Chapter Seven: Factorial ANOVA with Multiple Regression

The Kenton Example with dummy variables

First consider indicator dummy variable coding with an intercept. Here is the part of the data step that defines the dummy variables. Because we have an intercept, we'll represent the four categories of package design with three dummy variables. The table below shows the model for population mean sales.

Design	p1	p2	p3	$E[Y] = \beta_0 + \beta_1 p1 + \beta_2 p2 + \beta_3 p3$
3Colour Cartoon	1	0	0	$\beta_0 + \beta_1 \quad = \mu_1$
3Col No Cartoon	0	1	0	$\beta_0 + \beta_2 \quad = \mu_2$
5Colour Cartoon	0	0	1	$\beta_0 + \beta_3 \quad = \mu_3$
5Col No Cartoon	0	0	0	$\beta_0 \qquad = \mu_4$

To clarify the parallel between population parameters and sample statistics, the corresponding table of estimated sales figures is

Design	p1	p2	p3	$\hat{Y} = b_0 + b_1 p1 + b_2 p2 + b_3 p3$
3Colour Cartoon	1	0	0	$b_0 + b_1 \quad = \overline{Y}_1$
3Col No Cartoon	0	1	0	$b_0 + b_2 \quad = \overline{Y}_2$
5Colour Cartoon	0	0	1	$b_0 + b_3 \quad = \overline{Y}_3$
5Col No Cartoon	0	0	0	$b_0 \qquad = \overline{Y}_4$

One thing these tables show is something that is true of *any* valid dummy variable coding scheme (when there are only categorical independent variables): $\hat{Y} = \overline{Y}$ for each category or combination of categories.

It is also easy to see that to test for differences among means, we want to simultaneously test whether β_1, β_2 and β_3 are different from zero -- or equivalently, whether b_1, b_2 and b_3 are *significantly* different from zero. That is, we want to simultaneously test the dummy variables p1, p2 and p3. The overall F test of proc reg does the job.

```
proc reg;
     model sales = p1 p2 p3;
```

Yielding:

```
Model: MODEL1
Dependent Variable: SALES      Number of Cases Sold

                             Analysis of Variance

                                Sum of         Mean
       Source          DF      Squares       Square      F Value       Prob>F

       Model            3    588.22105    196.07368       18.591       0.0001
       Error           15    158.20000     10.54667
       C Total         18    746.42105
```

We got this same same F value for differences among the four means from `proc glm`. The next line does the 3 versus 5 color comparison.

```
     ncolour: test p1+p2 = p3; /* 3 vs 5 colours */
```

It works because we want to test whether $\frac{1}{2}(\mu_1 + \mu_2) = \frac{1}{2}(\mu_3 + \mu_4)$, and

$$\frac{1}{2}(\beta_0 + \beta_1 + \beta_0 + \beta_2) = \frac{1}{2}(\beta_0 + \beta_3 + \beta_0)$$

is algebraically equivalent to

$$\beta_1 + \beta_2 = \beta_3.$$

The estimate statement from `proc glm` yielded t = -6.25. Calculate $F = t^2 = 39.0625$, and compare the output of the `test` statement above:

```
Test: NCOLOUR   Numerator:     411.4000   DF:     1    F value:   39.0076
                Denominator:   10.54667   DF:    15    Prob>F:     0.0001
```

The difference is rounding error. It's the same test. But we'd rather avoid having to do algebra whenever we want to test a contrast. In **cell means coding**, we use an indicator dummy variable for each category (four, in this case), and omit the intercept. The tables that follow indicate why it's called cell mean coding.

Cell Means Coding for Package Design

Design	p1	p2	p3	p4	$E[Y] = \beta_1 p1 + \beta_2 p2 + \beta_3 p3 + \beta_4 p4$
3Colour Cartoon	1	0	0	0	$\beta_1 = \mu_1$
3Col No Cartoon	0	1	0	0	$\beta_2 = \mu_2$
5Colour Cartoon	0	0	1	0	$\beta_3 = \mu_3$
5Col No Cartoon	0	0	0	1	$\beta_4 = \mu_4$

Design	p1	p2	p3	p4	$\hat{Y} = b_1 p1 + b_2 p2 + b_3 p3 + b_4 p4$
3Colour Cartoon	1	0	0	0	$b_1 = \overline{Y}_1$
3Col No Cartoon	0	1	0	0	$b_2 = \overline{Y}_2$
5Colour Cartoon	0	0	1	0	$b_3 = \overline{Y}_3$
5Col No Cartoon	0	0	0	1	$b_4 = \overline{Y}_4$

Here is the `proc reg.`

```
proc reg;
     model sales = p1 p2 p3 p4 / noint;
     alleq:    test p1=p2=p3=p4;
     numcol:   test p1+p2 = p3+p4;
     cartoon:  test p1+p3 = p2+p4;
     inter1:   test p1-p2 = p3-p4; /* Effect of cartoon depends on ncolours */
     inter2:   test p1-p3 = p2-p4; /* Effect of ncolours depends on cartoon */
     Y3_N3:    test p1=p2;  /* All pairwise tests */
```

```
Y3_Y5:    test p1=p3;
Y3_N5:    test p1=p4;
N3_Y5:    test p2=p3;
N3_N5:    test p2=p4;
Y5_N5:    test p3=p4;
```

And the output. First, the overall F test, which is very different from what we had before.

```
Model: MODEL1
NOTE: No intercept in model. R-square is redefined.
Dependent Variable: SALES      Number of Cases Sold

                             Analysis of Variance

                                Sum of         Mean
       Source          DF      Squares       Square      F Value       Prob>F

       Model            4   7183.80000   1795.95000      170.286       0.0001
       Error           15    158.20000     10.54667
       U Total         19   7342.00000

           Root MSE       3.24756     R-square       0.9785
           Dep Mean      18.63158     Adj R-sq       0.9727
           C.V.          17.43042
```

With no intercept,

- Total sum of squares is now $\sum_{i=1}^{n} Y_i^2$. It's no longer corrected for the mean; U means uncorrected. R^2 is radically affected.

- The overall F-test is for whether ALL the betas are zero - usually uninteresting

Notice now the parameter estimates are exactly the cell means.

```
                              Parameter Estimates

                        Parameter      Standard    T for H0:
     Variable  DF        Estimate         Error   Parameter=0    Prob > |T|

     P1         1       14.600000    1.45235441        10.053        0.0001
     P2         1       13.400000    1.45235441         9.226        0.0001
     P3         1       19.500000    1.62378159        12.009        0.0001
     P4         1       27.200000    1.45235441        18.728        0.0001
```

Now the custom tests. I will repeat the test statement for each one, and provide some discussion.

The Statement

```
     alleq:    test p1=p2=p3=p4;
```

yields this output:

```
Dependent Variable: SALES
Test: ALLEQ    Numerator:    196.0737  DF:    3   F value:  18.5911
               Denominator:  10.54667  DF:   15   Prob>F:    0.0001
```

This really is the overall test for whether all four means are equal -- again. The F value is the same as we got earlier at least two times. But look at the test statement. As usual, it specifies restrictions on the betas that give us the reduced model. But this time, those restrictions are not of the simple form we saw before, setting a subset of the betas equal to zero. Now we're setting them all to be equal. This shows you two things:

° The `test` statement in `proc reg` is a little more general than it seemed at first. It lets you test simultaneously whether several linear combinations of betas equal zero. Here, we're testing three linear combinations: $\beta_1-\beta_2=0$, $\beta_2-\beta_3=0$, $\beta_3-\beta_4=0$. The test statement could have read: `alleq:    test p1-p2=0, p2-p3=0, p3-p4=p4;`

° The full versus reduced model business is also more general than you might think. In ordinary regression, "all" we can do is test collections linear restrictions on the parameters. But in the most general hypothesis testing framework, all one *ever* does is to compare the fit of a full model to the fit of a reduced model in which some restriction has been placed on the values of the parameters. Those restrictions are called the "null hypothesis." You didn't really need to know this.

To really understand the next several test statements, we need to recognize that the 4-category variable Package Design actually represents the combination of two independent variables: Number of Colours and Presence versus absence of cartoons. That is, we have a two-factor design. Consider the following table:

Population Cell Means and Marginal Means for the Kenton Example

	Cartoon	No Cartoon	
3 Colours	μ_1	μ_2	$\frac{\mu_1 + \mu_2}{2}$
5 Colours	μ_3	μ_4	$\frac{\mu_3 + \mu_4}{2}$
	$\frac{\mu_1 + \mu_3}{2}$	$\frac{\mu_2 + \mu_4}{2}$	

In addition to population mean sales for each package design (denoted by μ_1 through μ_4), the table above shows **marginal means** -- quantities like $\frac{\mu_2 + \mu_4}{2}$, which are obtained by averaging over rows or columns.

If there are differences among marginal means for a categorical independent variable in a two-way (or higher) layout like this, we say there is a **main effect** for that variable. Tests for main effects are of great interest; they can indicate whether, averaging over the values of the other categorical independent variables in the design, whether the independent variable in question is related to the dependent variable. Note that averaging over the values of other independent variables is not the same thing as controlling for them, but it can still be a valuable thing to do.

The population means in the preceding table are estimated by corresponding sample quantities. The numbers in the table below come from the `means` output of the first `proc glm`.

Sample Cell and Marginal Means for the Kenton Example

	Cartoon	No Cartoon	
3 Colours	14.6	13.4	14
5 Colours	19.5	27.2	23.35
	17.05	20.3	

(14.6+13.4)/2 = 14, and so on.

The next custom test is for the main effect of number of colours (3 vs. 5). It tests whether $\frac{\mu_1+\mu_2}{2}=\frac{\mu_3+\mu_4}{2}$. It's the same thing as asking whether the marginal mean for 2 Colours (14) is *significantly* different from the marginal mean for 5 colours (23.35).

The test command, obtained directly by multiplying both sides =f $\frac{\mu_1+\mu_2}{2}=\frac{\mu_3+\mu_4}{2}$ by 2 (this has no effect on the test), is

```
numcol:   test p1+p2 = p3+p4;
```

yielding this output:

```
Dependent Variable: SALES
Test: NUMCOL   Numerator:    411.4000  DF:    1   F value:  39.0076
               Denominator:  10.54667  DF:   15   Prob>F:    0.0001
```

So the answer is Yes. There is a significant main effect for number of colours, with 5-colour packages generating more sales when you average across Cartoon and No-cartoon designs. And notice how much more convenient the cell means coding makes this test. Recall

```
ncolour: test p1+p2 = p3; /* 3 vs 5 colours */
```

from Page 13.

Similarly, the main effect for presence versus absence of cartoons on the package is tested by asking whether $\frac{\mu_1+\mu_3}{2}=\frac{\mu_2+\mu_4}{2}$.

```
cartoon:  test p1+p3 = p2+p4;
```

```
Dependent Variable: SALES
Test: CARTOON   Numerator:     49.7059  DF:    1   F value:   4.7129
                Denominator:  10.54667  DF:   15   Prob>F:    0.0464
```

So the main effect for Cartoon is barely significant, with Non-cartoon designs doing better. With a Scheffee test, though, it's not significant. Fsch = `4.7129/3 = 1.57, which is less than the critical value of 3.29.`

The two-way design we have been looking at is called a factorial design. In a factorial design, there are two or more categorical independent variables (called factors, in this context) typically with data with for combinations of the factors being collected. Factorial designs are often found in experimental studies, but not always.

When Sir Ronald Fisher (in whose honour the F-test is named) dreamed up factorial designs, he pointed out that they enable the scientist to investigate the effects of several independent variables at much less expense than if a separate experiment had to be conducted to test each one. In addition, they allow one to ask systematically whether the effect of one independent variable *depends* on the value of another independent variable. If the effect of one independent variable depends on another, we will say there is an **interaction** between those variables. We talk about an A "by" B or A x B interaction. An interaction means "it depends."

Let's look at the table of population means again.

	Cartoon	No Cartoon	
3 Colours	μ_1	μ_2	$\frac{\mu_1 + \mu_2}{2}$
5 Colours	μ_3	μ_4	$\frac{\mu_3 + \mu_4}{2}$
	$\frac{\mu_1 + \mu_3}{2}$	$\frac{\mu_2 + \mu_4}{2}$	

The effect of Cartoons when the package has three colours is represented by μ_1-μ_2. The effect of Cartoons when the package has five colours is represented by μ_3-μ_4. Therefore, the interaction of

Cartoon by number of colours is a *difference between differences*, and we want to test whether $\mu_1-\mu_2=\mu_3-\mu_4$. That's what we're doing below:

```
     inter1:   test p1-p2 = p3-p4; /* Effect of cartoon depends on ncolours */

Dependent Variable: SALES
Test: INTER1   Numerator:     93.1882  DF:    1   F value:   8.8358
               Denominator:  10.54667  DF:   15   Prob>F:    0.0095
```

Another way to think about the interaction is to ask whether the effect of number of colours depends on presence versus absence of cartoon pictures. We are asking whether $\mu_1-\mu_3=\mu_2-\mu_4$. Here's the `test` statement and the output.

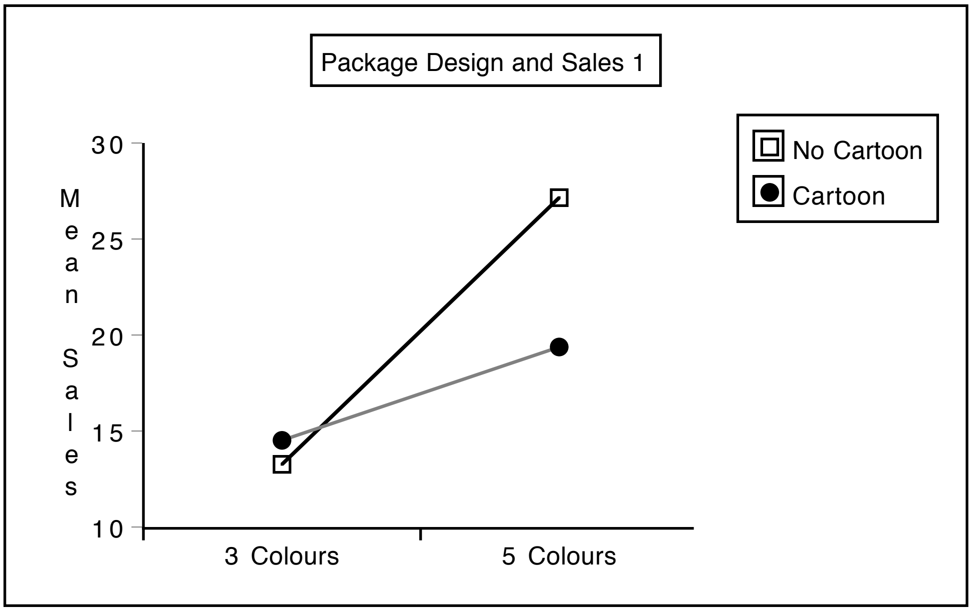
```
     inter2:   test p1-p3 = p2-p4; /* Effect of ncolours depends on cartoon */

Dependent Variable: SALES
Test: INTER2   Numerator:     93.1882  DF:    1   F value:   8.8358
               Denominator:  10.54667  DF:   15   Prob>F:    0.0095
```

Notice that this F test is identical to the last one? It happens because $\mu_1–\mu_2=\mu_3–\mu_4$ is algebraically equivalent to $\mu_1–\mu_3=\mu_2–\mu_4$. So the two ways of talking about the interaction are the same thing, mathematically. Fortunately, this *always* happens, no matter how big the design. If you express an interaction correctly as a collection of differences between differences, it is algebraically equivalent to all other correct ways of expressing the interaction. Choose the one that is easiest to think about.

Incidentally, p = `0.0095 seems impressive, but the test is not significant if it is considered as a Scheffe follow-up:  Fsch = 8.8358/3 = 2.945267 < 3.29.`

If an interaction is significant, you should graph it to figure out what it means. Here is one example:

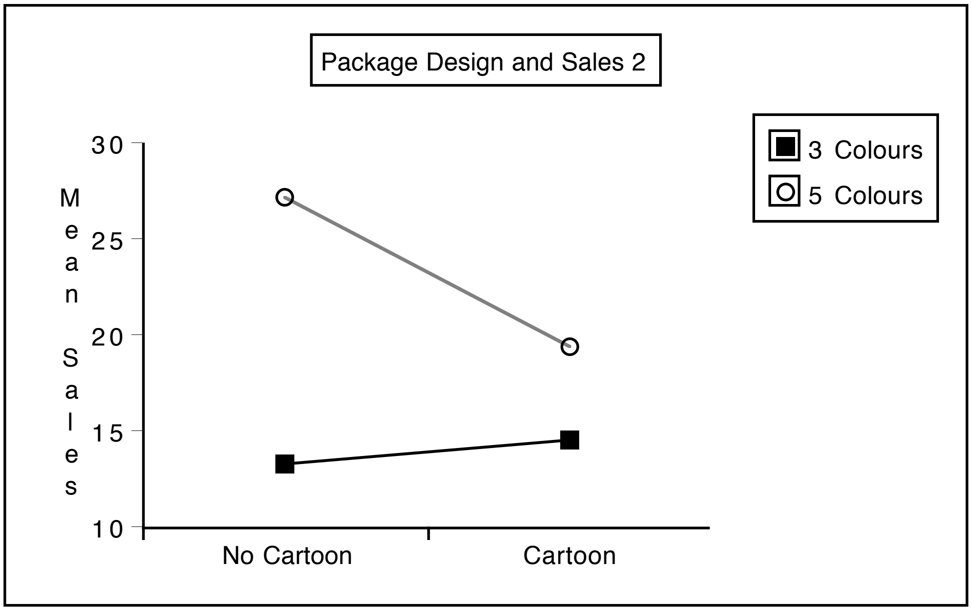


Whenever you have an interaction, such graphs will display non-parallel lines. Well actually, when you plot an interaction with real data, the lines will always be at least a little non-parallel. The question is whether they depart *significantly* from being parallel. Here, the advantage of 5 colours over 3 is significantly greater for designs without cartoons (unless you are a member of the Scheffé cult, as I am), and we can see it in the graph.

The post-hoc tests tell us that there is a significantly more sales with 5-colour designs, for both the cartoon and non-cartoon conditions. The interaction tells us that this effect is significantly greater when there are no cartoons.

Remember the significant main effect for cartoon? It was just barely significant: $p = 0.0464$. The graph above shows quite clearly that this effect is entirely due to the advantage of no-cartoon designs in the 5-colour condition. So here, we have a main effect that's significant, but we really should not interpret it, because of the interaction.

Some texts claim that if you have an interaction, you should *never* interpret the main effects. But look at the next figure, which graphs the same interaction in the other direction (there are only two ways to do it, because it is a two-factor interaction).



The picture that emerges here is that 5-colour designs are better overall, and the advantage is greater in the No-cartoon condition. Here, we can see that it makes sense to interpret both the main effect for number of colours *and* the interaction. This example shows why I disagree with the advice to never interpret main effects when there is an interaction.

The last six tests are the pairwise differences between means. Their value is that we can convert them easily to post-hoc Bonferroni or Scheffé tests. Personally, I like the idea of letting the tests for main effects, interactions and all pairwise differences as follow-ups to the initial oneway ANOVA. I prefer Scheffé, because I don't need to know in advance how many tests I'm going to do. I also love the Scheffé tests because of their 100% consistency with the initial tests. If the initial test is non-significant, no Scheffé follow-up can be significant, as a mathematical certainty. And if the initial test is significant, then there *must* be a significant Scheffé follow-up.

```
Dependent Variable: SALES
Test: Y3_N3    Numerator:      3.6000  DF:    1   F value:   0.3413
               Denominator:  10.54667  DF:   15   Prob>F:    0.5677


Dependent Variable: SALES
Test: Y3_Y5    Numerator:     53.3556  DF:    1   F value:   5.0590
               Denominator:  10.54667  DF:   15   Prob>F:    0.0399
```

```
Dependent Variable: SALES
Test: Y3_N5    Numerator:    396.9000  DF:    1   F value:  37.6327
               Denominator:  10.54667  DF:   15   Prob>F:    0.0001


Dependent Variable: SALES
Test: N3_Y5    Numerator:     82.6889  DF:    1   F value:   7.8403
               Denominator:  10.54667  DF:   15   Prob>F:    0.0135


Dependent Variable: SALES
Test: N3_N5    Numerator:    476.1000  DF:    1   F value:  45.1422
               Denominator:  10.54667  DF:   15   Prob>F:    0.0001


Dependent Variable: SALES
Test: Y5_N5    Numerator:    131.7556  DF:    1   F value:  12.4926
               Denominator:  10.54667  DF:   15   Prob>F:    0.0030
```

Sample Question: What p-value is required for significance if these 9 tests are to be protected with a Bonferroni correction at the 0.05 level? **Answer**: 0.05/9 = 0.0056

Effect	F	p	F_{sch} =F/3*	Significant with Bonferroni?	Significant with Scheffé?
Main Effect for Ncolours	39.0076	0.0001	13.0025	Yes	Yes
Main effect for Cartoon	4.7129	0.0464	1.57097	No	No
Interaction	8.8358	0.0095	2.9453	No	No
Cartoon3 vs NoCartoon3	0.3413	0.5677	0.1138	No	No
Cartoon3 vs Cartoon5	5.0590	0.0399	1.6863	No	No
Cartoon3 vs NoCartoon5	37.6327	0.0001	12.5442	Yes	Yes
NoCartoon3 vs Cartoon5	7.8403	0.0135	2.6134	No	No
NoCart3 vs Nocart5	45.1422	0.0001	15.0474	Yes	Yes
Cartoon5 vs NoCartoon5	12.4926	0.0030	4.1642	Yes	Yes

* Compare with critical value of F= 3.28738

The main thing to note here is that when you treat the test for interaction as a follow-up test instead of a one-at-a-time test, it's no longer significant. You are left with a simpler story. Five-colour designs work better than three-colour designs, and designs without cartoons work better in the 5-colour condition.

In general, if you go the multiple comparison route, it's going to make you more conservative. You will draw fewer conclusions. On the other hand, in terms of this particular example, the implications for *action* (marketing action) are the same whether or not you use multiple comparisons. The Kenton company should use a 5-colour design without cartoons.

We've seen how to do the tests above with dummy variables and `proc reg`. If you are only interested in testing single contrasts, the `estimate` command of `proc glm` is a bit more convenient, because proc glm sets up the dummy variables for you. All you have to do is give the coefficients of the contrast you want.

```
/* Single contrasts are just as convenient with the ESTIMATE
   statement of proc glm. Illustrate all pairwise.
   Note F = t-squared  */

proc glm;
     class package;
     model sales=package;
     estimate 'Y3_N3' package 1 -1  0  0;
     estimate 'Y3_Y5' package 1  0 -1  0;
     estimate 'Y3_N5' package 1  0  0 -1;
     estimate 'N3_Y5' package 0  1 -1  0;
     estimate 'N3_N5' package 0  1  0 -1;
     estimate 'Y5_N5' package 0  0  1 -1;
```

It's nice to have this degree of control, but not always necessary. In factorial analysis of variance, we commonly wish to test all main effects and interactions. Proc glm will compose the contrasts for you, as well as setting up the dummy variables:

```
proc glm;
     class ncolours cartoon;
     model sales = ncolours|cartoon;
/*   The model statement could have been
     model sales = ncolours cartoon ncolours*cartoon; */
```

In `proc glm`, if you separate a collection of classification variables with vertical bars, it means include all the main effects and interactions among the variables.

Here is the output:

```
                    General Linear Models Procedure
```

```
Dependent Variable: SALES   Number of Cases Sold
                                         Sum of           Mean
Source                   DF             Squares         Square  F Value    Pr > F

Model                     3        588.22105263   196.07368421    18.59    0.0001

Error                    15        158.20000000    10.54666667

Corrected Total          18        746.42105263

                   R-Square                C.V.       Root MSE         SALES Mean

                   0.788055            17.43042      3.2475632          18.631579


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

NCOLOURS                  1        452.86549708   452.86549708    42.94    0.0001
CARTOON                   1         42.16732026    42.16732026     4.00    0.0640
NCOLOURS*CARTOON          1         93.18823529    93.18823529     8.84    0.0095

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

NCOLOURS                  1        411.40000000   411.40000000    39.01    0.0001
CARTOON                   1         49.70588235    49.70588235     4.71    0.0464
NCOLOURS*CARTOON          1         93.18823529    93.18823529     8.84    0.0095
```

The output starts with an overall test that is 100% identical to the initial oneway ANOVA. It has the same R^2, the same F, the same p-value --- everything. This always happens. No matter how many independent variables you have or how many values each one has, simultaneously testing all the main effects and interactions is the same as defining a new independent variable whose values are the *combinations* of the variable values from the factorial ANOVA --- and then doing a one-way analysis of variance using that variable.

By default, SAS `proc glm` produces two sets of tests for the main effects and interaction(s). In the tests based on Type I Sums of Squares, each effect is controlled only for those before it in the table. In Type III Sums of Squares, each effect is controlled for all the others. That's why the last test is always identical for these two methods. When sample sizes are all equal or proportional, the independent variables are completely unrelated, and tests based on Type I and Type III sums of squares are all the same -- not just the last one.

The F and p values we get from Type III sums of squares match what we've done using `proc reg`. Most of the time, the tests from the Type III sums of squares are what we want.

Beyond the two-by-two Case

Methods for factorial ANOVA and testing interactions can easily be extended in several ways.

- More independent variables
- More than two values for an independent variable
- Interactions between continuous independent variables
- Interactions between categorical independent variables and continuous independent variables.

Extension to more than two factors is straightforward. Suppose we had grocery stores of three different sizes (small, medium and large), and within each size, the four package designs were randomly allocated to stores. We would have three factors -- store size, number of colours, and presence versus absence of cartoons.

- For each independent variable, averaging over the other two variables would give marginal means -- the basis for estimating and testing for main effects.
- Averaging over each of the independent variables in turn, we would have a two-way marginal table of means for the other two variables, and the pattern of means in that table could show a two-way interaction.

The full three-dimensional table of means would provide a basis for looking at a three-way, or three-factor interaction. The interpretation of a three-way interaction is that the nature of the two-way interaction depends on the value of the third variable. This principle extends to any number of factors, so we would interpret a six-way interaction to mean that the nature of the 5-way interaction depends on the value of the sixth variable.

- Fortunately, the order in which one considers the variables does not matter. For example, we can say that the A by B interaction depends on the value of C, or that the A by C interaction depends on B, or that the B by C interaction depends on the value of A. The translations of these statements into algebra are all equivalent to one another, always. This principle extends to any number of factors.

° As you might imagine, as the number of factors becomes large, *interpreting* higher-way interactions -- that is, figuring out what they mean -- becomes more and more difficult. For this reason, sometimes the higher-order interactions are deliberately omitted from the full model in big experimental designs; they are never tested. Is this reasonable? Most of my answers are just elaborate ways to say I don't know.

More than two values for an independent variable

Regardless of how many factors we have, or how many levels there are in each factor, we could always form a combination variable -- that is, a single categorical independent variable whose values represent all the combinations of independent variable values in the factorial design. We have seen that in a two-by-two design, the tests for both main effects and the interaction resolve themselves into tests for single contrasts -- contrasts of the means of the combination variable. When independent variables have more than two values, the same thing is true, except that tests for main effects and interactions appear as test for *collections* of contrasts on the combination variable.

It is useful to pursue this principle in detail, for three reasons.

° First, it thinking of an interaction as a collection of contrasts can really help you understand what an interaction is.

° Second, once you have seen the tests for main effects and interactions as collections of contrasts, you can easily compose a test for any collection of contrasts that is of interest.

° Third, seeing main effects and interactions in terms of contrasts makes it easy to see how they can be modified to become Bonferroni or Scheffe follow-ups to initial significant one-way ANOVA on the combination variable --- if you choose to follow this conservative data analytic strategy.

We'll start with an example.

The seeds of the canola plant yield a high-quality cooking oil. Canola is one of Canada's biggest cash crops. But each year, millions of dollars are lost because of a fungus that kills canola plants. Or is it just one fungus? All this stuff looks the same. It's a nasty black rot that grows fastest under moist, warm conditions. It looks quite a bit like the fungus that grows in between shower tiles.

A team of botanists recognized that although the fungus may look the same, there are actually several different kinds that are genetically distinct. There are also quite a few strains of canola plant, so the questions arose

- Are some strains of fungus more aggressive than others? That is, do they grow faster and overwhelm the plant's defenses faster?
- Are some strains of canola plant more vulnerable to infection than others?
- Are some strains of fungus more dangerous to certain strains of plant and less dangerous to others?

These questions can be answered directly by looking at main effects and the interaction, so a factorial experiment was designed in which canola plants of three different varieties were randomly selected to be infected with one of six genetically different types of fungus. The way they did it was to scrape a little patch at the base of the plant, and wrap the wound with a moist band-aid that had some fungus on it. Then the plant was placed in a very moist dark environment for three days. After three days the bandage was removed and the plant was put in a commercial greenhouse. On each of 14 consecutive days, various measurements were made on the plant. Here, we will be concerned with lesion length, the length of the fungus patch on the plant, measured in millimeters.

The dependent variable will be mean lesion length; the mean is over the 14 daily lesion length measurements for each plant. The independent variables are Cultivar (type of canola plant) and MCG (type of fungus). Type of plant is called cultivar because the fungus grows (is "cultivated") on the plant. MCG stands for "Mycelial Compatibility Group." This strange name comes from the way that the botanists decided whether two types of fungus were genetically distinct. The would grow two samples on the same dish in a nutrient solution, and if the two fungus patches stayed separate, they were genetically different. If they grew together into a single patch of fungus (that is, they were compatible), then they were genetically identical. Apparently, this phenomenon is well established.

Here is the SAS program `appgreen1.sas`. As usual, the entire program is listed first. Then pieces of the program are repeated, together with pieces of output and discussion.

```
/* appgreen1.sas */
%include 'gh91read.sas';
options pagesize=100;
proc freq;
     tables plant*mcg /norow nocol nopercent;
proc glm;
     class plant mcg;
     model meanlng = plant|mcg;
     means plant|mcg;
proc tabulate;
     class mcg plant;
     var meanlng ;
     table (mcg all),(plant all) * (mean*meanlng);

/* Replicate tests for main effects and interactions, using contrasts on a
   combination variable. This is the hard way to do it, but if you can do
   this, you understand interactions and you can test any collection of
   contrasts. The definition of the variable combo could have been in
   gh91read.sas   */

data slime;
     set mould; /* mould was created by ghread91.sas */
     if         plant=1 and mcg=1 then combo =  1;
        else if plant=1 and mcg=2 then combo =  2;
        else if plant=1 and mcg=3 then combo =  3;
        else if plant=1 and mcg=7 then combo =  4;
        else if plant=1 and mcg=8 then combo =  5;
        else if plant=1 and mcg=9 then combo =  6;
        else if plant=2 and mcg=1 then combo =  7;
        else if plant=2 and mcg=2 then combo =  8;
        else if plant=2 and mcg=3 then combo =  9;
        else if plant=2 and mcg=7 then combo = 10;
        else if plant=2 and mcg=8 then combo = 11;
        else if plant=2 and mcg=9 then combo = 12;
        else if plant=3 and mcg=1 then combo = 13;
        else if plant=3 and mcg=2 then combo = 14;
        else if plant=3 and mcg=3 then combo = 15;
        else if plant=3 and mcg=7 then combo = 16;
        else if plant=3 and mcg=8 then combo = 17;
        else if plant=3 and mcg=9 then combo = 18;
     label combo = 'Plant-MCG Combo';
```

```
/* Getting main effects and the interaction with CONTRAST statements */
proc glm;
     class combo;
     model meanlng = combo;
     contrast 'Plant Main Effect'
         combo 1  1  1  1  1  1  -1 -1 -1 -1 -1 -1   0  0  0  0  0  0,
         combo 0  0  0  0  0  0   1  1  1  1  1  1  -1 -1 -1 -1 -1 -1;
     contrast 'MCG Main Effect'
         combo 1 -1  0  0  0  0   1 -1  0  0  0  0   1 -1  0  0  0  0,
         combo 0  1 -1  0  0  0   0  1 -1  0  0  0   0  1 -1  0  0  0,
         combo 0  0  1 -1  0  0   0  0  1 -1  0  0   0  0  1 -1  0  0,
         combo 0  0  0  1 -1  0   0  0  0  1 -1  0   0  0  0  1 -1  0,
         combo 0  0  0  0  1 -1   0  0  0  0  1 -1   0  0  0  0  1 -1;
     contrast 'Plant by MCG Interaction'
         combo -1  1  0  0  0  0   1 -1  0  0  0  0   0  0  0  0  0  0,
         combo  0  0  0  0  0  0  -1  1  0  0  0  0   1 -1  0  0  0  0,
         combo  0 -1  1  0  0  0   0  1 -1  0  0  0   0  0  0  0  0  0,
         combo  0  0  0  0  0  0   0 -1  1  0  0  0   0  1 -1  0  0  0,
         combo  0  0 -1  1  0  0   0  0  1 -1  0  0   0  0  0  0  0  0,
         combo  0  0  0  0  0  0   0  0 -1  1  0  0   0  0  1 -1  0  0,
         combo  0  0  0 -1  1  0   0  0  0  1 -1  0   0  0  0  0  0  0,
         combo  0  0  0  0  0  0   0  0  0 -1  1  0   0  0  0  1 -1  0,
         combo  0  0  0  0 -1  1   0  0  0  0  1 -1   0  0  0  0  0  0,
         combo  0  0  0  0  0  0   0  0  0  0 -1  1   0  0  0  0  1 -1;

/* proc reg's test statement may be easier, but first we need to
   make 16 dummy variables for cell means coding. This will illustrate
   arrays and loops, too */

data yucky;
     set slime;
     array mu{18} mu1-mu18;
     do i=1 to 18;
          if combo=. then mu{i}=.;
          else if combo=i then mu{i}=1;
          else mu{i}=0;
     end;

proc reg;
     model meanlng = mu1-mu18 / noint;
     alleq:   test mu1=mu2=mu3=mu4=mu5=mu6=mu7=mu8=mu9=mu10=mu11=mu12
                   = mu13=mu14=mu15=mu16=mu17=mu18;

     plant:   test mu1+mu2+mu3+mu4+mu5+mu6    = mu7+mu8+mu9+mu10+mu11+mu12,
                   mu7+mu8+mu9+mu10+mu11+mu12 = mu13+mu14+mu15+mu16+mu17+mu18;

     fungus:  test mu1+mu7+mu13 = mu2+mu8+mu14 = mu3+mu9+mu15
                   = mu4+mu10+mu16 = mu5+mu11+mu17 = mu6+mu12+mu18;

     p_by_f:  test mu2-mu1=mu8-mu7=mu14-mu13,
                   mu3-mu2=mu9-mu8=mu15-mu14,
```

```
                   mu4-mu3=mu10-mu9=mu16-mu15,
                   mu5-mu4=mu11-mu10=mu17-mu16,
                   mu6-mu5=mu12-mu11=mu18-mu17;

/* Now illustrate effect coding, with the interaction represented by a
   collection of product terms.  */

data nasty;
     set yucky;
     /* Two dummy variables for plant */
        if plant=. then p1=.;
        else if plant=1 then p1=1;
        else if plant=3 then p1=-1;
        else p1=0;
     if plant=. then p2=.;
        else if plant=2 then p2=1;
        else if plant=3 then p2=-1;
        else p2=0;
     /* Five dummy variables for mcg */
     if mcg=. then f1=.;
        else if mcg=1 then f1=1;
        else if mcg=9 then f1=-1;
        else f1=0;
     if mcg=. then f2=.;
        else if mcg=2 then f2=1;
        else if mcg=9 then f2=-1;
        else f2=0;
     if mcg=. then f3=.;
        else if mcg=3 then f3=1;
        else if mcg=9 then f3=-1;
        else f3=0;
     if mcg=. then f4=.;
        else if mcg=7 then f4=1;
        else if mcg=9 then f4=-1;
        else f4=0;
     if mcg=. then f5=.;
        else if mcg=8 then f5=1;
        else if mcg=9 then f5=-1;
        else f5=0;
     /* Product terms for interactions */
        p1f1 = p1*f1; p1f2=p1*f2 ; p1f3=p1*f3 ; p1f4=p1*f4; p1f5=p1*f5;
        p2f1 = p2*f1; p2f2=p2*f2 ; p2f3=p2*f3 ; p2f4=p2*f4; p2f5=p2*f5;

proc reg;
     model meanlng = p1 -- p2f5;
     plant:  test p1=p2=0;
     mcg:    test f1=f2=f3=f4=f5=0;
     p_by_f: test p1f1=p1f2=p1f3=p1f4=p1f5=p2f1=p2f2=p2f3=p2f4=p2f5 = 0;
```

The SAS program starts with a `%include` statement that reads `ghread91.sas`. The file `ghread91.sas` consists of a single big data step. We'll skip it, because all we really need are the two independent variables `plant` and `mcg`, and the dependent variable `meanlng`.

Just to see what we've got, we do a `proc freq` to show the sample sizes.

```
proc freq;
     tables plant*mcg /norow nocol nopercent;
```

and we get

```
                                TABLE OF PLANT BY MCG

    PLANT(Type of Plant)     MCG(Mycelial Compatibility Group)

    Frequency|       1|       2|       3|       7|       8|       9|  Total
    ---------+--------+--------+--------+--------+--------+--------+
    GP159    |      6 |      6 |      6 |      6 |      6 |      6 |     36
    ---------+--------+--------+--------+--------+--------+--------+
    HANNA    |      6 |      6 |      6 |      6 |      6 |      6 |     36
    ---------+--------+--------+--------+--------+--------+--------+
    WESTAR   |      6 |      6 |      6 |      6 |      6 |      6 |     36
    ---------+--------+--------+--------+--------+--------+--------+
    Total          18       18       18       18       18       18      108
```

So it's a nice 3 by 6 factorial design, with 6 plants in each treatment combination. The `proc glm` for analyzing this is straightforward. Again, we get all main effects and interactions for the factor names separated by vertical bars.

```
proc glm;
     class plant mcg;
     model meanlng = plant|mcg;
     means plant|mcg;
```

And the output is

General Linear Models Procedure
Class Level Information

Class	Levels	Values
PLANT	3	GP159 HANNA WESTAR
MCG	6	1 2 3 7 8 9

Number of observations in data set = 108

General Linear Models Procedure

Dependent Variable: MEANLNG Average Lesion length

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	328016.87350	19295.11021	19.83	0.0001
Error	90	87585.62589	973.17362		
Corrected Total	107	415602.49939			

R-Square	C.V.	Root MSE	MEANLNG Mean
0.789256	48.31044	31.195731	64.573479

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PLANT	2	221695.12747	110847.56373	113.90	0.0001
MCG	5	58740.26456	11748.05291	12.07	0.0001
PLANT*MCG	10	47581.48147	4758.14815	4.89	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PLANT	2	221695.12747	110847.56373	113.90	0.0001
MCG	5	58740.26456	11748.05291	12.07	0.0001
PLANT*MCG	10	47581.48147	4758.14815	4.89	0.0001

Notice that the Type I and Type III tests are the same. This always happens when the sample sizes are equal.

```
                    1991 Greenhouse Study                               4
                                 10:42 Tuesday, February 19, 2002

          General Linear Models Procedure

   Level of        -----------MEANLNG-----------
   PLANT       N       Mean              SD

   GP159      36      14.055159       12.1640757
   HANNA      36      55.700198       30.0137912
   WESTAR     36     123.965079       67.0180440

   Level of        -----------MEANLNG-----------
   MCG         N       Mean              SD

   1          18     41.4500000       33.6183462
   2          18     92.1333333       78.3509451
   3          18     87.5857143       61.7086751
   7          18     81.7603175       82.6711755
   8          18     50.8579365       39.3417859
   9          18     33.6535714       39.1480830

Level of   Level of        -----------MEANLNG-----------
PLANT      MCG         N       Mean              SD

GP159      1           6      12.863095       12.8830306
GP159      2           6      21.623810       17.3001296
GP159      3           6      14.460714        7.2165396
GP159      7           6      17.686905       16.4258441
GP159      8           6       8.911905        7.3162618
GP159      9           6       8.784524        6.5970501
HANNA      1           6      45.578571       26.1430472
HANNA      2           6      67.296429       30.2424997
HANNA      3           6      94.192857       20.2877876
HANNA      7           6      53.621429       24.8563497
HANNA      8           6      47.838095       12.6419109
HANNA      9           6      25.673810       17.1723150
WESTAR     1           6      65.908333       35.6968616
WESTAR     2           6     187.479762       45.1992178
WESTAR     3           6     154.103571       26.5469183
WESTAR     7           6     173.972619       79.1793105
WESTAR     8           6      95.823810       22.3712022
WESTAR     9           6      66.502381       52.5253101
```

The main effects are fairly easy to look at, and we definitely can construct a plot from the 18 cell means (or copy them into a nicer-looking table. But the following `proc tabulate` prints a table that is much easier to look at.

```
proc tabulate;
     class mcg plant;
     var meanlng ;
     table (mcg all),(plant all) * (mean*meanlng);
```

The syntax of `proc tabulate` is fairly elaborate, and at times it's worth the effort. Any reader who has seen the type of stub-and-banner tables favoured by professional market researchers will be impressed to hear that `proc tabulate` can come close to that. I figured out how to make the table below by looking in the manual. I then promptly forgot the overall principles, because it's not a tool I use a lot -- and the syntax is rather arcane. However, this example is easy to follow if you want to produce good-looking two-way tables of means. Here's the output.

	Type of Plant			
	GP159	HANNA	WESTAR	ALL
	MEAN	MEAN	MEAN	MEAN
	Average Lesion length	Average Lesion length	Average Lesion length	Average Lesion length
Mycelial Compatibility Group				
1	12.86	45.58	65.91	41.45
2	21.62	67.30	187.48	92.13
3	14.46	94.19	154.10	87.59
7	17.69	53.62	173.97	81.76
8	8.91	47.84	95.82	50.86
9	8.78	25.67	66.50	33.65
ALL	14.06	55.70	123.97	64.57

The proc tabulate output makes it easy to graph the means. But before we do so, let's look at the main effects and interactions as collections of contrasts. This will actually make it easier to figure out what the results mean, once we see what they are.

We have a three by six factorial design that looks like this. Population means are shown in the cells. The single-subscript notation encourages us to think of the combination of MCG and cultivar as a single categorical independent variable with 18 categories.

	MCG (Type of Fungus)					
Cultivar (Type of Plant)	1	2	3	7	8	9
GP159	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
Hanna	μ_7	μ_8	μ_9	μ_{10}	μ_{11}	μ_{12}
Westar	μ_{13}	μ_{14}	μ_{15}	μ_{16}	μ_{17}	μ_{18}

Next is the part of the SAS program that creates the combination variable. Notice that it involves a data step that comes *after* the `proc glm`. This usually doesn't happen. I did it by creating a new data set called `slime` that starts by being identical to `mould`, which was created in the file `gh91read.sas`. The `set` command is used to read in the data set `mould`, and then we start from there. This is done just for teaching purposes. Ordinarily, I would not create multiple data sets that are mostly copies of each other. I'd put the whole thing in one data step. Here's the code.

```
data slime;
     set mould; /* mould was created by ghread91.sas */
     if         plant=1 and mcg=1 then combo =  1;
        else if plant=1 and mcg=2 then combo =  2;
        else if plant=1 and mcg=3 then combo =  3;
        else if plant=1 and mcg=7 then combo =  4;
        else if plant=1 and mcg=8 then combo =  5;
        else if plant=1 and mcg=9 then combo =  6;
        else if plant=2 and mcg=1 then combo =  7;
        else if plant=2 and mcg=2 then combo =  8;
        else if plant=2 and mcg=3 then combo =  9;
        else if plant=2 and mcg=7 then combo = 10;
        else if plant=2 and mcg=8 then combo = 11;
        else if plant=2 and mcg=9 then combo = 12;
        else if plant=3 and mcg=1 then combo = 13;
        else if plant=3 and mcg=2 then combo = 14;
        else if plant=3 and mcg=3 then combo = 15;
        else if plant=3 and mcg=7 then combo = 16;
        else if plant=3 and mcg=8 then combo = 17;
        else if plant=3 and mcg=9 then combo = 18;
     label combo = 'Plant-MCG Combo';
```

	MCG (Type of Fungus)					
Cultivar (Type of Plant)	1	2	3	7	8	9
GP159	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
Hanna	μ_7	μ_8	μ_9	μ_{10}	μ_{11}	μ_{12}
Westar	μ_{13}	μ_{14}	μ_{15}	μ_{16}	μ_{17}	μ_{18}

It is clear that the absence of a main effect for Cultivar is the same as

$$\mu_1+\mu_2+\mu_3+\mu_4+\mu_5+\mu_6 = \mu_7+\mu_8+\mu_9+\mu_{10}+\mu_{11}+\mu_{12} = \mu_{13}+\mu_{14}+\mu_{15}+\mu_{16}.$$

There are two equalities here, and they are saying that two contrasts of the eighteen cell means are equal to zero. To see why this is true, consider the first equality

$$\mu_1+\mu_2+\mu_3+\mu_4+\mu_5+\mu_6 = \mu_7+\mu_8+\mu_9+\mu_{10}+\mu_{11}+\mu_{12}$$

Subtracting the quantity on the right-hand side from both sider of the equation, we get

$$\mu_1+\mu_2+\mu_3+\mu_4+\mu_5+\mu_6 - (\mu_7+\mu_8+\mu_9+\mu_{10}+\mu_{11}+\mu_{12}) = 0,$$

and then distributing the minus sign to get rid of the parentheses yields

$$\mu_1+\mu_2+\mu_3+\mu_4+\mu_5+\mu_6-\mu_7-\mu_8-\mu_9-\mu_{10}-\mu_{11}-\mu_{12} = 0. \qquad (4.2)$$

Recall that here, a contrast is a linear combination of the form

$$L = a_1\mu_1 + a_2\mu_2 + ... + a_{18}\mu_{18}.,$$

where the *a* weights add up to zero. Expression (4.2) fits this description, with the first 6 weights equal to one, the next six weights equal to minus one (so they add to zero), and the last 6 weights equal to zero.

The table below gives the weights of the contrasts defining the test for the main effect of plant, one set of weights in each row.

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}
1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1

This is the basis of the first contrast statement in `proc glm`. Notice how the contrasts are separated by commas. Also notice that the variable on which we're doing contrasts (`combo`) has to be repeated.

```
/* Getting main effects and the interaction with CONTRAST statements */
proc glm;
     class combo;
     model meanlng = combo;
     contrast 'Plant Main Effect'
         combo 1  1  1  1  1  1  -1 -1 -1 -1 -1 -1   0  0  0  0  0  0,
         combo 0  0  0  0  0  0   1  1  1  1  1  1  -1 -1 -1 -1 -1 -1;
```

If there is no main effect for MCG, we are saying

$$\mu_1+\mu_7+\mu_{13} = \mu_2+\mu_8+\mu_{14} = \mu_3+\mu_9+\mu_{15} = \mu_4+\mu_{10}+\mu_{16} = \mu_5+\mu_{11}+\mu_{17} = \mu_6+\mu_{12}+\mu_{18}.$$

There are 5 contrasts here, one for each equals sign; there is always an equals sign for each contrast. Here is the table showing the contrasts.

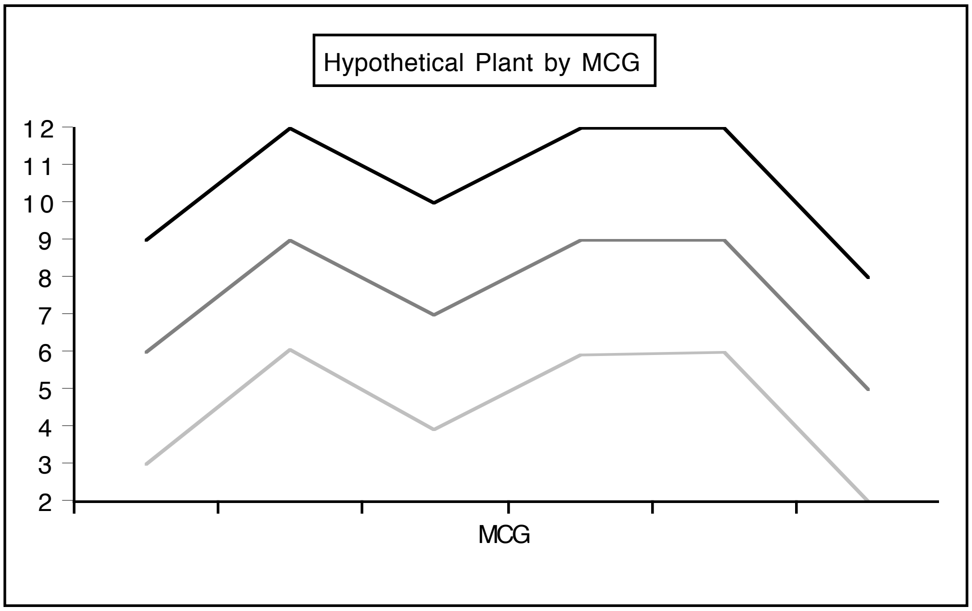
a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}
1	-1	0	0	0	0	1	-1	0	0	0	0	1	-1	0	0	0	0
0	1	-1	0	0	0	0	1	-1	0	0	0	0	1	-1	0	0	0
0	0	1	-1	0	0	0	0	1	-1	0	0	0	0	1	-1	0	0
0	0	0	1	-1	0	0	0	0	1	-1	0	0	0	0	1	-1	0
0	0	0	0	1	-1	0	0	0	0	1	-1	0	0	0	0	1	-1

And here is the corresponding `test` statement in `proc glm`.

```
contrast 'MCG Main Effect'
    combo 1 -1  0  0  0  0   1 -1  0  0  0  0   1 -1  0  0  0  0,
    combo 0  1 -1  0  0  0   0  1 -1  0  0  0   0  1 -1  0  0  0,
    combo 0  0  1 -1  0  0   0  0  1 -1  0  0   0  0  1 -1  0  0,
    combo 0  0  0  1 -1  0   0  0  0  1 -1  0   0  0  0  1 -1  0,
    combo 0  0  0  0  1 -1   0  0  0  0  1 -1   0  0  0  0  1 -1;
```

	MCG (Type of Fungus)					
Cultivar (Type of Plant)	1	2	3	7	8	9
GP159	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
Hanna	μ_7	μ_8	μ_9	μ_{10}	μ_{11}	μ_{12}
Westar	μ_{13}	μ_{14}	μ_{15}	μ_{16}	μ_{17}	μ_{18}

To compose the Plant by MCG interaction, consider the following hypothetical graph. You can think of the "effect" of MCG as a *profile*, representing a pattern of differences among means. If the three profiles are the same shape for each type of plant -- that is, if they are parallel, the effect of MCG does not depend on the type of plant, and there is no interaction.



For the profiles to be parallel, each set of corresponding line segments must be parallel. To start with the three line segments on the left, the rise represented by $\mu_2 - \mu_1$ must equal the rise $\mu_8 - \mu_7$, and $\mu_8 - \mu_7$ must equal $\mu_{14} - \mu_{13}$. This is two contrasts that equal zero:

$$\mu_2 - \mu_1 - \mu_8 + \mu_7 = 0 \text{ and } \mu_8 - \mu_7 - \mu_{14} + \mu_{13} = 0.$$

There are two contrasts for each of the four remaining sets of three line segments, for a total of ten contrasts. They appear directly in the `contrast` statement of `proc glm`. Notice how each row adds to zero; these are *contrasts*, not just linear combinations.

```
contrast 'Plant by MCG Interaction'
      combo -1  1  0  0  0  0   1 -1  0  0  0  0   0  0  0  0  0  0,
      combo  0  0  0  0  0  0  -1  1  0  0  0  0   1 -1  0  0  0  0,
      combo  0 -1  1  0  0  0   0  1 -1  0  0  0   0  0  0  0  0  0,
      combo  0  0  0  0  0  0   0 -1  1  0  0  0   0  1 -1  0  0  0,
      combo  0  0 -1  1  0  0   0  0  1 -1  0  0   0  0  0  0  0  0,
      combo  0  0  0  0  0  0   0  0 -1  1  0  0   0  0  1 -1  0  0,
      combo  0  0  0 -1  1  0   0  0  0  1 -1  0   0  0  0  0  0  0,
      combo  0  0  0  0  0  0   0  0  0 -1  1  0   0  0  0  1 -1  0,
      combo  0  0  0  0 -1  1   0  0  0  0  1 -1   0  0  0  0  0  0,
      combo  0  0  0  0  0  0   0  0  0  0 -1  1   0  0  0  0  1 -1;
```

Now we can compare the tests we get from these contrast statements with what we got from a two-way ANOVA. For easy reference, here is part of the two-way output.

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

PLANT                   2     221695.12747    110847.56373   113.90    0.0001
MCG                     5      58740.26456     11748.05291    12.07    0.0001
PLANT*MCG              10      47581.48147      4758.14815     4.89    0.0001
```

And here is the output from the `contrast` statements.

```
Contrast               DF      Contrast SS     Mean Square  F Value    Pr > F

Plant Main Effect       2     221695.12747    110847.56373   113.90    0.0001
MCG Main Effect         5      58740.26456     11748.05291    12.07    0.0001
Plant by MCG Interac   10      47581.48147      4758.14815     4.89    0.0001
```

So it worked. Here are some comments.

- Of course this is *not* the way you'd want to test for main effects and interactions. On the contrary, it makes you appreciate all the work that `glm` does for you when you say `model meanlng = plant|mcg;`

- These contrasts are supposed to be an aid to understanding --- understanding what main effects and interactions really are, and understanding how you can test nearly any hypothesis you can think of in a multi-factor design. Almost without exception, what you want to do is test whether some collection of contrasts are equal to zero. Now you can do it, whether the collection you're interested in happens to be standard, or not.

- On the other hand, this was brutal. Even though I am comfortable with high school algebra, the size of the design made specifying those contrasts an unpleasant experience. There is an easier way.

An Easier Way to test Sets of Contrasts in Factorial ANOVA

Because the `test` statement of `proc reg` has a more flexible syntax than the `contrast` statement of `proc glm`, it's a lot easier if you use cell means dummy variable coding, fit a model with no intercept in `proc reg`, and use `test` statements. In the following example, the indicator dummy variables are named `mu1` to `mu18`. This choice makes it possible to directly transcribe statements about the population cell means into test statements. I highly recommend it. Of course if you really hate Greek letters, you could always name them `m1` to `m18` or something.

First, we need to define 18 dummy variables. In general, it's a bit more tedious to define dummy variables than to make a combination variable. Here, I use the combination variable `combo` (which has already been created) to make the task a bit easier -- and also to illustrate the use of arrays and loops in the data step.

```
/* proc reg's test statement may be easier, but first we need to
   make 16 dummy variables for cell means coding. This will illustrate
   arrays and loops, too */

data yucky;
     set slime;
     array mu{18} mu1-mu18;
     do i=1 to 18;
          if combo=. then mu{i}=.;
          else if combo=i then mu{i}=1;
          else mu{i}=0;
     end;

proc reg;
     model meanlng = mu1-mu18 / noint;
     alleq:   test mu1=mu2=mu3=mu4=mu5=mu6=mu7=mu8=mu9=mu10=mu11=mu12
                   = mu13=mu14=mu15=mu16=mu17=mu18;

     plant:   test mu1+mu2+mu3+mu4+mu5+mu6    = mu7+mu8+mu9+mu10+mu11+mu12,
                   mu7+mu8+mu9+mu10+mu11+mu12 = mu13+mu14+mu15+mu16+mu17+mu18;

     fungus:  test mu1+mu7+mu13 = mu2+mu8+mu14 = mu3+mu9+mu15
                   = mu4+mu10+mu16 = mu5+mu11+mu17 = mu6+mu12+mu18;

     p_by_f:  test mu2-mu1=mu8-mu7=mu14-mu13,
                   mu3-mu2=mu9-mu8=mu15-mu14,
                   mu4-mu3=mu10-mu9=mu16-mu15,
                   mu5-mu4=mu11-mu10=mu17-mu16,
                   mu6-mu5=mu12-mu11=mu18-mu17;
```

Looking again at the table of means, it's easy to see how natural the syntax is.

	MCG (Type of Fungus)					
Cultivar (Type of Plant)	1	2	3	7	8	9
GP159	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
Hanna	μ_7	μ_8	μ_9	μ_{10}	μ_{11}	μ_{12}
Westar	μ_{13}	μ_{14}	μ_{15}	μ_{16}	μ_{17}	μ_{18}

And again, the tests are correct. First, repeat the output from the `contrast` statements of `proc glm` (which matched the `proc glm` two-way ANOVA output).

```
Contrast                  DF      Contrast SS     Mean Square  F Value    Pr > F

Plant Main Effect          2     221695.12747    110847.56373   113.90    0.0001
MCG Main Effect            5      58740.26456     11748.05291    12.07    0.0001
Plant by MCG Interac      10      47581.48147      4758.14815     4.89    0.0001
```

Then, compare output from the test statements of proc reg.

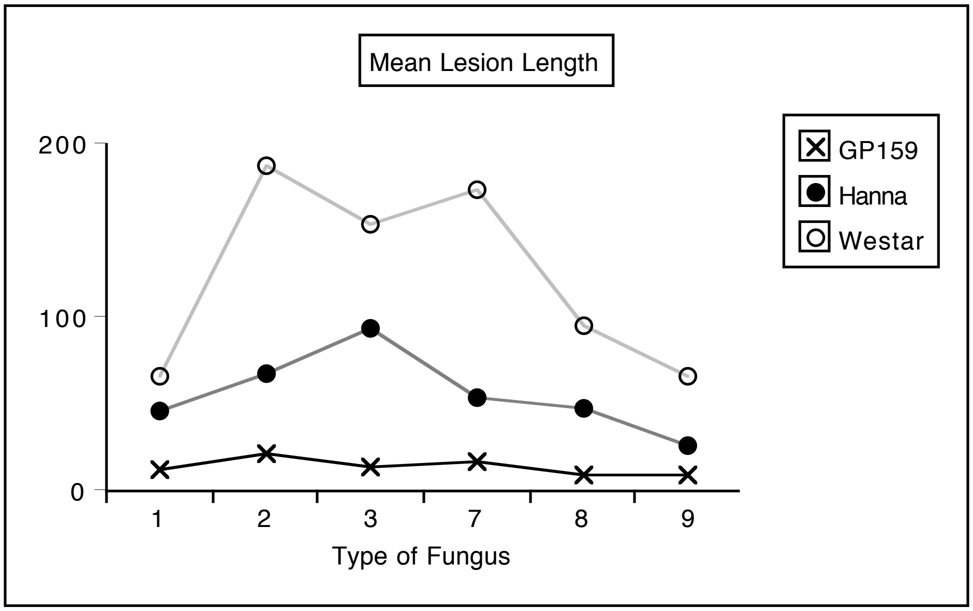
```
Dependent Variable: MEANLNG
Test: ALLEQ    Numerator:  19295.1102  DF:   17   F value:  19.8270
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: PLANT    Numerator: 110847.5637  DF:    2   F value: 113.9032
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: FUNGUS   Numerator:  11748.0529  DF:    5   F value:  12.0719
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: P_BY_F   Numerator:   4758.1481  DF:   10   F value:   4.8893
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001
```

Okay, now we know how to do anything. Finally, it is time to graph the interaction, and find out what these results mean!



First, we see a sizable and clear main effect for Plant. In fact, going back to the analysis of variance summary tables and dividing the Sum of Squares explained by Plant by the Total Sum of Squares, we observe that Plant explains around 53% of the variation in mean lesion length. That's huge. We will definitely want to look at pairwise comparisons of marginal means, too; we'll get back to this later.

Looking at the pattern of means, it's clear that while the main effect of fungus type is statistically significant, this is not something that should be interpreted, because which one is best (worst) depends on the type of plant. That is, we need to look at the interaction.

The profiles really look different. In particular, GP159 not only has a smaller average lesion length, but it seems to exhibit less responsiveness to different strains of fungus. A test for the equality of μ_1 through μ_6 would be valuable. Pairwise comparisons of the 6 means for Hanna and the 6 means for Westar look promising, too.

A Brief Consideration of Multiple Comparisons

The mention of pairwise comparisons brings up the issue of formal multiple comparison follow-up tests for this problem. The way people often do follow-up tests for factorial designs is to make a combination variable and then do all pairwise comparisons. It seems like they do this because they think it's the only thing the software will let them do. Certainly it's better than nothing. Some comments:

With SAS, pairwise comparisons of cell means are *not* the only thing you can do. `Proc glm` will do all pairwise comparisons of *marginal* means quite easily. This means it's easy to follow up a significant and meaningful main effect.

For the present problem, there are 120 possible pairwise comparisons of the 16 cell means. If we do all these as one-at-a-time tests, the chances of false significance are certainly mounting. There is a strong case here for doing multiple comparisons.

Since the sample sizes are equal, Tukey tests are most powerful for all pairwise comparisons. But it's not so simple. Pairwise comparisons within plants (for example, comparing the 6 means for Westar) are interesting, and pairwise comparisons within fungus types (for example, comparison of Hanna, Westar and GP159 for fungus Type 1) are interesting, but the remaining 57 pairwise comparisons are a lot less so.

Also, pairwise comparisons of cell means are not all we want to do. We've already mentioned the need for pairwise comparisons of the marginal means for plants, and we'll soon see that other, less standard comparisons are of interest.

Everything we need to do will involve testing collections of contrasts. The approach we'll take is to do everything as a one-at-a-time custom test initially, and then figure out how we should correct for the fact that we've done a lot of tests.

It's good to be guided by the data. Here we go. The analyses will be done in the SAS program `appgreen2.sas`. As usual, the entire program is given first. But you should be aware that the program was written one piece at a time and executed many times, with later analyses being suggested by the earlier ones.

The program starts by reading in the file `gh91bread.sas`, which is just `gh91read.sas` with the additional variables defined (especially `combo` and `mu1` through `mu18`) that were defined in `appgreen1.sas`.

```
/* appgreen2.sas:  */
%include 'gh91bread.sas';
options pagesize=100;


proc glm;
     title 'Repeating initial Plant by MCG ANOVA, full design';
     class plant mcg;
     model meanlng = plant|mcg;
     means plant|mcg;

/*  A.  Pairwise comparisons of marginal means for plant, full design
    B.  Test all GP159 means equal, full design
    C.  Test profiles for Hanna & Westar parallel, full design         */

proc reg;
     model meanlng = mu1-mu18 / noint;
     A_GvsH:   test  mu1+mu2+mu3+mu4+mu5+mu6 = mu7+mu8+mu9+mu10+mu11+mu12;
     A_GvsW:   test  mu1+mu2+mu3+mu4+mu5+mu6 = mu13+mu14+mu15+mu16+mu17+mu18;
     A_HvsW:   test  mu7+mu8+mu9+mu10+mu11+mu12 = mu13+mu14+mu15+mu16+mu17+mu18;
     B_G159eq: test  mu1=mu2=mu3=mu4=mu5=mu6;
     C_HWpar:  test  mu8-mu7=mu14-mu13,  mu9-mu8=mu15-mu14,
                     mu10-mu9=mu16-mu15, mu11-mu10=mu17-mu16,
                     mu12-mu11=mu18-mu17;


/*  D.  Oneway on mcg, GP158 subset  */

data just159;  /* This data set will have just GP159 */
     set mould;
     if plant=1;

proc glm data=just159;
     title 'D.  Oneway on mcg, GP158 subset';
     class mcg;
     model meanlng = mcg;

/*  E.  Plant by MCG, Hanna-Westar subset   */

data hanstar;  /* This data set will have just Hanna and Westar */
     set mould;
     if plant ne 1;

proc glm data=hanstar;
     title 'E.  Plant by MCG, Hanna-Westar subset';
     class plant mcg;
     model meanlng = plant|mcg;
```

```
/*  F.  Plant by MCG followup, Hanna-Westar subset
                     Interaction:  Follow with all pairwise differences of
                     Westar minus Hanna differences
   G.   Differences within Hanna?
   H.   Differences within Westar?  */

proc reg;
     model meanlng = mu7-mu18 / noint;
     F_inter: test   mu13-mu7=mu14-mu8=mu15-mu9
                   = mu16-mu10=mu17-mu11=mu18-mu12;
     F_1vs2:   test   mu13-mu7=mu14-mu8;
     F_1vs3:   test   mu13-mu7=mu15-mu9;
     F_1vs7:   test   mu13-mu7=mu16-mu10;
     F_1vs8:   test   mu13-mu7=mu17-mu11;
     F_1vs9:   test   mu13-mu7=mu18-mu12;
     F_2vs3:   test   mu14-mu8=mu15-mu9;
     F_2vs7:   test   mu14-mu8=mu16-mu10;
     F_2vs8:   test   mu14-mu8=mu17-mu11;
     F_2vs9:   test   mu14-mu8=mu18-mu12;
     F_3vs7:   test   mu15-mu9=mu16-mu10;
     F_3vs8:   test   mu15-mu9=mu17-mu11;
     F_3vs9:   test   mu15-mu9=mu18-mu12;
     F_7vs8:   test   mu16-mu10=mu17-mu11;
     F_7vs9:   test   mu16-mu10=mu18-mu12;
     F_8vs9:   test   mu17-mu11=mu18-mu12;
     G_Hanaeq: test   mu7=mu8=mu9=mu10=mu11=mu12;
     H_Westeq: test   mu13=mu14=mu15=mu16=mu17=mu18;


proc iml; /* Critical values for Scheffe tests */
     interac = finv(.95,5,60) ; print interac;
     oneway = finv(.95,11,60); print oneway;
```

After reading and defining the data with a `%include` statement, the program repeats the initial three by six ANOVA from `appgreen1.sas`. This is just for completeness.

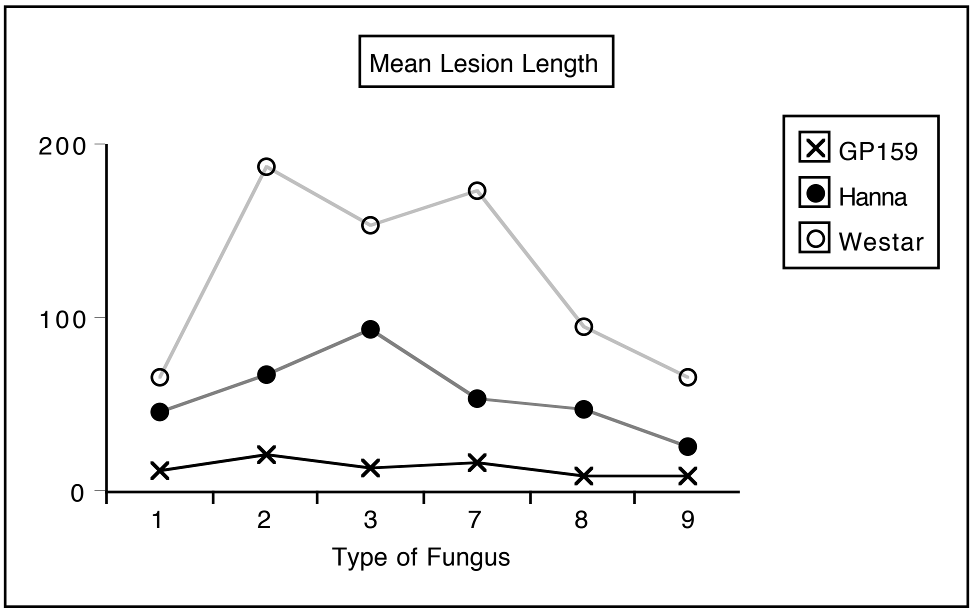
A. It then uses `proc reg` to fit a cell means model, and then tests for all three pairwise differences among Plant means. They are all significantly different from each other, confirming what appears visually in the interaction plot.

```
proc reg;
     model meanlng = mu1-mu18 / noint;
     A_GvsH: test  mu1+mu2+mu3+mu4+mu5+mu6 = mu7+mu8+mu9+mu10+mu11+mu12;
     A_GvsW: test  mu1+mu2+mu3+mu4+mu5+mu6 = mu13+mu14+mu15+mu16+mu17+mu18;
     A_HvsW: test  mu7+mu8+mu9+mu10+mu11+mu12 = mu13+mu14+mu15+mu16+mu17+mu18;
```

```
Dependent Variable: MEANLNG
Test: A_GVSH   Numerator:  31217.5679  DF:    1   F value:  32.0781
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001

Dependent Variable: MEANLNG
Test: A_GVSW   Numerator: 217443.4318  DF:    1   F value: 223.4374
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001

Dependent Variable: MEANLNG
Test: A_HVSW   Numerator:  83881.6915  DF:    1   F value:  86.1940
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001
```



As mentioned earlier, GP159 not only has a smaller average lesion length, but it seems to exhibit less variation in its vulnerability to different strains of fungus. Part of the significant interaction must come from this, and part from differences in the profiles of Hanna and Westar. Two questions arise:

1. Are μ_1 through μ_6 (the means for GP159) actually different from each other?
2. Are the profiles for Hanna and Westar different?

There are two natural ways to address these questions. The naive way is to subset the data --- that is, do a one-way ANOVA to compare the 6 means for GP159, and a two-way (2 by 6) on the Hanna-Westar subset. In the latter analysis, the interaction of Plant by MCG would indicate whether the two profiles were different.

A more sophisticated approach is not to subset the data, but to recognize that both questions can be answered by testing collections of contrasts of the entire set of 18 means; it's easy to do with the `test` statement of `proc reg`., or with `contrast` statements in `proc glm` -- see Chapter 6.

The advantage of the sophisticated approach is this. Remember that the model specifies a conditional normal distribution of the dependent variable for each combination of independent variable values (in this case there are 18 combinations of independent variable values), and that each conditional distribution has the *same variance*. The test for, say, the equality of μ_1 through μ_6 would use only $\overline{Y}_1$ through $\overline{Y}_6$ (that is, just GP159 data) to estimate the 5 contrasts involved, but it would use *all* the data to estimate the common error variance. From both a commonsense viewpoint and the deepest possible theoretical viewpoint, it's better not to throw information away. This is why the sophisticated approach should be better.

However, this argument is convincing only if it's really true that the dependent variable has the same variance for every combination of independent variable values. Repeating some output from the `means` command of the very first `proc glm`,

```
Level of    Level of          -----------MEANLNG-----------
PLANT       MCG          N       Mean                SD

GP159       1            6       12.863095       12.8830306
GP159       2            6       21.623810       17.3001296
GP159       3            6       14.460714        7.2165396
GP159       7            6       17.686905       16.4258441
GP159       8            6        8.911905        7.3162618
GP159       9            6        8.784524        6.5970501
HANNA       1            6       45.578571       26.1430472
HANNA       2            6       67.296429       30.2424997
HANNA       3            6       94.192857       20.2877876
HANNA       7            6       53.621429       24.8563497
HANNA       8            6       47.838095       12.6419109
HANNA       9            6       25.673810       17.1723150
WESTAR      1            6       65.908333       35.6968616
WESTAR      2            6      187.479762       45.1992178
WESTAR      3            6      154.103571       26.5469183
WESTAR      7            6      173.972619       79.1793105
WESTAR      8            6       95.823810       22.3712022
WESTAR      9            6       66.502381       52.5253101
```

we see that the sample standard deviations for GP159 look quite a bit smaller on average. Without bothering to do a formal test, we have some reason to doubt the equal variances assumption.

It's easy to see *why* GP159 would have less plant-to-plant variation in lesion length. It's so resistant to the fungus that there's just not that much fungal growth, period. So there's less *opportunity* for variation.

Note that the equal variances assumption is essentially just a mathematical convenience. Here, it's clearly unrealistic. But what's the consequence of violating it? It's well known that the equal variance assumption can be safely violated if the cell sample sizes are equal and large. Well, here they're equal, but n=6 is not large. So this is not reassuring.

In general, it's not easy to say HOW the tests will be affected when the equal variance assumption is violated, but for the two particular cases we're interested in here (are the GP159 means equal and are the Hanna and Westar profiles parallel), we can figure it out. Recall Formula (3.3) for the F-test.

$$\mathrm{F} = \frac{(SSR_F - SSR_R)\ /\ s}{MSE_F}.$$

The denominator --- Mean Squared Error from the full model --- is the estimated population error variance. That's the variance that's supposed to be the same for each conditional distribution. Since

$$\mathrm{MSE_F} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p},$$

and the predicted value $\hat{Y}_i$ is always the cell mean, we can draw the following conclusions.

1. When we test for equality of the GP159 means, using the Hanna-Westar data to help compute MSE will make the denominator of F bigger than it should be -- so F is made smaller, and the test is too conservative.

2. When we test whether the Hanna and Westar profiles are parallel, use of the GP159 data to help compute MSE will make the denominator of F *smaller* than it should be -- so F is made bigger, and the test is not conservative enough. That is, the chance of significance if the effect is absent will be greater than 0.05.

This makes me inclined to favour the "naive" subsetting approach. Because the GP159 means LOOK so equal, and I want them to be equal, I'd like to give the test for difference among them the best possible chance. And because it looks like the profiles for Hanna and Westar are not parallel (and I want them to be non-parallel, because it's more interesting for the effect of Fungus type to depend on type of Plant), I want a more conservative test.

Another argument in favour of subsetting is based on botany rather than statistics. Hanna and Westar are commercial canola crop varieties, but while GP159 is definitely in the canola family, it is more like a hardy weed than a food plant. It's just a different kind of entity, and so analyzing its data separately makes a lot of sense.

You may wonder, if it's so different, why was it included in the design in the first place? Well, taxonomically it's quite similar to Hanna and Westar; really no one knew it would be such a vigorous monster in terms of resisting fungus. That's why people do research -- to find out things they didn't already know.

Anyway, we'll do the analysis both ways -- both the seemingly naive way which is probably better once you think about it, and the sophisticated way that uses the complete set of data for all analyses.

Parts B and C represent the "sophisticated" approach that does not subset the data.

B. Test all GP159 means equal, full design

C. Test profiles for Hanna & Westar parallel, full design

```
proc reg;
     model meanlng = mu1-mu18 / noint;
     A_GvsH:   test  mu1+mu2+mu3+mu4+mu5+mu6 = mu7+mu8+mu9+mu10+mu11+mu12;
     A_GvsW:   test  mu1+mu2+mu3+mu4+mu5+mu6 = mu13+mu14+mu15+mu16+mu17+mu18;
     A_HvsW:   test  mu7+mu8+mu9+mu10+mu11+mu12 = mu13+mu14+mu15+mu16+mu17+mu18;
     B_G159eq: test  mu1=mu2=mu3=mu4=mu5=mu6;
     C_HWpar:  test  mu8-mu7=mu14-mu13,  mu9-mu8=mu15-mu14,
                     mu10-mu9=mu16-mu15, mu11-mu10=mu17-mu16,
                     mu12-mu11=mu18-mu17;


Dependent Variable: MEANLNG
Test: B_G159EQ Numerator:    151.5506  DF:    5   F value:   0.1557
               Denominator:  973.1736  DF:   90   Prob>F:    0.9778

Dependent Variable: MEANLNG
Test: C_HWPAR  Numerator:   5364.0437  DF:    5   F value:   5.5119
               Denominator:  973.1736  DF:   90   Prob>F:    0.0002
```

This confirms the visual impression of no differences among means for GP159, and non-parallel profiles for Hanna and Westar. Now compare the subsetting approach. Notice the creation of SAS data sets with subsets of the data.

D. Oneway on mcg, GP158 subset

E. Plant by MCG, Hanna-Westar subset

```
data just159; /* This data set will have just GP159 */
     set mould;
     if plant=1;

proc glm data=just159;
     title 'D.  Oneway on mcg, GP158 subset';
     class mcg;
     model meanlng = mcg;
```

```
                        D.  Oneway on mcg, GP158 subset                          2
                                                10:52 Friday, February 22, 2002

                        General Linear Models Procedure

Dependent Variable: MEANLNG   Average Lesion length
                                     Sum of             Mean
Source                  DF          Squares           Square  F Value    Pr > F

Model                    5     757.75319161     151.55063832     1.03    0.4189

Error                   30    4421.01258503     147.36708617

Corrected Total         35    5178.76577664

                  R-Square             C.V.         Root MSE       MEANLNG Mean

                  0.146319         86.37031        12.139485          14.055159


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

MCG                      5     757.75319161     151.55063832     1.03    0.4189

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

MCG                      5     757.75319161     151.55063832     1.03    0.4189
```

This analysis is consistent with what we got without subsetting the data. That is, it does not provide evidence that the means for GP159 are different. But when we didn't subset the data, we had p = 0.9778. This happened exactly because including Hanna and Westar data made MSE larger, F smaller, and hence p bigger.

```
data hanstar; /* This data set will have just Hanna and Westar */
     set mould;
     if plant ne 1;

proc glm data=hanstar;
     title 'E.  Plant by MCG, Hanna-Westar subset';
     class plant mcg;
     model meanlng = plant|mcg;
```

```
-------------------------------------------------------------------------------

                     E.  Plant by MCG, Hanna-Westar subset                    3
                                            10:52 Friday, February 22, 2002

                        General Linear Models Procedure
                            Class Level Information

                        Class    Levels    Values

                        PLANT         2    HANNA WESTAR

                        MCG           6    1 2 3 7 8 9

                    Number of observations in data set = 72

-------------------------------------------------------------------------------

                     E.  Plant by MCG, Hanna-Westar subset                    4
                                            10:52 Friday, February 22, 2002

                        General Linear Models Procedure

Dependent Variable: MEANLNG   Average Lesion length
                                      Sum of            Mean
Source                  DF           Squares          Square  F Value    Pr > F

Model                   11      189445.68433     17222.33494    12.43    0.0001

Error                   60       83164.61331      1386.07689

Corrected Total         71      272610.29764

                  R-Square              C.V.        Root MSE       MEANLNG Mean

                  0.694932          41.44379       37.230054          89.832639
```

```
Source                      DF        Type I SS     Mean Square  F Value    Pr > F

PLANT                        1     83881.691486    83881.691486    60.52    0.0001
MCG                          5     78743.774570    15748.754914    11.36    0.0001
PLANT*MCG                    5     26820.218272     5364.043654     3.87    0.0042

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

PLANT                        1     83881.691486    83881.691486    60.52    0.0001
MCG                          5     78743.774570    15748.754914    11.36    0.0001
PLANT*MCG                    5     26820.218272     5364.043654     3.87    0.0042

=========
```

The significant interaction indicates that the profiles for Hanna and Westar are non-parallel, confirming the visual impression we got from the interaction plot. But the p-value is larger this time. When all the data were used to calculate the error term, we had p = `0.0002`. This is definitely due to the low variation in GP159.

Further analyses will be limited to the Hanna-Westar subset.

Now think of the interaction in a different way. Overall, Hanna is more vulnerable than Westar, but the interaction says that the degree of that greater vulnerability depends on the type of fungus. Look at all pairwise comparisons of the DIFFERENCE between Hanna and Westar. First, verify that the interaction can be expressed this way. Of course it can.

F. Plant by MCG followup, Hanna-Westar subset

All pairwise differences of Westar minus Hanna differences

```
proc reg;
     model meanlng = mu7-mu18 / noint;
     F_inter: test   mu13-mu7=mu14-mu8=mu15-mu9
                   = mu16-mu10=mu17-mu11=mu18-mu12;
     F_1vs2:  test   mu13-mu7=mu14-mu8;
     F_1vs3:  test   mu13-mu7=mu15-mu9;
     F_1vs7:  test   mu13-mu7=mu16-mu10;
     F_1vs8:  test   mu13-mu7=mu17-mu11;
     F_1vs9:  test   mu13-mu7=mu18-mu12;
     F_2vs3:  test   mu14-mu8=mu15-mu9;
     F_2vs7:  test   mu14-mu8=mu16-mu10;
     F_2vs8:  test   mu14-mu8=mu17-mu11;
     F_2vs9:  test   mu14-mu8=mu18-mu12;
     F_3vs7:  test   mu15-mu9=mu16-mu10;
     F_3vs8:  test   mu15-mu9=mu17-mu11;
```

```
     F_3vs9:  test   mu15-mu9=mu18-mu12;
     F_7vs8:  test   mu16-mu10=mu17-mu11;
     F_7vs9:  test   mu16-mu10=mu18-mu12;
     F_8vs9:  test   mu17-mu11=mu18-mu12;


Dependent Variable: MEANLNG
Test: F_INTER  Numerator:   5364.0437  DF:    5   F value:   3.8699
               Denominator:  1386.077  DF:   60   Prob>F:    0.0042


Dependent Variable: MEANLNG
Test: F_1VS2   Numerator:  14956.1036  DF:    1   F value:  10.7902
               Denominator:  1386.077  DF:   60   Prob>F:    0.0017


Dependent Variable: MEANLNG
Test: F_1VS3   Numerator:   2349.9777  DF:    1   F value:   1.6954
               Denominator:  1386.077  DF:   60   Prob>F:    0.1979


Dependent Variable: MEANLNG
Test: F_1VS7   Numerator:  15006.4293  DF:    1   F value:  10.8265
               Denominator:  1386.077  DF:   60   Prob>F:    0.0017


Dependent Variable: MEANLNG
Test: F_1VS8   Numerator:   1147.2776  DF:    1   F value:   0.8277
               Denominator:  1386.077  DF:   60   Prob>F:    0.3666


Dependent Variable: MEANLNG
Test: F_1VS9   Numerator:    630.3018  DF:    1   F value:   0.4547
               Denominator:  1386.077  DF:   60   Prob>F:    0.5027


Dependent Variable: MEANLNG
Test: F_2VS3   Numerator:   5449.1829  DF:    1   F value:   3.9314
               Denominator:  1386.077  DF:   60   Prob>F:    0.0520


Dependent Variable: MEANLNG
Test: F_2VS7   Numerator:      0.0423  DF:    1   F value:   0.0000
               Denominator:  1386.077  DF:   60   Prob>F:    0.9956


Dependent Variable: MEANLNG
Test: F_2VS8   Numerator:   7818.7443  DF:    1   F value:   5.6409
               Denominator:  1386.077  DF:   60   Prob>F:    0.0208


Dependent Variable: MEANLNG
Test: F_2VS9   Numerator:   9445.7674  DF:    1   F value:   6.8147
```

```
                Denominator:  1386.077  DF:   60   Prob>F:    0.0114


Dependent Variable: MEANLNG
Test: F_3VS7    Numerator:   5479.5767  DF:    1   F value:   3.9533
                Denominator:  1386.077  DF:   60   Prob>F:    0.0513


Dependent Variable: MEANLNG
Test: F_3VS8    Numerator:    213.3084  DF:    1   F value:   0.1539
                Denominator:  1386.077  DF:   60   Prob>F:    0.6962


Dependent Variable: MEANLNG
Test: F_3VS9    Numerator:    546.1923  DF:    1   F value:   0.3941
                Denominator:  1386.077  DF:   60   Prob>F:    0.5326


Dependent Variable: MEANLNG
Test: F_7VS8    Numerator:   7855.1432  DF:    1   F value:   5.6672
                Denominator:  1386.077  DF:   60   Prob>F:    0.0205



Dependent Variable: MEANLNG
Test: F_7VS9    Numerator:   9485.7704  DF:    1   F value:   6.8436
                Denominator:  1386.077  DF:   60   Prob>F:    0.0112


Dependent Variable: MEANLNG
Test: F_8VS9    Numerator:     76.8370  DF:    1   F value:   0.0554
                Denominator:  1386.077  DF:   60   Prob>F:    0.8147
```

These analyses are summarized in the table below. Westar-Hanna differences marked with the same letter are not significantly different.

MCG	Westar-Hanna Difference		
7	120.35	A	
2	120.18	A	
3	59.91	A	B
8	47.98		B
9	40.83		B
1	20.33		B

The last two tests investigate whether there are significant differences in response to type of fungus, separately within Hanna and within Westar. We see that they are statistically significant for Westar, and almost reach significance for Hanna.

```
     G_Hanaeq: test   mu7=mu8=mu9=mu10=mu11=mu12;
     H_Westeq: test   mu13=mu14=mu15=mu16=mu17=mu18;
```

```
Dependent Variable: MEANLNG
Test: G_HANAEQ Numerator:   3223.5872  DF:    5   F value:   2.3257
               Denominator:  1386.077  DF:   60   Prob>F:    0.0536


Dependent Variable: MEANLNG
Test: H_WESTEQ Numerator:  17889.2114  DF:    5   F value:  12.9064
               Denominator:  1386.077  DF:   60   Prob>F:    0.0001
```

It makes sense to follow up with pairwise comparisons of the means with Westar, but first let's review what we've done so far, limiting the discussion to just the Hanna-Westar subset of the data. We've tested

- Overall difference among the 12 means
- Main effect for PLANT
- Main effect for MCG
- PLANT*MCG interaction
- 15 pairwise comparisons of the Hanna-Westar difference, following up the interaction
- One comparison of the 6 means for Hanna
- One comparison of the 6 means for Westar

That's 21 tests in all, and we really should do at least 15 more, testing for pairwise differences among the Westar means. Somehow, we should make this into a set of proper post-hoc tests, and correct for the fact that we've done a lot of them. But how?

Tukey tests are only good for pairwise comparisons, and a Bonferroni correction is very ill-advised, since these tests were not all planned before seeing the data. This pretty much leaves us with Scheffé or nothing. The earlier discussion of Scheffé tests was limited to testing single contrasts. Here, some of our involve testing collections of contrasts, so we need a little more generality.

General Scheffé Tests Assume a multifactor design. Create a combination independent variable whose values are all combinations of factor levels. All the tests we do will be tests for collections consisting of one or more contrasts of the cell means.

Start with an *initial* test, an F-test for s contrasts. A Scheffé follow-up test will be a test for d contrasts, *not* necessarily a subset of the contrasts of the initial test. The follow-up test must obey these rules:

- $d < s$
- If all s contrasts of the initial test are zero in the population, then all d contrasts of the follow-up test must be zero in the population. In other words, the null hypothesis of the follow-up test must be implied by the null hypothesis of the initial test.

Next, compute the ordinary one-at-a-time F statistic for the follow-up test (it will have d and n-p degrees of freedom). Then, use a calculator to compute

$$F_{sch} = \frac{d}{s} F, \tag{4.2}$$

and if F_{sch} is bigger than the critical value of F for the initial test, the Scheffé follow-up is significant.

Actually, Formula (4.2) is more general. It applies to testing linear combinations of regression coefficients in a multiple regression setting. The initial test is a test of s linear constraints on the regression coefficients, and the follow-up test is a test of d linear constraints, where $d < s$ and the linear constraints of the initial test imply the linear constraints of the follow-up test. This is very nice because it allows, for example, Scheffé follow-ups to a significant analysis of covariance.

Before applying Scheffé follow-ups to the greenhouse data, a few comments are in order.

- The term "linear constraints" sounds imposing, but a linear constraint is just a statement that some linear combination equals a constant. Almost always, the constant is zero. So for example, saying that a contrast of cell means is equal to zero is the same as specifying a linear constraint on the betas of a multiple regression model (with cell means coding).

° If you're testing 6 independent variables controlling for some other set of independent variables, the null hypothesis says that 6 regression coefficients are equal to zero. That's six linear constraints on the regression coefficients.

° In the initial one-way ANOVA setting where we were testing single contrasts of *p* cell means, the Scheffe F statistic was defined by $F_{sch} = F/(p-1)$. This was a special case of formula (4.2). The initial test for equality of *p* means involved *p-1* contrasts, so $s = p-1$. The followup tests were all for single contrasts, so $d=1$.

° As in the case of testing single contrasts in a one-way design, it is impossible for a followup to be significant if the initial test is not. And if the initial test is significant, there is always something to find in the family of Scheffé follow-ups.

° Suppose we have a follow-up test for *d* linear constraints, and it's not significant. Then *every* follow-up test whose null hypothesis is implied by those constraints will also be non-significant. To use the metaphor of data fishing, once you've looked for fish in a particular region of the lake and determined that there's nothing there, further detailed exploration in that region is a waste of time.

Formula (4.2) is very simple to apply. There are only two potential complications, and they are related to one another.

° First, you have to know what significance test you are following up. For example, if your initial test is the test for equality of *all* cell means, then the test for a given main effect could be carried out as a Scheffé followup, and a pairwise comparison of marginal means would be another followup to the same initial test. Or, you could start with the test for the main effect. Then, the pairwise comparison of marginal means would be a follow-up to the one-at-a-time test for the main effect. You could do it either way, and the conclusions might differ. Where you start is a matter of data-analytic philosophy. But starting with the standard tests for main effects and interactions is more traditional.

° The second potential complication is that you really have to be sure that the null hypothesis of the initial test implies the null hypothesis of the follow-up test. In terms of `proc reg` syntax, it means that the `test` statement of the initial test implies the `test` statements of all the follow-up tests. Sometimes this is easy to check, and sometimes it is tricky. To a large extent, how easy it is to check depends on what the initial test is.

a. If the initial test is a test for all cell means being equal (a one-way ANOVA on the combination variable), then it's easy, because if all the cell means are equal, then any possible contrast of the cell means equals zero. The proof is one line of High School algebra.

b. Similarly, suppose we are using a regression model with an intercept, and the initial test is for all the regression coefficients except β_0 simultaneously. This means that the null hypothesis of the initial test is $\beta_1 = ... = \beta_{p-1} = 0$, and therefore any linear combination of those quantities is zero. This means that you can test any subset of independent variables controlling for all the others as a proper Scheffé follow-up to the first test SAS prints.

c. If you're following up tests for main effects, then the standard test for any contrast of marginal means is a proper follow-up to the test for the main effect.

Beyond these principles, the logical connection between initial and follow-up tests really needs to be checked on a case-by-case basis. Often, the initial test can be expressed more than one way in the `test` statement of `proc reg`, and one of those statements will make things clear enough so you don't need to do any algebra. This is what I did with the significant Plant by Fungus interaction for the Hanna-Westar subset. When the interaction was written as

```
F_inter: test   mu13-mu7=mu14-mu8=mu15-mu9
              = mu16-mu10=mu17-mu11=mu18-mu12;
```

it was clear that all the pairwise comparisons of Westar-Hanna differences were implied.

```
F_1vs2:  test   mu13-mu7=mu14-mu8;
F_1vs3:  test   mu13-mu7=mu15-mu9;
F_1vs7:  test   mu13-mu7=mu16-mu10;
F_1vs8:  test   mu13-mu7=mu17-mu11;
F_1vs9:  test   mu13-mu7=mu18-mu12;
F_2vs3:  test   mu14-mu8=mu15-mu9;
F_2vs7:  test   mu14-mu8=mu16-mu10;
F_2vs8:  test   mu14-mu8=mu17-mu11;
F_2vs9:  test   mu14-mu8=mu18-mu12;
F_3vs7:  test   mu15-mu9=mu16-mu10;
F_3vs8:  test   mu15-mu9=mu17-mu11;
F_3vs9:  test   mu15-mu9=mu18-mu12;
F_7vs8:  test   mu16-mu10=mu17-mu11;
F_7vs9:  test   mu16-mu10=mu18-mu12;
F_8vs9:  test   mu17-mu11=mu18-mu12;
```

Sometimes it is easy to get this wrong. Just note that SAS will do all pairwise comparisons of marginal means (in the `means` statement of `proc glm`) as Scheffé follow-ups, but don't trust it unless the sample sizes are equal. Do it yourself. This warning applies up to SAS version 6.10. Is it a real error, or was it done deliberately to minimize calls to technical support? It's impossible to tell.

Now let's proceed, limiting the analysis to the Hanna-Westar subset. Just for fun, we'll start in two places. Our initial test will be either the test for equality of all 12 cell means, or the test for the Plant by Fungus interaction. Thus, we need two critical values.

```
proc iml; /* Critical values for Scheffe tests */
     interac = finv(.95,5,60) ; print interac;
     oneway = finv(.95,11,60); print oneway;

                                          INTERAC
                                        2.3682702

                                           ONEWAY
                                        1.9522119
```

Initial Test is for Difference Among 12 Cell Means

Let's start by treating the tests for main effects and the interaction as follow-ups to the significant ANOVA on the combination variable (F = 12.43; df=11,71; p < .0001). The table below is based on numbers displayed earlier.

Effect	One-at-a-time F	$F_{sch} = \frac{d}{s} F$	d	Significant with Scheffé?
PLANT	60.52	5.50	1	Yes
MCG	11.36	5.16	5	Yes
PLANT*MCG	3.87	1.75	5	No
All Hanna Equal?	2.33	1.06	5	No
All Westar Equal?	12.91	5.87	5	Yes

The main effect for Plant is still significant; it means that Westar is more vulnerable than Hanna. The main effect for Fungus (MCG) is significant, but as mentioned earlier, it should not be interpreted.

The interesting Plant by MCG interaction is no longer significant as a Scheffe test. This means that all the pairwise comparisons among Westar-Hanna differences will also be non-significant, as Scheffe follow-ups to the oneway ANOVA on the combination variable. There are no fish in that part of the lake. Just to check, the biggest Westar-Hanna difference was 120.35 for MCG 7, and the smallest was 20.33 for MCG 1. Comparing these two differences yielded a one-at-a-time F of 10.83. But d=1 here and s=11, so that F_{sch}=.98. This falls short of the 1.95 required for significance, and as expected, none of the proper follow-ups to a non-significant follow-up are significant.

Pairwise comparisons of the Westar means are of interest, and the easiest way to get them is to ask `proc glm` for all pairwise comparisons of cell means.

```
proc glm data=hanstar;
     class combo;
     model meanlng = combo;
     means combo / scheffe;
```

```
                   Scheffe's test for variable: MEANLNG

          NOTE: This test controls the type I experimentwise error rate but
                generally has a higher type II error rate than REGWF for all
                pairwise comparisons

                      Alpha= 0.05  df= 60  MSE= 1386.077
                         Critical Value of F= 1.95221
                     Minimum Significant Difference= 99.608

          Means with the same letter are not significantly different.

               Scheffe Grouping              Mean       N  COMBO

                             A             187.48       6  14
                             A
                             A             173.97       6  16
                             A
                     B       A             154.10       6  15
                     B       A
                     B       A    C         95.82       6  17
                     B       A    C
                     B       A    C         94.19       6  9
                     B            C
                     B            C         67.30       6  8
                     B            C
                     B            C         66.50       6  18
                     B            C
                     B            C         65.91       6  13
                                  C
                                  C         53.62       6  10
                                  C
                                  C         47.84       6  11
                                  C
                                  C         45.58       6  7
                                  C
                                  C         25.67       6  12
```

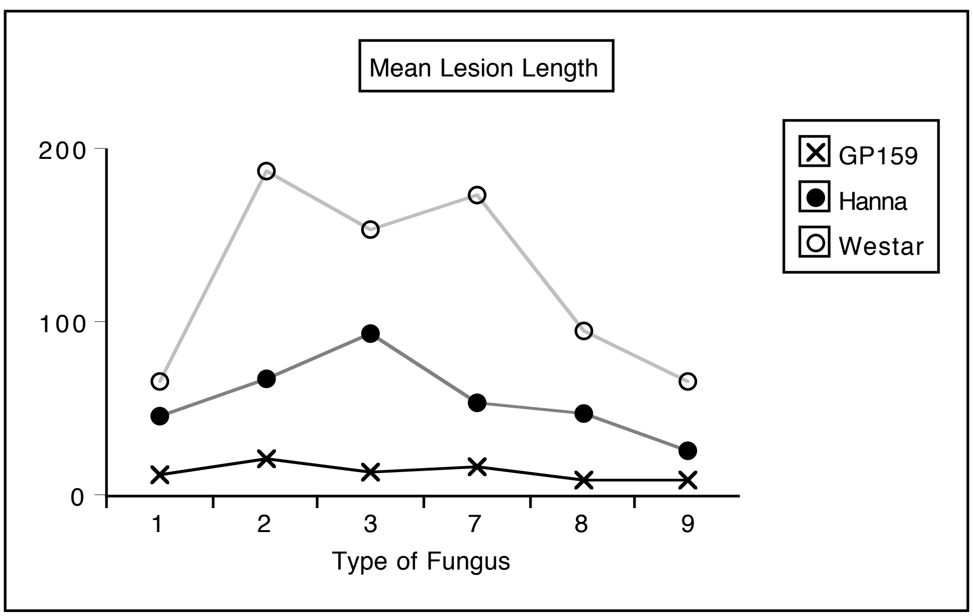
On Westar, Fungus types 2, 3 and 7 grow significantly faster than types 1 and 9, while type 8 is not significantly different from either group. As expected, there are no significant differences among Fungus types for Hanna.

Starting with the Interaction

Logically, a test for interaction can be a follow-up test, but almost no one ever does this in practice. It's much more traditional to start with a one-at a time test for interaction and then, if you're very sophisticated, do Scheffe follow-ups to that initial test. Now s = 5 and the critical value is `2.3682702`.

Again, the biggest Westar-Hanna difference was 120.35 for MCG 7, and the smallest was 20.33 for MCG 1. Comparing these two differences yielded a one-at-a-time F of 10.83. This yields $F_{sch} = \frac{d}{s} F = \frac{1}{5} * 10.83 = 2.16$. But this falls short of the critical value of 2.37, so none of the pairwise comparisons of Westar-Hanna differences reaches significance as a Scheffe follow-up -- even though they look very promising.

As a mathematical certainty, there *is* a single-contrast Scheffe follow-up to the interaction that is significant, but I am still looking for it. The next place I will look is: pairwise comparisons of the differences of line-segment slopes from the interaction plot.



Interactions as Products of Independent Variables

Categorical by Quantitative

An interaction between a quantitative variable and a categorical variable means that differences in E[Y] between categories depend on the value of the quantitative variable, or (equivalently) that the slope of the lines relating x to E[Y] are different, depending on category membership. Such an interaction is represented by **products** of the quantitative variable and the dummy variables for the categorical variable.

For example, consider the metric cars data (mcars.dat). It has length, weight, origin and fuel efficiency in kilometers per litre, for a sample of cars. The three origins are US, Japanese and Other. Presumably these refer to the location of the head office, not to where the car was manufactured.

Let's use indicator dummy variable coding for origin, with an intercept. In an Analysis of Covariance (ANCOVA), we'd test country of origin controlling, say, for weight. Letting x represent weight and c1 and c2 the dummy variables for country of origin, the model would be

$$E[Y] = b_0 + b_1 x + b_2 c_1 + b_3 c_2.$$

This model assumes no interaction between country and weight. The following model includes product terms for the interaction, and would allow you to test it.

$$E[Y] = \beta_0 + \beta_1 x + \beta_2 c_1 + \beta_3 c_2 + \beta_4 c_1 x + \beta_5 c_2 x$$

Country	c1	c2	Expected KPL (let x = weight)
U. S.	1	0	$(\beta_0 + \beta_2) + (\beta_1+\beta_4)x$
Japan	0	0	$\beta_0 + \beta_1 x$
European	0	1	$(\beta_0 + \beta_3) + (\beta_1+\beta_5)x$

It's clear that the slopes are parallel if and only if $\beta_4=\beta_5=0$, and that in this case the relationship of fuel efficiency to country would not depend on weight of the car.

As the program below shows, interaction terms are created by literally multiplying independent variables, and using products as additional independent variables in the regression equation.

```
/********************* mcars.sas *************************/
options linesize=79 pagesize=100 noovp formdlim='-';
title 'Metric Cars Data: Dummy Vars and Interactions';

proc format; /* Used to label values of the categorical variables */
     value carfmt    1 = 'US'
                     2 = 'Japanese'
                     3 = 'European' ;
data auto;
     infile 'mcars.dat';
     input id country kpl weight length;
/* Indicator dummy vars: Ref category is Japanese */
     if country = 1 then c1=1;  else c1=0;
     if country = 3 then c2=1;  else c2=0;
     /* Interaction Terms */
     cw1 = c1*weight; cw2 = c2*weight;
     label country = 'Country of Origin'
           kpl = 'Kilometers per Litre';
     format country carfmt.;

proc means;
     class country;
     var weight kpl;

proc glm;
     title 'One-way ANOVA';
     class country;
     model kpl = country;
     means country / tukey;

proc reg;
     title 'ANCOVA';
     model kpl = weight c1 c2;
     country: test c1 = c2 = 0;

proc reg;
     title 'Test parallel slopes (Interaction)';
     model kpl = weight c1 c2 cw1 cw2;
     interac: test cw1 = cw2 = 0;
     useuro:  test cw1=cw2;
     country: test c1 = c2 = 0;
     eqreg:   test c1=c2=cw1=cw2=0;

proc iml; /* Critical value for Scheffe tests */
     critval = finv(.95,4,94) ; print critval;
```

```
/* Could do most of it with proc glm: ANCOVA, then test interaction */

proc glm;
     class country;
     model kpl = weight country;
     lsmeans country;

proc glm;
     class country;
     model kpl = weight country weight*country;
```

Let's take a look at the output. First, proc means indicates that the US cars get lower gas mileage, and that weight is a potential confounding variable.

```
 COUNTRY  N Obs  Variable  Label                      N          Mean
 -------------------------------------------------------------------
 US          73  WEIGHT                              73       1540.23
                 KPL       Kilometers per Litre      73     8.1583562

 Japanese    13  WEIGHT                              13       1060.27
                 KPL       Kilometers per Litre      13     9.8215385

 European    14  WEIGHT                              14       1080.32
                 KPL       Kilometers per Litre      14    11.1600000
 -------------------------------------------------------------------

 COUNTRY  N Obs  Variable  Label                        Std Dev       Minimum
 ---------------------------------------------------------------------------
 US          73  WEIGHT                             327.7785402   949.5000000
                 KPL       Kilometers per Litre       1.9760813     5.0400000

 Japanese    13  WEIGHT                             104.8370989   891.0000000
                 KPL       Kilometers per Litre       2.3976719     7.5600000

 European    14  WEIGHT                             240.9106607   823.5000000
                 KPL       Kilometers per Litre       4.2440764     5.8800000
 ---------------------------------------------------------------------------

 COUNTRY  N Obs  Variable  Label                           Maximum
 -------------------------------------------------------------
 US          73  WEIGHT                                2178.00
                 KPL       Kilometers per Litre     12.6000000

 Japanese    13  WEIGHT                                1237.50
                 KPL       Kilometers per Litre     14.7000000

 European    14  WEIGHT                                1539.00
                 KPL       Kilometers per Litre     17.2200000
 -------------------------------------------------------------
```

The one-way ANOVA indicates that fuel efficiency is significantly related to country of origin; country explains 17% of the variation in fuel efficiency.

```
                         General Linear Models Procedure

Dependent Variable: KPL   Kilometers per Litre
                                   Sum of            Mean
Source                  DF        Squares          Square  F Value    Pr > F

Model                    2     121.59232403     60.79616201    10.09    0.0001
Error                   97     584.29697197      6.02368012
Corrected Total         99     705.88929600

                  R-Square             C.V.        Root MSE           KPL Mean
                  0.172254         27.90648       2.4543187          8.7948000
```

The Tukey follow-ups are not shown, but they indicate that only the US-European difference is significant. Maybe the US cars are less efficient because they are big and heavy. So let's do the same test, controlling for weight of car. Here's the SAS code. Note this is a standard Analysis of Covariance, and we're *assuming* no interaction.

```
proc reg;
     title 'ANCOVA';
     model kpl = weight c1 c2;
     country: test c1 = c2 = 0;
```

```
Dependent Variable: KPL        Kilometers per Litre

                              Analysis of Variance

                               Sum of         Mean
       Source          DF     Squares       Square      F Value       Prob>F

       Model            3   436.21151    145.40384       51.761       0.0001
       Error           96   269.67779      2.80914
       C Total         99   705.88930

           Root MSE       1.67605     R-square       0.6180
           Dep Mean       8.79480     Adj R-sq       0.6060
           C.V.          19.05728
```

```
                              Parameter Estimates

                       Parameter      Standard    T for H0:
     Variable  DF       Estimate         Error   Parameter=0    Prob > |T|

     INTERCEP   1      16.226336    0.76312281        21.263        0.0001
     WEIGHT     1      -0.006041    0.00057080       -10.583        0.0001
     C1         1       1.236147    0.57412989         2.153        0.0338
     C2         1       1.459591    0.64565633         2.261        0.0260

--------------------------------------------------------------------------------

Dependent Variable: KPL
Test: COUNTRY  Numerator:      8.6168  DF:    2   F value:   3.0674
               Denominator:  2.809144  DF:   96   Prob>F:    0.0511
```

First notice that by including weight, we're now explaining 61% of the variation, while before we explained just 17%. Also, while the effect for country was comfortably significant before we controlled for weight, now it narrowly fails to reach the traditional criterion (p = `0.0511`). But to really appreciate these results, we need to make a table.

Country	c1	c2	$E[Y] = \beta_0 + \beta_1 x + \beta_2 c_1 + \beta_3 c_2$
U. S.	1	0	$(\beta_0 + \beta_2) + \beta_1 x$
Japan	0	0	$\beta_0 \quad + \beta_1 x$
European	0	1	$(\beta_0 + \beta_3) + \beta_1 x$


```
                              Parameter Estimates

                       Parameter      Standard    T for H0:
     Variable  DF       Estimate         Error   Parameter=0    Prob > |T|

     INTERCEP   1      16.226336    0.76312281        21.263        0.0001
     WEIGHT     1      -0.006041    0.00057080       -10.583        0.0001
     C1         1       1.236147    0.57412989         2.153        0.0338
     C2         1       1.459591    0.64565633         2.261        0.0260
```


Observe that both b_2 and b_3 are positive -- and significant. Before we controlled for weight, Japanese gas mileage was a little better than US, though not significantly so. Now, because b_2 estimates β_2, and β_2 is the population difference between U.S. and Japanese mileage (for any fixed weight), a positive value of b_2 means that once you control for weight, the U.S. cars are getting better gas mileage than the Japanese -- significantly better, too, if you believe the t-test and not the F-test.

The *direction* of the results has changed because we controlled for weight. This can happen.

Also, may seem strange that the tests for β_2 and β_3 are each significant individually, but the simultaneous test for both of them is not. But this the simultaneous test implicitly includes a comparison between U.S. and European cars, and they are very close, once you control for weight.

The best way to summarize these results would be to calculate Y-hat for each country of origin, with weight set equal to its mean value in the sample. Instead of doing that, though, let's first test the interaction, which this analysis is *assuming* to be absent.

```
proc reg;
     title 'Test parallel slopes (Interaction)';
     model kpl = weight c1 c2 cw1 cw2;
        interac: test cw1 = cw2 = 0;
        useuro:  test cw1=cw2;
        country: test c1 = c2 = 0;
        eqreg:   test c1=c2=cw1=cw2=0;
```

```
   Dependent Variable: KPL        Kilometers per Litre

                                    Sum of         Mean
          Source          DF       Squares       Square      F Value       Prob>F

          Model            5     489.27223     97.85445       42.463       0.0001
          Error           94     216.61706      2.30444
          C Total         99     705.88930

              Root MSE       1.51804     R-square       0.6931
              Dep Mean       8.79480     Adj R-sq       0.6768
              C.V.          17.26062

                                  Parameter Estimates

                         Parameter      Standard    T for H0:
        Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

        INTERCEP   1     29.194817    4.45188417         6.558        0.0001
        WEIGHT     1     -0.018272    0.00418000        -4.371        0.0001
        C1         1    -12.973668    4.53404398        -2.861        0.0052
        C2         1     -4.891978    4.85268101        -1.008        0.3160
        CW1        1      0.013037    0.00421549         3.093        0.0026
        CW2        1      0.006106    0.00453064         1.348        0.1810
```

```
-------------------------------------------------------------------------------------

Dependent Variable: KPL
Test: INTERAC   Numerator:     26.5304  DF:    2   F value:  11.5127
                Denominator:  2.304437  DF:   94   Prob>F:    0.0001

Dependent Variable: KPL
Test: USEURO    Numerator:     33.0228  DF:    1   F value:  14.3301
                Denominator:  2.304437  DF:   94   Prob>F:    0.0003

Dependent Variable: KPL
Test: COUNTRY   Numerator:     24.4819  DF:    2   F value:  10.6238
                Denominator:  2.304437  DF:   94   Prob>F:    0.0001

Dependent Variable: KPL
Test: EQREG     Numerator:     17.5736  DF:    4   F value:   7.6260
                Denominator:  2.304437  DF:   94   Prob>F:    0.0001
```

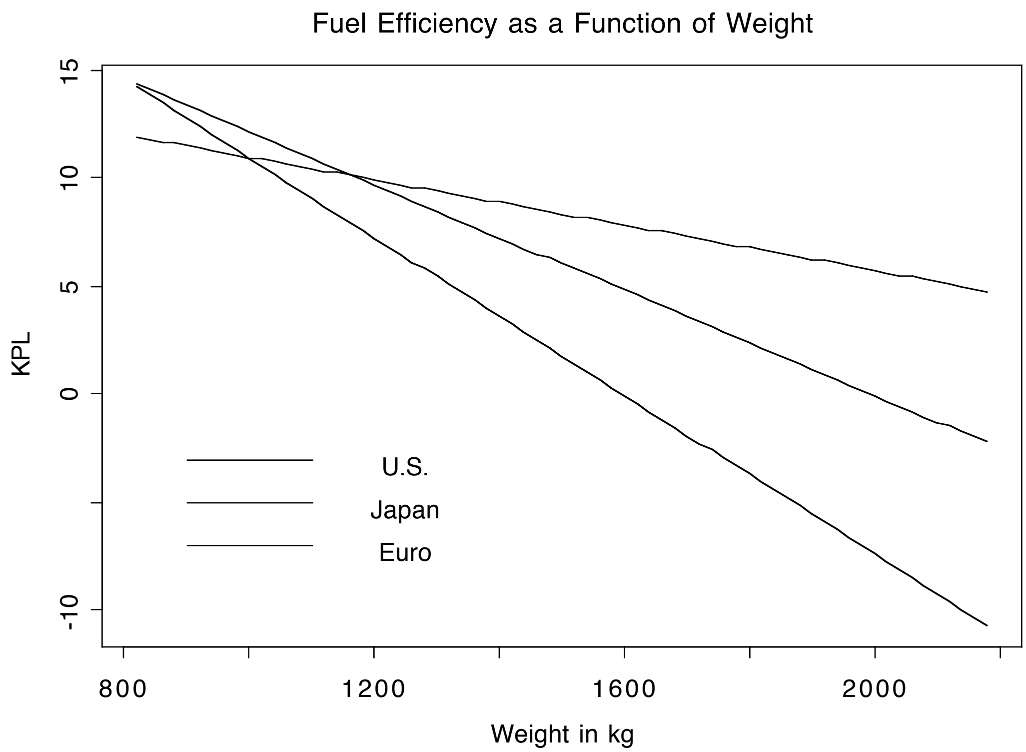
Now the coefficients for the dummy variables are both negative, and the coefficients for the interaction terms are positive. To see what's going on, we need a table *and* a picture -- of $\hat{Y}$.

$$\hat{Y} = b_0 + b_1x + b_2c_1 + b_3c_2 + b_4c_1x + b_5c_2x$$

$$= 29.194817 - 0.018272x - 12.973668c_1 - 4.891978c_2 + 0.013037c_1x + 0.006106c_2x$$

Country	c1	c2	Predicted KPL (let x = weight)
U. S.	1	0	$(b_0 + b_2) + (b_1+b_4)x$ = 16.22 - 0.005235 x
Japan	0	0	$b_0 + b_1 x$ = 29.19 - 0.018272 x
European	0	1	$(b_0 + b_3) + (b_1+b_5)x$ = 24.30 - 0.012166 x

From the proc means output, we find that the lightest car was 823.5kg, while the heaviest was 2178kg. So we will let the graph range from 820 to 2180.



When there were no interaction terms, b2 and b3 represented a main effect for country. What do they represent now?

From the picture, it is clear that the most interesting thing is that the slope of the line relating weight to fuel efficiency is least steep for the U.S. Is it significant? 0.05/3 = 0.0167.

Repeating earlier material, ...

```
                              Parameter Estimates

                       Parameter      Standard    T for H0:
Variable  DF            Estimate         Error   Parameter=0    Prob > |T|

INTERCEP   1           29.194817    4.45188417         6.558        0.0001
WEIGHT     1           -0.018272    0.00418000        -4.371        0.0001
C1         1          -12.973668    4.53404398        -2.861        0.0052
C2         1           -4.891978    4.85268101        -1.008        0.3160
CW1        1            0.013037    0.00421549         3.093        0.0026
CW2        1            0.006106    0.00453064         1.348        0.1810


useuro:  test cw1=cw2;
```

```
Dependent Variable: KPL
Test: USEURO   Numerator:     33.0228  DF:    1   F value:  14.3301
               Denominator:  2.304437  DF:   94   Prob>F:    0.0003
```

The conclusion is that with a Bonferroni correction, the slope is less (less steep) for US than for either Japanese or European, but Japanese and European are not significantly different from each other.

Another interesting follow-up would be to use Scheffé tests to compare the heights of the regression lines at many values of weight; infinitely many comparisons would be protected simultaneously. This is not a proper follow-up to the interaction. What is the initial test?

Quantitative by Quantitative

An interaction of two quantitative variables is literally represented by their product. For example, consider the model

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Hold x_2 fixed at some particular value, and re-arrange the terms. This yields

$$E[Y] = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1.$$

so that there is a linear relationship between x_1 and E[Y], with both the slope and the intercept depending on the value of x_2. Similarly, for a fixed value of x_1,

$$E[Y] = (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) x_2,$$

and the (linear) relationship of x_2 to E[Y] depends on the value of x_1. We always have this kind of symmetry.

Three-way interactions are represented by 3-way products, etc. Its interpretation would be "the 2-way interaction depends ..."

Product terms represent interactions ONLY when all the variables involved and all lower order interactions involving those variables are also included in the model!

Categorical by Categorical

It is no surprise that interactions between categorical independent variables are represented by products. If A and B are categorical variables, IVs representing the A by B interaction are obtained by multiplying each dummy variable for A by each dummy variable for B. If there is a third IV cleverly named C and you want the 3-way interaction, multiply each of the dummy variables for C by each of the products representing the A by B interaction. This rule extends to interactions of any order.

Up till now, we have represented categorical independent variables with indicator dummy variables, coded 0 or 1. If interactions between categorical IVs are to be represented, it is much better to use "effect coding," so that the regression coefficients for the dummy variables correspond to main effects. (In a 2-way design, products of indicator dummy variables still correspond to interaction terms, but if an interaction is present, the interpretation of the coefficients for the indicator dummy variables is not what you might guess.)

Effect coding. There is an intercept. As usual, a categorical independent variable with k categories is represented by k-1 dummy variables. The rule is

Dummy var 1: First value of the IV gets a 1, last gets a minus 1, all others get zero.
Dummy var 2: Second value of the IV gets a 1, last gets a minus 1, all others get zero.

. . .

Dummy var k-1: k-1st value of the IV gets a 1, last gets a minus 1, all others get zero.

Here is a table showing effect coding for Plant from the Greenhouse data.

Country	p1	p2	`E[Y] =` $\beta_0 + \beta_1 p_1 + \beta_2 p_2$
GP159	1	0	$\mu_1 = \beta_0 + \beta_1$
Hanna	0	1	$\mu_2 = \beta_0 + \beta_2$
Westar	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

It is clear that $\mu_1 = \mu_2 = \mu_3$ if and only if $\beta_1=\beta_2=0$, so it's a valid dummy variable coding scheme even though it looks strange.

Country	p1	p2	$E[Y] = \beta_0 + \beta_1 p_1 + \beta_2 p_2$
GP159	1	0	$\mu_1 = \beta_0 + \beta_1$
Hanna	0	1	$\mu_2 = \beta_0 + \beta_2$
Westar	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

Effect coding has these properties, which extend to any number of categories.

- $\mu_1 = \mu_2 = \mu_3$ if and only if $\beta_1=\beta_2=0$.
- The average population mean (grand mean) is $(\mu_1+\mu_2+\mu_3)/3 = \beta_0$.
- β_1, β_2 and $-(\beta_1+\beta_2)$ are deviations from the grand mean.

The real advantage of effect coding is that the dummy variables behave nicely when multiplied together, so that main effects correspond to collections of dummy variables, and interactions correspond to their products -- in a simple way. This is illustrated for Plant by MCG analysis, using the full greenhouse data set).

```
data nasty;
     set yucky;
     /* Two dummy variables for plant */
        if plant=. then p1=.;
        else if plant=1 then p1=1;
        else if plant=3 then p1=-1;
        else p1=0;
     if plant=. then p2=.;
        else if plant=2 then p2=1;
        else if plant=3 then p2=-1;
```

```
      else p2=0;
   /* Five dummy variables for mcg */
   if mcg=. then f1=.;
      else if mcg=1 then f1=1;
      else if mcg=9 then f1=-1;
      else f1=0;
   if mcg=. then f2=.;
      else if mcg=2 then f2=1;
      else if mcg=9 then f2=-1;
      else f2=0;
   if mcg=. then f3=.;
      else if mcg=3 then f3=1;
      else if mcg=9 then f3=-1;
      else f3=0;
   if mcg=. then f4=.;
      else if mcg=7 then f4=1;
      else if mcg=9 then f4=-1;
      else f4=0;
   if mcg=. then f5=.;
      else if mcg=8 then f5=1;
      else if mcg=9 then f5=-1;
      else f5=0;
   /* Product terms for the interaction */
      p1f1 = p1*f1; p1f2=p1*f2 ; p1f3=p1*f3 ; p1f4=p1*f4; p1f5=p1*f5;
      p2f1 = p2*f1; p2f2=p2*f2 ; p2f3=p2*f3 ; p2f4=p2*f4; p2f5=p2*f5;

proc reg;
   model meanlng = p1 -- p2f5;
   plant:  test p1=p2=0;
   mcg:    test f1=f2=f3=f4=f5=0;
   p_by_f: test p1f1=p1f2=p1f3=p1f4=p1f5=p2f1=p2f2=p2f3=p2f4=p2f5 = 0;
```

Here is the output from the test statement. For comparison, it is followed by `proc glm` output from `model meanlng = plant|mcg`.

```
Dependent Variable: MEANLNG
Test: PLANT    Numerator: 110847.5637  DF:    2   F value: 113.9032
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: MCG      Numerator:  11748.0529  DF:    5   F value:  12.0719
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001


Dependent Variable: MEANLNG
Test: P_BY_F   Numerator:   4758.1481  DF:   10   F value:   4.8893
               Denominator:  973.1736  DF:   90   Prob>F:    0.0001

-------------------------------------------------------------------------------


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

PLANT                    2     221695.12747    110847.56373   113.90    0.0001
MCG                      5      58740.26456     11748.05291    12.07    0.0001
PLANT*MCG               10      47581.48147      4758.14815     4.89    0.0001
```

It worked.

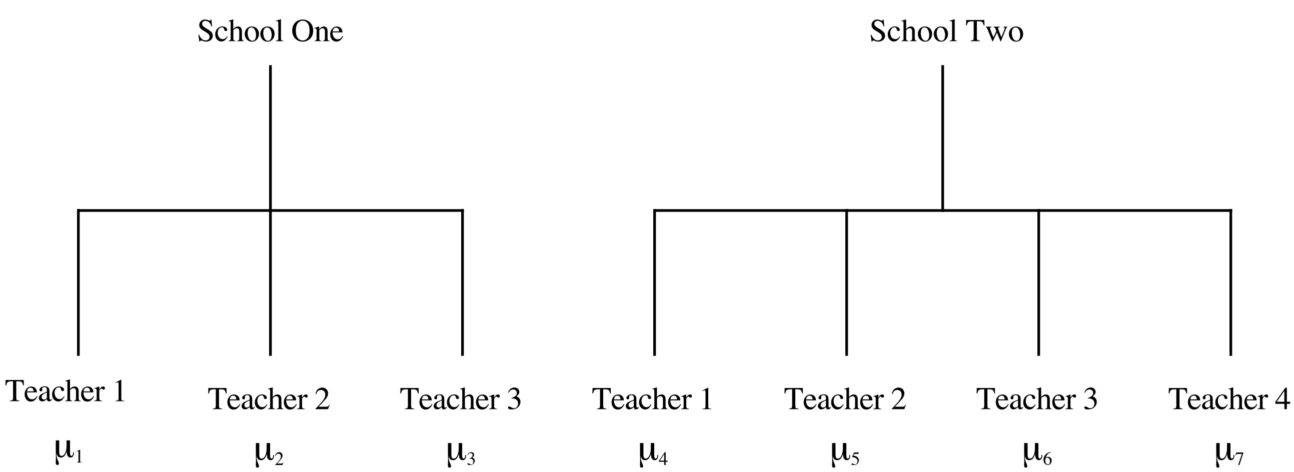
Effect coding works as expected in conjunction with quantitative independent variables. In particular, products of quantitative and indicator variables still represent interactions. In fact, the big advantage of effect coding is that you can use it to test categorical independent variables, and interactions between categorical independent variables -- in a bigger multiple regression context.

Nested and Random Effect models

Nested Designs

Suppose a chain of commercial business colleges is teaching a software certification course. After 6 weeks of instruction, students take a certification exam and receive a score ranging from zero to 100. The owners of the business school chain want to see whether performance is related to which school students attend, or which instructor they have -- or both. They compare two schools; one of the schools has three instructors teaching the course, and the other school has 4 instructors teaching the course. A teacher only works in one school.

There are two independent variables, school and teacher. But it's not a factorial design, because ``Teacher 1" does not mean the same thing in School 1 and School 2; it's a different person. This is called a **nested** design. By the way, it's also **unbalanced**, because there are different numbers of teachers withing each school. We say that *teacher is nested within school*. The diagram below shows what is going on, and give a clue about how to conduct the analysis.



To compare schools, we want to test $\frac{1}{3}(\mu_1+\mu_2+\mu_3) = \frac{1}{4}(\mu_4+\mu_5+\mu_6+\mu_7)$.

To compare instructors within schools, we want to test $\mu_1=\mu_2=\mu_3$ and $\mu_4=\mu_5=\mu_6=\mu_7$ simultaneously.

The first test involves one contrast of μ_1 through μ_7; the second test involves five contrasts. There really is nothing to it.

You can do it with `proc reg` and cell means coding, or you can take advantage of `proc glm`'s syntax for nested models.

```
proc glm;
     class school teacher;
     model score = school teacher(school);
```

The notation `teacher(school)` should be read ``teacher within school."

- It's easy to extend this to more than one level of nesting. You could have climate zones, lakes within climate zones, fishing boats within lakes, ...

- There is no problem with combining nested and factorial structures. You just have to keep track of what's nested within what. Factors that are not nested are sometimes called ``crossed."

Random Effect Models The preceding discussion (and indeed, the entire course to this point) has been limited to ``fixed effects" models. In a **random effects** model, *the values of the categorical independent variables represent a random sample from some population of values*. For example, suppose the business school had 200 branches, and just selected 2 of them at random for the investigation. Also, maybe each school has a lot of teachers, and we randomly sampled teachers within schools. Then, teachers within schools would be a random effects factor too.

It's quite possible to have random effect factors and fixed effect factors in the same design; such designs are called ``mixed." SAS `proc mixed` is built around this, but it does a lot of other things too.

Nested models are often viewed as random effects models, but there is no necessary connection between the two concepts. It depends on how the study was conducted. Were the two schools randomly selected from some population of schools, or did someone just pick those two (maybe because there are just two schools)?

Of course lots of the time, nothing is randomly selected -- but people use random effects models anyway. Why pretend? Well, sometimes they are thinking that in a better world, lakes *would* have been randomly selected. Or sometimes, the scientists are thinking that they really would like to generalize to the entire population of lakes, and therefore should use statistical tools that support such generalization -- even if there was no random sampling. (By the way, no statistical method can compensate for a biased sample.) Or sometimes it's just a tradition in certain sub-areas of research, and everybody expects to see random effects models.

In the traditional analysis of models with random or mixed effects and a normal assumption, F-tests are often possible, but they don't always use Mean Squared Error in the denominator of the F statistic. Often, it's the Mean Square for some interaction term or other. The choice of what error term to use is relatively mechanical for balanced models with equal sample sizes, but even then, sometimes (especially when it's a mixed model) a valid F-test for an effect of interest just doesn't exist.

When the design is unbalanced or has unequal sample sizes, a valid F-test rarely exists. It's a real pain. Sometimes, you can find an error term that produces a valid F-test *assuming* that some interaction (or maybe more than one interaction) is absent. Usually, you can't test for that interaction either. But people do it anyway and hope for the best.

SAS `proc mixed` goes a long way toward solving these problems. It's a great piece of software, based on recent, state-of the-art research as well as more venerable stuff. But we're running out of time. Goodbye, `proc mixed`. Goodbye, random effects.

Choosing Sample Size

The purpose of this section is to describe three related methods for choosing sample size before data are collected -- the classical power method, the sample variation method and the population variation method. The classical power method applies to almost any statistical test. After presenting general principles, the discussion zooms in on the important special case of factorial analysis of variance with no covariates. The sample variation method and the population variation methods are limited to multiple linear regression, including the analysis of variance and covariance. Throughout, it will be assumed that the person designing the study is a scientist who will only be allowed to discuss results if a null hypothesis is rejected at some conventional significance level such as $\alpha = 0.05$ or $\alpha = 0.01$. Thus, it is vitally important that the study be designed so that scientifically interesting effects are likely to be be detected as statistically significant.

The classical power method. The term "null hypothesis" has mostly been avoided until now, but it's much easier to talk about the classical power method if we're allowed to use it. Most statistical tests are based on comparing a full model to a reduced model. Under the reduced model, the values of population parameters are constrained in some way. For example, in a one-way ANOVA comparing three treatments, the parameters are μ_1, μ_2, μ_3 and σ^2. The reduced model says that $\mu_1=\mu_2=\mu_3$. This is a *constraint* on the parameter values. The **null hypothesis** (symbolized H_0) is a statement of how the parameters are constrained under the reduced model. When a test of a null hypothesis yields a small p-value, it means that the data are quite unlikely if the null hypothesis is true. We then reject the null hypothesis -- that is, we conclude it's not true, and therefore that some effect of interest is present in the population.

The following definition applies to hypothesis tests in general, not just those associated with common multiple regression. Assume that data are drawn from some population with parameter θ -- that's the Greek letter theta. Theta is typically a vector; for example, in simple linear regression with normal errors, $\theta = (\beta_0, \beta_1, \sigma^2)$.

The **power** of a statistical test is the probability of obtaining signific
the true parameter values. That is, it is a function of θ.

The **power** of a statistical test is the probability of obtaining significant results. Power is a function of the true parameter values. That is, it is a function of θ.

a) The common statistical tests have infinitely many power values.

b) If the null hypothesis is true, power cannot exceed å; in fact, this is the technical definition of α. Usually, $\alpha = 0.05$.

c) If the null hypothesis is false, more power is good.

d) For a good test, power $\rightarrow$ 1 (for fixed n) as the true parameter values get farther from those specified by the null hypothesis.

e) For a good test, power $\rightarrow$ 1 as $n \rightarrow \infty$ for any combination of fixed parameter values, provided the null hypothesis is false.

Classical power analysis is used to select a sample size n as follows. Choose an effect -- a particular combination of parameter values that makes the null hypothesis false. If possible, select the weakest effect that would still be scientifically important if it were present in the population. If the null hypothesis is false in this way, we would like to have a high probability of rejecting it and obtaining significance. Choose a sample size n, and calculate the probability of significance (that is, calculate power) for that sample size and that set of parameter values. Increase (or decrease) n, calculating power each time. Stop when the power is what you want. A common target value for power is 0.80. My guess is that it would be higher, except that, for common tests and effect sizes, the sample would have to be prohibitively large.

There are only two difficulties with carrying out a classical power analysis in practice; one is conceptual, the other technical. The conceptual problem is that scientists often have difficulty choosing a configuration of parameter values corresponding to an effect that is scientifically interesting. Maybe that's not too surprising, because scientists usually think in terms of data rather than in terms of statistical models. It could be different if the statistical models were serious scientific models of what the scientists are studying, but usually they're quite generic.

The technical problem is that sometimes -- especially for statistical methods other than those based on common multiple regression -- it can be difficult to calculate the probability of significance when the null hypothesis is false. This problem is not really serious; it can always be overcome with some effort and the right software. Once you move beyond multiple regression, SAS is not the right software.

Power for Factorial ANOVA. Considering this special case will provide a concrete example of the classical power method. It is also the most common example of power analysis.

The distributions commonly used for practical hypothesis testing (mainly the chi-square, t and F) are ones that hold when the null hypothesis is true. When the null hypothesis is false, these are no longer the distributions of the common test statistics; instead, they have probability distributions that migrate more into the rejection region (tail area, above the critical value) of the statistical test. The F distribution used for testing hypotheses in multiple regression is the central F distribution. If the null hypothesis is *false*, the F statistic has a non-central F distribution with parameters s, n-p and ϕ. The quantity ϕ is a kind of squared distance between the reduced model and the true model. It is called the **non–centrality parameter** of the non-central F distribution; $\phi \geq 0$, and $\phi = 0$ gives the usual central F distribution. The larger the non-centrality parameter, the greater the chance of significance -- that is, the greater the power.

The general formula for ϕ is best written in the notation of matrix algebra; it will not be given here. But the general idea, and some of its essential properties, are shown by the special case where we are comparing two treatment means (as in a two-sample t-test, or a simple regression with a binary independent variable). In this situation, the general formula for the non-centrality parameter of the non-central F distribution reduces to

$$\phi = \frac{(\mu_1 - \mu_2)^2}{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})} = \frac{\delta^2}{(\frac{1}{n_1} + \frac{1}{n_2})}, \qquad (4.3)$$

where $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$. Right away, it is possible to make some useful comments.

$$\phi = \frac{(\mu_1 - \mu_2)^2}{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})} = \frac{\delta^2}{(\frac{1}{n_1} + \frac{1}{n_2})}, \tag{4.3}$$

where $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$.

- The quantity δ is called **effect size**. It specifies how wrong the statement $\mu_1=\mu_2$ is, by expressing the absolute difference between μ_1 and μ_2 in units of the common within-cell standard deviation σ.

- For any statistical test, power is a function of the parameter values. Here, the non-centrality parameter (and hence, power) depends on the three parameters μ_1, μ_2 and σ^2 *only* through the effect size. This is quite wonderful; it does not always happen, even in the analysis of variance.

- The larger the effect size (that is, the more wrong the reduced model is -- in this metric), the larger the non-centrality parameter ϕ, and therefore the larger the probability of significance.

- If $\mu_1=\mu_2$, then δ=0, ϕ=0,the non-central F distribution becomes the usual central F distribution, and the probability of significance becomes exactly α=0.05.

- The size of the non-centrality parameter depends on another quantity involving *both* n_1 and n_2, not just the total sample size $n = n_1+n_2$.

This last point can be illuminated by a bit of algebra. Let

- $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$
- $n = n_1+n_2$
- $q = \frac{n_1}{n}$, the proportion of the sample allocated to Group One.

Then expression (4.3) can be re-written

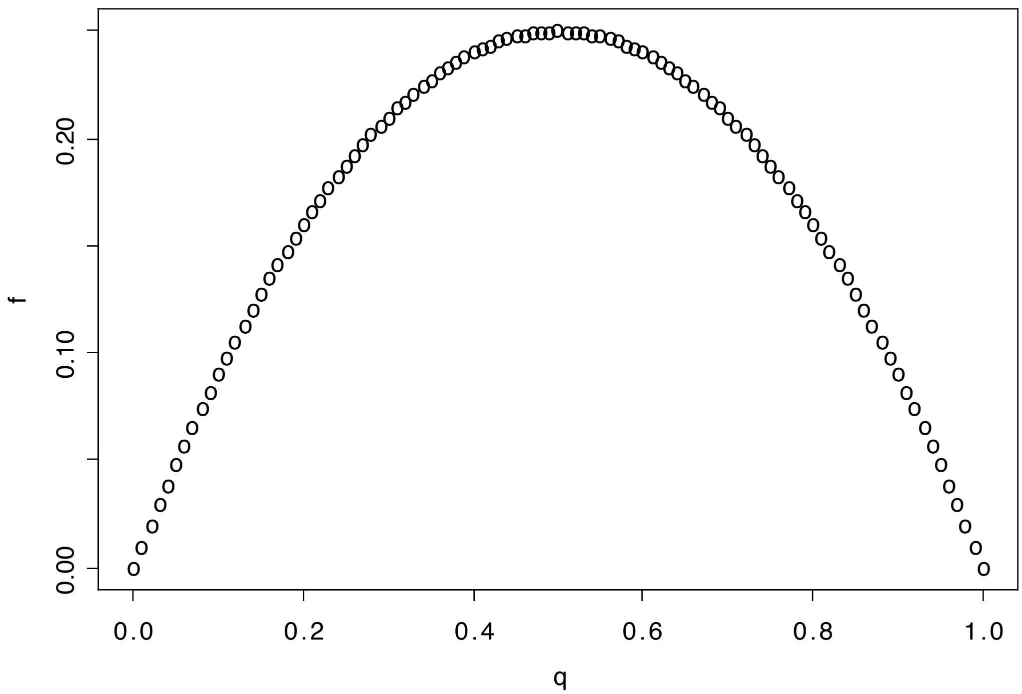
$$\phi = n \; q(1-q) \, \delta^2. \qquad (4.4)$$

Now it's clear.

- For any non-zero effect size and any (?) allocation of sample size to the two treatments, the greater the total sample size, the greater the power.
- For any sample size and any (?) allocation of sample size to the two treatments, the greater the effect size, the greater the power.
- Power depends not just on sample size and effect size, but on an aspect of *design* -- the allocation of sample size to the two treatments. This is a general feature of power in the analysis of variance and other statistical methods. It is important, but usually not mentioned.

Let's continue to pursue this interesting special case. For any given sample size and any non-zero effect size, we can maximize power by choosing q (the proportion of cases allocated to Group One) so that the function $f(q) = q(1-q)$ is as large as possible. What's the best value of q?

This is a simple calculus exercise, but the following plot gives the answer by brute force. I just computed $f(q) = q(1–q)$ for 100 equally spaced values of q ranging from zero to one.



So the best value of q is 1/2. That is, for comparing two means using the classical normal model, power is highest when the sample sizes are equal -- and this holds regardless of the total sample size or the magnitude of the effect.

This is a clear, simple example of something that holds for *any* classical ANOVA. The non-centrality parameter, and hence the power, depends on the total sample size, the effect, *and* the allocation of the sample to treatment combinations.

Equal sample sizes do not always yield the highest power. In general, the optimal allocation depends on the hypothesis being tested *and* the nature of the true effect. For example, suppose you have a design with 18 treatment combinations, and the test in question is to compare μ_1 with the average of μ_2 and μ_3. Further, suppose that $\mu_2 = \mu_3 \neq \mu_1$ (σ^2 can be anything); this is the effect. The optimal allocation is to give half the sample to Treatment One, split the other half any way at all between Treatments 2 and 3, and let n=0 for the other 15 treatments. This is why observations are not usually allocated to treatments based on a power analysis; it often advises you to put all your eggs in one basket.

In the analysis of variance, power analysis is used to select a sample size n as follows.

1. Choose an allocation of observations to treatments; usually, this is done without formal analysis, equal sample sizes being the most common choice.

2. Choose an effect. Your null hypothesis says that some collection of contrasts (of the treatment combination means) are all zero in the population. The "effect" you need to specify is that one or more of those contrasts is *not* zero. You must provide exact non-zero values, in units of the common within-treatment population standard deviation σ -- like, the difference between μ_1 and the average of μ_2 and μ_3 is minus 0.75σ. You don't need to know the numerical value of σ (thank goodness!), but you do need to be able to express differences between population means in units of σ. If possible, select the weakest effect that is still scientifically important.

3. Choose a desired power; again, a common choice is 0.80, but it's up to you.

4. Start with a modest but realistic value for the total sample size n. Increase it, each time determining the critical value of F, calculating the non-centrality parameter ϕ (you have enough information), and using ϕ to compute the probability that F will exceed the critical value. When that power becomes high enough, stop.

This is a rational strategy for choosing sample size. In practice, the hard part is selecting an effect. Scientists often can say what's a scientifically meaningful difference between means, but they usually have no clue about σ. Statisticians respond with the suggestion that σ^2 be estimated by MSE_F from similar studies. Scientists respond that there are no "similar" studies; the investigation being planned is new -- that's why we're doing it. In the end, the whole thing is based on so much guesswork that everyone feels uncomfortable. In my experience, this is what happens most of the time when people try to do a classical power analysis. Of course, there are exceptions; sometimes, everyone is happy.

The Sample Variation Method (Note STA442f05 has better sas programs. Fix this up!)

There are at least two main meanings of ``significance." One is statistical significance, and another is *explanatory* significance in the sense of explained variation. Formula (4.4) from Chapter 4 is relevant. It is reproduced here.

$$F = \left(\frac{n-p}{s}\right) \frac{a}{1-a}, \tag{4.4}$$

where, after controlling for the effects in a reduced model, a is the proportion of the *remaining* variation that is explained by the full model.

Formula (4.4) tells us that the two meanings of ``significance" need not coincide, since statistical significance can come from either strong results or from a large sample. The sample variation method can be viewed as a way of bringing the two types of significance into agreement. It's not really a power analysis, but it is a rational way to decide on sample size.

In equation (4.4), F is an increasing function of both n and a, so its p-value (the tail area beyond F) is a decreasing function of both n and a. The sample variation method is to choose a value of a that is just large enough to be interesting, and increase n, calculating F and its p-value each time until $p < 0.05$; then stop. The final value of n is the smallest sample size for which an effect explaining that much of the remaining variation will be significant. With that sample size, the effect will be significant if and only if it explains a or more of the remaining variation.

That's all there is to it. You tell me a proportion of remaining variation that you want to be significant, and I'll tell you a sample size. In exchange, you agree not to moan and complain and go fishing for more covariates if your results are almost significant, because they were too weak to be interesting anyway.

There are two questions you might want to ask.

- For a given proportion of the remaining variation, what sample size do I need for statistical significance?

- For a given sample size, what proportion of the remaining variation do I need for statistical significance?

To make things more definite, let us suppose we are contemplating a 2x3x4 analysis of covariance, with two covariates and factors cleverly named A, B and C. We are setting it up as a regression model, with one dummy variable for A, 2 dummy variables for B, and 3 for C. Interactions are represented by product terms, and there are 2 products for the AxB interaction, 3 for AxC, 6 for BxC, and 1*2*3 = 6 for AxBxC. The regression coefficients for these plus two for the covariates and one for the intercept give us $p = 26$. The null hypothesis is that of no BxC interaction, so $s = 6$. The "other effects in the model" for which we are "controlling" are represented by 2 covariates and 17 dummy variables and products of dummy variables.

First, let's find out what sample size we need for the interaction to be significant provided it explains at least 10% of the remaining variation after controlling for other effects in the model. This is accomplished by the program `sampvar1.sas`. It is a little unusual in that it uses the SAS `put` statement to write results to the *log* file. It never produces a list file, because there is no `proc` step.

```
/************************  sampvar1.sas ************************/
/*      Finds n needed for significance, for a given proportion of */
/*      remaining variation                                        */
/***************************************************************/

options linesize = 79 pagesize = 200;
data explvar;      /* Can replace alpha, s, p, and a below.   */
   alpha = 0.05;  /* Significance level.                     */
   s = 6;         /* Numerator df = # IVs being tested.      */
   p = 26;        /* There are p beta parameters.            */
   a = .10  ;     /* Proportion of remaining variation after */
                  /* controlling for all other variables.    */

   /* Initializing ... */  pval = 1; n = p+1;
   do until (pval <= alpha);
      F = (n-p)/s * a/(1-a);
      df2 = n-p;
      pval = 1-probf(F,s,df2);
      n = n+1 ;
   end;
   /* When finished, n is one too many */
   n = n-1; F = (n-p)/s * a/(1-a); df2 = n-p;
   pval = 1-probf(F,s,df2);

   put ' ******************************************************';
   put ' ';
   put '  For a multiple regression model with ' p 'betas, ';
   put '  testing ' s ' variables controlling for the others,';
   put '  a sample size of ' n 'is needed for significance at the';
   put '  alpha = ' alpha 'level, when the effect explains a = ' a ;
   put '  of the remaining variation after allowing for all other ';
   put '  variables in the model. ';
   put '  F = ' F ',df = (' s ',' df2 '), p = ' pval;
   put ' ';
   put ' ******************************************************';
```

Here is the part of the log file produced by the put statements.

```
 ******************************************************

  For a multiple regression model with 26 betas,
  testing 6  variables controlling for the others,
  a sample size of 144 is needed for significance at the
  alpha = 0.05 level, when the effect explains a = 0.1
  of the remaining variation after allowing for all other
  variables in the model.
  F = 2.1851851852 ,df = (6 ,118 ), p = 0.0491182815

 ******************************************************
```

Suppose you were considering n=120, and you wanted to know what proportion of the remaining variation the interaction must explain in order to be significant. This is accomplished by `sampvar2.sas`.

```
/************************  sampvar2.sas **************************/
/*  Finds proportion of remaining variation needed for significance, */
/*  given sample size n                                               */
/*******************************************************************/

options linesize = 79 pagesize = 200;
data explvar;      /* Replace alpha, s, p, and a below.  */
   alpha = 0.05;  /* Significance level.                */
   s = 6;         /* Numerator df = # IVs being tested. */
   p = 26;        /* There are p beta parameters.       */
   n = 120  ;     /* Sample size                        */

   /* Initializing ... */  pval = 1; a = 0; df2 = n-p;
   do until (pval <= alpha);
      F = (n-p)/s * a/(1-a);
      pval = 1-probf(F,s,df2);
      a = a + .001 ;
     end;
  /* When finished, a is .001 too much */
   a = a-.001; F = (n-p)/s * a/(1-a); pval = 1-probf(F,s,df2);

   put ' ***************************************************';
   put ' ';
   put '  For a multiple regression model with ' p 'betas, ';
   put '  testing ' s ' variables at significance level ';
   put '  alpha = ' alpha ' controlling for the other variables,';
   put '  and a sample size of ' n', the variables need to explain';
   put '  a = ' a ' of the remaining variation to be significant.';
   put '  F = ' F ', df = (' s ',' df2 '), p = ' pval;
   put '   ';
   put ' ***************************************************';
```

And here is the output.

```
 *****************************************************

  For a multiple regression model with 26 betas,
  testing 6  variables at significance level
  alpha = 0.05  controlling for the other variables,
  and a sample size of 120 , the variables need to explain
  a = 0.123  of the remaining variation to be significant.
  F = 2.1972633979 , df = (6 ,94 ), p = 0.0499350803

 ******************************************************
```

It's worth mentioning that the Sample Variation method is so simple that lots of people must know about it -- but I have never seen it described in print.

The Population Variation Method

This is a method of sample size selection for multiple regression due to Cohen (1988). It combines elements of classical power analysis and the sample variation method. Cohen does not call it the ``Population Variation Method;" he calls it ``Statistical Power Analysis." For most research psychologists, the population variation method *is* statistical power analysis, period.

The basic idea is this. Looking closely at the formula for the non-centrality parameter ϕ, Cohen decides that it is based on something he interprets as a *population* version of the quantity we are denoting by a. That is, one thinks of it as the proportion of remaining variation (Cohen uses the term variance instead of variation) that is explained by the effect in question -- in the population. He calls it ``effect size."

Just a comment: Of course the problem of comparing two means is a special case of multiple regression, but ``effect size" in the population variation method does not reduce to the traditional definition of effect size for the two-sample t-test with equal variances. In fact, effect size in the population variation method mixes the effect together with the design in such a way that they cannot be separated (by the way, this is true of the sample variation method too).

Still, from a so-called ``effect size" and a sample size, it's easy to calculate a non-centrality parameter, and then you can compute power, and increase the sample size until the power is as high as you wish. For most people, most of the time, it's a lot easier to think about proportions of explained variation than to think about collections of non-zero contrasts in units of σ. Plus, it applies to regression models in general, not just factorial ANOVA. To do a classical power analysis with observational data, you need the joint probability distribution of all the observed independent variables (which are presumably independent of any manipulated independent variables). Cohen's method is a lot easier. Here's a program that does it.

```
/**********************  popvar.sas ***************************/
options linesize = 79 pagesize = 200;
data fpower;        /* Replace alpha, s, p, and wantpow below    */
     alpha = 0.05;  /* Significance level                        */
     s = 6;         /* Numerator df = # IVs being tested         */
     p = 26;        /* There are p beta parameters               */
     a = .10  ;     /* Effect size                               */
     wantpow = .80; /* Find n to yield this power.               */
     power = 0; n = p+1; oneminus = 1-alpha; /* Initializing ... */
     do until (power >= wantpow);
        ncp = (n-p)*a/(1-a);
        df2 = n-p;
        power = 1-probf(finv(oneminus,s,df2),s,df2,ncp);
        n = n+1 ;
     end;
     n = n-1;
     put ' *********************************************************';
     put '   ';
     put '   For a multiple regression model with ' p 'betas, ';
     put '   testing ' s 'independent variables using alpha = ' alpha ',';
     put '   a sample size of ' n 'is needed';
     put '   in order to have probability ' wantpow 'of rejecting H0';
     put '   for an effect of size a = ' a ;
     put '   ';
     put ' *********************************************************';
```

```
 *********************************************************

   For a multiple regression model with 26 betas,
   testing 6 independent variables using alpha = 0.05 ,
   a sample size of 155 is needed
   in order to have probability 0.8 of rejecting H0
   for an effect of size a = 0.1

 *********************************************************
```

For comparison, when we specified a *sample* proportion of remaining variation equal to 10%, a sample size of 144 was required for significance. Getting into the spirit of the population variation method, we can talk about it like this. If the *population* effect size is 0.10 and n=155, then with 80% probability we'll get a *sample* effect size large enough for significance. How big does the sample effect size have to be? Running `sampvar2.sas`, it turns out that with n=155, you need a sample a=0.092 for significance. So if a=0.10 in the population and n=155, the probability that the sample a exceeds 0.092 is equal to 0.80.