EXPECTED MARGINAL MEANS IN THE LINEAR MODEL

by

BU-672-M April, 1979

S. R. Searle[1], G. A. Milliken[2] and F. M. Speed[3]

Abstract

The term "least squares mean" is replaced by the more meaningful "expected marginal mean" and its estimation discussed.

1. Introduction

An expression that is about to make an appearance in the statistical literature arising from its use in statistical computer packages is the "least squares mean". It is first mentioned in Harvey [1975], is discussed at length in Goodnight and Harvey [1978], occurs in Goodnight [1979] and is available as output in the current update of the widely-used SAS GLM (Statistical Analysis System, General Linear Model) computing procedure; it is in the user-supplied SAS HARVEY procedure and, for some years, it has been a standard feature of the LSMLGP computing package of Harvey [1968], the precursor to SAS HARVEY.

The words "least squares" have no intrinsic or logical meaning when used in apposition to "mean", at least not in any traditional manner. Thus "least squares mean" of itself conveys no implicit meaning of what it is supposed to represent. Hence, and also because it is a phrase that has its origins in computing rather than statistics, it seems appropriate to describe its meaning, especially since

---

[1] Biometrics Unit, Cornell University.

[2] Statistics Department, Kansas State University.

[3] Statistics Department, Mississippi State University.

Paper No. BU-672-M in the Biometrics Unit, Cornell University.

its occurrence in computer output is sparking a demand for knowing what that meaning is. (We note in passing the quixotic nature of this situation: statisticians having to provide a definition for certain computer output, in contrast to the usual situation of using a computer to get output they want.) What is even more important, we feel, is to suggest both a clear definition for the concept presently embodied in the least squares mean but nowhere available, and also a name that is descriptive and meaningful of itself. We also discuss its estimation.

As a framework for discussion we use the 2-way crossed (rows-by-columns) classification having, say, a rows and b columns. Let y_{ijk} be the k'th observation in the i'th row and j'th column of such data, with expected value

$$E(y_{ijk}) = \mu_{ij} , \tag{1}$$

where $k = 1, \cdots, n_{ij}$ when the cell in row i, denoted by α_i, and column j, denoted by β_j, has n_{ij} observations in it. When that cell is empty, i.e., has no data, we take $n_{ij} = 0$.

2. The Concept and A Descriptive Name

There are two basic ideas involved in the least squares mean. The first is that it is the expected value of a marginal mean. And this is the name by which we suggest it should be known: expected marginal mean, EMM. The second is that there is one EMM for each row, one for each column and, in the presence of interactions, one for each interaction. In general, there is an EMM for each level of each factor of a linear model. Furthermore, although an EMM is a function of the parameters of the model, it is not a function of the numbers of observations. Thus the EMM for the i'th row for example, is the arithmetic average of the cell means for <u>all</u> cells in row i:

$$\mathrm{EMM}(\alpha_i) = \bar{\mu}_{i.} = \sum_{j=1}^{b} \mu_{ij}/b \ . \tag{2}$$

Note that this is not the expected value of the observed row mean

$$E(\bar{y}_{i..}) = \sum_{j=1}^{b} n_{ij}\mu_{ij} / \sum_{j=1}^{b} n_{ij} \ . \tag{3}$$

Expression (2) does not involve the n_{ij}'s whereas (3) does. And (2) includes μ_{ij} for every cell in row i whereas, if some cells are empty, (3) does not; the empty cells are not represented in (3). Thus in the case of empty cells the definition "average of the cell means for all cells in row i" that precedes (2) must not be interpreted as referring to just the cells having data, but it always refers to all cells. To emphasize this we could use the definition "expected value of an observed marginal mean as if there were one observation in every cell".

Expression (2) is the EMM for the i'th row of the model (1). The comparable expression for the j'th column is

$$\mathrm{EMM}(\beta_j) = \bar{\mu}_{.j} = \sum_{i=1}^{a} \mu_{ij}/a \tag{4}$$

and that for the i'th row and j'th column is

$$\mathrm{EMM}(\mu_{ij}) = \mu_{ij} \ . \tag{5}$$

The concept of the EMM is that, for example, $\mathrm{EMM}(\alpha_i)$ is an expression for the population mean of the i'th row, unencumbered by the numbers of observations and how they are distributed throughout the row. This is presumably what Goodnight and Harvey [1978] have in mind with their definition: "estimates of the class and subclass arithmetic means that would be expected had equal subclass numbers been obtainable". An exception to this is that it misleadingly introduces the idea of an estimate into the definition of a parametric function; and further, it is not consistently adhered to in examples in that same paper, e.g., equations (26) and (28).

3. Estimability

Defining EMM in terms of estimable functions is impossible because in some models expressions like (2), (4) and (5) are estimable and in some they are not. This is the cause of an inconsistency in Goodnight and Harvey [1978] where on page 8 they have the phrase "all LSM's [EMM's] are defined in terms of estimable functions" whereas those of (26) and (28) are not. It seems to us advantageous to have a definition that stands on its own, independent of data, and which does not depend upon estimability which is an outcome of the data structure, in particular of the pattern of numbers of observations in the data. The definition given in (2) is of this nature: it is a function of parameters, without regard to data; it is also a function that might often be of use and interest. The matter of estimability then arises only when one comes to estimate that function.

4. Estimation

It is clear from (2), (4) and (5) that the b.l.u.e. (best linear unbiased estimator) of an EMM is the mean of the b.l.u.e.'s of the corresponding μ_{ij}'s, and exists if and only if <u>all</u> those μ_{ij}'s in the EMM are estimable. For the 2-way cross-classification of (1) we distinguish two cases, without interaction, and with.

4.1. Without interaction

The over-parameterized form of (1) for the no-interaction case is

$$E(y_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j \tag{6}$$

where μ is a general mean and α_i and β_j are as already defined. In all cases of this model, whether all cells are filled or whether some of them are empty, every μ_{ij} is estimable with b.l.u.e.

$$\hat{\mu}_{ij} = \mu^o + \alpha_i^o + \beta_j^o \quad \text{for } i = 1, \cdots, a \text{ and } j = 1, \cdots, b , \tag{7}$$

where μ^o, α_i^o and β_j^o are solutions to the normal equations as given, for example, in Searle [1971, Sec. 7.1d]. Hence the EMM's of (2) and (4) are estimable with b.l.u.e.'s

$$\widehat{\text{EMM}(\alpha_i)} = \mu^o + \alpha_i^o + \sum_{j=1}^{b} \beta_j^o/b, \quad \text{for } i = 1, \cdots, a$$

and (8)

$$\widehat{\text{EMM}(\beta_j)} = \mu^o + \sum_{i=1}^{a} \alpha_i^o/a + \beta_j^o, \quad \text{for } j = 1, \cdots, b .$$

4.2. With interaction

The with-interaction model equation is

$$E(y_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{9}$$

where μ, α_i and β_j are as before and γ_{ij} is the interaction term corresponding to row i and column j. In all cases, when μ_{ij} is estimable its b.l.u.e. is

$$\hat{\mu}_{ij} = \bar{y}_{ij.} \ . \tag{10}$$

But for the estimability, and hence estimation, of EMM's we must distinguish two cases; one is that of all cells filled, and the other is that of some cells empty.

(a) All cells filled. In this case, every μ_{ij} is estimable and therefore so is every EMM of (2), (4) and (5), with b.l.u.e.'s as follows:

$$\widehat{\text{EMM}(\alpha_i)} = \sum_{j=1}^{b} \hat{\mu}_{ij}/b = \sum_{j=1}^{b} \bar{y}_{ij.}/b \tag{11}$$

$$\widehat{\text{EMM}(\beta_j)} = \sum_{i=1}^{a} \hat{\mu}_{ij}/a = \sum_{i=1}^{a} \bar{y}_{ij.}/a \tag{12}$$

$$\widehat{\text{EMM}(\gamma_{ij})} = \hat{\mu}_{ij} = \bar{y}_{ij.} \,. \tag{13}$$

(b) Some cells empty. When some cells have no data in them, the only rows (and columns) having estimable EMM's are those that do have observations in every cell therein; and the corresponding b.l.u.e.'s are then exactly as shown in (11) and (12); and (13) applies to all the filled cells. For rows and (columns) that have even just one cell being empty, the EMM's are not estimable; and b.l.u.e.'s of those rows and columns do not exist.

4.3. Example

Consider the case of just two rows and two columns, with one empty cell:

Row	Column 1	Column 2
1	data	data
2	data	empty

For the no interaction model it is easily established that solutions to the normal equations corresponding to the methods of Searle [loc. cit.] are

$$\mu^o = 0, \quad \alpha_1^o = \bar{y}_{12.}, \quad \beta_1^o = \bar{y}_{11.} - \bar{y}_{12.} \,,$$

$$\alpha_2^o = \bar{y}_{12.} + \bar{y}_{21.} - \bar{y}_{11.}, \text{ and } \beta_2^o = 0 \,.$$

The EMM's and their b.l.u.e.'s are then:

EMM	b.l.u.e. of EMM
$EMM(\alpha_1) = \mu + \alpha_1 + \frac{1}{2}(\beta_1 + \beta_2)$	$\frac{1}{2}(\bar{y}_{11.} + \bar{y}_{12.})$
$EMM(\alpha_2) = \mu + \alpha_2 + \frac{1}{2}(\beta_1 + \beta_2)$	$\bar{y}_{21.} - \frac{1}{2}(\bar{y}_{11.} - \bar{y}_{12.})$
$EMM(\beta_1) = \mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_1$	$\frac{1}{2}(\bar{y}_{11.} - \bar{y}_{21.})$
$EMM(\beta_2) = \mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_2$	$\bar{y}_{12.} - \frac{1}{2}(\bar{y}_{11.} - \bar{y}_{21.})$.

For the with-interaction model, the b.l.u.e.'s of the μ_{ij}'s corresponding to cells containing data are

$$\hat{\mu}_{11} = \bar{y}_{11.}, \qquad \hat{\mu}_{12} = \bar{y}_{12.} \quad \text{and} \quad \hat{\mu}_{21} = \bar{y}_{21.} \ .$$

The EMM's for this model, and the b.l.u.e.'s of those that are estimable, are as follows:

EMM	b.l.u.e. of EMM	
$EMM(\alpha_1) = \mu + \alpha_1 + \frac{1}{2}(\beta_1 + \beta_2) + \frac{1}{2}(\gamma_{11} + \gamma_{12})$	$\frac{1}{2}(\hat{\mu}_{11} + \hat{\mu}_{12}) = \frac{1}{2}(\bar{y}_{11.} + \bar{y}_{12.})$	
$EMM(\alpha_2) = \mu + \alpha_2 + \frac{1}{2}(\beta_1 + \beta_2) + \frac{1}{2}(\gamma_{21} + \gamma_{22})$	Not estimable	
$EMM(\beta_1) = \mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_1 + \frac{1}{2}(\gamma_{11} + \gamma_{21})$	$\frac{1}{2}(\hat{\mu}_{11} + \hat{\mu}_{21}) = \frac{1}{2}(\bar{y}_{11.} + \bar{y}_{21.})$	
$EMM(\beta_2) = \mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_2 + \frac{1}{2}(\gamma_{12} + \gamma_{22})$	Not estimable	(14)
$EMM(\gamma_{11}) = \mu + \alpha_1 + \beta_1 + \gamma_{11}$	$\hat{\mu}_{11} = \bar{y}_{11.}$	
$EMM(\gamma_{12}) = \mu + \alpha_1 + \beta_2 + \gamma_{12}$	$\hat{\mu}_{12} = \bar{y}_{12.}$	
$EMM(\gamma_{21}) = \mu + \alpha_2 + \beta_1 + \gamma_{21}$	$\hat{\mu}_{21} = \bar{y}_{22.}$	
$EMM(\gamma_{22}) = \mu + \alpha_2 + \beta_2 + \gamma_{22}$	Not estimable .	

Notice that the definitions of $EMM(\alpha_2)$ and of $EMM(\beta_2)$ are different from those of equations (26) and (28) of Goodnight and Harvey [1978] but are consistent with the general EMM definition. Non-estimability is apparent in both cases.

5. A Modified EMM

There might sometimes be utility in having a modified and estimable EMM whenever data have empty cells. For row i it would be an average of the cell means for cells containing data, and could be formally expressed as

$$EMM^*(\alpha_i) = \sum_{j=1}^{b} \delta_{ij}\mu_{ij} / \sum_{j=1}^{b} \delta_{ij}$$

for δ_{ij} being an indicator variable having value unity for filled cells and zero for empty cells:

$$\delta_{ij} = 1 \text{ for } n_{ij} > 0, \qquad \delta_{ij} = 0 \text{ for } n_{ij} = 0.$$

Clearly, whenever all cells of a row are filled, EMM* for that row will be identical to EMM. In all cases EMM* will be estimable, with b.l.u.e.

$$\widehat{EMM^*(\alpha_i)} = \sum_{j=1}^{b} \delta_{ij}\hat{\mu}_{ij} / \sum_{j=1}^{b} \delta_{ij} .$$

As illustration, for the with-interaction example of the preceding section, $EMM^*(\alpha_2) = \mu + \alpha_2 + \beta_1 + \gamma_{21}$ with b.l.u.e. $\hat{\mu}_{21} = \bar{y}_{21.}$. This is the mean of the cell means on which there are data in row 2. It differs from $EMM(\alpha_2)$ of (14), which is the mean of all cell means in row 2; and both $EMM^*(\alpha_2)$ and $EMM(\alpha_2)$ differ from (26) of Goodnight and Harvey [1978], namely $LSM(\alpha_2) = \mu + \alpha_2 + \frac{1}{2}(\beta_1 + \beta_2) + \gamma_{21}$ which is not a mean of cell means at all.

6. Extensions

Extension of the ideas described herein in terms of the 2-way classification to higher-order classifications is obviously easy to make. It can be based on a fully-parameterized cell means model $E(y_{ijk\ell\cdots st}) = \mu_{ijk\ell\cdots s}$ which can be written in matrix notation as $E(\underset{\sim}{y}) = X\underset{\sim}{\mu}$

where $\underset{\sim}{y}$ is N × 1, the vector of observations, $\underset{\sim}{\mu}$ is p × 1, the vector of p cell means, and $\underset{\sim}{X}$ is N × p, a direct sum of $\underset{\sim}{1}$-vectors, vectors having every element equal to unity. If the model includes restrictions on $\underset{\sim}{\mu}$, to take account of the absence of interactions for example, they can be expressed as

$$\underset{\sim}{H}\underset{\sim}{\mu} = \underset{\sim}{0}$$

for $\underset{\sim}{H}$ of full row rank. Then $\underset{\sim}{\mu}$ is estimable with, as in Speed et al. [1978],

$$\begin{aligned}\hat{\underset{\sim}{\mu}} &= (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y} - (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{H}'[\underset{\sim}{H}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{H}']^{-1}\underset{\sim}{H} \\ &= (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\{\underset{\sim}{I} - \underset{\sim}{H}'[\underset{\sim}{H}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{H}']^{-1}\underset{\sim}{H}\}\underset{\sim}{X}'\underset{\sim}{y}\end{aligned}$$

with $\underset{\sim}{H} \equiv \underset{\sim}{0}$ if there are no restrictions. Then, for any subscript in $\mu_{ijk\ell \ldots s}$, k say,

$$EMM(\mu_k) = \bar{\mu}_{..k...}$$

and

$$\widehat{EMM(\mu_k)} = \hat{\bar{\mu}}_{..k...} .$$

References

Goodnight, J. H. [1979]. New features in GLM and VARCOMP. Proceedings of the Fourth Annual SAS Users' Group International Conference, SAS Institute, Raleigh, N. C., 208-217.

Goodnight, J. H. and Harvey, W. R. [1978]. Least squares means in the fixed effects general linear model. Technical Report No. R-103, SAS Institute, Raleigh, N. C.

Harvey, W. R. [1968]. LSMLGP, Least squares maximum likelihood general purpose program. Dairy Science Department, Ohio State University, Columbus, Ohio.

Harvey, W. R. [1975]. Least-squares analysis of data with unequal subclass numbers. Report No. ARS H-4, A.R.S., U. S. Dept. of Agriculture, Beltsville, Maryland (ARS 20-8, 1960, Revised).

Searle, S. R. [1971]. Linear Models. John Wiley and Sons, New York.

Speed, F. M., Hocking, R. R. and Hackney, O. P. [1978]. Methods of analysis of linear models with unbalanced data. J. Amer. Statist. Assoc., 73, 105-112.