**Unbalanced data with empty cells**(missing treatment combinations)

**Yuqin Liao**

There are four types of Sum squares which are Type I Sequential SS, Type II Hierarchical or Partially Sequential SS, type III Marginal or Orthogonal SS, and a variation of type III-type IV SS, which is specifically developed for designs with missing cells and that’s the one I’m going to focus on.

Testing an ANOVA Ho is equivalent to testing a set of ***νn*** linearly independent contrasts, where ***νn*** is the degrees of freedom of Ho. These contrasts are formed by averaging over the levels of other factors in the experiment. With data from design without missing cells, all linearly independent sets of contrasts yield the same SS for Ho. But with data from a design missing some treatment combinations (having empty cells), not all contrasts can average over all levels of other factors. Since missing levels cannot be averaged, different sets of contrasts yield different SS for Ho.

The contrasts that are chosen by SAS to compute Type IV SS are dependent on the order of input of the data, so that by relabeling or inputting data in a different fashion, we can get a different SS for the same null hypothesis. Contrast statement in PROC GLM using effects model coding can ensure that the set of contrasts tested for Ho is the set that we wish to test. The E4 option in the model statement can have SAS print out the contrasts that it is using to test Ho. When we are computing least-squares means, if all levels of other factors are not observed with a particular level of an effect, then that level is said to be non-estimable.

**SAS example:**

**DATA:**

**data miss1 ;**

**input a b n ;**

**do i= 1 to n ;**

**input y @@;**

**c = 4 - b ;**

**output;**

**end;**

**title 'ANOVA with missing cells';**

**cards;**

**1 1 2**

**5 6**

**1 2 5**

**2 3 5 6 7**

**2 1 2**

**2 3**

**2 2 3**

**8 8 9**

**2 3 5**

**4 4 6 6 7**

**; run ;**

**proc print data=miss1 ; run;**

Obs a b n i y c

1 1 1 2 1 5 3

2 1 1 2 2 6 3

3 1 2 5 1 2 2

4 1 2 5 2 3 2

5 1 2 5 3 5 2

6 1 2 5 4 6 2

7 1 2 5 5 7 2

8 2 1 2 1 2 3

9 2 1 2 2 3 3

10 2 2 3 1 8 2

11 2 2 3 2 8 2

12 2 2 3 3 9 2

13 2 3 5 1 4 1

14 2 3 5 2 4 1

15 2 3 5 3 6 1

16 2 3 5 4 6 1

17 2 3 5 5 7 1

Class Level Information

Class Levels Values

a 2 1 2

b 3 1 2 3

Number of observations 17

**Code 1:**

**proc** **glm** ;

class a b;

model y=a b a\*b /e3 e4 ss3 ss4 ;

lsmeans a\*b ;

contrast 'u21-u23 and u22-u23'\* Specify the contrast;

b **0** **1** -**1** a\*b **0** **0** **0** **1** -**1** ,

b **1** **0** -**1** a\*b **0** **0** **1** **0** -**1** ;

**run**;

Type IV Estimable Functions

------Coefficients------

Effect a b a\*b

Intercept 0 0 0

a 1 L2 0 0

a 2 -L2 0 0

b 1 0 L4 0

b 2 0 L5 0

b 3 0 -L4-L5 0

a\*b 1 1 0.5\*L2 0 L7

a\*b 1 2 0.5\*L2 0 -L7

a\*b 2 1 -0.5\*L2 L4 -L7

a\*b 2 2 -0.5\*L2 L5 L7

a\*b 2 3 0 -L4-L5 0

NOTE: Other Type IV estimable functions exist.

**For b: since** ****

**Set L4=0, L5=1, then we can get the contrast “ ”,**

**Set L4=1, L5=0, then we can get another contrast: “ ”.**

Source DF Type III SS Mean Square F Value Pr > F

a 1 0.35072464 0.35072464 0.16 0.6949

b 2 16.07330642 8.03665321 3.70 0.0560

a\*b 1 29.56811594 29.56811594 13.61 0.0031

Source DF Type IV SS Mean Square F Value Pr > F

a 1\* 0.35072464 0.35072464 0.16 0.6949

b 2\* 41.73333333 20.86666667 9.61 0.0032

a\*b 1 29.56811594 29.56811594 13.61 0.0031

Dependent Variable: y

Contrast DF Contrast SS Mean Square F Value Pr > F

u21-u23 and u22-u23 2 41.73333333 20.86666667 9.61 0.0032

Code 2:

proc glm ;

class a c;

model y=a c a\*c /e3 e4 ss3 ss4 ;

lsmeans a\*c ;

contrast '(u12+u22)-(u13-u23)and u21-u23'

c 0 1 -1 a\*c .5 -.5 0 .5 -.5 ,

c 1 0 -1 a\*c 0 0 1 0 -1 ;

run;

Type IV Estimable Functions

--------Coefficients--------

Effect a c a\*c

Intercept 0 0 0

a 1 L2 0 0

a 2 -L2 0 0

c 1 0 L4 0

c 2 0 L5 0

c 3 0 -L4-L5 0

a\*c 1 2 0.5\*L2 0.5\*L5 L7

a\*c 1 3 0.5\*L2 -0.5\*L5 -L7

a\*c 2 1 0 L4 0

a\*c 2 2 -0.5\*L2 0.5\*L5 -L7

a\*c 2 3 -0.5\*L2 -L4-0.5\*L5 L7

**For c:**

Set L4=0, L5=1, then we can get the contrast: “(u12+u22)-(u13-u23)”.

Set L4=1, L5=0, then we can get another contrast: “u21-u23”.

Source DF Type III SS Mean Square F Value Pr > F

a 1 0.35072464 0.35072464 0.16 0.6949

c 2 16.07330642 8.03665321 3.70 0.0560

a\*c 1 29.56811594 29.56811594 13.61 0.0031

Source DF Type IV SS Mean Square F Value Pr > F

a 1\* 0.35072464 0.35072464 0.16 0.6949

c 2\* 18.97786775 9.48893387 4.37 0.0376

a\*c 1 29.56811594 29.56811594 13.61 0.0031

Dependent Variable: y

Contrast DF Contrast SS Mean Square F Value Pr > F

(u12+u22)-(u13-u23)and u21-u23 2 18.97786775 9.48893387 4.37 0.0376

The results show that when the order of input of the data is changed, the type III SS keeps the same, but type IV SS doesn’t do this. And we could see what contrast the SAS is using by e3 e4 options. Different contrasts have an influence on type IV SS but not on type III SS since this is an unbalanced data with empty cells.