SAS® Technical Report R-103

Least-Squares Means in the Fixed-Effects General Linear Models

SAS Institute Inc.

5982

Least Squares Means in the Fixed Effects General Model

by

J. H. Goodnight Walter R. Harvey

SAS® Technical Report R-103

SAS Institute Inc.
SAS Circle, Box 8000
Cary, NC 27512-8000

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS® Technical Report R-103, Least-Squares Means in the Fixed-Effects General Linear Models*, Cary, NC: SAS Institute Inc., 1997. 9 pp.

SAS® Technical Report R-103, Least-Squares Means in the Fixed-Effects General Linear Models

ISBN 1-55544-967-7

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1. Introduction

The basic definition of Least Squares Means for unbalanced designs is given by Harvey (1975). Simply put, they are estimates of the class or subclass arithmetic means that would be expected had equal subclass numbers been obtainable. Although Harvey's work on Least Squares Means (LSM's) is in the framework of the reparameterized model with the "usual restrictions" imposed, his basic definition of LSM's needs little modification for the model without the usual restrictions imposed. This paper will attempt to point out the usefulness of LSM's, how they can be computed for the general linear model and show that in fact they do not always exist where missing cells prohibit their estimability.

2. The Model

Let the general linear model be

$$y = X\beta + e \tag{1}$$

where y is an nxl vector of individual observations; X is an nxk matrix of 0's and 1's and continuous independent variables; β is a kxl vector of constant but unknown parameters, and e is an nxl vector of random variables normally and independently distributed with common variance, σ_e^2.

3. Examples

To illustrate the nature of the expected values of class and subclass means normally computed for balanced designs, consider the following two factor main effects model:

$$y_{ijk} = u + \alpha_i + \beta_j + \varepsilon_{ijk} \tag{2}$$

where $i,j = 1,2$ and $k = 1,2,\ldots,n_{ij}$.

For balanced designs (all n_{ij}'s equal) the arithmetic means normally computed are: $\bar{y}_{..}$, $\bar{y}_{1.}$, $\bar{y}_{2.}$, $\bar{y}_{.1}$ and $\bar{y}_{.2}$. The overall mean has the following expected value:

$$E(\bar{y}_{..}) = \mu + \tfrac{1}{2}(\alpha_1+\alpha_2) + \tfrac{1}{2}(\beta_1+\beta_2) \tag{3}$$

The α main effect means have

$$E(\bar{y}_{1.}) = \mu + \alpha_1 + \tfrac{1}{2}(\beta_1+\beta_2) \tag{4}$$

$$E(\bar{y}_{2.}) = \mu + \alpha_2 + \tfrac{1}{2}(\beta_1+\beta_2) \tag{5}$$

and the β main effect means have

$$E(\bar{y}_{.1}) = \mu + \tfrac{1}{2}(\alpha_1+\alpha_2) + \beta_1 \tag{6}$$

$$E(\bar{y}_{.2}) = \mu + \tfrac{1}{2}(\alpha_1+\alpha_2) + \beta_2 \tag{7}$$

Given the expectations of the α main effect means in (4) and (5) it is clear that a test of significance of the difference between $\bar{y}_{1.}$ and $\bar{y}_{2.}$ is equivalent to a test of significance between α_1 and α_2. The same rational also applies to the β main effect means.

For unbalanced designs this equivalence no longer holds as is illustrated by letting the cell frequencies for (2) be: $n_{11} = n_{21} = n_{22} = 2$ and $n_{12} = 1$. The expectations of the class and subclass means are now:

$$E(\bar{y}_{..}) = \mu + 3/7\ \alpha_1 + 4/7\ \alpha_2 + 4/7\ \beta_1 + 3/7\ \beta_2 \tag{8}$$

$$E(\bar{y}_{1.}) = \mu + \alpha_1 + 2/3\ \beta_1 + 1/3\ \beta_2 \tag{9}$$

$$E(\bar{y}_{2.}) = \mu + \alpha_2 + \tfrac{1}{2}\beta_1 + \tfrac{1}{2}\beta_2 \tag{10}$$

$$E(\bar{y}_{.1}) = \mu + \tfrac{1}{2}\alpha_1 + \tfrac{1}{2}\alpha_2 + \beta_1 \tag{11}$$

$$E(\bar{y}_{.2}) = \mu + 1/3\alpha_1 + 2/3\alpha_2 + \beta_2 \tag{12}$$

A test of significance of the difference between $\bar{y}_{1.}$ and $\bar{y}_{2.}$ given in (9) and (10) above no longer represents a test of significance between α_1 and α_2 since the comparison of the means now involves β_1 and β_2. Thus a casual inspection of the raw cell means in an unbalanced design may well lead to incorrect inferences about the main effects in question. For model (2) with unequal n's, computing B.L.U.E.'s which have the same expected value as (3)-(7) would be far more informative than computing the raw cell means.

4. LSM's Defined

By defining the LSM's for an effect as a linear combination of the parameters of the model, (in other words: a "super" parameter), then the general theory of estimability may be used to decide whether or not these "super" parameters are estimable. In this context, the LSM of μ in model (2) or simply LSM(μ) may be defined as:

$$\text{LSM}(\mu) = \mu + \tfrac{1}{2}(\alpha_1+\alpha_2) + \tfrac{1}{2}(\beta_1+\beta_2)$$

The LSM's for α_i may be defined as:

$$\text{LSM}(\alpha_i) = \mu + \alpha_i + \tfrac{1}{2}(\beta_1+\beta_2). \qquad i=1,2$$

The LSM's for β_j in model (2) may be defined as:

$$\text{LSM}(\beta_j) = \mu + \tfrac{1}{2}(\alpha_1+\alpha_2) + \beta_j. \qquad j=1,2$$

For the general linear model defined in (1), each LSM is a linear combination of the β parameters and can be expressed as $L\beta$ where L is a lxk vector. The actual construction of the elements of L given below makes use only of the parameters present for a given set of data and can lead to the construction of an L such that $L\beta$ is not estimable. This merely implies that for data with missing cells it is not always possible to construct a statistic which has the same expected value as the corresponding class or subclass arithmetic mean in the balanced case.

To construct LSM's for models, involving many different effects, such as:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_k + \gamma X_{ijkl} + \varepsilon_{ijkl} \qquad (13)$$

the LSM super parameters for any effect in the model may be constructed by first writing down the expected value of the dependent variable given that all covariables are at their overall mean; such as:

$$E(y_{ijkl}|x_{ijkl} = \bar{x}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_k + \gamma\bar{x} \qquad (14)$$

Once this has been done, the LSM for any effect may be written by holding that effects subscripts constant and summing each other effect in the model over its remaining subscripts, if any. For the above example:

$$\mathrm{LSM}(\mu) = \mu + \frac{1}{N_i}\sum_i \alpha_i + \frac{1}{N_j}\sum_j \beta_j + \frac{1}{N_{ij}}\sum (\alpha\beta)_{ij} + \frac{1}{N_k}\sum \delta_k + \gamma\bar{x} \qquad (15)$$

$$\mathrm{LSM}(\alpha_i) = \mu + \alpha_i + \frac{1}{N_j}\sum_j \beta_j + \frac{1}{N_{j/i}}\sum_{j/i} (\alpha\beta)_{ij} + \frac{1}{N_k}\sum \delta_k + \gamma\bar{x} \qquad (16)$$

$$\mathrm{LSM}(\beta_j) = \mu + \frac{1}{N_i}\sum_i \alpha_i + \beta_j + \frac{1}{N_{i/j}}\sum_{i/j} (\alpha\beta)_{ij} + \frac{1}{N_k}\sum \delta_k + \gamma\bar{x} \qquad (17)$$

$$\mathrm{LSM}(\alpha\beta_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \frac{1}{N_k}\delta_k + \gamma\bar{x} \qquad (18)$$

$$LSM(\delta_k) = \mu + \frac{1}{N_i}\sum_i \alpha_i + \frac{1}{N_j}\sum_j \beta_j + \frac{1}{N_{ij}}\Sigma(\alpha\beta)_{ij} + \delta_k + \gamma\bar{x} \tag{19}$$

where,

N_i = the total number of α_i terms,

N_j = the total number of β_j terms,

N_k = the total number of δ_k terms,

N_{ij} = the total number of $(\alpha\beta)_{ij}$ terms,

$N_{j/i}$ = the number of $(\alpha\beta)_{ij}$ terms for a given i,

$N_{i/j}$ = the number of $(\alpha\beta)_{ij}$ terms for a given j.

In equation (14) the overall mean of x was used. Although any constant could be used in place of $\bar{x}$, its use will produce LSM estimates, which correspond to adjusted means as defined by most authors. For models involving both a covariate and deviations from the average slope such as:

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + \gamma_i x_{ij} + \varepsilon_{ij} \tag{20}$$

$$LSM(\mu) = \mu + \frac{1}{N_i}\sum_i \alpha_i + \gamma\bar{x} + \frac{1}{N_i}\Sigma\, \gamma_i\bar{x} \tag{21}$$

$$LSM(\alpha_i) = \mu + \alpha_i + \gamma\bar{x} + \gamma_i\bar{x} \tag{22}$$

5. Relationship to Adjusted Means

For the model,

$$y_{ij} = \mu + \alpha_i + \beta\, x_{ij} + \varepsilon_{ij}, \tag{23}$$

main effect means are usually adjusted for the covariable as:

$$\bar{y}_{i.}\ adj = \bar{y}_{i.} - b(\bar{x}_{i.} - \bar{x}_{..}) \tag{24}$$

where b is the L.S. estimate of β in model (23). Taking the expected value of (24)yields:

$$E(\bar{y}_{i.}\ adj) = E(\bar{y}_i) - \beta(\bar{x}_{i.} - \bar{x}_{..})$$
$$= \mu + \alpha_i + \beta\ \bar{x}_{i.} - \beta\ \bar{x}_{i.} + \beta\ \bar{x}_{..}$$
$$= \mu + \alpha_i + \beta\ \bar{x}_{..}$$

which corresponds to the LSM (α_i).

6. Non-Estimability

The basic definition of an LSM, as given in section 4, sometimes leads to non-estimable "super" parameters. For example, consider the model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

with the observed cell frequencies: $n_{11}, n_{12}, n_{21} > 0$ and $n_{22} = 0$. Because of the missing cell, the list of parameters $\beta' =$

$[\mu\ \alpha_1\ \alpha_2\ \beta_1\ \beta_2\ (\alpha\beta)_{11}\ (\alpha\beta)_{12}\ (\alpha\beta)_{21}]$

and thus

$$LSM(\alpha_1) = \mu + \alpha_1 + \tfrac{1}{2}\beta_1 + \tfrac{1}{2}\beta_2 + \tfrac{1}{2}(\alpha\beta)_{11} + \tfrac{1}{2}(\alpha\beta)_{12} \tag{25}$$

$$LSM(\alpha_2) = \mu + \alpha_2 + \tfrac{1}{2}\beta_1 + \tfrac{1}{2}\beta_2 + (\alpha\beta)_{21} \tag{26}$$

$$LSM(\beta_1) = \mu + \tfrac{1}{2}\alpha_1 + \tfrac{1}{2}\alpha_2 + \beta_1 + \tfrac{1}{2}(\alpha\beta)_{11} + \tfrac{1}{2}(\alpha\beta)_{21} \tag{27}$$

$$LSM(\beta_2) = \mu + \tfrac{1}{2}\alpha_1 + \tfrac{1}{2}\alpha_2 + \beta_2 + \alpha\beta_{12} \tag{28}$$

$$LSM(\alpha\beta_{11}) = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11} \tag{29}$$

$$LSM(\alpha\beta_{12}) = \mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12} \tag{30}$$

$$LSM(\alpha\beta_{21}) = \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21} \tag{31}$$

It is easily verified that neither LSM(α_2) nor LSM(β_2) is estimable for the particular set of data. Simply put, this is the price one pays for having missing data.

7. Variances and Covariances

Since all LSM's are defined in terms of estimable functions their variances and covariances are easily computed. For example if the β vector of (1) contained the three parameters α_1, α_2 and α_3. Then

$$\text{LSM}(\alpha_1) = L_1\beta \tag{32}$$

$$\text{LSM}(\alpha_2) = L_2\beta \tag{33}$$

$$\text{LSM}(\alpha_3) = L_3\beta \tag{34}$$

where each vector L_1, L_2, and L_3 is 1xk and whose elements are constructed as described in section 4.

Letting $L = \begin{bmatrix} L_1 \\ \hline L_2 \\ \hline L_3 \end{bmatrix}$

then the variance-covariance matrix of the LSM's for α is

$$L(X'X)^{-} L' \sigma_e^2$$

where $(X'X)^{-}$ is any generalized (g^2) inverse of $X'X$.

Since LSM's are to unbalanced designs as class and subclass arithmetic means are to balanced designs, they should be used accordingly. However, care must be taken in any direct comparison of LSM's since they are usually correlated.

References

(1975) Harvey, Walter R. "Least-Squares Analysis of Data with Unequal Subclass Numbers," ARS H-4, Agricultural Research Service, U. S. Dept. of Agriculture, Room 13, NAL Bldg., Beltsville, Maryland 20705 (Formerly ARS 20-8 July 1960).

Index

SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513

ISBN 1-55544-969-7
9 781555 449698