ANOVA on Unbalanced Design with Unequal Cells Using SAS

George C. Chao
Abbott Laboratories
North Chicago, Illinois

Due to unequal cell frequency in the experimental design, the variations of overall main effects and interactions are nonadditive. Several different least squares methods, e.g., Graybill (1961), Searle (1971) and Winer (1971), can be shown to yield identical results in the balanced design situation but will yield substantially different results when applied to data involving unequal cell frequency. Three least squares methods for the analysis of experimental data are of particular interest.

Method I General Linear Model	Method II Experimental Design	Method III Fitting Constants Method
A|B, AB	A|B	A
B|A, AB	B|A	B|A
AB|A, B	AB|A, B	AB|A, B

Selection of one of the three methods should depend upon conceptualization of the problem and the nature of questions to be asked. Actually, the selection depends upon who is "the statistician", or which computer program is available.

The goal of this presentation is not to argue which model should be used. There are a lot of papers of this nature, such as Overall & Spiegel (1969), Francis (1973), Kutner (1974), Winer (1971), etc. There are different favored methods in these papers and even different method names. Rather, I consider how to use SAS '72 to analyze the unbalanced design with varied methods.

In SAS procedure REGR, the "Sequential SS" is equivalent to "Fitting Constant Method", i.e., Method III in the previous table, (also called the "Stepwise Forward Method") and the "Partial SS" is equivalent to "General Linear Model", i.e., Method I in the previous table, (also called "Backward Approach Method").

To my knowledge, there is no major statistical package that will do Method II, directly. Contrary to Francis' negative comment on SAS, SAS-REGR maybe the only package that will provide all three methods without too much trouble.

The generalization of Method II from the above two way factorial design is as follows:

(1) Each covariate is adjusted by the rest of the covariates.

(2) Each main effect is adjusted by all covariates and the rest of the main effects.

(3) Each 2nd order interaction is adjusted by all covariates, main effects and the rest of the 2nd order interactions.

(4) In general any order interaction is adjusted by all covariates, main effects, lower order interactions and the rest of same order interactions.

The corresponding SAS-REGR models are as follows:

(1) MODEL Y = X1 X2 X3;

(2) MODEL Y = X1 X2 X3 A B C;

(3) MODEL Y = X1 X2 X3 A B C AB AC BC;

(4) MODEL Y = X1 X2 X3 A B C AB AC BC ABC;

For the computation of F-value, use the Partial SS of each additional variable in every model as numerator and the error of the last model as denominator. The F-value can be calculated "by hand".

Example

Consider the following set of data taken from Kutner (1974), which were adopted from Afifi and Azen (1972). The measurment to be analyzed was the increase in systolic pressure (mmHg) due to the treatment applied to six dogs.

	Disease			
Drug	1	2	3	Num. Mean
1	44,42,36 22,19,13	33,33,26 21	31,25,25 24,-3	15 26.07
2	42,28,24 23,13	36,34,33 31	32,28,26 16, 4, 3	15 24.87
3	29,19, 1	11, 9, 7 1,-6	21, 9, 3 1	12 8.75
4	24,22,15 9,-2	27,16,15 12,12,-5	25,22,12 7, 5	16 13.50
Num.	19	19	20	58
Mean	22.26	18.21	15.80	18.71

SAS-REGR

MODEL (1) Y = DISEASE DRUG;

SOURCE	DF	SEQ SS	PARTIAL SS
DISEASE	2	413.975	359.503
DRUG	3	2950.629	2950.629
ERROR	52	5771.413	
TOTAL	57	9136.017	

MODEL (2) Y = DISEASE DRUG DISEASE*DRUG;

SOURCE	DF	SEQ SS	PARTIAL SS
DISEASE	2	413.975	366.179
DRUG	3	2950.629	2861.920
DISEASE*DRUG	6	730.597	730.597
ERROR	46	5040.817	
TOTAL	57	9136.017	

The above result does not agree with Kutner's; his TOTAL SS = 9340.155 with 57 d.f. (calculated using BMD10V).[1]

From SAS-REGR Model (2) - PARTIAL SS, we have the ANOVA table of Model I as follows:

SOURCE	DF	SS
DISEASE|DRUG, DRUG*DISEASE	2	366.179
DRUG|DISEASE, DRUG*DISEASE	3	2861.920**
DRUG*DISEASE|DRUG, DISEASE	6	730.597
ERROR	46	5040.817

[1] It was noted in the presentation that there was a misprint in Kutner's data.

From SAS-REGR Model (2) - SEQ SS we have the ANOVA table of Model III as follows.

SOURCE	DF	SS
DISEASE	2	413.975
DRUG|DISEASE	3	2950.629**
DRUG*DISEASE|DISEASE, DRUG	6	730.597
ERROR	46	5040.817

From SAS-REGR Model (1) - PARTIAL SS, we have DISEASE|DRUG and DRUG|DISEASE. From Model (2) - PARTIAL SS, we have DRUG*DISEASE|DRUG, DISEASE. The ANOVA table of Model II is as follows.

SOURCE	DF	SS
DISEASE|DRUG	2	359.503
DRUG|DISEASE	3	2950.629**
DISEASE*DRUG|DRUG, DISEASE	6	730.597
ERROR	46	5040.817

The adjusted means of the increase in systolic pressure for each disease on each model are obtained from SAS-REGR and are as follows:

Disease	N	Unadj	Adj. by Drug	Adj. by Drug & Interaction
1	19	22.2632	20.9317	21.3167
2	19	18.2105	19.2393	19.7458
3	20	15.8000	14.9928	15.3167
Total	58	18.7069	18.3294	18.7331

Most statisticians expect the weighted average of adjusted means to equal the unadjusted overall mean. The differences noted in the table above may be due to the underlying assumptions related to the unequal sample size; however, this point is not documented in the manual. Hopefully, the new GLM procedure will provide a clear explanation of the adjusted means.

REFERENCE

1. Afifi, A. A. and Azen, S. P. (1972). Statistical Analysis A Computer Oriented Approach, Academic Press, N.Y.

2. Francis, I. (1973). Comparison of Several Analysis of Variance Programs, JASA 68, p. 860-865.

3. Graybill, F. A. (1961). An Introduction to Linear Statistical Models, Vol. I, McGraw-Hill Book Co., N.Y.

4. Kutner, M. H. (1974). Hypothesis Testing in Linear Models (Eisenhart Model I), The American Statistician 28, p. 98-100.

5. Overall, J. R. and Spiegel, D. K. (1969). Concerning Least Squares Analysis of Experimental Data, Psychological Bulletin 72, p. 311-322.

6. Searle, S. R. (1971). Linear Models, John Wiley & Sons, N.Y.

7. Winer, B. J. (1971). Statistical Principles in Experimental Design, 2nd Edition, McGraw-Hill Book Co., N.Y.