The Annals of Statistics
1993, Vol. 21, No. 3, 1093–1114

MATRIX WEIGHTING OF SEVERAL REGRESSION COEFFICIENT VECTORS

BY ALAN T. JAMES[1] AND WILLIAM N. VENABLES

University of Adelaide

For small sample random effects models, results are derived which show in certain cases, and indicate in general, that an estimated random effects variance matrix may be used in the weight matrices without causing undue error in the empirically weighted mean. Exact error variances are derived mathematically for the empirically weighted mean for the two sample case in one and two dimensions. Simulation is used to determine errors for a practical example of six five-variate samples. For estimation of their mean, the differences between the samples are ancillary. The biases of the average and weighted mean estimators conditional on these ancillaries is illustrated in a diagram plotting values obtained by simulation. A curious range anomaly is illustrated which arises if random effects are ignored when present.

1. Introduction. A general problem which occurs throughout a wide spectrum of statistical practice is as follows. A series of data sets analysed by the same linear model has led to independent estimates of the same regression coefficient parameter vector $\boldsymbol{\beta}$. An efficient estimate of $\boldsymbol{\beta}$ is needed, allowing for both within data set and between data set matrix components of variation, as in the random effects models introduced by Henderson, Kempthorne, Searle and von Krosigk (1959).

The simplest such random effects model specifies $p + 1$ independent sample regression coefficient vectors $\mathbf{b}_i \in R^n$, $i = 1, 2, \ldots, p + 1$, conditionally distributed as

$$\mathbf{b}_i | \boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}_i, \Gamma_i)$$

with the $\boldsymbol{\beta}_i$ independently marginally distributed as $N(\boldsymbol{\beta}, \boldsymbol{\Delta})$. Hence, marginally, the $\mathbf{b}_i$ are independently distributed as $N(\boldsymbol{\beta}, \boldsymbol{\Delta} + \Gamma_i)$. Models for comparison of two or more mean vectors, or more general linear models can be built up, but the inferential issues with which we are concerned are the same as for the estimation of a single mean vector $\boldsymbol{\beta}$ and between regressions variance component matrix $\boldsymbol{\Delta}$.

We shall assume that the Γ_i are known or accurately estimated, but that $\boldsymbol{\Delta}$ is not only unknown, but that there is no prior information concerning it. It may be zero but this cannot be assumed.

Received May 1991; revised March 1992.
[1]Supported by Australian Research Council Grant A68931380
AMS 1991 *subject classifications*. Primary 62H12; secondary 62J10.
Key words and phrases. Matrix weighting, random effects model, small sample random effects model, estimated generalized least squares, unbalanced data, simulation, residual maximum likelihood, moment estimator, conditional bias, range anomaly, cutoff function, efficiency.

This problem of similar regressions has arisen in so many different contexts that different terminologies have grown up. Feldman (1988) classifies four different methods of analysis which have been used:

1. Pool all the original data and do a combined analysis ignoring random effects. The above author calls this method naively pooled data (NPD) but other authors such as Gumpertz and Pantula (1989) simply call it ordinary least squares (OLS).
2. Standard two stage (STS). Ignore the estimated errors of the regression coefficient vectors $\mathbf{b}_i$ and treat them as raw homoscedastic data. The maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ will then simply be the average, $\bar{\mathbf{b}}_.$.
3. Weighted least squares (WLS). Assume a random effects model and form a matrix weighted mean using a simple but inefficient moment estimator $\boldsymbol{\Delta}^{(m)}$ of $\boldsymbol{\Delta}$. Equivalently, however, one can go back to the original data by incorporating $\boldsymbol{\Delta}$, or its estimate $\boldsymbol{\Delta}^{(m)}$ into the variance matrix of the pooled original data to give a generalized least squares (GLS) analysis of it. When an estimate of $\boldsymbol{\Delta}$ subject to appreciable sampling error is used instead of $\boldsymbol{\Delta}$ itself, the method is called *estimated* GLS (EGLS).
4. Residual maximum likelihood (REML) estimates can be made of the mean vector $\boldsymbol{\beta}$ and variance matrix $\boldsymbol{\Delta}$. This is iterative. We shall refer to the use of the REML estimate, $\hat{\boldsymbol{\Delta}}$ of $\boldsymbol{\Delta}$ in weights as *empirical* weighting.

If the original $p + 1$ regressions are linear, then for normal data the ML estimates $\mathbf{b}_i$ together with the REML estimates s_i^2 of the error variances σ_i^2 are sufficient. It is therefore not necessary to go back to the original data because a weighted mean will supply an algebraically identical estimate. For nonlinear or nonnormal data this will be approximately so if efficient ML estimation is used. Besides going back to the original observations, which was unnecessary in the context of our example to be described in Section 2, much of the literature deals with best linear unbiased prediction (BLUP) of the random effects or coefficients $\boldsymbol{\beta}_i$. BLUP is important in many areas such as the plant and animal breeding discussed by Henderson, Kempthorne, Searle and von Krosigk (1959), but not in our work, though our results do impinge on BLUP.

Data or models with equal Γ_i, that is, $\Gamma_1 = \Gamma_2 = \cdots = \Gamma$ are called *balanced*. The average $\bar{\mathbf{b}}_.$ is then the ML estimator of $\boldsymbol{\beta}$ as given by method 2. The REML variance estimator involves cutoff as described and proved by Ameniya (1985). Let $\mathbf{B}$ be the $n \times (p + 1)$ matrix of vectors $\mathbf{b}_i$. If their conditional variance matrix $\boldsymbol{\Gamma}$ is ignored as in method 2, then the REML estimator of their variance matrix is the matrix of mean sums of squares and products,

$$\mathbf{V}_b = \left(\mathbf{B} - \bar{\mathbf{b}}_.\mathbf{1}'_{p+1}\right)\left(\mathbf{B} - \bar{\mathbf{b}}_.\mathbf{1}'_{p+1}\right)'/p,$$

where $\bar{\mathbf{b}}_. = \mathbf{B}\mathbf{1}_{p+1}/(p + 1)$. The variance matrix of $\bar{\mathbf{b}}_.$ would then be estimated

by $\mathbf{V}_b/(p+1)$. The standard errors of $\bar{\mathbf{b}}_.$ given in Table 1 of Section 2 are estimated in this way.

If $\mathbf{\Gamma}$ is taken into account, however, then $E[\mathbf{V}_b] = \mathbf{\Delta} + \mathbf{\Gamma}$ and $\mathbf{V}_b - \mathbf{\Gamma}$ is the REML estimator $\hat{\mathbf{\Delta}}$ of $\mathbf{\Delta}$, provided it is positive or positive semidefinite. If not, then the REML estimator requires cutoff, because, if for $\mathbf{a} \in R^n$, the estimator of $\mathbf{a}'(\mathbf{\Delta} + \mathbf{\Gamma})\mathbf{a}$ is less than $\mathbf{a}'\mathbf{\Gamma}\mathbf{a}$, then the latter value must be used, since $\mathbf{a}'\mathbf{\Delta}\mathbf{a}$ cannot be negative. This requirement is achieved as follows.

Provided $\mathbf{\Gamma}$ is positive definite, there exists a nonsingular matrix $\mathbf{L}$ and a diagonal matrix $\operatorname{diag}(\lambda_i)$, such that $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}'$ and $\mathbf{V}_B = \mathbf{L}\operatorname{diag}(\lambda_i)\mathbf{L}'$. Then

$$\hat{\mathbf{\Delta}} = \mathbf{L}\operatorname{diag}(\max((\lambda_i - 1), 0))\mathbf{L}'.$$

If for unbalanced data, we replace $\mathbf{\Gamma}$ by $\bar{\mathbf{\Gamma}}_. = \Sigma\mathbf{\Gamma}_i/(p+1)$, then we obtain the moment estimator $\mathbf{\Delta}^{(m)}$. For $p = 1$, it is proved in Theorem 1 of Section 4 that this is the REML estimator, but for $p > 1$, $\hat{\mathbf{\Delta}} \neq \mathbf{\Delta}^{(m)}$, for unbalanced data with which we shall be concerned.

The moment estimator is easy to simulate but the REML estimator, being iterative, is not.

It has been widely assumed that with a small number $p + 1$ of random effects, the use in weights of an estimator of $\mathbf{\Delta}$ with large sampling error would render the weighted mean $\tilde{\boldsymbol{\beta}}$ much less accurate than other possible estimators. Our main result in this paper is that the estimator $\tilde{\boldsymbol{\beta}}$ is more accurate than the average $\bar{\mathbf{b}}_.$ for small $\mathbf{\Delta}$ and generally not much less accurate for large $\mathbf{\Delta}$.

On account of the large sampling error of an estimate of $\mathbf{\Delta}$ from a small sample of random effects, one is tempted for convenience to assume $\mathbf{\Delta}$ is zero, or simply ignore it, and use method 1. We show that in two or more dimensions this can produce a strange range anomaly.

The paper aims to overcome a prejudice against the use of empirically weighted means in the *small sample* random effects model that has little foundation and establish it as a viable statistical method.

1.1. *Outline.* Section 2 describes a practical example in which the problem arose. Section 3 shows by an artificial example how the range anomaly arises, and how an empirically weighted mean cures it. Exact mathematical theory is developed and discussed in Sections 4–6. Simulation is treated in Section 7. The paper closes with a discussion and succinct conclusions.

2. An example from mitochondrial experiments. Our interest in the problem of matrix weighting came from our attempts to summarize the results of six experiments designed to produce a quantitative model of mitochondrial performance. Mitochondria are numerous organelles within the cells of plants and animals which generate the aerobic power. During nine months of preliminary experiments, James, Wiskich and Conyers (1989, 1993) perfected experimental technique and developed a thermokinetic model to describe quantitatively how the potential of mitochondrial power decreases with load in

TABLE 1

Six independent row vectors of five regression coefficients with their standard errors, and four types of mean vectors. e_1, e_2, e_3 are potentials and r_1, r_2 are resistances measuring mitochondrial performance

Rat number, i	e_1 ln units	e_2 ln units	e_3 ln units	$100r_1$ ln units $\%^{-1}$	$100r_2$ ln units $\%^{-1}$
Hexokinase					
1	4.249 ± 0.055	4.786 ± 0.065	5.538 ± 0.080	0.775 ± 0.144	1.078 ± 0.056
2	4.380 ± 0.048	4.985 ± 0.057	5.401 ± 0.064	1.408 ± 0.158	1.072 ± 0.078
3	4.556 ± 0.090	5.145 ± 0.095	5.913 ± 0.112	1.943 ± 0.303	1.176 ± 0.107
4	4.424 ± 0.066	4.976 ± 0.073	5.567 ± 0.079	0.864 ± 0.172	1.023 ± 0.068
5	4.537 ± 0.047	5.016 ± 0.057	5.700 ± 0.069	1.292 ± 0.118	0.794 ± 0.044
6	4.391 ± 0.038	4.899 ± 0.043	5.370 ± 0.048	0.861 ± 0.118	1.076 ± 0.046
Mean Vector Weight					
Average I	4.423 ± 0.046	4.968 ± 0.049	5.582 ± 0.082	1.191 ± 0.184	1.037 ± 0.053
Simulated error	0.042	0.043	0.071	0.164	0.057
Scalar $(\text{diag}(\boldsymbol{\Gamma}_i))^{-1}$	4.410 ± 0.021	4.945 ± 0.024	5.509 ± 0.028	1.080 ± 0.060	0.992 ± 0.024
Matrix $\boldsymbol{\Gamma}_i^{-1}$	4.283 ± 0.019	4.780 ± 0.022	5.348 ± 0.025	0.842 ± 0.056	0.945 ± 0.023
REML $(\boldsymbol{\Gamma}_i + \hat{\boldsymbol{\Delta}})^{-1}$	4.407 ± 0.041	4.946 ± 0.041	5.557 ± 0.070	1.144 ± 0.161	1.027 ± 0.055
Moment $(\boldsymbol{\Gamma}_i + \boldsymbol{\Delta}^{(m)})^{-1}$	4.415 ± 0.046	4.954 ± 0.049	5.567 ± 0.082	1.164 ± 0.182	1.031 ± 0.052
Simulated error	0.042	0.043	0.071	0.161	0.056

a manner analogous to the fall of voltage of a car battery when the starter motor is engaged.

Six final independent experiments were done as a test of the model. The results were fitted by nonlinear regression giving rise to six sample regression vectors each of 10 components published in James, Wiskich and Conyers (1993). Five of the components of each regression vector pertain to the mitochondria and five to the load. Three potentials and two resistances comprise the five components pertaining to the mitochondria. They are given as the row vectors $\mathbf{b}_i'$, $i = 1, 2, \ldots, 6$, of Table 1 with their standard errors. Since the standard errors are estimated on the equivalent of 43 degrees of freedom, they are reasonably accurate compared to the uncertainty of the between regressions component $\boldsymbol{\Delta}$. The six 5×5 correlation matrices of the $\mathbf{b}_i$ obtained from determining variables X_i are given in Table 2. The imbalance of the data lies in the large differences between the correlation matrices, and between the standard errors for the six experiments.

A biological interpretation of these results requires the six regression coefficient vectors to be summarized by a single vector of potentials and resistances. The variation between and within experiments is of secondary interest, but must be acknowledged by the analysis. In other situations hypothesis tests on the components of the mean vector may likewise be important.

With hindsight, we can now see that a simple average would have been adequate biologically, with some cutoff for its estimated error variance based upon $\bar{\boldsymbol{\Gamma}}$. We were concerned, however, that there were big differences in errors,

TABLE 2

The six sample correlation matrices given as upper or lower halves. The variance matrices Γ_i can be reconstituted from them and the standard errors given in Table 1

	0.8010	0.8216	0.6393	0.1616		0.8697	0.8835	0.7127	0.2215
0.7288		0.8638	0.6729	0.1689	0.7677		0.9407	0.7576	0.2381
0.7454	0.7900		0.6943	0.1677	0.7905	0.9119		0.7735	0.2366
0.4781	0.5370	0.5491		−0.5317	0.5918	0.6754	0.6931		−0.3969
0.1321	0.1025	0.1050	−0.7146		0.0244	0.0318	0.0424	−0.6656	
	0.8109	0.8251	0.6731	0.1462			samples		
0.7198		0.8441	0.6910	0.1458		1			3
0.7535	0.8294		0.7063	0.1436	2			4	
0.5619	0.6144	0.6402		−0.5216		5			
0.0541	0.0653	0.0725	−0.6705		6				

and at an early stage considered omitting experiment number 3 from the average on the grounds that it had much higher error than the other results. Such an arbitrary procedure, however, seemed difficult to justify.

It nevertheless seemed desirable to downweight the less accurate experiments by weighting by the inverses of their variances. By dealing with each component marginally, we obtained the estimates shown in Table 1 opposite the heading "Scalar" that were in accord with our intuition. The standard errors are low because they do not contain the between experiments variance component which we now recognize to be important.

Although the results look reasonable, they do not necessarily provide the best estimate of a function of the components, nor its error. For a coherent multivariate analysis, matrix weights given by the inverses of the variance matrices should be used. When we did this using matrix weights Γ_i^{-1}, because the estimate of Δ was so much subject to error, we obtained the matrix weighted means shown in Table 1 whose second and third components lay outside the range of the six values of which they are supposed to form a summary. This is what we refer to as the potential *range anomaly* of method 1 or OLS.

In Section 3, we illustrate by an artificial example just how it comes about, and how it does not appear to occur under the more realistic random effects models of methods 3 and 4.

From now on we change our notation for the vectors and their marginal distributions from $\mathbf{b}_i \sim N(\boldsymbol{\beta}_i, \Delta + \Gamma_i)$ to $\mathbf{y}_i \sim N(\boldsymbol{\mu}, \Delta + \Gamma_i)$, $i = 1, 2, \ldots, p + 1$.

3. Illustration of matrix-weighted mean displacement. Suppose two independent bivariate sample vectors had the following values with known variance matrices given by

$$\mathbf{y}_1 = \begin{bmatrix} 0 \\ a \end{bmatrix}, \quad \Gamma_1 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \quad \mathbf{y}_2 = \begin{bmatrix} 0 \\ -a \end{bmatrix}, \quad \Gamma_2 = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}.$$

It looks highly artificial to take both observed abscissae as zero, but it makes

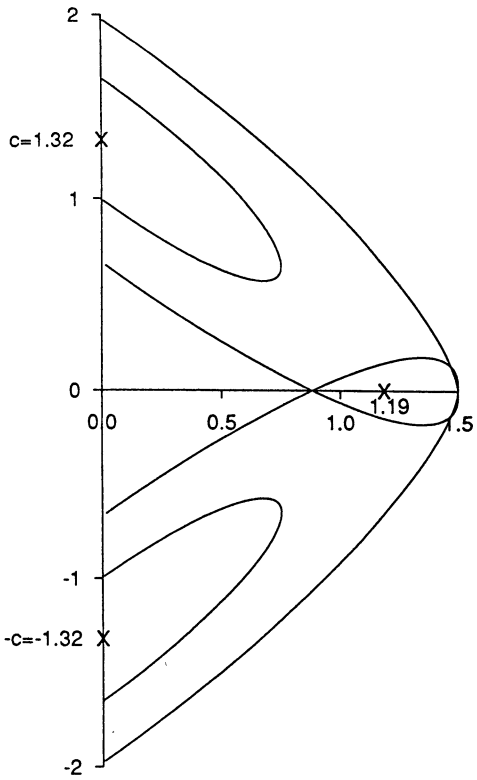

FIG. 1. *Likelihood contours for two separate bivariate likelihoods.*

the figure described below easier to comprehend. One could rotate it a little and displace it from zero without altering the essential argument.

We begin by treating this "data" by method 1 which ignores random effects, to illustrate the consequences.

Figure 1 shows the contours of the separate likelihood functions for the expectation vector, obtained from each observed vector. One sees how two ridges of high likelihood extend out from the observed vectors when the correlations are numerically large, and how, when the correlations are very unequal, the ridges will intersect in a region of high product of their likelihoods. If it can be assumed that the two observed vectors have a common expectation vector, its likelihood function is the product of their likelihoods. This function is illustrated graphically by the surface shown in Figure 2. Hence one can see how a matrix weighted mean can be well away from a scalar weighted mean, due to the first of the two variates having a strong covariance on the second.

If it is reasonable to assume that the expectations of the two vectors are equal, and if the region of the intersection of the two ridges of high likelihood is well within the two confidence intervals of the expectation obtained from the

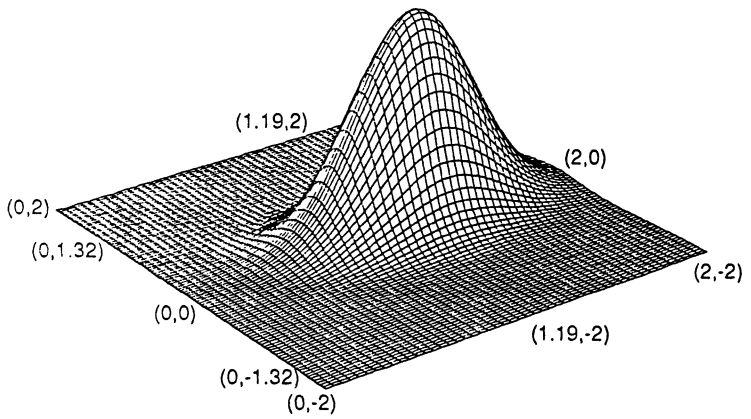

FIG. 2. *Likelihood surface given by the product of the likelihoods in Figure* 1.

individual observed vectors, then the matrix weighted mean is acceptable. The considerable displacement and high accuracy are due to the extra information from the strong covariance.

On the other hand, if the observed vectors differ significantly as tested by

$$\chi_2^2 = \mathbf{d}'(\mathbf{\Gamma}_1 + \mathbf{\Gamma}_2)^{-1}\mathbf{d} = 2a^2,$$

where $\mathbf{d} = \mathbf{y}_2 - \mathbf{y}_1$, then the assumption underlying method 1 is significantly rejected. A random effects model, however, will fit the "data." A variance component matrix must be estimated even though it is only on one degree of freedom. It is given for the n dimensional case by Theorem 1 stated and proved in Section 4.

We now come to a common situation in which random effects may possibly be zero or negligible but this cannot be assumed. Suppose there is no a priori certainty that the expectations are equal, but χ_2^2 is below significance as, for example, $\chi_2^2 = 3.5$ when $a = \sqrt{3.5/2} = 1.32$. If, in an attempt to use method 1, we persist in specifying a model of equal expectations on the basis that the estimates do not differ significantly at the 5% level, then the ML estimate of the mean vector, with standard errors, is

$$\boldsymbol{\mu}^+ = \begin{bmatrix} 1.19 \pm 0.31 \\ 0 \pm 0.31 \end{bmatrix}.$$

On this specification, the first element is highly significantly different from zero. But this inference depends entirely upon the assumption of no random effect. If the assumption is doubtful, the inference is correspondingly dubious.

If on the other hand one specifies a random effects model, then the estimated variance component matrix is

$$\hat{\mathbf{\Delta}} = \frac{1}{2}\left(1 - \frac{1}{2a^2}\right)\begin{bmatrix} 0 & 0 \\ 0 & 4a^2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{5}{2} \end{bmatrix}$$

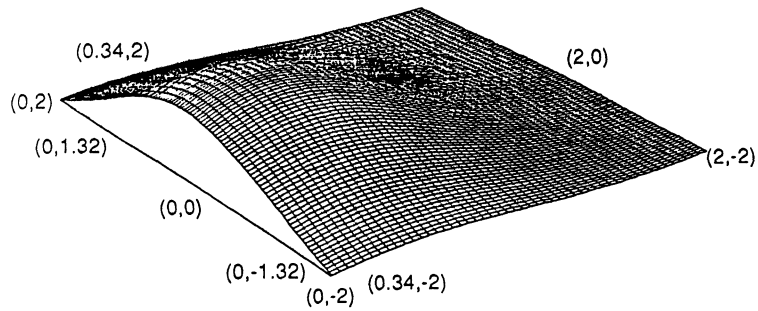

FIG. 3. *Likelihood surface for the parameter vector in the random coefficients model.*

giving a weighted mean of $\hat{\boldsymbol{\mu}} = \begin{bmatrix} 0.34 \pm 0.62 \\ 0 \pm 1.16 \end{bmatrix}$ if we ignore the sampling error of $\hat{\boldsymbol{\Delta}}$. If we allow for the sampling error of $\hat{\boldsymbol{\Delta}}$ as developed in Section 5, but evaluate the error at $\boldsymbol{\Delta} = \hat{\boldsymbol{\Delta}}$, we obtain

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 0.34 \pm 0.67 \\ 0 \pm 1.25 \end{bmatrix}.$$

To us, this seems a far more reasonable inference. This weighted mean is the value at which the likelihood surface for the random coefficients model, shown in Figure 3, has its maximum. The surface also shows high likelihood at $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$. At $\boldsymbol{\Delta} = \hat{\boldsymbol{\Delta}}$, this estimate has a smaller error than the average

$$\bar{\mathbf{y}}_{\cdot} = \begin{bmatrix} 0 \pm 0.71 \\ 0 \pm 1.32 \end{bmatrix}.$$

The range anomaly comes about by treating data with a nonzero $\boldsymbol{\Delta}$ as if it were zero. But $\boldsymbol{\Delta}$ is not the direct cause of the range anomaly, it is an indirect cause. $\boldsymbol{\Delta}$ operates only through a large difference, $\mathbf{d} = \mathbf{y}_2 - \mathbf{y}_1$, which is reflected in $\hat{\boldsymbol{\Delta}}$ as a function of $\mathbf{d}$. If $\mathbf{d}$ is fortuitously small, then there is no range anomaly when NPD or OLS is applied. Likewise, when the NPD produces a range anomaly, it is $\hat{\boldsymbol{\Delta}}$ as a function $\mathbf{d}$ which rectifies it. In dimensions in which $\hat{\boldsymbol{\Delta}}$ is singular, there are no differences to create a range anomaly.

On the other hand, if $\hat{\boldsymbol{\Delta}}$ fortuitously overestimates $\boldsymbol{\Delta}$, then the weights move conservatively towards the equality which gives the average, $\bar{\mathbf{y}}_{\cdot}$.

These considerations led us to question a widely held belief that large errors in $\hat{\boldsymbol{\Delta}}$ would produce excessive errors if it was used in weighting of means. The situation should not be confused with a common one in which errors in weights are independent of the observations they weight.

Consequently, we set out to investigate mathematically what would otherwise be regarded as the hopelessly inaccurate situation of two samples. Fortunately this is mathematically tractable.

4. Theory for two vectors. For two vectors with $p = 1$, we can give an explicit formula for the REML estimated variance component $\hat{\mathbf{\Delta}}$.

THEOREM 1. *For two independently normally distributed vectors,* $\mathbf{y}_1, \mathbf{y}_2 \in R^n$, *with a common expectation vector* $\boldsymbol{\mu}$, *and variance matrices* $\mathbf{\Gamma}_1 + \mathbf{\Delta}, \mathbf{\Gamma}_2 + \mathbf{\Delta}$, *where* $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2$ *are known, the REML estimator of the variance component matrix* $\mathbf{\Delta}$ *between them is the moment estimator,*

$$\hat{\mathbf{\Delta}} = \begin{cases} 0, & \text{if } \chi^2 \le 1, \\ (\chi^2 - 1)\mathbf{d}\mathbf{d}'/(2\chi^2), & \text{if } \chi^2 \ge 1, \end{cases} \tag{4.1}$$

where $\mathbf{d} = \mathbf{y}_2 - \mathbf{y}_1$, $\chi^2 = \mathbf{d}'\mathbf{\Gamma}^{-1}\mathbf{d}$ *and* $\mathbf{\Gamma} = \mathbf{\Gamma}_1 + \mathbf{\Gamma}_2$.

PROOF. The residual likelihood must be based on the distribution of $\mathbf{d}$, because the probability density of $\mathbf{y}_1, \mathbf{y}_2$ factorizes into the probability densities of the weighted mean $\hat{\boldsymbol{\beta}}$ of $\mathbf{y}_1$ and $\mathbf{y}_2$ with weights $(\mathbf{\Gamma}_1 + \mathbf{\Delta})^{-1}, (\mathbf{\Gamma}_2 + \mathbf{\Delta})^{-1}$, and of $2^{-1/2}\mathbf{d} \sim N(\mathbf{0}, (1/2)\mathbf{\Gamma} + \mathbf{\Delta})$. Hence the residual log likelihood is

$$\text{constant} - \tfrac{1}{2}\log\det\left(\tfrac{1}{2}\mathbf{\Gamma} + \mathbf{\Delta}\right) - \tfrac{1}{2}\operatorname{tr}\left\{\left(\tfrac{1}{2}\mathbf{\Gamma} + \mathbf{\Delta}\right)^{-1}\tfrac{1}{2}\mathbf{d}\mathbf{d}'\right\}.$$

The rest of the proof is an adaptation of the Appendix to Ameniya (1985).

Choose a matrix L not depending upon $\mathbf{\Delta}$ such that $L(\mathbf{\Gamma}/2)L' = I$, $L(\mathbf{d}\mathbf{d}'/2)L' = S$, where S is zero except for $s_{11} = \chi^2$ in its top left corner. Put $L\mathbf{\Delta}L' = U$ and $W = (I + U)^{-1} \le I$. Apart from an additive constant involving $\log\det(L)$, the log likelihood to be maximized with respect to W is then

$$\tfrac{1}{2}(\log\det(W) - w_{11}\chi^2).$$

The maximum occurs when W is diagonal and $w_{ii} = 1$ for $i = 2, \ldots, n$, leaving

$$\tfrac{1}{2}(\log\det(w_{11}) - w_{11}\chi^2),$$

which is maximized at

$$w_{11} = \begin{cases} 1, & \text{if } \chi^2 \le 1, \\ 1/\chi^2, & \text{if } \chi^2 \ge 1, \end{cases}$$

that is,

$$u_{11} = \begin{cases} 0, & \text{if } \chi^2 \le 1, \\ \chi^2 - 1, & \text{if } \chi^2 \ge 1, \end{cases}$$

and all other elements $u_{ij} = 0$. Hence, the estimator is either zero or

$$U = \frac{\chi^2 - 1}{\chi^2} S$$

and on detransforming with L^{-1}, we have either zero or

$$\hat{\mathbf{\Delta}} = \frac{\chi^2 - 1}{\chi^2}\frac{1}{2}\mathbf{d}\mathbf{d}'. \qquad \square$$

Since for $p = 1$, the range of $\mathbf{\Delta}$ lies in the range of $\mathbf{d} = \mathbf{y}_2 - \mathbf{y}_1$, it might likewise be conjectured that with three or more vectors, $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \cdots, \mathbf{y}_{p+1} \in R^n$ with $p < n$, the REML estimate of the $(n \times n)$ variance matrix would have its range in the subspace of dimension $\leq p < n$, spanned by the differences $\mathbf{y}_i - \mathbf{y}_j$ of the vectors; but a numerical example showed that this is not so in general.

If $\hat{\boldsymbol{\mu}}$ is the least squares weighted mean of $\mathbf{y}_1$ and $\mathbf{y}_2$ with weights $(\mathbf{\Gamma}_i + \mathbf{\Delta})^{-1}$ and $\tilde{\boldsymbol{\mu}}$ is their empirically weighted mean with weights $(\mathbf{\Gamma}_i + \hat{\mathbf{\Delta}})^{-1}$, then by substituting for the $\mathbf{y}_i$ in terms of their average $\bar{\mathbf{y}}.$ and $\mathbf{d}$ one has the following:

THEOREM 2. *For two vectors as in Theorem* 1, *the least squares weighted mean is*

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}. - (\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1)\mathbf{\Gamma}^{-1}\mathbf{\Lambda}^{-1}\mathbf{d}/2, \tag{4.2}$$

where $\mathbf{\Gamma}_d = \operatorname{var}(\mathbf{d}) = \mathbf{\Gamma} + 2\mathbf{\Delta}$ *and* $\mathbf{\Lambda} = \mathbf{\Gamma}_d\mathbf{\Gamma}^{-1}$, *and the empirically weighted mean is*

$$\tilde{\boldsymbol{\mu}} = \bar{\mathbf{y}}. - (\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1)\mathbf{\Gamma}^{-1}\mathbf{c}(\mathbf{d}) \tag{4.3}$$

where, what we shall call the cutoff function is

$$\mathbf{c}(\mathbf{d}) = \begin{cases} \mathbf{d}/2, & \text{if } \chi^2 \leq 1, \\ \mathbf{d}/(2\chi^2), & \text{if } \chi^2 \geq 1, \end{cases} \tag{4.4}$$

and its variance is

$$\operatorname{var}(\tilde{\boldsymbol{\mu}}) = \operatorname{var}(\bar{\mathbf{y}}.) + (\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1)\mathbf{\Gamma}^{-1}\mathbf{V}_k\mathbf{\Gamma}^{-1}(\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1), \tag{4.5}$$

where $\mathbf{V}_k$ *is the variance kernel, given by*

$$\mathbf{V}_k = \operatorname{var}(\mathbf{c}) - \tfrac{1}{2}\mathbf{\Lambda}^{-1}E[\mathbf{d}\mathbf{c}'] - \tfrac{1}{2}E[\mathbf{c}\mathbf{d}']\mathbf{\Lambda}^{-1} \tag{4.6}$$

and $\operatorname{var}(\bar{y}.) = \mathbf{\Gamma}_d/4$.

COMMENT. The correction subtracted from the average $\bar{\mathbf{y}}.$ to give the empirically weighted mean $\tilde{\boldsymbol{\mu}}$ is the imbalance factor $(\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1)(\mathbf{\Gamma}_2 + \mathbf{\Gamma}_1)^{-1}$ times the cutoff function, and the difference of their variances has the variance kernel $\mathbf{V}_k$ premultiplied by the imbalance factor and post multiplied by its transpose. The variance kernel $\mathbf{V}_k$ is the variance of the cutoff function $\mathbf{c}(\mathbf{d})$ minus terms which come from covariances of $\bar{\mathbf{y}}.$ with $\mathbf{c}$ and $\mathbf{c}$ with $\bar{\mathbf{y}}.$.

The variance kernel expresses the difference of the variances of $\tilde{\boldsymbol{\mu}}$ and $\bar{\mathbf{y}}.$ in the most extreme cases of imbalance. Its positive or negative definiteness and magnitude therefore express their relative accuracies.

PROOF OF THEOREM 2. Formula (4.3) follows from formula (4.1) by use of the identity

$$(I + \lambda\,\mathbf{d}\mathbf{a}')^{-1} = I - \lambda\,\mathbf{d}\mathbf{a}'/(1 + \lambda\mathbf{a}'\,\mathbf{d})$$

with $\lambda = (\chi^2 - 1)/\chi^2$ and $\mathbf{a} = \mathbf{\Gamma}^{-1}\mathbf{d}$.

In formula (4.5) there are terms for

$$\operatorname{cov}(\bar{\mathbf{y}}_\cdot, \mathbf{c}) = E\big[(\bar{\mathbf{y}}_\cdot - \boldsymbol{\mu})\mathbf{c}'\big] = E\big[E\big[(\bar{\mathbf{y}}_\cdot - \boldsymbol{\mu})|\mathbf{d}\big]\mathbf{c}'\big]$$

and its transpose. Since $\bar{\mathbf{y}}_\cdot$ regresses on $\mathbf{d}$, we have the conditional expectation

$$E\big[(\bar{\mathbf{y}}_\cdot - \boldsymbol{\mu})|\mathbf{d}\big] = \operatorname{cov}(\bar{\mathbf{y}}_\cdot, \mathbf{d})\boldsymbol{\Gamma}_d^{-1}\mathbf{d} = \tfrac{1}{2}(\boldsymbol{\Gamma}_2 - \boldsymbol{\Gamma}_1)\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{d}$$

and from this the covariance is obtained. □

4.1. *The scalar case.* In the scalar case, $n = 1$, we can judge magnitudes of the error relative to $\operatorname{var}(\bar{y}_\cdot) = \Gamma_d/4$, by studying the function $f(\Delta/\Gamma) = V_k/\operatorname{var}(\bar{y}_\cdot)$ such that

$$\frac{\operatorname{var}(\tilde{\mu}) - \operatorname{var}(\bar{y}_\cdot)}{\operatorname{var}(\bar{y}_\cdot)} = \left(\frac{\Gamma_2 - \Gamma_1}{\Gamma}\right)^2 f\left(\frac{\Delta}{\Gamma}\right).$$

THEOREM 3. *When* $n = 1$,

$$\tilde{\mu} = \bar{y}_\cdot - \frac{\Gamma_2 - \Gamma_1}{\Gamma} c(d),$$

where the cutoff function is

$$c(d) = \begin{cases} d/2, & \text{if } |d| \le \sqrt{\Gamma}, \\ \Gamma/(2d), & \text{if } |d| \ge \sqrt{\Gamma}. \end{cases}$$

Also

$$\operatorname{var}(\tilde{\mu}) = \operatorname{var}(\bar{y}_\cdot) + \left(\frac{\Gamma_2 - \Gamma_1}{\Gamma}\right)^2 \left\{\operatorname{var}(c) - \frac{E[dc]}{\Lambda}\right\}$$

and

$$f(\Delta/\Gamma) = g_1 - g_2,$$

where

$$\begin{aligned} g_1 &= \operatorname{var}(c)/\operatorname{var}(\bar{y}_\cdot) \\ &= \sqrt{2/(\pi\Lambda)}\,(\Lambda^{-1} - 1)\exp\{-1/(2\Lambda)\} + 2\Phi(\Lambda^{-1/2}) \\ &\quad - 1 - 2\{1 - \Phi(\Lambda^{-1/2})\}/\Lambda^2 \end{aligned}$$

and

$$\begin{aligned} g_2 &= \Lambda^{-1}E[dc]/\operatorname{var}(\bar{y}_\cdot) \\ &= -(2/\Lambda)^{3/2}\exp\{-1/(2\Lambda)\}/\sqrt{\pi} + 2\{2\Phi(\Lambda^{-1/2}) - 1\}/\Lambda \\ &\quad + 4\{1 - \Phi(\Lambda^{-1/2})\}/\Lambda^2, \end{aligned}$$

where $\Lambda = 1 + 2\Delta/\Gamma$ *and* $\Phi(z)$ *is the standard normal distribution function.*

The integrals were evaluated by hand, checked algebraically with the computer program MAPLE and checked numerically by simulation.

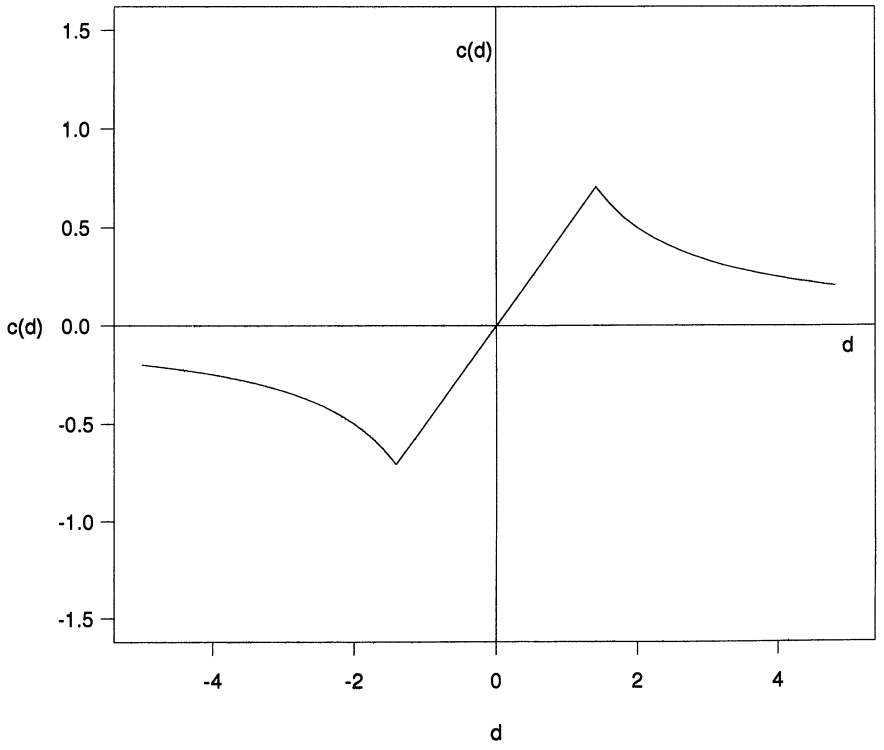

FIG. 4. *The cutoff function*, $\mathbf{c(d)}$, *for two sample scalar random effects models*, $\mathbf{c}$, *multiplied by the imbalance factor*, $(\Gamma_2 - \Gamma_1)/(\Gamma_2 + \Gamma_1)$, *gives the correction subtracted from the average*, $\bar{\mathbf{y}}_\cdot$, *to give the empirically weighted mean*, $\tilde{\boldsymbol{\mu}}$.

4.2. *The cutoff function.* The cutoff function, shown in Figure 4, is analogous to the influence function used by Hampel, Ronchetti, Rousseeuw and Stahel (1986) in robust regression, as shown in Figure 5. When the difference d is small, the estimate of Δ is zero by cutoff, and one corrects $\bar{y}_\cdot$ linearly in d for the variance imbalance of $\Gamma_1 \neq \Gamma_2$. When the effect of Δ appears in d from under the shadow of $\Gamma = \Gamma_1 + \Gamma_2$, however, the effect of $\hat{\Delta} > 0$, offsets the variance imbalance and gradually brings the weights back to the equality of those in $\bar{y}_\cdot$. To ignore Δ as in NPD, is to extrapolate the line to infinity both ways. This is analogous to accepting the full effect of an outlier in a least squares analysis with its consequent distortion, instead of downgrading its influence by the influence function in a robust regression. Not to apply cutoff is to extrapolate the hyperbola to $\pm\infty$ at $d = \pm 0$.

4.3. *Comparison of variances of estimators.* By the Gauss theorem, the variance of the truly weighted mean $\hat{\mu}$ with random effects, constitutes a lower bound, but such an estimator is not available due to ignorance of Δ. We shall

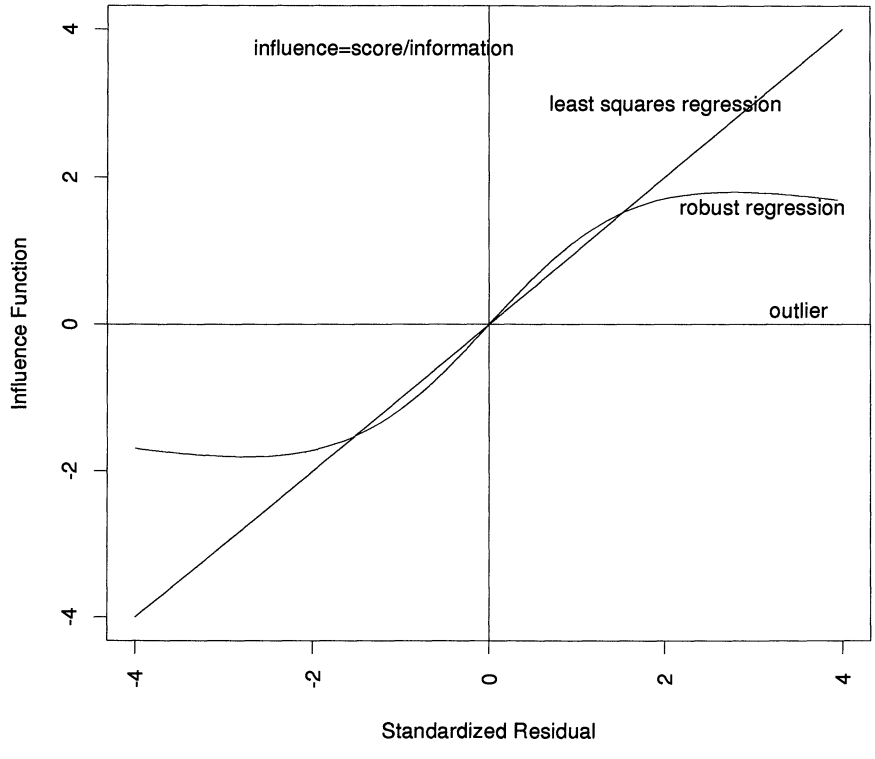

FIG. 5. *The influence function.*

compare variances of the empirically weighted mean with random effects $\tilde{\mu}$ the weighted mean without random effects μ^+ and the average $\bar{y}_.$.

For the various estimators, $m = \hat{\mu}, \tilde{\mu}, \mu^+$ of the mean, Figure 6 shows the function

$$\left(\left(\frac{\text{var}(m) - \text{var}(\bar{y}_.)}{\text{var}(\bar{y}_.)}\right) \Big/ \left(\frac{\Gamma_2 - \Gamma_1}{\Gamma_2 + \Gamma_1}\right)^2\right) \times 100\%$$

plotted against Δ/Γ. When $m = \tilde{\mu}$, the function is f and has a maximum of $f = 0.0080$ at $\Delta/\Gamma = 4.2$.

The abscissa constitutes the curve for $\bar{y}_.$.

For $\Delta/\Gamma \leq 2$, the empirically weighted mean $\tilde{\mu}$ has lower variance than the equally weighted mean $\bar{y}_.$ and the difference is substantial for $\Delta/\Gamma < 1$. For $\Delta/\Gamma > 2$, $\tilde{\mu}$ has a slightly higher variance than $\bar{y}_.$ but from a practical point of view, the difference is negligible. Hence, the empirically weighted mean is to be recommended.

For Δ/Γ from 0 to 0.3, one would do best by weighting without random effects. Unless one has definite information that $\Delta/\Gamma \leq 0.3$, however, one cannot be sure of this and empirical weighting should be used.

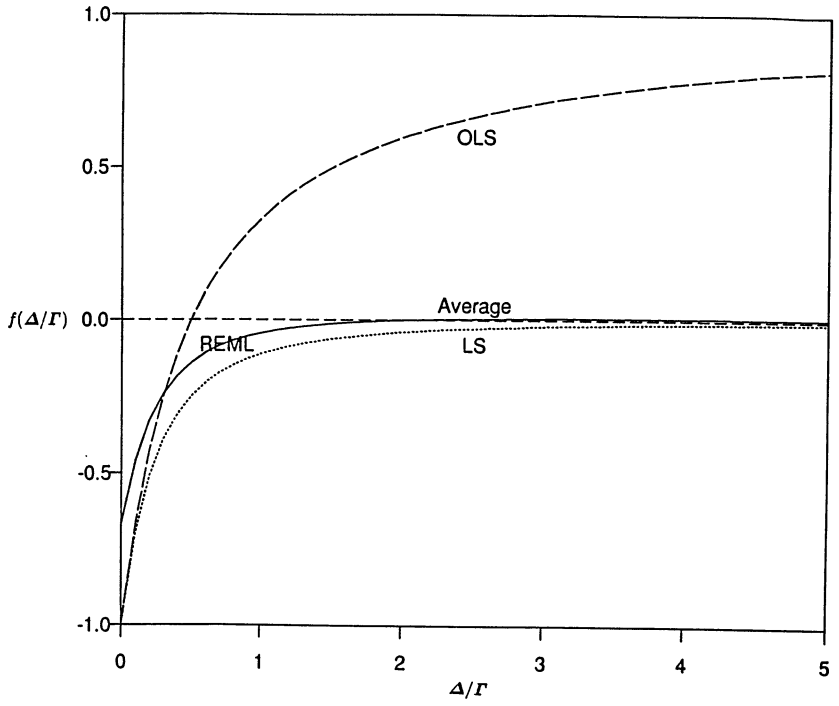

FIG. 6. *Difference of variance estimators as a percentage of the variance of the average.*

At about $\Delta/\Gamma = 1.5$, there is 50:50 chance that random effects will be significant, that is, $\Delta > 0$. Hence above 1.5, it is clear that random effects must be specified, but in this range, $\bar{y}_\cdot$ has about the same accuracy as $\tilde{\mu}$.

For $0.5 < \Delta/\Gamma < 1.5$, there is low power in the significance test, but the weighted mean without random effects μ^+ is highly inefficient. This is the dangerous region if one ignores Δ in the hopes that there are no random effects. Such a procedure is highly nonrobust. If the data is completely unbalanced with Γ_2/Γ_1 large, then if one is certain of no random effect, that is, that $\Delta = 0$, one can ignore y_2 and use y_1 as the estimator of μ. It has a negligible variance, and hence is 100% below the average, $\bar{y}_\cdot = (y_1 + y_2)/2$. In the absence of certain knowledge of no random effect, it is dangerous to ignore its possibility. If one allows for a possible random effect when in fact $\Delta = 0$, one pays a penalty that the variance of $\tilde{\mu}$ is only decreased to 70% of $\bar{y}_\cdot$.

Within the range $0 < \Delta/\Gamma < 1.5$, the empirically weighted mean $\tilde{\mu}$ has less error variance then the average $\bar{y}_\cdot$ and substantially less in the lowest part of this range. The results confirm the use of the empirically weighted mean.

5. The two dimensional case. In the case of two vectors $\mathbf{y}_1, \mathbf{y}_2 \in R^2$, $p = 1$, $n = 2$, we can compare the error variance matrices of the empirically

weighted mean $\tilde{\boldsymbol{\mu}}$ with the average $\bar{\mathbf{y}}_.$ by making a nonsingular linear transformation:

$$\mathbf{y}_i \to L\mathbf{y}_i, \qquad \tilde{\boldsymbol{\mu}} \to L\tilde{\boldsymbol{\mu}}, \qquad \bar{\mathbf{y}}_. \to L\bar{\mathbf{y}}_.$$

such that the induced congruence transformations map $\boldsymbol{\Gamma}$ to I and diagonalize $\boldsymbol{\Delta}$:

$$\boldsymbol{\Gamma} \to L\boldsymbol{\Gamma}L' = I, \qquad \boldsymbol{\Delta} \to L\boldsymbol{\Delta}L' = \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix}.$$

The variance kernel V_k given by (4.6) undergoes the same congruence transformation and becomes a diagonal matrix

$$\begin{bmatrix} z_1(\delta_1, \delta_2) & 0 \\ 0 & z_2(\delta_1, \delta_2) \end{bmatrix},$$

which is a function of δ_1, δ_2 with $z_2(\delta_1, \delta_2) = z_1(\delta_2, \delta_1)$.

There are two extreme cases of imbalance:

(a) Different but proportionate variance matrices,

$$\boldsymbol{\Gamma}_1 = 0, \qquad \boldsymbol{\Gamma}_2 = I, \qquad \boldsymbol{\Gamma}_2 - \boldsymbol{\Gamma}_1 = I;$$

this is analogous with the scalar case. Then

$$f_1(\delta_1, \delta_2) = \frac{\operatorname{var}(\tilde{\mu}_1) - \operatorname{var}(\bar{y}_{.1})}{\operatorname{var}(\bar{y}_{.1})} = \frac{4z_1(\delta_1, \delta_2)}{(1 + 2\delta_1)}. \tag{5.1}$$

(b) Difference of correlation,

$$\boldsymbol{\Gamma}_1 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}, \qquad \boldsymbol{\Gamma}_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \qquad \boldsymbol{\Gamma}_2 - \boldsymbol{\Gamma}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix};$$

this is the case where NPD leads to range anomaly. We have

$$\begin{aligned} f_2(\delta_1, \delta_2) &= \frac{\operatorname{var}(\tilde{\mu}_1) - \operatorname{var}(\bar{y}_{.1})}{\operatorname{var}(\bar{y}_{.1})} = \frac{4z_2(\delta_1, \delta_2)}{1 + 2\delta_1} \\ &= \frac{4z_1(\delta_2, \delta_1)}{1 + 2\delta_1}. \end{aligned} \tag{5.2}$$

If we put $d_1 = r\cos(\theta)$ and $d_2 = r\sin(\theta)$, we have the following:

THEOREM 4. *For the two sample two dimensional case, we have*

$$z_1(\delta_1, \delta_2) = \frac{1}{2}\zeta_1 + \zeta_3 - \frac{\zeta_1 + \zeta_2}{1 + 2\delta_1},$$

where

$$\zeta_1 = \frac{1}{\pi\sqrt{\delta_1\delta_2}} \int_{\theta=0}^{x/2} \int_{r=0}^{1} \exp\left(-\frac{1}{2} r^2 \left\{ \frac{\cos^2(\theta)}{\delta_1} + \frac{\sin^2(\theta)}{\delta_2} \right\} \right) r^3 \cos^2(\theta)\, dr\, d\theta,$$

$$\zeta_2 = \frac{1}{\pi\sqrt{\delta_1\delta_2}} \int_{\theta=0}^{\pi/2} \int_{r=1}^{\infty} \exp\left(-\frac{1}{2} r^2 \left\{ \frac{\cos^2(\theta)}{\delta_1} + \frac{\sin^2(\theta)}{\delta_2} \right\} \right) r \cos^2(\theta)\, dr\, d\theta,$$

$$\zeta_3 = \frac{1}{2\pi\sqrt{\delta_1\delta_2}} \int_{\theta=0}^{\pi/2} \int_{r=1}^{\infty} \exp\left(-\frac{1}{2} r^2 \left\{ \frac{\cos^2(\theta)}{\delta_1} + \frac{\sin^2(\theta)}{\delta_2} \right\} \right) r^{-1} \cos^2(\theta)\, dr\, d\theta.$$

The integrals have been evaluated by MAPLE and checked by simulation.

Contour plots for the functions in (5.1) and (5.2) are shown in Figure 7 and 8. As in the univariate case, one sees that $\text{var}(\tilde{\mu}_1)$ is substantially less than $\text{var}(\bar{y}_{.1})$ for small $\boldsymbol{\Delta}$, and generally only negligibly greater for large $\boldsymbol{\Delta}$. Usually f_1 and f_2 are negative or less than 0.01, but in case (b), f_2 rises to a maximum of 0.0575 at $\delta_1 = 0$ and $\delta_2 = 18.5$. This appreciable value comes about in an extreme case in which $\bar{y}_{.1}$ has a small variance of 0.25 but the

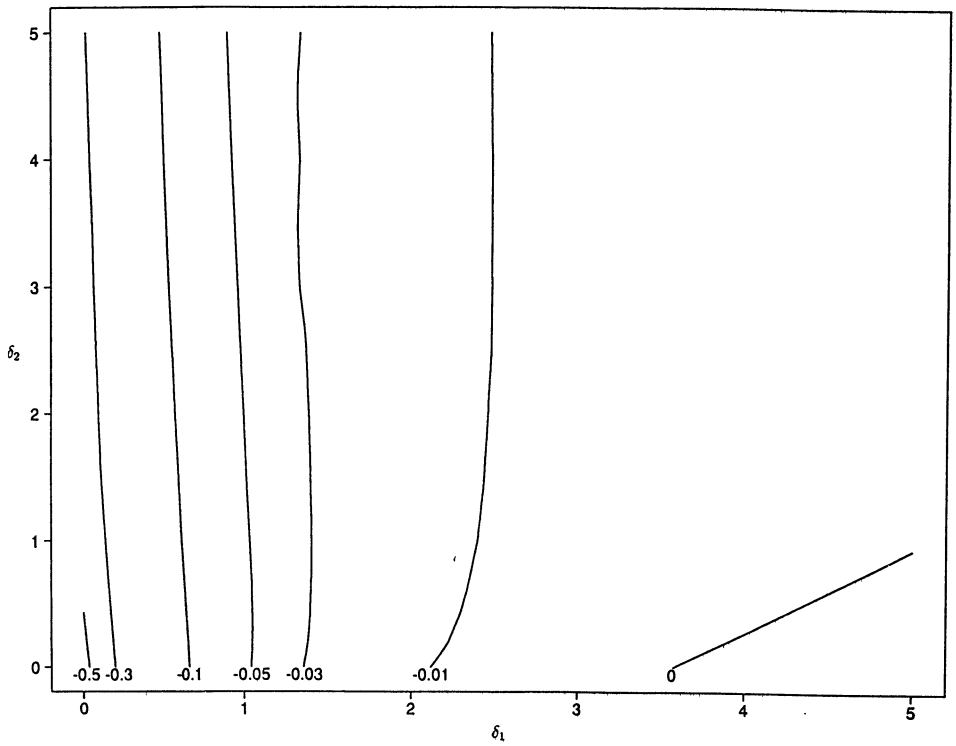

FIG. 7. *Bivariate case (a).*

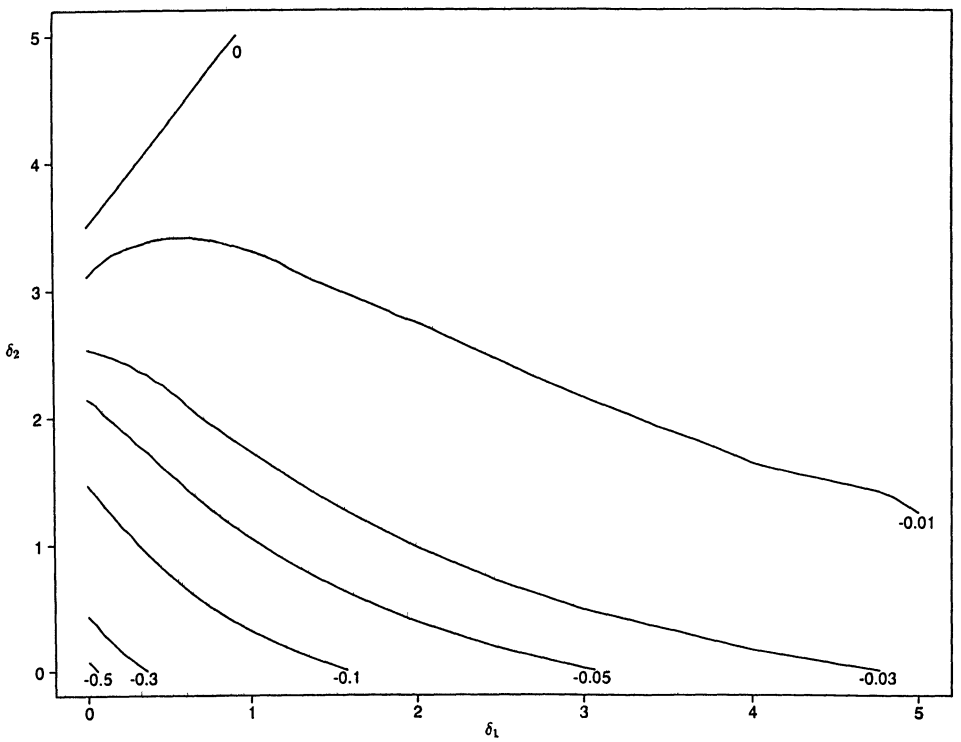

FIG. 8. *Bivariate case (b).*

correction to $\bar{y}_{.1}$ to obtain the weighted mean involves d which has a variance of 37, giving a variance of 0.27 for the weighted mean component $\boldsymbol{\mu}_{.1}$.

For bivariate data, the dangers of ignoring a random effect when present are worse than for univariate. If, on the other hand, one allows for a possible random effect when in fact $\boldsymbol{\Delta} = \mathbf{0}$, the penalties for both completely unbalanced cases (a) and (b) are a variance reduction of 56% instead of 100% below the variance of the average $\bar{\mathbf{y}}$.

6. Comment. For the example in Section 2, the 5×5 matrix REML estimator $\hat{\boldsymbol{\Delta}}^{(m)}$ has a log likelihood which is 0.83 higher than that for the moment estimator $\boldsymbol{\Delta}^{(m)}$. An increase of 0.83 in log likelihood is an appreciable indicator of a better estimator. Both $\hat{\boldsymbol{\Delta}}$ and $\boldsymbol{\Delta}^{(m)}$ have rank 3 and, as one would expect, have practically identical ranges. Indeed, the cosines of the critical angles between their range spaces, which are given by the canonical correlations between the first three columns of the respective matrices, are 0.9839, 0.9995 and 1. But the 3×3 nonsingular matrices $\hat{\boldsymbol{\Delta}}_{11}$ and $\boldsymbol{\Delta}_{11}^{(m)}$ in their top left-hand corners, differ appreciably, as is shown by the eigenvalues of $\boldsymbol{\Delta}_{11}^{-1}\boldsymbol{\Delta}_{11}^{(m)}$, which are 0.71, 1.75 and 1.05. This may be due to the following situation.

Since the moment estimator $\Delta^{(m)}$ of Δ involves the average $\bar{\Gamma}_.$ of the Γ_i, it is strongly affected by the most inaccurate experiment, number 3, with a variance matrix Γ_3 much greater than the other Γ_i. By contrast, the efficient REML estimator has a score involving the information matrices Γ_i^{-1} and is hence predominantly influenced by the accurate experiments with low error variance Γ_i, $i \neq 3$.

6.1. *Marginal inefficiency and conditional bias.* For normally distributed observations $\mathbf{y}_i$ for which the variance matrices $\Delta + \Gamma_i$ are known, at least to the point of having correct relative weights, there is no point in considering other estimators than the least squares estimator, because this is the unbiased minimum variance estimator. Furthermore, since it is sufficient, it has no nontrivial ancillary statistics.

When Δ has to be estimated and consequently the least squares estimator is not known exactly, it is then useful to compare one estimator such as a weighted mean, with estimated weights $(\Gamma_i + \hat{\Delta})^{-1}$ with another estimator such as an average $\bar{\mathbf{y}}_.$ by expressing the weighted mean as the simple average plus multiples of differences of the observations $\mathbf{y}_i - \mathbf{y}_j$ which are ancillary to the estimation of the mean. Inference concerning the mean should be conditional on the differences because they are ancillary. In the conditional dis-

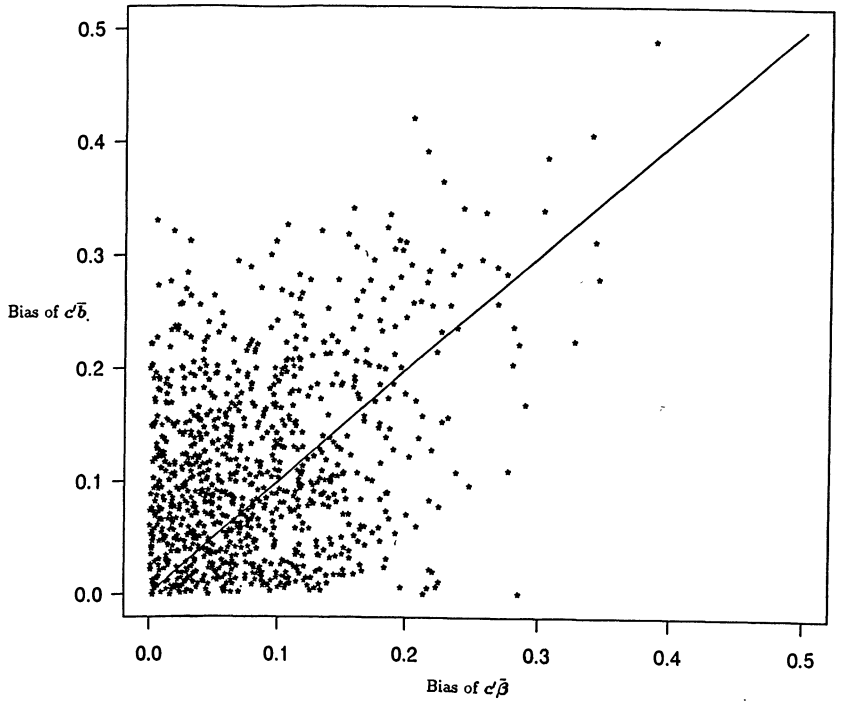

FIG. 9. *Absolute value of conditional biases from a simulation of* 1000.

tribution, the differences are constants with no variance. Consequently, all marginally unbiased weighted means have the same conditional variances irrespective of their weights, and this must be the minimum variance $(\sum_{i=1}^{p+1}(\boldsymbol{\Delta} + \boldsymbol{\Gamma}_i)^{-1})^{-1}$ of the least squares estimator.

Of such conditionally minimum variance estimators, however, only the least squares estimator is conditionally unbiased. The conditional bias of any other estimator is its difference from the least squares estimator, and in fact the marginal expected mean square conditional bias of an estimator equals the increase of its marginal variance over the variance of the least squares estimator.

For data with unknown variances, these considerations are of no use, but in simulations for which a variance matrix has been set, the least squares estimates and hence the conditional biases can be computed and used to illustrate what effect inefficiency of an estimator, having a marginal variance above the least squares estimator, produces in conditional bias. Figure 9 shows the biases of a certain contrast $\mathbf{a}'\bar{\mathbf{y}}_.$ of $\bar{\mathbf{y}}_.$ plotted against those of the same contrast of $\tilde{\boldsymbol{\mu}}$. The contrast will be explained in the next section.

7. Simulation. Since the data in Section 2 has highly significant differences between experiments $\chi^2_{25} = 913$, a random effects model is required. Attempts to avoid it have been shown to lead to serious anomalies.

For the least favourable case $p = 1$ the mathematical theory developed in Sections 4–6 shows, at least for $n = 1, 2$, that REML will achieve lower error variance than the average for small $\boldsymbol{\Delta}$, and only negligibly larger variance for large $\boldsymbol{\Delta}$. One would expect the same to hold for the more favourable cases of $p > 1$. To confirm this we turn to simulation.

An approximation to the error variance matrix of the moment estimator $\boldsymbol{\beta}^{(m)}$ can be found as the mean square sum and product matrix $\sum_{i=1}^{N}\boldsymbol{\beta}_i^{(m)}\boldsymbol{\beta}^{(m)\prime}/N$ of moment estimates $\boldsymbol{\beta}_i^{(m)}$ of the zero vector $\boldsymbol{\beta} = \mathbf{0}$, obtained from a large number N of samples generated by random numbers. The REML estimator could also be simulated. It tended to agree with the moment estimator relative to the average, but being iterative, was too slow for large scale simulation.

At $\boldsymbol{\Delta} = \hat{\boldsymbol{\Delta}}$, $N = 600{,}000$ simulations were done. The coefficient of variation of the variance should be $\sqrt{2/N} \times 100\% = 0.18\%$. Two variance matrices V_1, V_2 can be compared by taking the eigenvalues of $V_1^{-1}V_2$. The simulated variance matrices of the average and least squares estimators were compared with their theoretical values $(\boldsymbol{\Delta} + \bar{\boldsymbol{\Gamma}}_.)/6$ and $(\Sigma(\boldsymbol{\Delta} + \boldsymbol{\Gamma}_i)^{-1})^{-1}$. In each case the eigenvalues came out close to one, the average discrepancies being about 0.25%, a figure comparable with 0.18%.

The variances are given in Table 3.

From component parts of the simulation, we have noticed some correlation in the simulated variances. Hence it may be slightly more accurate to make comparisons with the moment estimator between simulated values even when theoretical values are available for the error variance matrices of the least squares estimator and the average.

TABLE 3
Variances of estimators × 100

Theoretical least squares	0.1672	0.1709	0.4942	2.5988	0.3040
Average	0.1777	0.1817	0.5089	2.6778	0.3241
Simulated least squares	0.1677	0.1714	0.4945	2.6057	0.3050
Moment	0.1734	0.1776	0.5031	2.6498	0.3151
Average	0.1784	0.1825	0.5097	2.6871	0.3254

TABLE 4
Eigenvalues measuring efficiencies and relative efficiencies

$\text{var}(\bar{\mathbf{b}}_\cdot)^{-1}\,\text{var}(\boldsymbol{\beta}^{(m)})$	0.869	0.965	0.985	1.000	0.999
$\text{var}(\boldsymbol{\beta}^{(m)})^{-1}\,\text{var}(\hat{\boldsymbol{\beta}})$	0.853	0.950	0.980	0.998	0.999
$\text{var}(\bar{\mathbf{b}}_\cdot)^{-1}\,\text{var}(\hat{\boldsymbol{\beta}})$	0.741	0.916	0.965	0.997	0.999

One sees that the variances of the averages are about 2–3% higher than for the moment estimator which in turn is about the same amount higher than for the least squares estimator.

The variance matrices, $\text{var}(\bar{\mathbf{b}}_\cdot)$ and $\text{var}(\boldsymbol{\beta}^{(m)})$ can be compared by the eigenvalues of the inverse of the first times the second, because if $\mathbf{a}$ is an eigenvector corresponding to the eigenvalue λ, that is, $\text{var}(\bar{\mathbf{b}}_\cdot)\mathbf{a} = \text{var}(\boldsymbol{\beta}^{(m)})\mathbf{a}\lambda$, then $\lambda = (\mathbf{a}'\,\text{var}(\boldsymbol{\beta}^{(m)})\mathbf{a})/(\mathbf{a}'\,\text{var}(\bar{\mathbf{b}}_\cdot)\mathbf{a})$ is the ratio of the variances of the estimates of the contrast $\mathbf{a}'\boldsymbol{\beta}$, that is, the efficiency of $\mathbf{a}'\bar{\mathbf{b}}_\cdot$ relative to $\mathbf{a}'\boldsymbol{\beta}^{(m)}$. The absolute efficiencies are found by eigenvalues relative to the least squares estimator $\hat{\boldsymbol{\beta}}$ as in Table 3.

One sees that all the eigenvalues are close to 1 except for the first, which has as an eigenvector for $\text{var}(\bar{\mathbf{b}}_\cdot)^{-1}\,\text{var}(\boldsymbol{\beta}^{(m)})$, the contrast

$$\mathbf{a}' = [\,0.83 \quad -0.19 \quad 0.13 \quad -0.10 \quad 0.51\,]$$

The first eigenvectors of the second and third matrices are practically identical with $\mathbf{a}$, having cosines of angles with it greater than 0.998 relative to the inner product given by the least squares variance matrix $\text{var}(\hat{\boldsymbol{\beta}})$. The closeness of the first eigenvectors show why the efficiency 0.74 of $\bar{\mathbf{b}}_\cdot$ is practically equal to its efficiency 0.87 relative to $\boldsymbol{\beta}^{(m)}$ times the efficiency 0.85 of $\boldsymbol{\beta}^{(m)}$ (relative to $\hat{\boldsymbol{\beta}}$).

As explained in subsection 6.2, the inefficiencies of $\boldsymbol{\beta}^{(m)}$ and $\bar{\mathbf{b}}_\cdot$ show up in their conditional biases which for the contrasts, $\mathbf{a}'\boldsymbol{\beta}^{(m)}$ and $\mathbf{a}'\bar{\mathbf{b}}_\cdot$ are plotted in Figure 9 for a simulation of $N = 1000$.

8. Discussion. For small sample random effects, the notion that the large sampling error of the estimated random effects variance matrix, $\hat{\boldsymbol{\Delta}}$ must produce much larger sampling error $\text{var}(\hat{\boldsymbol{\beta}})$ in the empirically weighted mean $\hat{\boldsymbol{\beta}}$ than the error $\text{var}(\bar{\mathbf{b}}_\cdot)$ of the average $\bar{\mathbf{b}}_\cdot$ is contradicted in the special case of two samples, $p = 1$ and $n = 1$ and 2 dimensions. Our mathematical theory of

Sections 4 and 5 shows that var($\hat{\boldsymbol{\beta}}$) can be considerably less than var($\bar{\mathbf{b}}_.$) for small $\boldsymbol{\Delta}$, and is barely greater for large $\boldsymbol{\Delta}$.

Since $p = 1$ is the most unfavourable case in which $\hat{\boldsymbol{\Delta}}$ has maximum error, it strongly indicates that var($\hat{\boldsymbol{\beta}}$) will compare even more favourably with var($\bar{\mathbf{b}}_.$) when $p > 1$. The range anomaly of NPD or OLS, illustrated in Section 3, and its correction by inclusion of $\hat{\boldsymbol{\Delta}}$ in the weights, supplied heuristic evidence of this. Furthermore, near the end of Section 3, it was pointed out that it is the difference $\mathbf{d}$ not $\boldsymbol{\Delta}$, which is the direct cause of range anomaly, and the fact that $\hat{\boldsymbol{\Delta}}$ is a function of $\mathbf{d}$ is the reason why the inclusion of $\hat{\boldsymbol{\Delta}}$ in weights rectifies the range anomaly. Such arguments also apply for more than two samples.

Namely, if in some direction, $\hat{\boldsymbol{\Delta}}$ seriously underestimates $\boldsymbol{\Delta}$, leaving the weighted mean subject to highly different variance matrices, $\hat{\boldsymbol{\Delta}} + \boldsymbol{\Gamma}_i$ and $\hat{\boldsymbol{\Delta}} + \boldsymbol{\Gamma}_j$, then this is because the differences $\mathbf{b}_i - \mathbf{b}_j$ are small in that direction, that is, the $\mathbf{b}_i$ in that direction are close together and hence errors in the weights in that direction will have little effect on the weighted mean.

If $\hat{\boldsymbol{\Delta}}$ is singular, then the components of the differences $\mathbf{b}_i - \mathbf{b}_j$ in its null space tend to be small. Hence the empirically weighted mean is little affected by part of $\boldsymbol{\Delta}$ being missing from its estimator $\hat{\boldsymbol{\Delta}}$.

If on the other hand, $\hat{\boldsymbol{\Delta}}$ overestimates $\boldsymbol{\Delta}$, then the weights will become more equal than they should thereby pushing $\hat{\boldsymbol{\beta}}$ towards the conservative average $\bar{\mathbf{b}}_.$. This also explains why, when $\boldsymbol{\Delta}$ is zero or negligible but one cannot be sure of this, there can be considerable loss of accuracy in incorporating $\hat{\boldsymbol{\Delta}}$ in the weights. In spite of this, the use of $\hat{\boldsymbol{\Delta}}$ is still to be recommended because if $\boldsymbol{\Delta}$ were appreciable, the omission of $\hat{\boldsymbol{\Delta}}$ would lead to a high error which would appear as a conditional bias.

To put the matter to the test for more than two random samples $p > 1$, however, in Section 5 we simulated the situation of the practical example in Section 2. Although the standard errors of the moment estimator $\mathbf{b}^{(m)}$ are barely less than those of the average $\bar{\mathbf{b}}_.$ one of the five canonical components has an efficiency for $\bar{\mathbf{b}}_.$ of only 87% of that for $\mathbf{b}^{(m)}$. The effect of this in the simulated conditional biases of the corresponding contrasts $\mathbf{a}'\boldsymbol{\beta}^{(m)}$ and $\mathbf{a}'\bar{\mathbf{b}}_.$ are shown in Figure 9.

The simulations are also valuable for producing standard errors for the moment estimators which allow for the error of $\boldsymbol{\Delta}^{(m)}$. We believe that in this case, they will be close to the standard errors of the REML estimator, $\hat{\boldsymbol{\beta}}$.

Even if one can assure oneself that whatever the value of $\boldsymbol{\Delta}$, the weighted mean compares favourably with the sample average, their estimated errors are only on low degrees of freedom p.

The integration and simulation prove the validity of empirically matrix weighted means in the cases to which they apply, and they indicate this is generally so, though when $\boldsymbol{\Delta}$ is clearly large relative to $\bar{\boldsymbol{\Gamma}}_.$ the method 2 of simple averages will work just as well.

9. Conclusions. Estimated weights can always be used in unbalanced small random effects models.

There is never any appreciable loss of accuracy compared with a simple average, but there can be a considerable gain if the random effects are small.

The NPD or OLS of method 1 should not be used unless one has definite prior knowledge that there are no random effects.

Simulation can be used to find the error variance of a moment estimator of a mean.

Acknowledgments. We gratefully acknowledge considerable help from Dr. Arunas P. Verbyla and also the work of Clarence Tiong in the mathematics, statistics, computing and preparation of this paper.

REFERENCES

AMENIYA, Y. (1985). What should be done when an estimated between-group covariance matrix is non negative definite? *Amer. Statist.* **39** 112–117.

FELDMAN, H. A. (1988). Families of lines: Random effects in linear regression analysis. *Journal of Applied Physiology* **64** 1721–1732.

GUMPERTZ, M. and PANTULA, S. G. (1989). A simple approach to inference in random coefficient models. *Amer. Statist.* **43** 203–210.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

HENDERSON, C. A., KEMPTHORNE, O., SEARLE, S. R. and VON KROSIGK, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15** 192–218.

JAMES, A. T., WISKICH, J. T. and CONYERS, R. A. J. (1989). The statistical modelling of mitochondrial power supply. *J. Math. Biol.* **27** 29–48.

JAMES, A. T., WISKICH, J. T. and CONYERS, R. A. J. (1993). *t*-REML for robust heteroscedastic regression analysis of mitochondrial power. *Biometrics* **49**. To appear.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ADELAIDE
NORTH TERRACE
ADELAIDE, SA 5001
AUSTRALIA