A Predictive Approach to the Random Effect Model

by Seymour Geisser

Technical Report No. 205

April, 1973

University of Minnesota

Minneapolis, Minnesota

A Predictive Approach to the Random Effect Model

by Seymour Geisser

University of Minnesota

1. Introduction

Suppose one has K_j observations on the j^{th} treatment, batch or group, $j=1,\ldots,J$ and one wishes to estimate the treatment means, then there are basically two "Bayesian" ways of handling the problem. The first is to use $\bar{x}_{\cdot j}$, the sample treatment mean as the estimator of θ_j the population mean as in the classical case $j=1,\ldots,J$, or secondly to utilize a weighted combination of the j^{th} treatment mean and a grand mean a la Stein. From a Bayesian point of view and slightly different prior distributions both of these estimators are given a justification by Box and Tiao (1968) for $K_j=K$. The first is designated the fixed effect model and is based on a linear normal model with "uninformative" priors for the θ_j's and the common variance. The second, termed the random effect model, is derived from a linear normal hierarchical model with the location parameters θ_j having independent prior normal distributions with common mean and variance. The latter mean and the logarithm of a linear function of the two variances along with the logarithm of the basic observational variance are all assigned uniform priors. We note that the priors having to do with the variances are found objectionable by Lindley (1971) for several reasons chief among them is that both variances should be assigned proper priors, independent of the sample size K. We shall here examine this problem from the point of view of predicting new observations one from each of the J groups. If one were to utilize the results presented by Box-Tiao and use the mean of the predictive distribution of the set of J

future observations then one would be using the posterior means of θ_j as predictors under each model. This is essentially true whenever $E(x_{kj}|\theta_j) = \theta_j$ as this implies that posterior expectation of θ_j is equivalent to the predictive expectation of a new observation when both latter expectations exist. In the first case this is $\bar{x}_{\cdot j}$ and in the second $(1-\bar{\mu})\bar{x}_{\cdot j} + \bar{\mu}\bar{x}_{\cdot\cdot}$ where for $0 \leq \bar{\mu} \leq 1$

$$\bar{\mu} = \frac{B(\frac{J+1}{2}, \frac{J(K-1)-2}{2})}{B(\frac{J-1}{2}, \frac{J(K-1)}{2})} \frac{I_y(\frac{J+1}{2}, \frac{J(K-1)-2}{2})}{I_y(\frac{J-1}{2}, \frac{J(K-1)}{2})} \frac{J(K-1)m_1}{(J-1)m_2} ; \quad (1.1)$$

$$y = \frac{(J-1)m_2}{(J-1)m_2 + J(k-1)m_1} ; \quad B(p,q) \text{ and } I_y(p,q) \text{ are,}$$

respectively, the complete and the incomplete beta functions and

$$m_1 = J^{-1}(k-1)^{-1} \sum_{k=1}^{K}\sum_{j=1}^{J}(x_{kj} - \bar{x}_{\cdot j})^2 , \quad \bar{x}_{\cdot j} = K^{-1} \sum_{k=1}^{K} x_{kj} ; \quad (1.2)$$

$$m_2 = K(J-1)^{-1} \sum_{j=1}^{J} (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2 , \quad \bar{x}_{\cdot\cdot} = J^{-1} \sum_{j=1}^{J} x_{\cdot j} .$$

We shall focus on the problem of prediction from a heuristic data analytic point of view and incidentally suggest that the methods of prediction can also be used to generate estimates of the θ_j if this is desirable. Before doing this we shall digress to present the "correct" solution.

Suppose one wished to infer which prior distribution was more appropriate for predicting from a given set of data and a given likelihood. Under each model π_i we would compute

$$p_1 f(X|\pi_1)/p_2 f(X|\pi_2) \quad (1.3)$$

where p_i represents the prior probability the Model π_i is correct where X represents the data set and

$$f(X|\pi_i) = \int f(X|\alpha_i)dG_i(\alpha_i),$$

where $G_i(\alpha_i)$ represents the prior distribution of α_i under model π_i. Strictly speaking this method will only be useful when p_i is assumed known and $G_i(\alpha_i)$ is a completely proper prior that depends only on known constants! When the priors are not proper it would be necessary to utilize this procedure on a new data set Y where one would insert in (1.3), $f(Y|X,\pi_i)$, the predictive density of Y given X and the model.

We shall assume here that p_i is not known and we have only the original data set. Further in the Box and Tiao development the prior models are completely or partially improper. Hence none of the aforementioned possibilities exist.

We propose here two data analytic methods that may be of some value in this case of discriminating between the models. They are also capable of producing "best" estimates with regard to a least squares type criteria.

2. Method I.

The first modified predictive method for evaluating alternative procedures giving rise to estimators of the form $(1-\mu)\bar{x}._j + \mu\bar{x}..$ with respect to μ involves omitting say the k^{th} observation in the j^{th} group and computing the estimator $(1-\mu)c_{kj} + \mu\bar{c}_{kj}$ from the $N-1$ remaining observations where $N = \sum_{j=1}^{J} K_j$, $K_j \geq 2$ and

$$c_{kj} = (K_j\bar{x}._j - x_{kj})/(K_j - 1) \; ; \; \bar{x}._j = K_j^{-1}\sum_k x_{kj}$$

$$c_{kj} = (N\bar{x}.. - x_{kj})/(N-1); \; \bar{x}.. = N^{-1}\sum_{k,j} x_{kj} \; . \tag{2.1}$$

This produces a predictor for x_{kj}, namely $(1-\mu)c_{kj} + \mu\bar{c}_{kj}$, then repeating this for all k and j we evaluate the mean squared prediction error

$$s_\mu^2 = N^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K_j}[(1-\mu)c_{kj} + \mu\bar{c}_{kj} - x_{kj}]^2 \; . \tag{2.2}$$

It is to be noted that this need not be a sensible measure unless the variance for each observation is the same. We shall assume that this holds. Use of (2.1) above and the usual identities of the analysis of variance leads to

$$s_\mu^2 = \frac{\mu^2 N}{(N-1)^2}\sum_{j=1}^{J} K_j(\bar{x}._j - \bar{x}..)^2 + \sum_{j=1}^{J}\left[\frac{(1-\mu)^2K_j^2}{N(K_j-1)} + \frac{\mu^2N(K_j-1)}{(N-1)^2}\right.$$

$$\left. + \frac{2\mu(1-\mu)K_j}{N-1}\right] s_j^2 \tag{2.3}$$

where $s_j^2 = (K_j - 1)^{-1}\sum_{j=1}^{K_j}(x_{kj} - \bar{x}_{ij})^2$.

Hence (2.3) could be appropriate for evaluating various values of μ, in particular if $\mu=0$ then

$$s_0^2 = N^{-1}\sum_{j=1}^{J}\frac{K_j^2 s_j}{K_j - 1} \; . \tag{2.4}$$

If it were desirable to find a "best" μ from this technique then minimization of s_μ^2 with respect to μ yields

$$\hat{\mu}_1 = \frac{(N-1)\sum_{j=1}^{J} \frac{K_j(N-K_j)s_j^2}{K_j - 1}}{N^2 \sum_{j=1}^{J} K_j(\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_{j=1}^{J} \frac{(N-K_j)^2}{K_j-1} s_j^2} \quad . \tag{2.5}$$

Due to the constraint the estimator is $\min(\hat{\mu}_1, 1)$ which we designate as the Method I estimator.

For the special case $K_j=K$ treated by Box and Tiao (1968) we obtain the following simplifications

$$s_\mu^2 = \frac{K[JK-1-\mu(J-1)]^2}{(K-1)(JK-1)^2} m_1 + \frac{\mu^2 JK(J-1)}{(JK-1)^2} m_2 \tag{2.6}$$

where m_1 and m_2 are as defined in (1.2) and

$$\hat{\mu}_1 = \frac{(JK-1)m_1}{(J-1)m_1 + (K-1)Jm_2} \tag{2.7}$$

so that the estimator is $\min[\hat{\mu}_1, 1]$. As $K(J)$ increases for $J(K)$ held constant, the estimator tends to $\min(m_1 m_2^{-1}, 1)$. In particular we note for $\mu = 0$

$$s_0^2 = \frac{K}{K-1} m_1 \quad . \tag{2.8}$$

We refer to this as a modified method since if a model or a method for deriving μ is proposed and it depends on the data then μ will, strictly speaking, vary for each observation omitted. Computation of the mean squared

error of prediction would serve as one means of evaluating the various methods of estimating μ for the data set. This can easily be accomplished by computer for alternative estimators. However, the Box-Tiao method is not immediately adaptable to cases where the K_j's are different so one could not produce estimates when a single $K_j = K-1$ as this assessment requires. Hence we utilize the μ computed from all the data in the comparison assuming that the variation in μ for any particular method when an observation is omitted will in general minimally influence the comparison for reasonable sample sizes. The assumption that μ may for all intensive purpose, be considered constant permits us also to use the least squares type procedure to generate a "best" estimate for μ. It is also quite likely that for any method of estimating μ, s^2_μ will be smaller for a μ computed from all the data than computing μ separately for each omitted data point. One then could also utilize $(1-\hat{\mu}_1)\bar{x}._j + \hat{\mu}_1\bar{x}..$ as the estimate of θ_j, working backwards.

3. Method II

For the second method we assume $K_j=K$ from the start as the more general case involves heavier algebra. Here we simultaneously omit a single observation from each of the J groups and assume a predictor of the same form $(1-\mu)\bar{x}._j + \mu\bar{x}..$ $j=1,\ldots,J$. From the reduced data set we obtain the predictor $(1-\mu)m_{kj} + \mu\bar{m}_{kj}$ by arbitrarily reordering where m_{kj} and $\bar{m}_{kj}$ are the j^{th} group mean and the grand mean, based respectively on K-1 and J(K-1) observations. We then compute the squared deviation of the predicted value from the actual value for every possible configuration of the KJ data set i.e. permuting the observational values within the groups where we are predicting simultaneously the entire omitted row. This is then divided by the total number of such configurations. The algebra

again generally follows the traditional break down of sum of squares in the analysis of variance, and we obtain with m_1 and m_2 as defined previously

$$t_\mu^2 = \left[\frac{(k-\mu)^2 + J^{-1}(2K - \mu)\mu}{K(K-1)}\right] m_1 + (JK)^{-1}(J-1)\mu^2 m_2 \qquad (3.1)$$

Hence the relative evaluation of particular μ's given a set of data is possible by computing t_μ^2. For example, if $\mu=0$, which is essentially the estimator for the fixed effect model we have as before

$$t_0^2 = \frac{K}{K-1} m_1 . \qquad (3.2)$$

If we wish to derive from this data analytic procedure the "best" estimator for μ, we can minimize t_μ^2 with respect to μ. This yields

$$\hat{\mu}_2 = \frac{Km_1}{(K-1)m_2 + m_1} \qquad (3.3)$$

so that the estimator is $\min(\hat{\mu}_2,1)$. As k increases the estimator tends to $\min(m_1 m_2^{-1},1)$. The estimator for the full set of data for the prediction of a new set of observations one from each of the J groups is the $(1-\hat{\mu}_2)\bar{x}_{\cdot j} + \hat{\mu}_2\bar{x}_{\cdot\cdot}$ for $j=1,\ldots,J$. Similarly one can use this to estimate θ_j as well.

We note that this case, omitting simultaneously a single observation from each group, does permit use of Box-Tiao estimators since K has been reduced by 1, however the number of separate analyses involved is of the order of K^J. Hence again we have opted for the modified version which keeps μ constant.

4. Example

Consider the problem of estimating θ_j for the example given by Box and Tiao (1968)

Dyestuff Data

	Batch					
	1	2	3	4	5	6
	145	140	195	45	195	120
	40	155	150	49	230	55
	40	90	205	195	115	50
	120	160	110	65	235	80
	180	95	160	145	225	45
Average	105	128	164	98	200	70

$J = 6$, $K = 5$, $m_1 = 2{,}451.25$, $m_2 = 11{,}271.50$, $\bar{x}.. = 127.5$

	Predicted Value or Estimate of θ_i										
Method	1	2	3	4	5	6	μ	s^2_μ	s_μ	t^2_μ	t_μ
"fixed effect"	105	128	164	98	200	70	0	3064	55.4	3064	55.4
"random effect"	110	128	156	105	183	83	.233	2932	54.2	2934	54.2
"Predictive" I	111	128	155	105	182	84	.251	2931	54.1	2932	54.1
"Predictive" II	111	128	155	106	181	85	.258	2931	54.1	2932	54.1
Asymptotic $\mu=m_1m_2^{-1}$	110	128	156	104	184	82	.217	2934	54.2	2936	54.2

It is to be noted that the two predictive methods yield very close estimates. We cannot discriminate between them, whether we use s^2_μ or t^2_μ to compare them. The random effect estimator while exhibiting an 8% difference from the predictive estimates for the weighting coefficient again exhibits s^2_μ and t^2_μ to be extraordinarily close to the minima for both. This indicates that both s^2_μ and t^2_μ are, for this set of data, fairly flat near their minima and variations in the weighting coefficient do not yield appreciable changes in the comparison functions.

This is further borne out by the asymptotic case where a change in μ of close to 20% from the "best" results in practically no change in the mean squared predictive error. In fact, even the fixed effect model increases the mean squared predictive error by only about 4%.

Lindley (1971) has also put forth a Bayesian procedure for the analysis of this type of data. We have hesitated to compute Lindley's estimators since it would require arbitrarily assigning some numbers to the priors and the modal iterations look prohibitive. It would be of great interest to see how it would fare under this type of assessment.

5. Mixed Model

Suppose we are in the mixed model case i.e. we have K vector observations each of J components. Let us assume that we are considering predictors of the components of a future $x_k' = (x_{k1},\ldots,x_{kj})$ of the form $(1-\mu)\bar{x}_{.j} + \mu\bar{x}_{..}$ $j=1,\ldots J$. Here of course there is a natural order i.e. the k^{th} vector $x_k' = (x_{k1},\ldots,x_{kJ})$ is the fundamental sampling unit. Again as in method II we omit the k^{th} row using $(1-\mu)y_{kj} + \mu\bar{y}_{kj}$ where

$$y_{kj} = (K\bar{x}_{.j} - x_{kj})/(K-1)$$

$$\bar{y}_{kj} = (KJ\bar{x}_{..} - x_{kj})/(JK-1) \qquad (5.1)$$

to predict the missing row x_k'. Repeating this for $k=1,\ldots,K$ we compute the mean squared prediction error

$$v_\mu^2 = J^{-1}K^{-1}\sum_{j=1}^{J}\sum_{k=1}^{K}((1-\mu)y_{kj} + \mu\bar{y}_{kj} - x_{kj})^2 \qquad (5.2)$$

Again by simple algebra, evaluation yields

$$v_\mu^2 = \frac{(K-\mu)^2}{K(K-1)} m_1 + (JK)^{-1}(J-1)\mu^2 m_2 + \frac{\mu(2K-\mu)}{JK(K-1)} m_3 \tag{5.3}$$

where

$$m_3 = J(K-1)^{-1} \sum_{k=1}^{K} (\bar{x}_{k\cdot} - \bar{x}_{\cdot\cdot})^2 ; \quad \bar{x}_{k\cdot} = J^{-1} \sum_{j=1}^{J} x_{kj} \tag{5.4}$$

(We note that method II is derived in this manner for an arbitrary reordering and then is averaged over all possible permutations within each column. It is clear that m_1 and m_2 are unaffected by such permutations and we need only average m_3 over these permutations to obtain t_μ^2. This average with regard to m_3 is easily shown to be m_1. Hence substitution of m_3 by m_1 in v_μ^2 yields t_μ^2.)

Again

$$v_0^2 = \frac{K}{K-1} m_1$$

and minimization of v_μ^2 w.r.t. $\mu \in [0,1]$ yields $\min(\mu^*,1)$ where

$$\mu^* = \frac{K(m_1 J - m_3)}{(J-1)(K-1)m_2 + Jm_1 - m_3} \geq 0 \tag{5.5}$$

since $m_1 J \geq m_3$.

6. Remarks

In summary we have proposed for the random effect model, two bootstrap predictive methods for evaluating different estimates possibly generated from different prior models for a set of data such that columns are independent. In addition the methods themselves are capable of yielding on their own terms a "best" predictor. This of course is only with respect to mean squared prediction error but on the other hand they are fairly free of distributional assumptions.

It can also be considered as a possible means of estimating the parameter in question if this is of interest. Of course these mean squared predictive errors are just one of a host of possibilities that one may use for guidance as to which model is more appropriate for a set of data. It is also of some interest that the estimate of the coefficient μ is considerably simpler in form than the Box-Tiao or Lindley estimate.

Further we have also presented the same type of method for the case where the columns are not necessarily independent which is sometimes termed the "mixed model" or "multivariate model."

Box, G.E.P. and Tiao, G. (1968). Bayesian estimation of means for the random effect model. JASA, Vol. 63, pp. 174-181.

Lindley, D.V. (1971). Bayesian statistics, a review. SIAM, Regional Conference Series in Applied Mathematics. No. 2.