

## An Easy and Convenient Method for Constructing Contrasts

Robert Vogel, Georgia Southern University, Statesboro, Georgia

### ABSTRACT

In many experiment design and observational study settings the researcher is not interested in the omnibus tests of hypotheses for "main effects" or "interaction effects." When confronted with this situation, SAS™ software provides an easy and convenient way to test the "testable" hypotheses of interest via the CONTRAST and ESTIMATE statements. This paper will demonstrate an easy way to construct interpretable and testable hypotheses.

### INTRODUCTION

In many situations arising in experimental design and observational studies, researchers are interested in "testable" hypotheses that are not associated with the ANOVA table provided as part of the default output. To acquire the appropriate sums of squares and p-values for these testable hypotheses researcher need to use either the CONTRAST or ESTIMATE statements that are provided in several SAS™ procedures. Although these statements have been available for many years, it has been my experience that researchers frequently have difficulty in writing "testable" hypotheses and "estimable" functions. In this paper, we will provide several examples as to how to write an "estimable" function of the parameters that represent the "testable" hypothesis of interest. The examples come from Applied Linear Statistical Models, 5th edition by Kutner, Nachtsheim, Neter and Li (KNNL).

### MODELS

The key to writing "estimable" functions of the model parameters is to realize that most people can visualize differences between means as opposed to differences between effects. For example, consider a two-factor (factors A and B) design with interaction:

	B=1	B=2	Marginal
A=1	$\mu_{11}$	$\mu_{12}$	$\mu_{1.}$
A=2	$\mu_{21}$	$\mu_{22}$	$\mu_{2.}$
A=3	$\mu_{31}$	$\mu_{32}$	$\mu_{3.}$
Marginal	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

The means model, also known as the full rank model is written as:

$$y_{ijk} = \mu_{ij} + e_{ijk}$$

where i is an index for the levels of the first factor A, j is the index for the second factor B and k is the index for subjects within each factor combination;  $y_{ijk}$  is the observed value for kth subject within the ith level of A and , jth level of B;  $\mu_{ij}$  is the mean of the ith level of A and jth level of B; and  $e_{ijk}$  is the random error associated with the observation.

SAS™ procedures use what is called the over parameterized model which is written as:

$$y_{ijk} = \mu + A_i + B_j + (A*B)_{ij} + e_{ijk}$$

In the over parameterized model, each observation is a sum of the overall mean, the "effect" of factor A, the "effect" of factor B, the interaction "effect" of (A\*B) and a random error term. SAS™ procedures also use the "last equals zero" side condition to solve the normal equations. As a results, the individual parameters are not uniquely estimable and are called "non-estimable." However, some functions of the parameters are estimable and this is where the problems seem to arise. To easily solve the problem of estimable functions, all we need to do is equate the two models which yields:

$$\mu_{ij} = \mu + A_i + B_j + (A*B)_{ij}$$

and express all hypotheses of interest in terms of the means  $\mu_{ij}$ .

**Example 1:** Consider the following problem taken from KNNL:

A=Age, B=Gender	B=1	B=2
A=1 (young)	21, 23, 19, 22, 22, 23	21, 22, 20, 21, 19, 25
A=2 (middle)	30, 29, 26, 28, 27, 27	26, 29, 27, 28, 27, 29
A=3 (elderly)	25, 22, 23, 21, 22, 21	23, 19, 20, 21, 20, 20

The main hypothesis is to test equality of the average result of the young and elderly against that of the middle age. In terms of the means, this is expressed as:

$$L = (\mu_1 + \mu_3) / 2 - \mu_2.$$

To construct the contrast, we rewrite  $\mu_1 = (\mu_{11} + \mu_{12}) / 2$ ;  $\mu_2 = (\mu_{21} + \mu_{22}) / 2$ ; and  $\mu_3 = (\mu_{31} + \mu_{32}) / 2$ . Next, we equate each  $\mu_{ij} = \mu + A_i + B_j + (A*B)_{ij}$ , for example  $\mu_{11} = \mu + A_1 + B_1 + (A*B)_{11}$ . The final step is to replace each  $\mu_{ij}$  in the contrast with its respective "effects" equivalent and simplify the resultant algebraic solution to get  $2A_1 + 2A_3 - 4A_2 + A_{11} + A_{12} + A_{31} + A_{32} - 2A_{21} - 2A_{22}$  and finally, rewrite the expression in the correct lexicographic order:

$$2A_1 - 4A_2 + 2A_3 + A_{11} + A_{12} - 2A_{21} - 2A_{22} + A_{31} + A_{32}$$

The SAS™ code for this example is given as:

```

Titlel "Two factor study, KNNL, page 866";
Data Cash;
Input A B cash;
Datalines:
...data...
Run;
Proc GLM Data=Cash;
Class A B;
Model cash = A B A*B;
Contrast 'Average age 1 and 3 verse age 2'
A 2 -4 2 A*B 1 1 -2 -2 1 1;
Estimate 'same as above' A 2 -4 2 A*B 1 1 -2 -2 1 1 /divisor=4;
Run;

```

**Example 2:** A second example from KNNL deals with hay fever relief. The response is hours of relief from two different active ingredients. The data is given as:

Ingredient A \ Ingredient B	J=1, low	J=2, medium	J=3, high
I=1, low	2.4, 2.7, 2.3, 2.5	4.6, 4.2, 4.9, 4.7	4.8, 4.5, 4.4, 4.6
I=2, medium	5.8, 5.1, 5.5, 5.3	8.9, 9.1, 8.7, 9.0	9.1, 9.3, 8.7, 9.4
I=3, high	6.1, 5.7, 5.9, 6.2	9.9, 10.5, 10.6, 10.1	13.5, 13.0, 13.3, 13.2

In this example it was presumed there would be an interaction which was born out in the analysis and the primary concern was on the nature of the interaction presented by the following contrasts:

$$L_1 = (\mu_{12} + \mu_{13}) / 2 - \mu_{11}; \quad L_2 = (\mu_{22} + \mu_{23}) / 2 - \mu_{21}; \quad L_3 = (\mu_{32} + \mu_{33}) / 2 - \mu_{31}$$

As before, the solution to this problem is to equate the terms of the full rank model to the over-parameterized model:  $u_{ij} = \mu + A_i + B_j + (A*B)_{ij}$ , as in  $\mu_{11} = \mu + A_1 + B_1 + (A*B)_{11}$ . Once we have all of the expressions in terms of the "effects" for the  $\mu_{ij}$ , we substitute them into the contrasts, simply and voilà! our contrasts are:

$$L_1 = -2B_1 + B_2 + B_3 - 2(A*B)_{11} + (A*B)_{12} + (A*B)_{13}$$

$$L_2 = -2B_1 + B_2 + B_3 - 2(A*B)_{21} + (A*B)_{22} + (A*B)_{23}$$

$$L_3 = -2B_1 + B_2 + B_3 - 2(A*B)_{31} + (A*B)_{32} + (A*B)_{33}$$

The SAS™ code for this example is given as:

```
Title1 "Hay Fever Relief, KNNL, page 868";
Data Relief;
Input A B hours;
Datalines:
...data...
Run;
Proc GLM Data=Relief;
Class A B;
Model hours = A B A*B;
Contrast 'L1' B -2 1 1 A*B -2 1 1 0 0 0 0 0 0;
Contrast 'L2' B -2 1 1 A*B 0 0 0 -2 1 1 0 0 0;
Contrast 'L3' B -2 1 1 A*B 0 0 0 0 0 0 -2 1 1;
Run;
```

**Example 3:** This example comes from (KNNL) and involves the hay fever relief data with cell (2,2) as a missing cell. In this case, Type IV Sums of Squares are used and the output is not unique because there are many possible hypotheses to test. One of the great advantages to using the CONTRAST statement is that we can ignore the ANOVA output and write the hypotheses we wish to test rather than accepting the hypotheses tested by SAS™.

Ingredient A \ Ingredient B	J=1, low	J=2, medium	J=3, high
I=1, low	2.4, 2.7, 2.3, 2.5	4.6, 4.2, 4.9, 4.7	4.8, 4.5, 4.4, 4.6
I=2, medium	5.8, 5.1, 5.5, 5.3		9.1, 9.3, 8.7, 9.4
I=3, high	6.1, 5.7, 5.9, 6.2	9.9, 10.5, 10.6, 10.1	13.5, 13.0, 13.3, 13.2

In this example we are interested in testing the following hypotheses:

$$L_1 = \mu_{13} - \mu_{11};$$

$$L_2 = \mu_{23} - \mu_{21};$$

$$L_3 = \mu_{33} - \mu_{31}$$

To test these hypotheses we follow the formula that was established in example 1 and example 2. The only twist is that the cell (2,2) is empty which means there is no  $(A*B)_{22}$  term. This in turn means the vector of interaction terms does not have nine entries but only eight with the entry for  $(A*B)_{22}$  missing. The E4 option of the MODEL statement provides information about the terms that are present and their order in the vector of effects.

The SAS™ code for this example is given as:

```
Title1 "Hay Fever Relief, KNNL, page 868";
Data Relief;
Input A B hours;
Datalines:
...data...
Run;
Proc GLM Data=Relief;
Class A B;
Model hours = A B A*B/ss4 e4;
Contrast 'L1' B -1 0 1 A*B -1 0 1 0 0 0 0 0;
Contrast 'L2' B -1 0 1 A*B 0 0 0 -1 1 0 0 0;
Contrast 'L3' B -1 0 1 A*B 0 0 0 0 0 0 -1 0 1;
Run;
```

Selected Output:

Type IV Estimable Functions				
Effect		Coefficients		
		A	B	A*B
Intercept		0	0	0
A	1	L2	0	0
A	2	L3	0	0
A	3	-L2-L3	0	0
B	1	0	L5	0
B	2	0	L6	0
B	3	0	-L5-L6	0
A*B	1 1	0.3333*L2	0.3333*L5	L8
A*B	1 2	0.3333*L2	0.5*L6	L9
A*B	1 3	0.3333*L2	-0.3333*L5-0.5*L6	-L8-L9
A*B	2 1	<u>0.5*L3</u>	<u>0.3333*L5</u>	<u>L11</u>
A*B	2 3	<u>0.5*L3</u>	<u>-0.3333*L5</u>	<u>-L11</u>
A*B	3 1	-0.3333*L2-0.5*L3	0.3333*L5	-L8-L11
A*B	3 2	-0.3333*L2	0.5*L6	-L9
A*B	3 3	-0.3333*L2-0.5*L3	-0.3333*L5-0.5*L6	L8+L9+L11

Source	DF		Type IV SS	Mean Square	F Value	Pr > F
A	2	*	214.3657500	107.1828750	1612.78	<.0001
B	2	*	116.3505000	58.1752500	875.36	<.0001
A*B	3		28.3077083	9.4359028	141.98	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
u13-u11	1	8.8200000	8.8200000	132.71	<.0001
u23-u21	1	27.3800000	27.3800000	411.99	<.0001
u33-u31	1	105.8512500	105.8512500	1592.75	<.0001

## **Conclusion**

When using SAS™ Proc GLM to perform an analysis, representing the model in terms of the cell means provides an easy and convenient method for constructing contrasts of any interesting and testable hypothesis.

## **References:**

Kutner Michael H, Nachtsheim Christopher J, Neter John, Li William. 2005. Applied Linear Statistical Models. New York: McGraw Hill.

SAS and all other SAS Institute Inc. product or services names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

## **Contact Information**

Robert Vogel  
Jian Ping Hsu College of Public Health  
Georgia Southern University  
Statesboro, Georgia  
(912) 478-7423  
rvogel@georgiasouthern.edu