The Fundamental Theorem of Least Squares: Its Relevance to Experimental Design and Bayesian Inference

By Heng Li and Alan Zaslavsky

University of Rochester and Harvard Medical School

The necessary and sufficient condition for the least squares estimator to be best linear unbiased estimator is stated for a large class of common linear models. It is shown that this condition implies parallelism between the Bayesian and frequentist inferences under a non-informative reference prior for this class of models.

1. Introduction. It is difficult to find a theorem in statistics that has been stated in as many different ways in the literature as the necessary and sufficient condition for the ordinary least squares (OLS) estimator to be best linear unbiased estimator (BLUE). In a comprehensive review, Puntanen and Styan [9] collected at least twenty different statements of the condition, and claimed that seventeen more may easily be obtained. That this condition is a fundamental theorem that is not so well known in statistics is witnessed by its rediscoveries since its first appearance in [1] ([5, 9]). The class of models in terms of which the theorem has usually been stated may have made it sound more restrictive and less relevant than it really is, and contributed to its relative obscurity. In this paper we explicitly specify a large class of realistic models to which the theorem is applicable, and point out its relevance to Bayesian inference for this class of models.

2. The Fundamental Theorem. The necessary and sufficient condition for the OLS estimator to be BLUE is commonly stated, as in Kruskal's influential paper [6], and in Puntanen and Styan's [9] comprehensive review, for the linear models in which the covariance is known up to a multiplicative constant:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \quad E(\epsilon) = \mathbf{0}, \quad cov(\epsilon) = \sigma^2\mathbf{V}, \tag{1}$$

where $\mathbf{V}$ is fixed and known. One of the formulations of the condition, labeled as Z1 (where Z is for Zyskind) in [9], can be expressed as: "A subset of the eigenvectors of $\mathbf{V}$ span the column space of the design matrix $\mathbf{X}$". Notice that since the above statement refers only to the eigenvectors, $\mathbf{V}$ does not have to be completely known in order to verify the condition. In particular, the condition can always be verified if $\mathbf{V}$ has known eigenspaces. Hence condition Z1 is also meaningful and valid for the following model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \quad E(\epsilon) = \mathbf{0}, \quad cov(\epsilon) = \mathbf{\Sigma},$$

where

$$\mathbf{\Sigma} = \lambda_1\mathbf{E}_1 + \cdots + \lambda_K\mathbf{E}_K \tag{2}$$

with λ's being (unknown) non-negative real numbers and $\mathbf{E}$'s being *known* orthogonal projections. The change of the specification of covariance structure from the one-parameter family $\sigma^2\mathbf{V}$ to the multi-parameter family (2) is significant. While linear models with covariance structure (1) are unusual beyond the trivial case $\mathbf{V} = \mathbf{I}$, many realistic designs have covariance structure (2). According to Nelder [8], the covariance structure in model (2) encompasses an impressively long list of commonly used models in factorial designs, and this list has since been further expanded [13]. Essentially, all the covariance structures that are induced by the nesting and crossing of random factors in a completely balanced way are of the form in model (2). Therefore, the potential wide applicability of the necessary and sufficient condition for OLS to be BLUE becomes much more evident when the theorem is applied to model (2).

The conditions for OLS to be BLUE also imply *general balance*, a well-known concept in experimental design, as shown by the following straightforward and familiar proof. Let $\mathbf{M}$ denote the orthogonal projection onto the column space of the design matrix $\mathbf{X}$. Using the definition $(GB)^*$ in [12, p. 321], the condition of general balance can be stated in our notation as follows:

$$\mathbf{M}\mathbf{E}_i\mathbf{M} = \sum_j c_{ij}\mathbf{M}_j, \quad i = 1, \cdots, k,$$

where c_{ij}'s are scalars and $\mathbf{M}_j$'s are orthogonal projections such that $\sum_j \mathbf{M}_j = \mathbf{M}$. Suppose condition Z1 holds. Let $\mathbf{M}_j$ be the orthogonal projection onto the space spanned by those eigenvectors in the range of $\mathbf{E}_j$ among all the eigenvectors spanning the column space of the design matrix $\mathbf{X}$. Then we have $\sum_j \mathbf{M}_j = \mathbf{M}$ and

$$\mathbf{M}\mathbf{E}_i\mathbf{M} = \sum_j \delta_{ij}\mathbf{M}_j, \quad i = 1, \cdots, k,$$

where δ_{ij} is the Kronecker delta for i and j. Therefore condition Z1 for model (2) implies general balance. We introduce the term "strict balance" for the subclass of "generally balanced" linear models that also satisfy the necessary and sufficient condition for OLS to be BLUE.

Condition Z1 holds for many commonly used models with covariance structure (2). Examples include models used for the completely balanced layout of the following designs: randomized block and latin square designs [4, chap. 4], the split-plot design [4, chap. 7], and the nested design [11, sec. 13.9]. Instead of providing an exhaustive list of models satisfying Z1, it is perhaps more informative to describe formally the conventional analysis of variance procedure for models with covariance structure (2), and state Z1 in the context of such analysis.

Each eigenspace of the covariance (the range space of an $\mathbf{E}_i$) corresponds to a stratum [8, 13, 2] of an analysis of variance table, and usually has a suggestive name such as within, between, or block, etc. The quadratic form $\mathbf{y}'\mathbf{E}_i\mathbf{y}$ is the total sum of squares of stratum i, which, assuming $\mathbf{y}$ is normally distributed, has a non-central chi-square distribution in general, and a central chi-square distribution if the range of $\mathbf{E}_i$ is orthogonal to that of the design matrix $\mathbf{X}$. If $\mathbf{E}_i$ is not orthogonal to $\mathbf{X}$, the

ith stratum total sum of squares is then decomposed into two sums of squares. One is the squared length of the projection of data vector on the range of $\mathbf{E}_i\mathbf{X}$ (SSF$_i$), and the other is the remainder (SSE$_i$). The former captures the fixed effects that are not orthogonal to $\mathbf{E}_i$ and has a noncentral chi-square distribution when such fixed effects are present. The latter, the error sum of squares, has a central chi-square distribution and is the error term for inferences about fixed effects using the F (or t) distribution. Accordingly, each stratum of an analysis of variance table is divided into two sections, one for the fixed effects and the other for the error term. Parts of SSF$_i$ can be associated with one or more contrasts within the ith eigenspace. The value of a contrast is a linear combination $\mathbf{c}'\mathbf{y}$ of the observations where $\mathbf{c}$ is within the range of $\mathbf{X}$, and hence a contrast is represented by a vector $\mathbf{c}$ of the same dimension as $\mathbf{y}$.

The concept of homoscedasticity applies to this situation [2]: we call contrasts within the range of eigenspaces in covariance structure (2) "homoscedastic contrasts." With this terminology, another way of stating condition Z1 for model (2) is: The column space of the design matrix (the mean space) is spanned by homoscedastic contrasts. If we extend the terminology to call subspaces spanned by homoscedastic contrasts in stratum i "homoscedastic subspaces," then yet another way of stating Z1 is: "The mean space is a sum of homoscedastic subspaces." In completely balanced factorial designs, usually a factorial component of a fixed effect such as a main effect or an interaction corresponds to a homoscedastic subspace, thus making the mean space the sum of homoscedastic subspaces. This is not the case for incomplete block designs; in these designs, the treatment main effect does not correspond to a homoscedastic subspace and there may not be any homoscedastic contrasts between treatments.

To concretize the above discussion, consider a balanced split-plot design in which there is a whole-plot factor and a subplot factor, with the covariance induced by the whole-plots having compound symmetry. This covariance structure has two eigenspaces: within whole-plot (subplot) eigenspace and (between) whole-plot eigenspace. Correspondingly, the mean space can be divided into the whole-plot factor main effect, the subplot factor main effect, and their interaction, of which the first is a subspace of the whole-plot eigenspace and the latter two are subspaces of the subplot eigenspace. The mean space itself is of course the direct sum of those three subspaces which are homoscedastic, and therefore the balanced split-plot design satisfies condition Z1.

3. Relevance to Bayesian Inference.

Suppose that condition Z1 holds for model (2), in which $\mathbf{E}_i$ has rank n_i. We can parameterize the mean space so that the design matrix $\mathbf{X}$ has as its columns mutually orthogonal eigenvectors (of unit length) of the covariance $\mathbf{\Sigma}$. More explicitly, the design matrix can be expressed as $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_K)$, where the columns of $\mathbf{X}_i$ are $n_i - \nu_i$ mutually orthogonal eigenvectors of the ith eigenspace of $\mathbf{\Sigma}$ $(0 < \nu_i \leq n_i)$. Let β_i be the subvector of β corresponding to $\mathbf{X}_i$, with the convention that $\beta_i = 0$ if $\nu_i = n_i$. The likelihood can then be written as (see Appendix)

$$L = \lambda_1^{-\frac{n_1}{2}} \cdots \lambda_K^{-\frac{n_K}{2}} \exp\frac{1}{2}\left[-\sum_i \frac{\mathrm{SSE}_i + (\beta_i - \hat{\beta}_i)'(\beta_i - \hat{\beta}_i)}{\lambda_i}\right],$$

where $\hat{\beta}_i$ is the least squares estimate of β_i and SSE_i is the ν_i degrees of freedom error sum of squares for stratum i ($\hat{\beta}_i = 0$ if $\beta_i = 0$). If we put the non-informative reference prior [3, p. 480-481]

$$(3) \qquad p(\beta, \lambda_1, \cdots, \lambda_K) = \lambda_1^{-1} \cdots \lambda_K^{-1}$$

on the parameters, then the same integrations as in [3, p. 480-481] lead to

$$p(\beta|\mathbf{y}) \propto \prod_{i=1}^{K} \left(SSE_i + (\beta_i - \hat{\beta}_i)'(\beta_i - \hat{\beta}_i)\right)^{-\frac{n_i}{2}} \propto \prod_{i=1}^{K} \left(1 + \frac{(\beta_i - \hat{\beta}_i)'(\beta_i - \hat{\beta}_i)}{MSE_i \nu_i}\right)^{-\frac{n_i}{2}}.$$

That is to say, given data, β_i's are independent of each other; and for each i, $(\beta_i - \hat{\beta}_i)/\sqrt{MSE_i}$ has a spherical $n_i - \nu_i$ dimensional multivariate t distribution with ν_i degrees of freedom, which is precisely the same as the sampling distribution of $(\beta_i - \hat{\beta}_i)/\sqrt{MSE_i}$ given β_i. This implies that, since the multivariate t distribution is closed under marginalization, for each i a confidence set for β_i or its subvector based on F (or univariate t) distribution coincides with a Bayesian posterior region having a posterior probability content equal to the coverage probability of the confidence set, and vice versa. Since a traditional analysis of variance F test for sets of homoscedastic contrasts can be viewed as inversions of a confidence set for a β_i (or a subvector) based on the multivariate t distribution given appropriately chosen $\mathbf{X}_i$'s, a parallelism is established between Bayesian inference and frequentist inference on linear models with covariance (2) that satisfy condition Z1. The usual analysis of variance inference for fixed effects based on their proper error terms for this class of models is equivalent to Bayesian inference using the non-informative prior (3).

In many instances, the covariance structure in model (2) arises from variance component models. When the variance component model formulation is used, the nonnegativity of variance components may imply further constraints on the range of λ's beyond nonnegativity. If the prior honors those constraints, we obtain different posterior distributions from the ones derived in the previous section. Results in [7] indicate that, for the balanced nested designs they considered, combining the nonnegativity constraints for variance components with the non-informative prior (3) makes the inference for the fixed effect parameters conservative, in that the coverage probability of Bayesian intervals is larger than their posterior probability content. Whether or how the results in [7] generalize to the class of models considered in this paper is yet unknown. However, in any event, given the Bayesian-frequentist parallelism established in this paper, and the arguments made in [10], it does not seem unreasonable to use the non-informative reference prior (3) without incorporating additional constraints on the λ's for models with covariance structure (2), when inference is focused on the fixed effects.

4. Discussion. The necessary and sufficient condition for the ordinary least squares (OLS) estimator in a linear model to be best linear unbiased estimator (BLUE) is one of the most fundamental theorems in statistics. However, so far it seems to have remained an abstract mathematical fact with its relevance to applied statistics largely unexplored. This paper is an attempt to introduce the practical relevance of this theorem to the literature of applied statistics, by explicitly connecting it to a large class of commonly used models, and examining its implications in Bayesian inference. This theorem is almost certainly connected to many other aspects of the theory and application of linear models, and it may prove to deserve much more visibility than it currently enjoys.

Appendix: Likelihood for Balanced Linear Models. Let $\mathbf{P}_i$ be the orthogonal projection onto the column space of $\mathbf{X}_i$ and $\mathbf{Q}_i$ be the orthogonal projection onto the span of all the eigenvectors in the range of $\mathbf{E}_i$ orthogonal to those in bfX_i. Then $\mathbf{E}_i = \mathbf{P}_i + \mathbf{Q}_i$, and $\mathrm{SSE}_i = \mathbf{y}'\mathbf{Q}_i\mathbf{y}$. Furthermore, $\mathbf{P}_1 + \cdots + \mathbf{P}_K$ is the orthogonal projection onto the range of $\mathbf{X}$. So $(\mathbf{P}_1 + \cdots + \mathbf{P}_K)\mathbf{y} = \mathbf{X}\hat{\beta}$. Due to the orthogonality between the ranges of the $\mathbf{X}_i$'s, $\mathbf{P}_i\mathbf{y} = \mathbf{X}_i\hat{\beta}_i$. Hence

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{E}_i(\mathbf{y} - \mathbf{X}\beta) &= \mathbf{y}'\mathbf{Q}_i\mathbf{y} + (\mathbf{X}_i\hat{\beta}_i - \mathbf{X}_i\beta_i)'(\mathbf{X}_i\hat{\beta}_i - \mathbf{X}_i\beta_i)\\ &= \mathrm{SSE}_i + (\beta_i - \hat{\beta}_i)'\mathbf{X}_i'\mathbf{X}_i(\beta_i - \hat{\beta}_i).\end{aligned}$$

Note that $\mathbf{X}_i'\mathbf{X}_i = \mathbf{I}$, the identity matrix. The likelihood is thus obtained.

REFERENCES

[1] T.W. Anderson. On the theory of testing serial correlation. *Skandinavisk Aktuarietidskrift*, 31:88–116, 1948.

[2] R.A. Bailey. Strata for randomized experiments (with discussion). *Journal of Royal Statistical Society Series B*, 53:27–78, 1991.

[3] G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, MA, 1973.

[4] W.G. Cochran and G.M. Cox. *Experimental Designs.* Wiley, New York, 2 edition, 1957.

[5] M.H. DeGroot. A conversation with T.W. Anderson. *Statistical Science*, 1:97–105, 1986.

[6] W. Kruskal. When are Gauss-Markov and least squares estimators identical? a coordinate-free approach. *The Annals of Mathematical Statistics*, 39:70–75, 1968.

[7] H. Li and H. Stern. Bayesian inference for nested designs based on Jeffreys's prior. *The American Statistician*, 51:219–224, 1997.

[8] J.A. Nelder. The analysis of randomized experiments with orthogonal block structure, i and ii. *Proceedings of Royal Society London, Series A*, 283:147–178, 1965.

[9] S. Puntanen and G.P.H. Styan. The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, 43:153–164, 1989.

[10] D.B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–1172, 1984.

[11] G.W. Snedecor and W.G. Cochran. *Statistical Methods.* The Iowa State University Press, Ames, IA, 8 edition, 1989.

[12] T.P. Speed. General balance. In S. Kotz and N.L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 3, pages 320–326. Wiley, New York, 1983.

[13] T.P. Speed. What is an analysis of variance? (with discussion). *The Annals of Statistics*, 15:885–910, 1987.

FDA/CDRH
1350 Piccard Drive
Rockville MD 20850
hxl@cdrh.fda.gov

Department of Health Care Policy
Harvard Medical School
180 Longwood Ave
Boston, MA 02115-5899
zaslavsk@hcp.med.harvard.edu