^T$, $\boldsymbol{\mu}_\beta$, and $\mathbf{H}_\beta$ is shown to be multivariate normal, with mean vector

$$\tilde{\beta} = E(\beta \mid \mathbf{x}, \boldsymbol{\phi}, \boldsymbol{\mu}_\beta, \mathbf{H}_\beta) = (\mathbf{A}^T\mathbf{R}\mathbf{A} + \mathbf{H}_\beta)^{-1}(\mathbf{A}^T\mathbf{R}\mathbf{Z} + \mathbf{H}_\beta\boldsymbol{\mu}_\beta) \quad (2)$$

and precision matrix $\mathbf{A}^T\mathbf{R}\mathbf{A} + \mathbf{H}_\beta$ where

$$\mathbf{Z} = (x_{1.}, \cdots, x_{m.})^T \quad (3)$$

and

$$\mathbf{R} = \text{diag}\,(n_1\phi_1^{-1}, \cdots, n_m\phi_m^{-1}) \quad (4)$$

The expression in (2) adjusts the weighted least squares vector

$$\hat{\beta} = (\mathbf{A}^T\mathbf{R}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{R}\mathbf{Z} \quad (5)$$

by taking into account the prior information about β, as represented by $\boldsymbol{\mu}_\beta$ and $\mathbf{H}_\beta$.

We would like to make assumptions about $\boldsymbol{\phi}$ of a similar nature to those previously made about $\boldsymbol{\theta}$, since these would enable us to take into account prior information about relationships between the ϕ_i. We are able to do this by considering the logarithmic transformations

$$\alpha_i = \log \phi_i \qquad (i = 1, \cdots, m) \quad (6)$$

and assuming a linear model for $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_m)^T$. We suppose that

$$\boldsymbol{\alpha} = \mathbf{B}\gamma \quad (7)$$

where $\mathbf{B}$ is a full rank $n \times q$ matrix of known constants, and γ is a $q \times 1$ vector of unknown parameters.

We suppose that γ is a priori independent of β, and that, given $\boldsymbol{\mu}_\gamma$ and $\mathbf{H}_\gamma$, the distribution of γ in the prior assessment is multivariate normal, with mean vector $\boldsymbol{\mu}_\gamma$, and precision matrix $\mathbf{H}_\gamma$. Under this formulation it is very easy to allow for prior information about relationships between the log-variances, and hence between the variances. Our assumptions are much more flexible than those made in [7] using inverse chi-squared priors for the variances. It will be possible to add a suitable second stage to the prior model in some cases, by assigning distributions to the elements of $\boldsymbol{\mu}_\gamma$ and $\mathbf{H}_\gamma$, so that specific values need not be chosen for these elements.

The general idea of seeking suitable transformations of sets of the unknown parameters such that the new parameters may be considered to be a priori normally distributed is suggested by us in the discussion of [8]. In [3] we apply these ideas to the estimation of several binomial parameters, using logistic transformations, and the present paper provides another particular case.

In the next section we discuss the estimation of γ when β is known, and will later generalise our analysis to the situation where both β and γ are unknown.

2. Population Means Known

The arguments in the present section are only intended to hold conditionally on $\boldsymbol{\theta} = \mathbf{A}\beta$ being known. It is now well-known that the sums of squares

$$S_i(\theta_i) = \sum_{j=1}^{n_i} (x_{ij} - \theta_i)^2 \qquad (i = 1, \cdots, m) \quad (8)$$

comprise a set of jointly sufficient statistics for the ϕ_i. These expressions may be rearranged in the forms

$$S_i(\theta_i) = S_i^w + n_i(\theta_i - x_{i.})^2 \quad (9)$$

where

$$S_i^w = \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2 \quad (10)$$

We note that the expression in (9) is never less than that in (10). We now state, without proof, two standard results about the sampling distributions of the $S_i(\theta_i)$. These are

(i) Given the θ_i and ϕ_i, the quantities
$$\phi_1^{-1}S_1(\theta_1), \cdots, \phi_m^{-1}S_m(\theta_m)$$
are mutually independent, and possess chi-squared distributions on $n_1, \cdots, n_m$ degrees of freedom respectively.

(ii) Consider the vector $\mathbf{l} = (l_1, \cdots, l_m)^T$, and the matrix $\mathbf{U}$, such that
$$l_i = \log\{S_i(\theta_i)/n_i\} \qquad (i = 1, \cdots, m) \quad (11)$$
and

$$\mathbf{U} = \text{diag}\ (\tfrac{1}{2}n_1\ ,\ \cdots\ ,\ \tfrac{1}{2}n_m) \qquad (12)$$

Then unless any of the n_i are small, the distribution of $\mathbf{l}$, conditional on $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, is approximately multivariate normal with mean vector $\boldsymbol{\alpha}$ and precision matrix $\mathbf{U}$.

The second result follows from the first, together with the standard result that the log of a chi-squared variate on ν degrees of freedom is approximately normally distributed with mean log ν and variance $2\nu^{-1}$. Under our approximations, we notice that we may carry out the analysis for $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha} = B\boldsymbol{\gamma}$, conditionally on $\boldsymbol{\theta}$ being known, by analogy with the previous analysis for $\boldsymbol{\beta}$, and $\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\beta}$, conditional on $\boldsymbol{\phi}$ being known. We merely have to replace $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\mathbf{A}$, $\mathbf{Z}$, $\mathbf{R}$, $\boldsymbol{\mu}_\beta$, and $\mathbf{H}_\beta$, by $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\mathbf{B}$, $\mathbf{l}$, $\mathbf{U}$, $\boldsymbol{\mu}_\gamma$, and $\mathbf{H}_\gamma$ respectively.

The adequacy in the present Bayesian context of the above normal approximations to the distributions of the logs of the chi-squared variates should in principle be considered by examining the accuracy of the resultant approximation to the posterior distribution in regions of the parameter space to which it assigns high probability. However in Ch. 7 of an unpublished thesis [4] we indicate an exact approach for the resultant estimates, which is omitted here as it is rather complicated. Algebraic comparisons with the approximate estimates given below suggest to us that the latter should be adequate as long as they are not over-radically different from the maximum likelihood estimates i.e. very close to the prior means, in which case the corresponding exact estimates will be more conservative. We guess that for many choices of the prior distribution, these approximations may be reasonable whenever none of the n_i are less than about 5, though they might sometimes be adequate when some of the n_i are smaller. Further work would be necessary if we wished to examine the accuracy of the approximation to the whole posterior distribution. We of course prefer the approximations rather than the exact method, for reasons of simplicity.

Whenever the above approximations hold, we have, by analogy with (2), and using the substitutions described above, that the conditional posterior mean vector of $\boldsymbol{\gamma}$, given $\boldsymbol{\theta}$, $\boldsymbol{\mu}_\gamma$, and $\mathbf{H}_\gamma$ is approximated by

$$\tilde{\boldsymbol{\gamma}} = E(\boldsymbol{\gamma} \mid \mathbf{x},\ \boldsymbol{\theta},\ \boldsymbol{\mu}_\gamma\ ,\ \mathbf{H}_\gamma)$$

$$= (\mathbf{B}^T\mathbf{U}\mathbf{B} + \mathbf{H}_\gamma)^{-1}(\mathbf{B}^T\mathbf{U}\mathbf{l} + \mathbf{H}_\gamma\boldsymbol{\mu}_\gamma) \qquad (13)$$

where $\mathbf{l}$ has elements in (11) and $\mathbf{U}$ is given in (12).

If some of the non-diagonal elements of $\mathbf{H}_\gamma$ are non-zero then our estimate for $\boldsymbol{\gamma}$ will take account of prior relationships between the elements of $\boldsymbol{\gamma}$ or $\boldsymbol{\alpha}$. We suggest estimating the vector $\boldsymbol{\phi}$ of variances by

$$\tilde{\boldsymbol{\phi}} = (e^{\tilde{\alpha}_1},\ \cdots\ ,\ e^{\tilde{\alpha}_m})^T \qquad (14)$$

where $\tilde{\alpha}_i$ is the ith element of $\mathbf{B}\tilde{\boldsymbol{\gamma}}$. This is not the mean vector of $\boldsymbol{\phi}$, but it should still provide us with reasonable estimates.

3. Population Means and Variances Unknown

When $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (and hence $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$) are unknown it appears virtually impossible to obtain their unconditional posterior mean vectors, since the joint posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is rather complicated, and is not multivariate normal. Instead we propose an alternative method of estimation, which involves some simple iterations. We find approximations to the joint posterior *mode* vectors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ i.e. those vectors maximising the joint posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

In [8] posterior modes are also employed, and we cite an important general result on page 12 of this paper, which enables us to obtain joint modes by considering the conditional modes. The result may be paraphrased to the present situation by saying that the joint posterior mode vectors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are given by vectors $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ satisfying

$$\tilde{\boldsymbol{\beta}} = \mathbf{b}(\tilde{\boldsymbol{\gamma}})$$

and

$$\tilde{\boldsymbol{\gamma}} = \mathbf{c}(\tilde{\boldsymbol{\beta}})$$

where $\mathbf{b}(\boldsymbol{\gamma})$ is the conditional posterior mode vector of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$, and $\mathbf{c}(\boldsymbol{\beta})$ is the corresponding vector of $\boldsymbol{\gamma}$, given $\boldsymbol{\beta}$.

We use this in conjunction with the useful and commonly known result that for a multivariate normal distribution the mode vector is identical to the mean vector.

Since the conditional posterior distribution of $\boldsymbol{\beta}$, given $\boldsymbol{\gamma}$, is multivariate normal, the conditional posterior mode vector of $\boldsymbol{\beta}$, given $\boldsymbol{\gamma}$, may be obtained from the mean vector in (2) upon replacing the ϕ_i in the expression for $\mathbf{R}$ in (4) by the exponentials of the corresponding elements of $\mathbf{B}\boldsymbol{\gamma}$. The conditional posterior mode vector of $\boldsymbol{\gamma}$, given $\boldsymbol{\beta}$, may be approximated by the mean vector in (13) upon replacing the θ_i in the expressions for the elements of $\mathbf{l}$ in (11) by the corresponding elements of $\mathbf{A}\boldsymbol{\beta}$.

As a consequence of the above-cited result in [8], we therefore have that the joint posterior mode vectors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are approximated by the solutions for $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ to the equations

$$\tilde{\boldsymbol{\beta}} = (\mathbf{A}^T\tilde{\mathbf{R}}\mathbf{A} + \mathbf{H}_\beta)^{-1}(\mathbf{A}^T\tilde{\mathbf{R}}\mathbf{Z} + \mathbf{H}_\beta\boldsymbol{\mu}_\beta) \qquad (15)$$

and

$$\tilde{\boldsymbol{\gamma}} = (\mathbf{B}^T\mathbf{U}\mathbf{B} + \mathbf{H}_\gamma)^{-1}(\mathbf{B}^T\mathbf{U}\tilde{\mathbf{l}} + \mathbf{H}_\gamma\boldsymbol{\mu}_\gamma) \qquad (16)$$

where

$$\tilde{\mathbf{R}} = \text{diag}\,(n_1\tilde{\phi}_1^{-1}, \cdots, n_m\tilde{\phi}_m^{-1}) \quad (17)$$

and the ith element of $\tilde{\mathbf{l}}$ is denoted by

$$\tilde{l}_i = \log\{S_i(\tilde{\theta}_i)/n_i\} \quad (18)$$

where

$$S_i(\tilde{\theta}_i) = S_i^w + n_i(\tilde{\theta}_i - x_{i.})^2 \quad (19)$$

with

$$\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_m)^T = \mathbf{A}\tilde{\boldsymbol{\xi}} \quad (20)$$

$$\tilde{\phi}_i = e^{\tilde{\alpha}_i} \quad (21)$$

and

$$\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m)^T = \mathbf{B}\tilde{\boldsymbol{\gamma}} \quad (22)$$

where $\mathbf{Z}$, S_i^w, and $\mathbf{U}$ are given in (3), (10), and (12) respectively.

A simple substitution procedure for the solution of the above equations is described as follows:

(a) Use the elements $x_{i.}$ of $\mathbf{Z}$ as initial values for the corresponding elements of $\tilde{\boldsymbol{\theta}}$ and use them to calculate values for the $S_i(\tilde{\theta}_i)$ from (19), and hence for the $\tilde{l}_i$ from (18), the elements of $\tilde{\boldsymbol{\gamma}}$ from (16), and $\tilde{\boldsymbol{\alpha}}$ from (22).

(b) Use the latest values for the $\tilde{\alpha}_i$ to calculate values for the $\tilde{\phi}_i$ from (21), $\tilde{\mathbf{R}}$ from (17), and hence for the elements of $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\theta}}$ from (15) and (20) respectively.

(c) Return to (a), using the new values for the elements of $\tilde{\boldsymbol{\theta}}$ instead of the old values, and keep cycling until convergence.

The above procedure is extremely simple and unless the matrices are of high dimension it will converge in a few seconds of computer time. We hope to publish an algorithm at a future date.

4. Exchangeability of the Means and Variances

In [7] the situation is treated where the relationships between the θ_i are of a symmetric nature, so that these parameters are exchangeable. In this case there should be exactly the same prior information about each θ_i, and also about all subsets of $\{\theta_1, \cdots, \theta_m\}$ which are of the same size. This would for example be satisfied if the statistician were completely ignorant about the θ_i and the relationships between them, in which case he would possess a symmetric lack of knowledge about them.

In [7] a two-stage prior model is used for the θ_i. At the first stage the θ_i are taken to possess the probability structure of a random sample from a normal distribution with mean u_θ and variance σ_θ^2. Further distributions are then chosen for the first-stage parameters u_θ and σ_θ^2. In the special case where σ_θ^2 is known, the mean u_θ is integrated out to show that the joint distribution of the θ_i is given by

$$\pi(\tilde{\theta} \mid \sigma_\theta^2) \propto \exp\{-\tfrac{1}{2}\sigma_\theta^{-2}\sum_i(\theta_i - \theta_.)^2\} \quad (23)$$

With some algebraic manipulation, it is straightforward to show that this provides a special case of the formulation in section 1, but with $\mathbf{A} = \mathbf{I}_m$, $\boldsymbol{\mu}_\beta$ equal to the vector of zeros, and $\mathbf{H}_\beta = \sigma_\theta^{-2}(\mathbf{I}_m - m^{-1}\mathbf{J}_m)$. Here $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, and $\mathbf{J}_m$ denotes the $m \times m$ matrix, every element of which is unity.

The distribution in (23) is used in [7] to show that the conditional posterior mean vector of $\boldsymbol{\theta}$, given $\boldsymbol{\phi}$ and σ_θ^2, possesses elements $\tilde{\theta}_1, \cdots, \tilde{\theta}_m$ satisfying

$$\tilde{\theta}_i = \frac{n_i\phi_i^{-1}x_{i.} + \sigma_\theta^{-2}\tilde{\theta}_.}{n_i\phi_i^{-1} + \sigma_\theta^{-2}} \qquad (i = 1, \cdots, m) \quad (24)$$

where

$$\tilde{\theta}_. = \sum_i \rho_\theta^{(i)}x_{i.}/\sum_i \rho_\theta^{(i)} \quad (25)$$

with

$$\rho_\theta^{(i)} = n_i\phi_i^{-1}/(n_i\phi_i^{-1} + \sigma_\theta^{-2}) \quad (26)$$

The expressions in (24) provide the elements of the vector in (2) which reduces to the posterior mean vector of $\boldsymbol{\theta}$, given $\boldsymbol{\phi}$ and σ_θ^2, in this special case. The expression for $\tilde{\theta}_i$ takes the form of a weighted average of the standard estimate $x_{i.}$ and the central value $\tilde{\theta}_.$ which we see from (25) to take the form of a weighted average of all the sample means. The standard estimates are therefore shrunk towards a value based on collateral information, and this will hopefully smooth out some of the random fluctuation in the data.

It should be remarked that the estimates in (24) were proposed in [7] for the *fixed effects* situation. The assumption of exchangeability does not imply that the θ_i constitute a random sample from a hyperpopulation of θ's. Random effects models are discussed in [10].

When the ϕ_i, and hence the α_i, are also exchangeable, we employ the prior model for $\boldsymbol{\alpha}$ which is described in section 1, but with $\mathbf{B} = \mathbf{I}_m$, $\boldsymbol{\mu}_\gamma$ equal to the vector of zeros, and $\mathbf{H}_\gamma = \sigma_\alpha^{-2}(\mathbf{I}_m - m^{-1}\mathbf{J}_m)$. The vector in (13) now provides an approximation to the conditional posterior mean vector of $\boldsymbol{\alpha}$, given $\boldsymbol{\theta}$ and σ_α^2, and by analogy with (24)–(26) this has elements $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m$ satisfying

$$\tilde{\alpha}_i = \frac{\frac{1}{2}n_il_i + \sigma_\alpha^{-2}\tilde{\alpha}_.}{\frac{1}{2}n_i + \sigma_\alpha^{-2}} \qquad (i = 1, \cdots, m) \quad (27)$$

with

$$\tilde{\alpha}_. = \sum \rho_\alpha^{(i)}l_i/\sum \rho_\alpha^{(i)} \quad (28)$$

where

$$\rho_\alpha^{(i)} = \tfrac{1}{2}n_i/(\tfrac{1}{2}n_i + \sigma_\alpha^{-2}) \quad (29)$$

and l_i is given in (11).

We therefore have similar weighted average forms for the log-variances to those obtained in [7] for the means. The expression in (27) takes the form of a weighted average of l_i and $\tilde{\alpha}$ the weights being $\frac{1}{2}n_i$ and σ_α^{-2} respectively. When the θ_i are known the estimate $\tilde{\phi}_i = e^{\tilde{\alpha}_i}$ for ϕ_i shrinks the maximum likelihood estimate

$$\hat{\phi}_i = e^{l_i} = n_i^{-1} S_i(\theta_i) = n_i^{-1} \sum_{j=1}^{n_i} (x_{ij} - \theta_i)^2 \quad (30)$$

towards the geometric mean

$$e^{\tilde{\alpha}_\cdot} = (\prod_i \tilde{\phi}_i)^{1/m} \quad (31)$$

This contrasts with the estimates in [7] where the shrinkages are towards the harmonic mean.

When the θ_i and ϕ_i are unknown we need to replace the ϕ_i in (24) and (26) by the corresponding $e^{\tilde{\alpha}_i}$, and the l_i in (27) and (28) by the corresponding $\tilde{l}_i$ in (18). The resultant equations may be solved using the iterative procedure described below, or as a special case of the procedure described towards the end of the previous section.

When the θ_i are unknown, then the maximum likelihood estimate of ϕ_i is instead given by

$$\hat{\phi}_i = n_i^{-1} S_i^w = n_i^{-1} \sum_{j=1}^{n_i} (x_{ij} - x_{i\cdot})^2 \quad (32)$$

We see from (9) that when $\theta_i \neq x_{i\cdot}$ the expression in (30) is always greater than that in (32). Our prior assumptions about the θ_i therefore have the effect of expanding the estimates of the ϕ_i, as well as shrinking the estimates of the θ_i towards the expression in (25). The expansion is greatest when $\sigma_\theta^2 = 0$, so that the $\tilde{\theta}_i$ are all equal, and hence equal to the quantity in (25) which reduces to

$$\tilde{\theta}_\cdot = \bar{x} = \sum_i n_i \phi_i^{-1} x_{i\cdot} / \sum_i n_i \phi_i^{-1}$$

In this extreme situation, we have

$$n_i^{-1} S_i(\tilde{\theta}_i) = n_i^{-1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = n_i^{-1} S_i^w + (x_{i\cdot} - \bar{x})^2$$

and these are the same as the maximum likelihood estimates obtained upon taking the θ_i to be equal. We see that the expansions may be quite considerable in this case.

The prior assumptions about the ϕ_i have the effect of shrinking the expanded estimates towards the geometric mean in (31), as well as affecting the estimates of the θ_i.

We now indicate how to generalise our results to the situation where the first-stage prior variances σ_θ^2 and σ_α^2 of the θ_i and α_i are unknown. We assume that $g_\theta = \log \sigma_\theta^2$ is a priori normally distributed with mean ξ_θ and variance τ_θ^2, and that $g_\alpha = \log \sigma_\alpha^2$ is a priori independent of g_θ and possesses a similar distribution, but with parameters ξ_α and τ_α^2. Whenever there is strong prior information that the θ_i are close together the statistician should choose small values for e^{ξ_θ} and τ_θ^2. If there is information that the θ_i are likely to be only slightly related a large value should instead be chosen for e^{ξ_θ}. If there is not much prior information about the strength of the symmetric relationship between the θ_i, then a large value should be chosen for τ_θ^2. Similar considerations apply to the choices of e^{ξ_α} and τ_α^2.

A convenient method of estimation is given in chapter 7 of the unpublished thesis [4], and we take the liberty of omitting the details from the present paper. We show that, unless m is small, the joint posterior modes of the θ_i, α_i, g_θ, and g_α may be approximated using equations (24) and (27) for the $\tilde{\theta}_i$ and $\tilde{\alpha}_i$ respectively. We merely have to replace ϕ_i, θ_i, σ_θ^2, and σ_α^2 in these equations by $e^{\tilde{\alpha}_i}$, $\tilde{\theta}_i$, $e^{\tilde{g}_\theta}$, and $e^{\tilde{g}_\alpha}$ respectively, where $\tilde{g}_\theta$ and $\tilde{g}_\alpha$ denote the corresponding modes of g_θ and g_α, and satisfy

$$\tilde{g}_\theta = \frac{\nu_\theta \xi_\theta + (m-1) M(\tilde{\boldsymbol{\theta}})}{\nu_\theta + m - 1} \quad (33)$$

and

$$\tilde{g}_\alpha = \frac{\nu_\alpha \xi_\alpha + (m-1) M(\tilde{\boldsymbol{\alpha}})}{\nu_\alpha + m - 1} \quad (34)$$

where $\nu_\theta = 2\tau_\theta^{-2}$, $\nu_\alpha = 2\tau_\alpha^{-2}$, and for any $m \times 1$ vector $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_m)^T$ we have

$$M(\boldsymbol{\epsilon}) = \log \{ \sum_i (\epsilon_i - \epsilon_\cdot)^2 / (m-1) \} \quad (35)$$

The expression for $\tilde{g}_\theta$ in (33) is a weighted average of the prior mean ξ_θ and the iterated contribution $M(\tilde{\boldsymbol{\theta}})$, the weights being ν_θ and $m - 1$ respectively. If $\nu_\theta \gg 1$, so that the prior variance τ_θ^2 is small, the prior mean ξ_θ will predominate. If $m - 1 \gg \nu_\theta$, the iterated contribution predominates. Similar properties are satisfied by the expression in (34).

An iterative procedure for the solution of the resultant equations is described as follows:

(i) Use the $x_{i\cdot}$ as initial values for the corresponding $\tilde{\theta}_i$, and use them to calculate values for the $S_i(\tilde{\theta}_i)$ from (19), and hence for the $\tilde{l}_i$ from (18). Use these values for the $\tilde{l}_i$ as initial values for the corresponding $\tilde{\alpha}_i$.

(ii) Use the latest values for the $\tilde{\theta}_i$ and $\tilde{\alpha}_i$ to calculate values for the averages

$$\tilde{\theta}_\cdot = m^{-1} \sum_i \tilde{\theta}_i \text{ and } \tilde{\alpha}_\cdot = m^{-1} \sum_i \tilde{\alpha}_i$$

(iii) Use the latest values for the $\tilde{\theta}_i$, $\tilde{\alpha}_i$, $\tilde{\theta}_\cdot$, and $\tilde{\alpha}_\cdot$ to calculate values for $\tilde{g}_\theta$ and $\tilde{g}_\alpha$ from (33) and (34) respectively.

(iv) Use the latest values for the $\tilde{\theta}_i$ to calculate

new values for the $\tilde{l}_i$ from (18). Substitute the latest values for the $e^{\tilde{\alpha}_i}$, $\tilde{\theta}_.$, $e^{\tilde{\phi}_\theta}$, $\tilde{l}_i$, $\tilde{\alpha}_.$, and $e^{\tilde{\phi}_\alpha}$ for the corresponding ϕ_i, $\tilde{\theta}_.$, σ_θ^2, l_i, $\tilde{\alpha}_.$, and σ_α^2 in the right hand sides of (24) and (27) and obtain new values for the $\tilde{\theta}_i$ and $\tilde{\alpha}_i$ on the left hand sides.

(v) Return to (ii) and keep cycling until convergence.

We have always found the above procedure to converge in a few seconds of computer time. One advantage is that it does not involve the direct inversion of any matrices.

In this section we have provided an alternative to the method in [7] which was the pioneer work on exchange-ability of variances. This method also involves some approximations, but it should be perfectly viable, and we leave it to the reader to decide which method he prefers in this special case. We of course feel that the principle advantage of our own method is that it generalises to estimations where more complex prior relationships exist between the variances. Professor Lindley has in fact been kind enough to inform us (personal communication) that there is no computer program prvoiding the solutions to his equations when the variances are unequal. He has concentrated on the case where the variances are equal, since the other case does not show promise of capability of generalisation. We are therefore unable to compare the alternative methods numerically, but in the next section we illustrate our own method before proceeding to discuss other applications of our general approach.

5. Numerical Example

We consider data previously analysed on p. 145 of [1] and concerning the breaking strengths of six different fabrics. There are $n_i = 10$ observations $x_{i1}, \cdots, x_{i10}$ on fabric no. i for $i = 1, \cdots, 6$. We found that the maximum likelihood estimates of the theoretical breaking strengths θ_i were not substantially affected by our particular prior assumptions in this case. We therefore restrict our descriptions to the estimation of the corresponding variances ϕ_i. Their maximum likelihood estimates $\hat{\phi}_i$ were calculated using (32) and are given in the first row of Table 1.

In [1] it is shown that Bartlett's test for the homogeneity of the variances fails for any sensible significance level. As a point of interest we note that this test is based upon the logs of the variances, thus employing similar transformations to our own. Our approach avoids the need for such tests since our estimates compromise between those obtained via classical methods upon assuming the variances unequal, and those obtained upon assuming them equal.

The heterogeneity in the present case is primarily due to the high value of the estimated variance $\hat{\phi}_2$ for fabric no. 2. We are therefore particularly interested in the effect of our Bayesian assumptions on this estimate.

It is reasonable to assume exchangeability of the ϕ_i if there is the required symmetry of prior information about the various fabrics. Exchangeability would be inappropriate if, for example, the fabrics were known to have been produced in a particular order by the same machine. In this case the alternative assumptions indicated at the end of section 6 may be more reasonable, since the time-dependence destroys the symmetry. Exchangeability of the ϕ_i would be appropriate if there was no prior information about the fabrics, in which case it appears to be more plausible than independence. This is because information about a particular ϕ_i would surely give us some idea about the values of the other ϕ_i, thus implying that the ϕ_i are related.

Table I—*Estimates of the Fabric Variances* ($e^{\xi_\alpha} = 0{\cdot}1$)

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	
Maximum Likelihood	0·72	14·26	3·39	5·79	1·93	0·81	
$\nu_\alpha = 0$	0·91	10·61	3·24	5·07	2·04	1·00	(0·91)
$\nu_\alpha = 1$	1·04	8·83	3·15	4·65	2·11	1·33	(0·50)
$\nu_\alpha = 3$	1·30	6·57	3·01	4·04	2·22	1·39	(0·24)
$\nu_\alpha = 5$	1·46	5·67	2·94	3·77	2·28	1·53	(0·17)
$\nu_\alpha = 10$	1·65	4·80	2·87	3·48	2·34	1·72	(0·11)

We firstly assume that there is no prior information to suggest likely values for the first-stage prior variances σ_θ^2, and σ_α^2. The latter respectively measure the closeness to each other of the θ_i, and of the $\alpha_i = \log \phi_i$. In this case we set $\nu_\theta = \nu_\alpha = 0$ which provides uninformative log-uniform priors for σ_θ^2 and σ_α^2. As mentioned in [10] we should strictly choose small positive values for ν_θ and ν_α to keep the distributions of σ_θ^2 and σ_α^2 proper, but this would have a negligible effect on our numerical results.

The corresponding Bayesian estimates for the ϕ_i are given in the second row of Table 1, and the value obtained for $\tilde{\sigma}_\alpha^2 = e^{\tilde{\nu}\alpha}$ is given in brackets at the end of the row. The effects of our assumptions on the estimates are quite noticeable. For example, our estimate of 10.61 for ϕ_2 compares with $\hat{\phi}_2 = 14.26$. The prior assumptions about the θ_i in fact cause $\hat{\phi}_2$ to be marginally increased to $n_2^{-1}S_2(\tilde{\theta}_2) = 14.43$, where $n_2^{-1}S_2(\theta_2)$ is given in (30). The prior assumptions about the ϕ_i then cause this value to be shrunk about one third of the way towards 2.62, which is the geometric mean of the $\hat{\phi}_i$. Some of the other differences are also substantial e.g. the estimates of ϕ_1, and ϕ_6 are both increased by about 20%.

We next examine the effect of prior information about σ_α^2 on our estimates, though we still keep $\nu_\theta = 0$. Such information may for example be based on general experience, or on data from previous fabrics.

For illustrative purposes we suppose hypothetically that the prior information about the closeness to each other of the α_i suggests a value of 0.1 for σ_α^2, and we therefore set $e^{\xi\alpha} = 0.1$. We then examine our estimates of the ϕ_i and σ_α^2 for various choices of $\nu_\alpha = 2\tau_\alpha^{-2}$. These estimates are given in the last four rows of Table 1.

As ν_α increases from zero, $\tilde{\sigma}_\alpha^2$ decreases from 0.91, and in fact approaches the prior estimate of 0.1 as $\nu_\alpha \to \infty$. The relationship is not simple, owing to the dependence of the expression for $\tilde{g}_\alpha = \log \tilde{\sigma}_\alpha^2$ in (34) upon the α_i. As ν_α increases, the estimates for the ϕ_i become closer together, and further from the corresponding maximum likelihood estimates.

The estimates depend very much upon the prior information about σ_α^2 which happens to be available. If it is not possible to ascertain this precisely, we recommend simply setting $\nu_\alpha = 0$, in which case the estimates do not depend upon any specific prior parameter values, but still differ from the maximum likelihood estimates.

6. Other Applications of General Results

The results in sections 2 and 3 may be applied to a whole range of special cases, only a few of which are mentioned here.

In [7], [10], and [11] the authors discuss regression models as special cases of their general model for the means. In some situations it may be appropriate to assume the variances, as well as the means, to depend upon explanatory variables. In such cases the elements of the design matrix $\mathbf{B}$ in (1.7) may be chosen as suitable functions of the explanatory variables, and $\boldsymbol{\gamma}$ may be taken to represent a vector of regression coefficients. The authors in particular discuss exchangeability between the coefficients of several normal regression lines, and also exchangeability between the coefficients of one line in a multiple regression situation. Alternative assumptions for the variances could well lead to quite different results.

In [9] a two-way layout of normal observations is analysed. The means θ_{ij} are taken to satisfy a relationship of the form

$$\theta_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad (i = 1, \cdots, r; j = 1, \cdots, s)$$

where the λ_i^A, λ_j^B and λ_{ij}^{AB} respectively denote the row, column, and interaction effects. The variances are either assumed equal, or unequal but exchangeable. We feel that it would sometimes be appropriate to assume a similar structure for the log-variances to that previously assumed for the means. In [9] certain assumptions of exchangeability are made for the various effects, and the results are obtained as a special case of the analysis in [8]. It is possible to obtain similar results for the effects of the log-variances by using a special case of the result in (13). We may alternatively proceed directly using a similar method to that employed by us in a forthcoming paper on contingency tables. This provides us with the following explicit approximations to the estimates suggested implicitly in [9] for the various effects:

$$\tilde{\mu} = \sum_{kq} w_{kq} x_{kq.} / \sum_{kq} w_{kq} \tag{36}$$

$$\tilde{\lambda}_i^A = \frac{s\sigma_{AB}^{-2}}{s\sigma_{AB}^{-2} + \sigma_A^{-2}} \left\{ \frac{\sum_q w_{iq}(x_{iq.} - b^A)}{\sum_q w_{iq}} \right\} \tag{37}$$

$$\tilde{\lambda}_j^B = \frac{r\sigma_{AB}^{-2}}{r\sigma_{AB}^{-2} + \sigma_B^{-2}} \left\{ \frac{\sum_k w_{kj}(x_{kj.} - b^B)}{\sum_k w_{kj}} \right\} \tag{38}$$

and

$$\tilde{\lambda}_{ij}^{AB} = w_{ij}(x_{ij.} - \tilde{\mu} - \tilde{\lambda}_i^A - \tilde{\lambda}_j^B) \tag{39}$$

where

$$w_{ij} = n_{ij}\phi_{ij}^{-1}/(n_{ij}\phi_{ij}^{-1} + \sigma_{AB}^{-2}) \tag{40}$$

and b^A and b^B are complicated expressions which ensure that $\tilde{\lambda}_.^A = \tilde{\lambda}_.^B = 0$. Here we use the notation

$x_{ij.}$, n_{ij} , and ϕ_{ij} to respectively denote the average observation, the number of observations, and the variance corresponding to the (i, j)th cell. Also σ_A^2, σ_B^2, and σ_{AB}^2 respectively represent the common variances of the λ_i^A, the λ_j^B, and the λ_{ij}^{AB} at the first stage of the exchangeable prior model.

The estimates in (37), (38), and (39) roughly speaking shrink the corresponding maximum likelihood estimates towards zero. For a full discussion in the special case where all the n_{ij} and ϕ_{ij} are equal we refer the reader to [8]. Analogous approximations for the effects of the log-variances may be obtained directly via the substitutions mentioned in section 2. In our forthcoming contingency table paper the $x_{ij.}$ and $n_{ij}^{-1}\phi_{ij}$ in (37)–(40) will be replaced by the logs of the appropriate observed multinomial frequencies, and their reciprocals. The corresponding exact equations will provide a method for coping with zero frequencies.

Our methods are also applicable to linear models in time series analysis. Consider for example the stationary first-order autoregressive process for the θ_i where

$$\theta_{i+1} = \rho_\theta \theta_i + \eta_i \quad (i = 1, \cdots, m-1; |\rho_\theta| < 1) \quad (41)$$

and the η_i are uncorrelated normal errors

$$\eta_i \sim N(0, \tau_\theta^2)$$

with η_i independent of θ_i , θ_{i-1} , $\cdots$, θ_1 . As discussed in chapter 4 of [4] this in fact provides a special case of the formulation in section 1, but with $\mathbf{A}$ in (1) equal to the $m \times m$ identity matrix, the prior mean vector $\mathbf{u}_\beta$ equal to the zero vector, and the (i, k)th element of the covariance matrix $\mathbf{H}_\beta^{-1}$ equal to

$$\tau_\theta^2 \rho_\theta^{|i-k|}/(1 - \rho_\theta^2) \quad (42)$$

If the corresponding log-variances α_1 , $\cdots$, α_m are unequal it may very often be reasonable to assume ordered relationships of the α_i are of a similar nature to that assumed in (41) for the θ_i . We do not include the analysis here since it follows by analogy from a method in [5] for smoothing histograms, which also employs the covariance structure in (42). It is however possible to obtain simultaneous estimates for the θ_i and α_i , and also to estimate σ_θ^2, ρ_θ , and the corresponding parameters for the α_i . Such a method might find application in quality control, and may be used to detect whether the mean and variance corresponding to a particular time stage fall outside a designated region of the parameter space.

7. Acknowledgments

The author is extremely grateful to Professor D. V. Lindley for suggesting the problem, supervising the research, providing frequent advice and discussions about his own research, and suggesting the covariance structure in (42).

Thanks are also due to Professor M. R. Novick for his kind advice, and Professor P. J. Harrison for discussing the autoregressive process.

The research was partially carried out at University College London where it was financed by the Science Research Council, and comprised part of the seventh chapter of a Ph.D. thesis; and partially at the American College Testing Program, Iowa City, where it was financed by an A.C.T./I.T.P. predoctoral research fellowship, and where our exchangeability method is programmed conversa tionally on an interactive computer terminal.

References

[1] Fryer H. C. (1966). *Concepts and Methods of Experimental Statistics*, Allyn and Bacon

[2] James W. and Stein C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium*, Vol. 1 pp. 361–379 University of California Press.

[3] Leonard T. (1972). Bayesian methods for binomial data *Biometrika* Vol. 59, 581–589.

[4] Leonard T. (1973). Bayesian methods for the simultaneous estimation of several parameters. *Ph.D. thesis* (unpublished) University of London.

[5] Leonard T. (1973). A Bayesian method for histograms *Biometrika* Vol. 60, 297–308.

[6] Lindley D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part II: Inference*, Cambridge University Press.

[7] Lindley D. V. (1971). The estimation of many parameters. In *Foundations of Statistical Inference* (ed. by V. P. Godambe and D. A. Sprott) pp. 435–455 Toronto: Holt, Rinehart, and Winston.

[8] Lindley D. V. and Smith A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B.* Vol. 34, 1–41.

[9] Lindley D. V. (1973). A Bayesian solution for the two-way analysis of variance *A. C. T. Technical Bulletin No. 8*, Iowa City, Iowa: The American College Testing Program.

[10] Smith A. F. M. (1973). A general Bayesian linear model. *J. Roy. Statist. Soc. Ser. B.* Vol. 35, 67–75.

[11] Smith A. F. M. (1973). Bayes estimates for one-way and two-way models. *Biometrika* 60, 319–330.