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 following table 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.

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 μ_1-μ_2=μ_3-μ_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
```

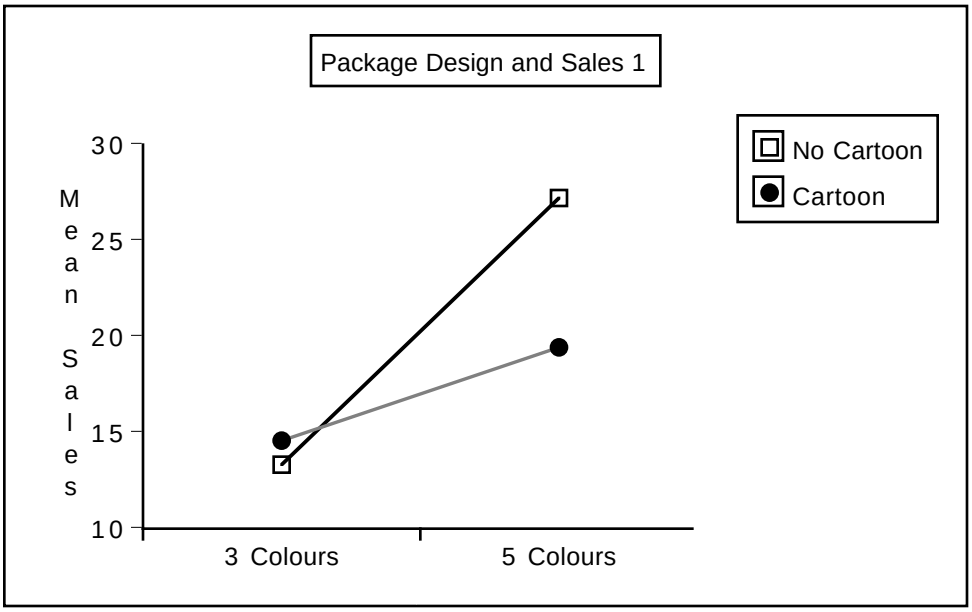
Another way to think about the interaction is to ask whether the effect of number of colours depends on presence versus absence of cartoon pictures. We are asking whether μ_1-μ_3=μ_2-μ_4. Here's the `test` statement and the output.

```
     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 μ_1–μ_2=μ_3–μ_4 is algebraically equivalent to μ_1–μ_3=μ_2–μ_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.

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

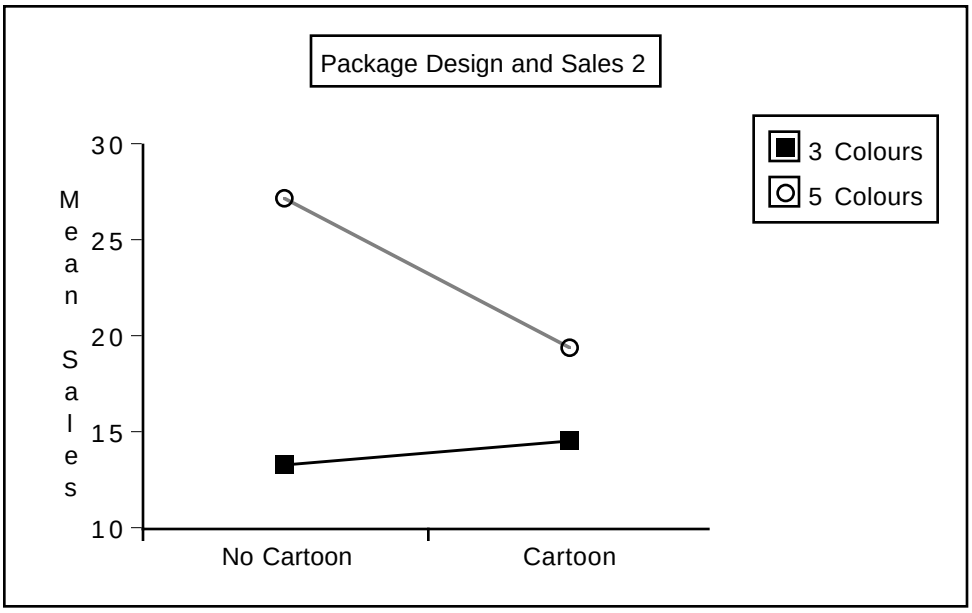


Whenever you have an interaction, such graphs will display non-parallel lines. Well actually, when you plot an interaction with real data, the lines will always be at least a little non-parallel. The question is whether they depart *significantly* from being parallel. Here, the advantage of 5 colours over 3 is significantly greater for designs without cartoons, 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 tests 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.

```
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 all 9 tests are to be protected with a Bonferroni correction?

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:

```
/* Actually it's a two-way ANOVA */

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

Source	DF	Sum of Squares	Mean 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.

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.