FIXED AND RANDOM EFFECTS AND BEST PREDICTION

Charles E. McCulloch

Biometrics Unit and Statistics Center

Cornell University

Ithaca, NY 14853

BU-1241-M

May, 1994

Abstract

Conventional wisdom holds that effects in a linear model be treated as fixed effects whenever interest focusses on inference for the levels of that effect. I trace the evolution of this notion and argue that it is incorrect. I offer suggestions on how to decide if a factor is fixed or random and explore the connections between fixed and random factors, best prediction and the Bayesian paradigm.

KEY WORDS: Best linear unbiased prediction, Hierarchical models

1. Introduction

It has long been suggested (Eisenhart, 1947, Scheffe, 1959) that two main assumptions can be made about the parameters describing the levels of a categorical variable (factor) in a linear model. They can be assumed to be fixed, unknown constants or to them can be attributed a distribution. If the parameters are assumed to be fixed, unknown constants the factor is described as a fixed effect; if a distribution is attributed to it, it is known as a random effect. This is well accepted. However deciding which of the two models to use in practice has been problematic. Conventional wisdom hold that a factor be treated as fixed if one is interested in drawing inferences about the specific levels included in the experiment (Searle, 1987, p.4; Snedecor and Cochran, 1989, p.320) or if, on repeated selection of the levels of that factor, the same levels are selected (Ott, 1984, p.638; Snedecor and Cochran, 1989, p.320). If inferences focus on the population from which the levels were selected or if, on repeated selections of the levels of that factor, the same levels are not selected, then the factor is declared to be a random factor.

I argue that both of these criteria are incorrect and that to decide in practice whether to treat a factor as fixed or random we need to make the decision based on a criterion more closely related to that assumption of a distribution for the parameters rather than the focus of the inferences. That we should not base our decision on the focus of inferences is made clear by consideration of the concept of best prediction, where we are willing to assume that the levels of the factor follow a distribution, but by calculating best predicted values we are making inferences about (and are "interested in") the specific levels included in the experiment.

2. Deciding Fixed or Random

In arguing that the conventional criteria are incorrect I consider three generic examples: a randomized blocks design, prediction of sire effects in animal breeding and spatial prediction (kriging). I first consider the randomized blocks design and the criterion of whether or not we would get the same levels (in this case blocks) if we were to conduct the experiment again. In many experiments the following facts are all true:

a) The same blocks would be used if the experiment were repeated (which blocks are used is often determined by the availability of experimental material),
b) The investigator wants to draw inferences beyond the blocks on hand in the experiment,
c) The investigator is willing to make inferences to a population of blocks similar to the ones in the experiment,
d) It is reasonable to assume the blocks in the experiment are an i.i.d. sample from the population described in c).

Points b), c) and d) mean that, by definition, blocks are a random factor. But a) argues that blocks should be declared fixed. Essentially the criterion fails because we do not have a physical sampling scheme which guarantees random sampling but we are willing to assume the blocks form an i.i.d. sample from some distribution.

I now consider the second criterion: interest in the levels actually included in the experiment. My primary example is prediction of sire effects in animal breeding, but I will also note the parallel in spatial prediction. Animal breeders often face the following problem. They wish to improve the genetic value (e.g., the ability of cows to produce protein in milk) by selective breeding of the population. The data used for the analysis often includes the daughters of all the sires whose data are available through a registry. The goal is to estimate the ability of a specfic sire to produce genetically superior offspring. On one hand it is easy to envision the sires included in the analysis as coming from a population (actual or conceptual) of sires; hence the argument for treating the effect as random. On the other hand, interest focusses specifically on the sires included in the analysis. Those are the only ones which would be considered for use in a breeding program. This, according to the second criterion, would argue for treating the sire effects as fixed.

A similar situation arises in the context of spatial prediction (kriging) where it is desirable to predict at one or more of the locations where data were gathered (Cressie, 1991, p.128). On one hand, interest focusses specifically on the locations at which data are gathered (analogous to the levels of a factor included in the experiment). On the other hand, the usual model for spatial statistics assumes that the observations are selected from a distribution and are not fixed, unknown quantities.

The basic idea is that we should not base our decision on the inferential goal because we can make inferences specifically about the parameters for levels of either a fixed (via the usual estimates) or a random (via best prediction) effect.

3. Best Prediction and the Bayesian Paradigm

When a factor is declared random the parameters representing the levels of that factor are considered random variables selected from a distribution. This is analogous to, though different from, the Bayesian approach. For concreteness consider the following simple mixed model, containing a single fixed and random factor (the usual convention is not to count the intercept):

$$Y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij}, \qquad i=1, ..., k \text{ and } j=1, ..., n \tag{1}$$

where μ and α_i are fixed, unknown constants and b_j are distributed i.i.d. $N(0,\sigma_b^2)$ independently of the ϵ_{ij}, which are i.i.d. $N(0,\sigma_\epsilon^2)$. In a Bayesian approach a distribution would usually be selected to reflect prior belief, which is not too far afield from assuming that the factor is random, though in my

experience, the random factor assumption is often based on past empirical evidence. The Bayesian would then go farther and attribute distributions to μ, α_i, the two σ^2s and any parameters found in the distribution of μ and α_i. So a Bayesian does not distinguish between fixed and random factors.

A Bayesian has a relatively automatic way to "estimate" the b_j. She would calculate, e.g., the posterior expected value of b_j:

$$\widehat{b_j} = E[b_j|\mathbf{Y}].$$

A frequentist needs to work a bit harder to find a predictor. It is well known that the best "predictor" (in a MSE sense) of a random variable based on data $\mathbf{Y}$ is the conditional expected value of that random variable. The catch is that once the conditional expectation is taken, it usually depends on unknown parameters and hence is not calculable from the data. For example, for (1) we have the "predictor" (Searle, Casella and McCulloch, 1992):

$$\widehat{b_j} = E[b_j|\mathbf{Y}] = \frac{k\sigma_b^2}{\sigma_\epsilon^2 + k\sigma_b^2}(\overline{Y}_{\cdot j} - \mu) \tag{2}$$

which depends on the unknown fixed effects and the variances.

To partially circumvent this problem, Henderson (1963) and others (e.g., Goldberger, 1963) developed the idea of Best Linear Unbiased Prediction (BLUP). They proved that replacing the fixed effects by their generalized least squares estimators gave the best (in a MSE sense) linear (in the data) unbiased predictor assuming the variances were known. Of course, in practice the variances usually are not known and so investigators would usually also plug in estimates of the variances also. This invalidates the optimal properties (e.g. it is not even linear any more), but seems like a quite reasonable procedure.

In any case, the basic point is that from a frequentist perspective there is a reasonable procedure to follow to construct predictions for the effects of a random factor. This allows one to both assume a distribution for the effects and "be interested in them." The same methodology is followed in spatial prediction, where many of the kriging estimates are BLUPs.

4. History

How did we get to this state of affairs? It seems to have been first crystallized with the influential work of Eisenhart (1947), where he first set out Model I (fixed) and Model II (random) for analysis of variance. He has a section entitled "Which Model - Model I or Model II?" in which he states,

> ...the answer is clear as soon as a decision is reached on whether the parameters of interest specify fixed relations or components of random variation.

This, of course, is exactly correct, but he gets in trouble as he tries to explain further:

(1) Are the conclusions to be confined to the things actually studied ... or expanded to apply to a more general population?

(2) In complete repetitions of the experiment would the same things be studied again ... or would new samples be drawn from the more general populations?

4. Conclusions

I think the key is to go back to Eisenhart's original recommendation and ask the question: "Are we willing to assume the the parameters representing the levels of the factor are a random sample from some population?" If the answer is yes then we treat the factor as random. If we are unwilling to make that assumption then we must treat them as fixed, unknown quantities and suffer the restriction of inference that goes along with not being able to infer to a larger population.

The key is to divorce the attribution of a distribution from the inferential goal since we can make inferences to the specific levels of a factor whether it is fixed or random. In the case of a random factor, that would then mean using some form of best prediction.

5. References

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics 3: 1-21.

Henderson, C. (1963). Selection index and expected genetic advance. In Statistical Genetics and Plant Breeding (W.D. Hanson and H.F. Robinson, eds.), 141-163. National Academy of Sciences and National Research Council Publication No. 982, Washington, D.C.

Goldberger, A.S. (1963). Best linear unbiased prediction in the generalized linear regression model. Journal of the American Statistical Association 57: 369-375.

Ott, L. (1984). An Introduction to Statistical Methods and Data Analysis, 2nd Ed. Duxbury, Boston.

Scheffe, H. (1959). The Analyis of Variance. Wiley, New York.

Searle, S.R. (1987). Linear Models for Unbalanced Data. Wiley, New York.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). Variance Components. Wiley, New York.

Snedecor, G.W. and Cochran, W.G. (1989). Statistical Methods, 8th Ed. Iowa State University Press, Ames, Iowa.