LSMEANS

A common question asked about GLM is the difference between the MEANS and LSMEANS statements. In some cases they are equivalent and at other times LSMEANS are more appropriate. The definition of each is as follows:

MEANS - These are what is usually meant by mean (average) and are computed by summing all the data points and dividing by the total # of points.
They are also referred to as arithmetic means and they are based on the data only.

LSMEANS - Least Squares Means can be defined as a linear combination (sum) of the estimated effects (means, etc) from a linear model. These means are based on the model used.

In the case where the data contains NO missing values, the results of the MEANS and LSMEANS statements are identical. When missing values do occur, the two will differ. In such a case the LSMEANS are preferred because they reflect the model that is being fit to the data. LSMEANS are also used when a covariate(s) appears in the model such as in ANCOVA (See handout # 4).

The following example illustrates the similarity and difference between theses two methods in balanced and unbalanced data.

EXAMPLE:

This data set has a factor A with 3 levels (1, 2, & 3) with 3 reps of each.

	Factor A		
Rep	1	2	3
1	4	7	4
2	6	3	2
3	2	5	3
GX_{ai}	12	15	9
$\bar{X}_{a.}$	4	5	3

Means for the 3 levels of factor A ($\bar{X}_{a.}$) are given below each respective column.

A MEANS statement would calculate the overall mean of factor A by summing all 9 data points & dividing by 9,

$$\bar{X}.. = (EEX_{ai})/n = (4 + 6 + 2 + ... + 4 + 2 + 3)/9 = 4.0 .$$

The LSMEANS statement would use a linear combination of the estimated factor A effects, which in this case are the factor A means, $\bar{X}_{a.}$,

$$\bar{X}.. = (E\bar{X}_{a.})/3 = (4 + 5 + 3)/3 = 4.0$$

Since the data were balanced the two methods produced the same result. If we delete a data point however, the results will change. Suppose the data were revised as follows:

	Factor A		
Rep	1	2	3
1	4	.	4
2	6	3	2
3	2	5	3
GX_{ai}	12	8	9
$\bar{X}_{a.}$	4	4	3

Note that the second level of factor A at rep 1 is now deleted and hence the sums and means are updated to reflect this change.

The MEANS statement now produces:

$$\bar{X}.. = (GG X_{ai})/n = (4 + 6 + 2 + ... + 4 + 2 + 3)/8 = 3.625,$$

whereas the LSMEANS gives:

$$\bar{X}.. = (G\bar{X}_{a.})/3 = (4 + 4 + 3)/3 = 3.667.$$

Thus, when the data includes missing values, the average of all the data will no longer equal the average of the averages. LSMEANS is the proper choice here because it imposes the treatment structure of factor A on the calculated mean $\bar{X}..$ There is no inherent structure implied by the MEANS statement. The exact difference between MEANS and LSMEANS becomes more obscure with increasingly complex treatment arrangements and experimental designs. When covariates are present in the model, the LSMEANS statement produces means which are adjusted for the average value of the specified covariate(s).