Multiple Factors

Matthew W Wheeler

From Blocks to Factors

Recall a Block is a variable that controls variability that is not associated with your factor.

To analyze a block we put it into proc glm using the class statement (this tells SAS how to 'code it' – more on this later)

We then put it into the model statement.

From the last example

```
proc glm data = chew;
       *Three class variables;
       class chef kitchen flour;
       *Only main effects for now;
       model chew = flour chef kitchen;
       lsmeans flour/cl adjust=bon;
       *above line adjusts using a Bonferroni
adjustment;
       run;
quit;
```

Here Chef and kitchen were two types of blocks in that we said They were things we COULD NOT ASSIGN.

A second factor is something I assign.

Suppose now that I control the kitchen.   That is, it has certain instruments/appliances that are different, and it is the experimenter that controls these.

The Block now becomes a Factor SAS does not treat it differently.

You are the one that treats it differently.

Analyzing Kitchen differences

```
proc glm data = chew;
        *Three class variables;
        class chef kitchen flour;
        *Only main effects for now;
        model chew = flour chef kitchen;
        contrast 'Kitchen 1 vs. Kitchen 2' kitchen 1 -1;
        estimate 'Kitchen 1 vs. Kitchen 2' kitchen 1 -1;
        run;
quit;
```

This is the same as analyzing the flour variable that we looked at in
The last lecture.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Kitchen 1 vs. Kitchen 2	-1.63030394	0.45546135	-3.58	0.0013

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Kitchen 1 vs. Kitchen 2	1	18.13985095	18.13985095	12.81	0.0013

So there is a difference in Chewiness between kitchen!

Main effects vs. interactions

A main effect is thought to be independent of the other variables.

A interaction is what happens when two factors act together to change the response, and that result is greater than it would be otherwise.

Example: Drinking is not known to cause esophageal cancer, neither is smoking, but together the increase the risk of esophageal cancer.

If we believe there is an interaction, we need to check before we conclude what is in the previous slide.

```
/*
 What do we do if there are interactions between
 kitchen AND flour.
*/
proc glm data = chew;
       *Three class variables;
       class chef kitchen flour;
       *Only main effects for now;
       model chew = flour kitchen flour*kitchen chef ; * the
'*' is an interaction;
run;
quit;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
flour	2	146.7662153	73.3831077	48.86	<.0001
kitchen	1	18.2425041	18.2425041	12.15	0.0018
kitchen*flour	2	0.5910631	0.2955315	0.20	0.8226
chef	1	0.0534003	0.0534003	0.04	0.8519

Unsurprisingly there is nothing going on between kitchen and flour
Type, we can drop this term from the model and just model main effects.

On interactions and Coding

We do not always get to model main effects, so if we have treatments, we need to check for interactions:

Types:

1. Main Effects – Direct effect on the response of interest.
2. Two way interaction- Interaction between two factors of interest.
3. Three way or higher- Interactions between three or more factors of interest (these are usually considered implausible and no people don't usually worry about them - we will not investigate them further in this course).

SAS and coding:

We have ignored how SAS is coding our factors and our treatments, because we have been able to get along just fine saying: Experimental Unit 1 gets treatment 1. As we get more factors, it is nice to understand how SAS is coding the variables. This will give insight into how to design a more complicated experiment and how to analyze complicated contrasts.

If we design our experiment wrong, we may not be able to estimate something after we have collected our data. We have to be carefull.

GLM Uses a type of dummy coding that estimates a 'global' mean by default. By dummy we say that the variable is 1 if the treatment is applied 0 otherwise. SAS is going to give you matrices like this so you are going to need to know what it is doing.

`	Flour			Kitchen		Flour*Kitchen					
Intercept	1	2	3	1	2	1*1	1*2	2*1	2*2	3*1	3*2
1	1	0	0	1	0	1	0	0	0	0	0
1	0	1	0	1	0	0	0	1	0	0	0
1	0	0	1	1	0	0	0	0	0	1	0
1	1	0	0	0	1	0	1	0	0	0	0
1	0	1	0	0	1	0	0	0	1	0	0
1	0	0	1	0	1	0	0	0	0	0	1

Multiple Factors + Interactions continued

The problem with the way SAS codes it is that it is based upon a model that has too many parameters. Unfortunately you must think of it in this way to use SAS (or any standard software package). You can never estimate all of these parameters, but you can estimate FUNCTIONS of these parameters, and you will have to be able to play with these functions to be able to do a contrast or estimate statement.

When you figure this out you will be able to do it for other types of data.

A different (in my opinion easier) way to think of it: you can estimate any of the boxes.

	Flour - 1	Flour – 2	Flour – 3
Kitchen – 1	$M + F_1 + K_1 + (FK)_{11}$	$M + F_2 + K_1 + (FK)_{21}$	$M + F_3 + K_1 + (FK)_{32}$
Kitchen – 2	$M + F_1 + K_2 + (FK)_{12}$	$M + F_2 + K_1 + (FK)_{22}$	$M + F_3 + K_2 + (FK)_{32}$

So you can estimate any box minus any other box or any combination of boxes.

Average effect of Flour-1 vs. Flour- 2:

$$0.5*[M + F_1 + K_1 + (FK)_{11} + M + F_1 + K_2 + (FK)_{12}]-0.5[M + F_2 + K_1 + (FK)_{21} + M + F_2 + K_1 + (FK)_{22}]$$

$$0.5*[F_1 + K_1 + (FK)_{11} + F_1 + (FK)_{12}]-0.5[F_2 + (FK)_{21} + F_2 + (FK)_{22}]$$

A quick note on the estimate function

When SAS has a class variable it codes it in ascending order. It does the same for interaction effects too. This coding goes through to the estimate and contrast statement.

For example: If I have the statement

```
model response =A B A*B;
```

where A and B have two levels

Then….

Then if I use an estimate statement SAS expects

A: (1) (2)

B: (1) (2)

AB: (11) (12) (21) (22)

Thus if I want to estimate $A_1 + B_1 + (AB)_{11} - [A_2 + B_2 + (AB)_{22}]$

I use the statement:

```
estimate 'name of estimate' A 1 -1  B 1 -1 A*B 1 0 0 -1;
```

```
proc glm data = chew;
        *remove chef this time class variables;
        class  kitchen flour;
        *Only main effects for now;
        model chew = flour kitchen flour*kitchen; * the '*' is an interaction;
        estimate 'Mean Flour 1 vs Flour 2' flour 1 -1 0 flour*kitchen 0.5   -0.5    0    0.5   -0.5   0;
                                               *1  2 3               (1,1) (1,2) (1,3) (2,1) (2,2) (2,3)
                                                                                              The above is
the treatment coding's of flour and flour*kitchen;
run;
```

The new effect with the interaction:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Mean Flour 1 vs Flour 2	-2.56055000	0.54930180	-4.66	<.0001

The one without the interaction:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Mean Flour 1 vs Flour 2	-2.45063051	0.50678097	-4.84	<.0001

Example 2

The effects of a variety of wheat and pesticide level were investigated. Three types of wheat (A,B and C) and three pesticides were used (‘None’,’Low’,’Heavy’). The yield in bushels is recorded for each plot:

	None	Low	Heavy
A	115 ,101	120, 127	136, 130
B	96, 94	113, 108	117, 124
C	98, 109	110, 122	130, 128

Factor: Wheat Type and Amount of Pesticide.

Block : None.

Experimental Unit: Plot of Land

Measurement: Crop Yield.

Tests of Interest: Differences between yield in wheat type.

Differences between Pesticide type.

Is there an interaction?

Experiment Wise Error Rate: $\alpha=0.05$

Uses Characters vs. Numbers

```
/*Example 2 Class 4*/

data crop;
        input pesticide $2. variety $2. yield;
        datalines;
N A 115
N A 101
L A 120
L A 127
H A 136
H A 130
N B 96
N B 94
L B 113
L B 108
H B 117
H B 124
N C 98
N C 109
L C 110
L C 122
H C 130
H C 128
;
```

```
*FIRST CHECK TO SEE IF THERE IS AN
INTERACTION
*;
proc glm data = crop;
        class pesticide variety;
        model yield = pesticide variety
variety*pesticide;
run;
quit;
```

Check for interaction:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
pesticide	2	1938.777778	969.388889	27.79	0.0001
variety	2	498.777778	249.388889	7.15	0.0138
pesticide*variety	4	8.888889	2.222222	0.06	0.9912

No interaction

Reduced model fit.

Note we can use Tukey because there are an equal amount of observations in each bin. That is we have a BALANCED design.

```
*NO INTERACTION, REDUCE THE MODEL AND
*MAKE ESTIMATES: NOTE THE WAY SAS CODES IT!;
proc glm data = crop;
      class pesticide variety;
      model yield = pesticide variety;
      lsmeans pesticide/cl adjust=tukey;
      lsmeans variety/cl adjust=tukey;
run;
quit;
```

Least Squares Means for Effect pesticide				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	10.833333	3.236010	18.430657
1	3	25.333333	17.736010	32.930657
2	3	14.500000	6.902677	22.097323

Least Squares Means for Effect variety				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	12.833333	5.236010	20.430657
1	3	5.333333	-2.263990	12.930657
2	3	-7.500000	-15.097323	0.097323

Multiple Comparisons Revisited

Each of the above technically controls at the α=0.05 rate for the given set of tests. We have 3 tests:

1. The test for the interaction.
2. The effect of pesticide.
3. The effect of variety.

We can have guarantee a α=0.05 with an ADITIONAL BF adjustment.

Least Squares Means for Effect pesticide				
i	j	Difference Between Means	Simultaneous 98.33% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	10.833333	1.524534	20.142132
1	3	25.333333	16.024534	34.642132
2	3	14.500000	5.191201	23.808799

Same conclusion, but this time it is at a guaranteed 0.05 error rate. CI are a wider.

Least Squares Means for Effect variety				
i	j	Difference Between Means	Simultaneous 98.33% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	12.833333	3.524534	22.142132
1	3	5.333333	-3.975466	14.642132
2	3	-7.500000	-16.808799	1.808799

Observed Covariates that are not Blocks

When we talked about blocking, we stated that these are variables that we can’t control but we can deal with in our experimental plan, e.g., Kroger/Harris Teeter.

There are times when we have an observed variable, that we can’t control, but it may impact what we are measuring. We should include these things in our model.

For Example:

Suppose I was looking at the spending increases due to a marketing campaign. Naturally, I would expect people with more money to spend more. What if, by chance, I assigned more wealthy people to the marketing campaign, as compared to the control.

	< $100,000	>$100,000
Control	70	30
New Campaign	30	70

If I don't control for the income, I might falsely conclude the new campaign will increase spending.

Example 3

A small college wants to compare the salaries of faculty in three areas: science, humanities and business. Their salaries are recorded as well as their years of experience (salary,experience)

Science	Humanities	Business
(35,2) (47,7) (65,22) (51,14) (45,4)	(68,28) (54,17) (38,6) (59,19) (47,10) (36,5) (32,4)	(46,5) (39,1) (47,7) (63,18) (68,22)

```
/*Example 3: Class 4*/
data salary;
*science = 1, humanities = 2, business=3 ;
       input dept exp sal @@;
       cards;
1 2 35 1 7 47 1 22 65 1 14 51 1 4 45
2 28 68 2 17 54 2 6 38 2 19 59 2 10 47 2 5 36
2 4 32 3 5 46 3 1 39 3 1 39 3 7 47 3 18 63
3 22 68
;
```

```
/*FIRST RUN IT JUST BY DEPARTMENT*/
proc glm data=salary;
       class dept;
       model sal = dept;
       lsmeans dept/ cl adjust=BON; *why
not Tukey?;
run;
```

If I don’t account for experience, there is no difference

Least Squares Means for Effect dept				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.885714	-18.611395	20.382824
1	3	-1.733333	-21.896064	18.429397
2	3	-2.619048	-21.144152	15.906057

First see if there is an interaction

```
/*NOW SEE IF THERE IS AN EXPERIENCE*DEPT
INTERACTION*/
proc glm data=salary;
       class dept;
       model sal = dept exp exp*dept;
run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
dept	2	97.366126	48.683063	8.18	0.0057
exp	1	2035.945157	2035.945157	342.17	<.0001
exp*dept	2	6.568497	3.284249	0.55	0.5898

```
/*FINAL MODEL*/
proc glm data=salary alpha=0.025; *why 0.025? ;
       class dept;
       model sal = dept exp;
       lsmeans dept/ cl adjust=BON; *why not Tukey?;
run;
quit;
```

Least Squares Means for Effect dept				
i	j	Difference Between Means	Simultaneous 97.5% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	4.935989	0.649846	9.222132
1	3	-2.845173	-7.233848	1.543501
2	3	-7.781162	-11.887910	-3.674414

Humanities prof’s are paid less than Business prof’s

In conclusion, If I didn’t include the covariate, I would have said the salaries were the same by discipline.

By chance, I sampled more experienced humanities professors.

Including the covariate, shows there is a difference in pay by discipline.

THIS MISTAKE HAPPENS ALL OF THE TIME!
DON’T MAKE THIS MISTAKE!