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.,


Copyright © 1997 by SAS Institute Inc., Cary, NC, USA.

ISBN 1-55544-967-7

All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

U.S. Government Restricted Rights Notice

Software and accompanying documentation are provided to the U.S. government in a transaction subject to the Federal Acquisition Regulations with Restricted Rights. Use, duplication, or disclosure of the software by the government is subject to restrictions as set forth in FAR 52.227-19 Commercial Computer Software-Restricted Rights (June 1987). The Contractor/Licensor is SAS Institute Inc., located at SAS Campus Drive, Cary, North Carolina 27513.


The SAS® System is an integrated system of software providing complete control over data access, management, analysis, and presentation. Base SAS software is the foundation of the SAS System. Products within the SAS System include SAS/ACCESS®, SAS/AF®, SAS/ASSIST®, SAS/CALC®, SAS/CONNECT®, SAS/CPE®, SAS/DMP®, SAS/EIS®, SAS/ENGLISH®, SAS/ETS®, SAS/FSP®, SAS/GRAPH®, SAS/IMAGE®, SAS/IML®, SAS/IMS-DL/I®, SAS/INSIGHT®, SAS/IntrNet®, SAS/LAB®, SAS/MDB®, SAS/NVISION®, SAS/OR®, SAS/PH-Clinical®, SAS/QC®, SAS/REPLAY-CICS®, SAS/SESSION®, SAS/SHARE®, SAS/SPECTRAVIEW®, SAS/STAT®, SAS/TOOLKIT®, SAS/TRADER®, SAS/TUTOR®, SAS/DB2®, SAS/GEOSAS®/GIS®, SAS/PH-Kinetics®, SAS/SHARE*NET®, and SAS/SQL-DS™ software. Other SAS Institute products are SYSTEM 2000* Data Management Software, with basic SYSTEM 2000, CREATE*, Multi-User*, QueX*, Screen Writer*, and CICS interface software; InfoTap* software; JAZZ* software; NeoVisuals* software; JMP*, JMP IN* and JMP Serve* software; SAS/RTERM* software; and the SAS/QC® Compiler and the SAS/CX* Compiler; Video Reality™ software; VisualSpace™ software; Budget Vision®, Campaign Vision®, CFO Vision®, Compensation Vision®, HR Vision®, and IT Service Vision™ software; Scalable Performance Data Server™ software; SAS OnlineTutor® software; and Emulus* software. MultiVendor Architecture® and MVA® are trademarks of SAS Institute Inc. SAS Institute also offers SAS Consulting® and SAS Video Productions® services. Authorline®, Books by Users®, The Encore Series®, ExecSolutions®, JMP® Cable®, Observations®, SAS Communications®, SAS OnlineDoc®, SAS Professional Services®, SAS Views®, the SASware Balloon®, SelectText®, and Solutions@Work® documentation are published by SAS Institute Inc. The SAS Video Productions logo, the Books by Users SAS Institute’s Author Service logo, and The Encore Series logo are registered service marks or registered trademarks of SAS Institute Inc. The Helpplus logo, the SelectText logo, the Video Reality logo, the SAS Online Samples logo, the Quality Partner logo, the SAS Business Solutions logo, the SAS Rapid Warehousing Program logo, the SAS Publications logo, the Instructor-based Training logo, the Online Training logo, the Trainer’s Kit logo, and the Video-based Training logo are service marks or trademarks of SAS Institute Inc. All trademarks above are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

The Institute is a private company devoted to the support and further development of its software and related services.

Other brand and product names are registered trademarks or trademarks of their respective companies.
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

\[ y = X\beta + e \]  

(1)

where \( y \) is an nx1 vector of individual observations; \( X \) is an nxk matrix of 0's and 1's and continuous independent variables; \( \beta \) is a kx1 vector of constant but unknown parameters, and \( e \) is an nx1 vector of random variables normally and independently distributed with common variance, \( \sigma^2_e \).

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:

\[ y_{ijk} = u + \alpha_i + \beta_j + \epsilon_{ijk} \]  

(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:

$$E(\bar{y}_{..}) = \mu + \frac{1}{4}(\alpha_1 + \alpha_2) + \frac{1}{4}(\beta_1 + \beta_2)$$

(3)

The $\alpha$ main effect means have

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

(4)

$$E(\bar{y}_{2.}) = \mu + \alpha_2 + \frac{1}{4}(\beta_1 + \beta_2)$$

(5)

and the $\beta$ main effect means have

$$E(\bar{y}_{.1}) = \mu + \frac{1}{4}(\alpha_1 + \alpha_2) + \beta_1$$

(6)

$$E(\bar{y}_{.2}) = \mu + \frac{1}{4}(\alpha_1 + \alpha_2) + \beta_2$$

(7)

Given the expectations of the $\alpha$ 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 $\alpha_1$ and $\alpha_2$. The same rational also applies to the $\beta$ 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:

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

(8)

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

(9)

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

(10)

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

(11)
\[ E(\bar{\bar{y}}_{2}) = \mu + 1/3\alpha_1 + 2/3\alpha_2 + \beta_2 \] (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 \( \alpha_1 \) and \( \alpha_2 \) since the comparison of the means now involves \( \beta_1 \) and \( \beta_2 \). 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 \( \mu \) in model (2) or simply LSM(\( \mu \)) may be defined as:

\[ \text{LSM}(\mu) = \mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \frac{1}{2}(\beta_1 + \beta_2) \]

The LSM's for \( \alpha_i \) may be defined as:

\[ \text{LSM}(\alpha_i) = \mu + \alpha_i + \frac{1}{2}(\beta_1 + \beta_2). \quad i=1,2 \]

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

\[ \text{LSM}(\beta_j) = \mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_j. \quad j=1,2 \]

For the general linear model defined in (1), each LSM is a linear combination of the \( \beta \) 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 Lβ 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:

$$y_{ijkl} = \mu + a_i + \beta_j + (a\beta)_{ij} + \delta_k + \gamma x_{ijkl} + \epsilon_{ijkl}$$

(13)

corresponding class or subclass arithmetic mean in the balanced case.

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:

$$E(y_{ijkl} | x_{ijkl} = \bar{x}) = \mu + a_i + \beta_j + (a\beta)_{ij} + \delta_k + \gamma \bar{x}$$

(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:

$$\text{LSM}(\mu) = \mu + \frac{1}{N_i} \sum_i a_i + \frac{1}{N_j} \sum_j \beta_j + \frac{1}{N_{ij}} \sum_{ij} (a\beta)_{ij} + \frac{1}{N_k} \sum_k \delta_k + \gamma \bar{x}$$

(15)

$$\text{LSM}(a_i) = \mu + a_i + \frac{1}{N_j} \sum_j \beta_j + \frac{1}{N_{ij}} \sum_{ij} (a\beta)_{ij} + \frac{1}{N_k} \sum_k \delta_k + \gamma \bar{x}$$

(16)

$$\text{LSM}(\beta_j) = \mu + \frac{1}{N_i} \sum_i a_i + \beta_j + \frac{1}{N_{ij}} \sum_{ij} (a\beta)_{ij} + \frac{1}{N_k} \sum_k \delta_k + \gamma \bar{x}$$

(17)

$$\text{LSM}(a\beta_{ij}) = \mu + a_i + \beta_j + (a\beta)_{ij} + \frac{1}{N_k} \delta_k + \gamma \bar{x}$$

(18)
\[
\text{LSM}(\delta_k) = \mu + \frac{1}{N_i} \sum_{i} \alpha_i + \frac{1}{N_j} \sum_{j} \beta_j + \frac{1}{N_{ij}} \sum_{ij} (\alpha\beta)_{ij} + \delta_k + \gamma \bar{x}
\]  \hspace{1cm} (19)

where,

\[\begin{align*}
N_i & = \text{the total number of } \alpha_i \text{ terms}, \\
N_j & = \text{the total number of } \beta_j \text{ terms}, \\
N_k & = \text{the total number of } \delta_k \text{ terms}, \\
N_{ij} & = \text{the total number of } (\alpha\beta)_{ij} \text{ terms}, \\
N_j/i & = \text{the number of } (\alpha\beta)_{ij} \text{ terms for a given } i, \\
N_i/j & = \text{the number of } (\alpha\beta)_{ij} \text{ terms for a given } j.
\end{align*}\]

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:

\[
y_{ij} = \mu + \alpha_i + \gamma x_{ij} + \gamma_i x_{ij} + \varepsilon_{ij}
\]  \hspace{1cm} (20)

\[
\text{LSM}(\mu) = \mu + \frac{1}{N_i} \sum_{i} \alpha_i + \gamma \bar{x} + \frac{1}{N_i} \sum_{i} \gamma_i \bar{x}
\]  \hspace{1cm} (21)

\[
\text{LSM}(\alpha_i) = \mu + \alpha_i + \gamma \bar{x} + \gamma_i \bar{x}
\]  \hspace{1cm} (22)

5. Relationship to Adjusted Means

For the model,

\[
y_{ij} = \mu + \alpha_i + \gamma x_{ij} + \varepsilon_{ij},
\]  \hspace{1cm} (23)

main effect means are usually adjusted for the covariable as:

\[
\bar{y}_{i. \text{ adj}} = \bar{y}_{i.} - b(\bar{x}_{i.} - \bar{x}_{..})
\]  \hspace{1cm} (24)
where $b$ is the L.S. estimate of $\beta$ in model (23). Taking the expected value of (24) yields:

$$E(\tilde{y}_{1, \text{adj}}) = E(\tilde{y}_{1}) - \beta(\bar{x}_1) - \bar{x}_2$$

$$= \mu + \alpha_1 + \beta \bar{x}_1 - \beta \bar{x}_1 + \beta \bar{x}_2$$

$$= \mu + \alpha_1 + \beta \bar{x}_2$$

which corresponds to the LSM ($\alpha_1$).

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:

$$y_{ijk} = \mu + \alpha_1 + \alpha_2 + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

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

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

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

$$\text{LSM}(\beta_1) = \mu + \frac{1}{2} \alpha_1 + \frac{1}{2} \alpha_2 + \beta_1 + \frac{1}{2} (\alpha\beta)_{11} + \frac{1}{2} (\alpha\beta)_{12} \quad (27)$$

$$\text{LSM}(\beta_2) = \mu + \frac{1}{2} \alpha_1 + \frac{1}{2} \alpha_2 + \beta_2 + (\alpha\beta)_{12} \quad (28)$$

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

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

$$\text{LSM}(\alpha\beta_{21}) = \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21} \quad (31)$$
It is easily verified that neither LSM($\alpha_2$) nor LSM($\beta_2$) 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 $\beta$ vector of (1) contained the three parameters $\alpha_1$, $\alpha_2$ and $\alpha_3$. Then

\[
\text{LSM}(\alpha_1) = L_1\beta \\
\text{LSM}(\alpha_2) = L_2\beta \\
\text{LSM}(\alpha_3) = L_3\beta
\]

(32)  
(33)  
(34)

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

Letting

\[
L = \begin{bmatrix}
L_1 \\
L_2 \\
L_3
\end{bmatrix}
\]

then the variance-covariance matrix of the LSM's for $\alpha$ is

\[
L(X'X)^{-1}L'\sigma^2_e
\]

where $(X'X)^{-1}$ 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

Index

F
fixed-effects linear model
  equation 2
  Least Squares Means 2

L
Least Squares Means
  computing variances and covariances 8
  defined 2, 4
  non-estimability of parameters 7
  unbalanced designs 2
  usefulness of 2

U
unbalanced designs
  Least Squares Means 2

V
variances
  computing for Least Squares Means 8