Chapter 6

Multiple Comparisons (Follow-up Tests)

6.1 A One-way Example

The following is a textbook example taken from Neter et al.'s (1996) *Applied linear statistical models* [9]. The Kenton Food Company is interested in testing the effect of different package designs on sales. Five grocery stores were randomly assigned to each of four package designs. The package designs used either three or five colours, and either had cartoons or did not. Because of a fire in one of the stores, there were only four stores in the 5-colour cartoon condition.

The dependent variable is sales, defined as number of cases sold. Actually, there are two independent variables: number of colours and presence versus absence of cartoons. But we will initially consider package design as a single categorical independent variable with four values.

Sample Question 6.1.1 *If there is a statistically significant relationship between package design and sales, would we be justified in concluding that differences in package design* caused *differences in sales?*

Answer to Sample Question 6.1.1 *Yes, if the study is carried out properly. It's an experimental study.*

Sample Question 6.1.2 *Is there a problem with external validity here?*

Answer to Sample Question 6.1.2 *It's impossible to tell for sure, but there easily could be. The behaviour of the sales force would have to be controlled somehow. A double blind would be ideal.*

The SAS program `kenton.sas` does a lot of things, starting with a one-way ANOVA using `proc glm`. The strategy will be to first present the entire program, and then go through it piece by piece and explain what is going on – with a few major digressions to explain the statistics.

```
/********************* kenton.sas ************************/
options linesize=79 pagesize=100 noovp formdlim=' ';
title 'Kenton Oneway Example From Neter et al.';

proc format;
     value pakfmt 1 = '3Colour Cartoon'   2 = '3Col No Cartoon'
                  3 = '5Colour Cartoon'   4 = '5Col No Cartoon';

data food;
     infile 'kenton.dat';
     input package sales;
     label package = 'Package Design'
           sales   = 'Number of Cases Sold';
     format package pakfmt.;

     /* Define ncolours and cartoon */
     if package = 1 or package = 2 then ncolours = 3;
        else if package = 3 or package = 4 then ncolours = 5;
     if package = 1 or package = 3 then cartoon = 'No ';
        else if package = 2 or package = 4 then cartoon = 'Yes';

     /* Indicator Coding for package: Use 3 at a time */
     if package = . then p1 = .; else if package = 1 then p1 = 1;
        else p1 = 0;
     if package = . then p2 = .; else if package = 2 then p2 = 1;
        else p2 = 0;
     if package = . then p3 = .; else if package = 3 then p3 = 1;
        else p3 = 0;
     if package = . then p4 = .; else if package = 4 then p4 = 1;
```

```
        else p4 = 0;

/* Basic one-way ANOVA -- well, not very basic */

proc glm;
     class package;
     model sales = package;
     means package;
     means package / bon tukey scheffe;
     /* Test some custom contrasts */
     contrast '3Colourvs5Colour' package 1  1 -1 -1;
     contrast 'Cartoon'          package 1 -1  1 -1;
     contrast 'CartoonDepends'   package 1 -1 -1  1;
     /* Test a collection of contrasts */
     contrast 'Overall F'        package 1 -1  0  0,
                                 package 0  1 -1  0,
                                 package 0  0  1 -1;
     /* Test effects of Colour and Cartoons simultaneously, allowing for
        a possible interaction */
     contrast 'ColorCartoon'     package 1  1 -1 -1,
                                 package 1 -1  1 -1;
     /* Get estimated value of a contrast along with a test (F=t-squared) */
     estimate '3Colourvs5Colour' package 1  1 -1 -1 / divisor = 2;

proc iml;
     title2 'Table of critical values for all possible Scheffe tests';
     numdf = 3;   /* Numerator degrees of freedom for initial test */
     dendf = 15;  /* Denominator degrees of freedom for initial test */
     alpha = 0.05;
     critval = finv(1-alpha,numdf,dendf);
     zero = {0 0}; S_table = repeat(zero,numdf,1);  /* Make empty matrix */
     /* Label the columns */
     namz = {"Number of Contrasts in followup test"
             "    Scheffe Critical Value"}; mattrib S_table colname=namz;
     do i = 1 to numdf;
        s_table(|i,1|) = i;
        s_table(|i,2|) = numdf/i *  critval;
     end;
```

```
     reset noname; /* Makes output look nicer in this case */
     print "Initial test has"  numdf " and " dendf "degrees of freedom."
           "Using significance level alpha = " alpha;
     print s_table;

proc reg;
     title2 'Using proc reg and dummy variables';
     model sales = p1 p2 p3;
     ncolour: test p1+p2 = p3; /* 3 vs 5 colours */

proc glm;
     title2 "Actually it's a two-way ANOVA";
     class ncolours cartoon;
     model sales = ncolours|cartoon;

/*   The model statement could have been
     model sales = ncolours cartoon ncolours*cartoon; */
```

The `proc format` statement provides labels for the package designs. After reading the data in a routine way, `if` statements are used to construct the categorical independent variables `ncolours` and `cartoon`. Notice the extra space in the 'No ' value of the alphanumeric variable `cartoon`. At first I didn't have a space, and `Yes` was truncated to `Ye`.

Now we'll look at what the first `proc glm` does. The complete `proc glm` statement is given above. Here, we will look at it a piece at a time, examining the output as we go. First, we have

```
proc glm;
     class package;
     model sales = package;
```

The class statement declares package to be categorical. Without it, `proc glm` would do a regression with package as a quantitative independent variable. The main F-test for equality of the four means is

```
                    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
```

We conclude that package design (or, if the study was poorly controlled, some variable confounded with package design) caused a difference in sales. The statement `means package;` produces mean sales for each value of the variable package.

```
Level of                  ------------SALES------------
PACKAGE            N        Mean               SD

3Col No Cartoon    5     13.4000000       3.64691651
3Colour Cartoon    5     14.6000000       2.30217289
5Col No Cartoon    5     27.2000000       3.96232255
5Colour Cartoon    4     19.5000000       2.64575131
```

Such a display is essential for seeing what is going on, but it still does not tell you which means are different from which other means. But before we lose control and start doing all possible t-tests, consider the following.

6.2 The Curse of a Thousand t-tests

Significance tests are supposed to help screen out random garbage, and help us ignore "trends" that could easily be due to chance. But all the common significance tests are designed in isolation, as if each one were the only test you would ever be doing. The chance of getting significant results when nothing is going on may be about 0.05 (more or less, depending on how well the assumptions are met), but if you do a *lot* of tests on a data set that is purely noise (no true relationships between any independent variable and any dependent variable), the chances of false significance mount up. It's like looking for your birthday in tables of stock market prices. If you look long enough, you will find it.

This problem definitely applies when you have a significant difference among more than two treatment means, and you want to know which ones are different from each other. For example, in an experiment with 10 treatment conditions (this is not an unusually large number, for real experiments), there are 45 pairwise differences among means.

You have to pity the poor scientist who learns about this and is honest enough to take this problem seriously (let's use the term "scientist" generously to apply to anyone trying to use significance test to learn something about a data set). On one hand, good scientific practice and common sense dictate that if you have gone to the trouble to collect data, you should explore thoroughly and try to learn something from the data. But at the same time, it appears that some stern statistical entity is scolding you, and saying that you're naughty if you peek.

There are two main ways to resolve the dilemma. One is to basically ignore the problem, while perhaps acknowledging that it is there. According to this point of view, well, you're crazy if you don't explore the data. Maybe the true significance level for the entire process is greater than 0.05, but still the use of significance tests is a useful way to decide which results might be real. Nothing's perfect; let's carry on.

The other reaction is to look for ways that significance tests can be modified to allow for the fact that we're doing a lot of them. What we want are methods for holding the chances of false significance to a single low level for a *set* of tests, simultaneously. The general term for such methods is **multiple comparison** procedures. Often, when a significance test (like a one-way ANOVA) tests several things simultaneously and turns out to be significant, multiple comparison procedures are used as a second step, to investigate where the effect came from. In cases like this, the multiple comparisons are called **follow-up** tests, or **post hoc** tests, or sometimes **probing**.

It is generally acknowledged that multiple comparison methods are often helpful (even necessary) for following up significant F-tests in order to see where an effect comes from. There is less agreement on how far the principle should be extended. Personally, I like the idea of limiting the chance of false significance to 0.05 for an entire study – say, for all the tests reported in a scientific paper, and all the ones that were not reported, too. This is a fairly radical view, shared by almost no one. But it can work in practice if you have enough data. More on this later. For now, let's concentrate on following up a significant F test in a one-way analysis of variance.

In the Kenton package design data, there are 4 treatment conditions, and 6 potential pairwise comparisons. The next line in the SAS program,

```
means package / bon tukey scheffe;
```

requests three kinds of multiple comparison tests for all pairwise differences among means.

6.2.1 Bonferroni

The Bonferroni method is very general, and extends far beyond pairwise comparisons of means. It is a simple correction that can be applied when you are performing multiple tests, and you want to hold the chances of false significance to a single low level for all the tests simultaneously. *It applies when you are testing multiple sets of independent variables, multiple dependent variables, or both.*

The Bonferroni correction consists of simply dividing the desired significance level (that's α, the maximum probability of getting significant results when actually nothing is happening, usually $\alpha = 0.05$) by the number of tests. In a way, you're splitting the alpha equally among the tests you do.

For example, if you want to perform 5 tests at joint significance level 0.05, just do everything as usual, but only declare the results significant at the *joint* 0.05 level if one of the tests gives you $p < 0.01$ (0.01=0.05/5). If you want to perform 20 tests at joint significance level 0.05, do the individual tests and calculate individual p-values as usual, but only believe the results of tests that give $p < 0.0025$ (0.0025=0.05/20). Say something like "Protecting the 20 tests at joint significance level 0.05 by means of a Bonferroni correction, the difference in reported liking between worms and spinach soufflé was the only significant food category effect."

The Bonferroni correction is conservative. That is, if you perform 20 tests, the probability of getting significance at least once just by chance is less than or equal to 0.0025 – almost always less. The big advantages of the

Bonferroni approach are simplicity and flexibility. It is the only way I know to analyze quantitative and categorical dependent variables simultaneously.

The main disadvantages of the Bonferroni approach are

1. *You have to know how many tests you want to perform in advance, and you have to know what they are.* In a typical data analysis situation, not all the significance tests are planned in advance. The results of one test will give rise to ideas for other tests. If you do this and then apply a Bonferroni correction to all the tests that you happened to do, it no longer protects all the tests simultaneously. On the other hand, you could randomly split your data into an exploratory sample and a replication sample. Test to your heart's content on the first sample. Then, when you think you know what your results are, perform only those tests on the replication sample, and protect them simultaneously with a Bonferroni correction. This could be called "Bonferroni-protected cross-validation." It sounds good, eh?

2. *The Bonferroni correction can be too conservative,* especially when the number of tests becomes large. For example, to simultaneously test all 780 correlations in a 40 by 40 correlation matrix at joint $\alpha = 0.05$, you'd only believe correlations with $p < 0.0000641 = 0.05/780$.

   Is this "too" conservative? Well, with $n = 200$ in that 40 by 40 example, you'd need $r = 0.27$ for significance (compared to $r = .14$ with no correction). With $n = 100$ you'd need $r = .385$, or about 14.8% of one variable explained by another *single* variable. Is this too much to ask? You decide.

6.2.2 Tukey

This is Tukey's Honestly Significant Difference (HSD) method. It is not his Least Significant Different (LSD) method, which has a better name but does not really get the job done. Tukey tests apply only to pairwise differences among means in ANOVA. It is based on a deep study of the probability distribution of the difference between the largest sample mean and the smallest sample mean, assuming the population means are in fact all equal.

- If you are interested in all pairwise differences among means and nothing else, and if the sample sizes are equal, Tukey is the best (most powerful) test, period.

- If the sample sizes are unequal, the Tukey tests still get the job of simultaneous protection done, but they are a bit conservative. When sample sizes are unequal, Bonferroni or Scheff can sometimes be more powerful.

6.2.3 Scheffé

Suppose there are p treatments (groups, values of the categorical independent variable, whatever you want to call them). A **contrast** is a special kind of linear combination of means in which the weights add up to zero. A population contrast has the form

$$\ell = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p$$

where $a_1 + a_2 + \cdots + a_p = 0$. The case where all of the a values are zero is uninteresting, and is excluded. A population contrast is estimated by a sample contrast:

$$L = a_1\overline{Y}_1 + a_2\overline{Y}_2 + \cdots + a_p\overline{Y}_p.$$

By setting $a_1 = 1$, $a_2 = -1$, and the rest of the a values to zero we get $L = \overline{Y}_1 - \overline{Y}_2$, so it's easy to see that any pairwise difference is a contrast. Also, the average of one set of means minus the average of another set is a contrast.

The initial F test for equality of p means can be viewed as a simultaneous test of $p-1$ contrasts. For example, suppose there are four treatments, and the null hypothesis of the initial test is $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. The table gives the a_1, a_2, a_3, a_4 values for three contrasts; if all three contrasts equal zero then the four population means are equal, and *vice versa*.

a_1	a_2	a_3	a_4
1	-1	0	0
0	1	-1	0
0	0	1	-1

The way you read this table is

$$\begin{array}{ccccccccc}
\mu_1 & - & \mu_2 & & & & & = & 0 \\
 & & \mu_2 & - & \mu_3 & & & = & 0 \\
 & & & & \mu_3 & - & \mu_4 & = & 0
\end{array}$$

Clearly, if $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ and $\mu_3 = \mu_4$, then $\mu_1 = \mu_2 = \mu_3 = \mu_4$, and if $\mu_1 = \mu_2 = \mu_3 = \mu_4$, then $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ and $\mu_3 = \mu_4$. The simultaneous F test for the three contrasts is 100% equivalent to a one-way ANOVA; it yields the same F statistic, the same degrees of freedom, and the same p-value.

There is always more than one way to set up the contrasts to test a given hypothesis. Staying with the example of testing differences among four means, we could have specified

a_1	a_2	a_3	a_4
1	0	0	-1
0	1	0	-1
0	0	1	-1

so that all the means are equal to the last one. These contrasts (differences between means) are actually *equal* to the regression coefficients in a multiple regression with indicator dummy variables, in which the last category is the reference category. But no matter how you set up collection of contrasts, if you do it correctly you always get the same answer.

The Scheffé tests allow testing whether *any* contrast (or set of contrasts) of treatment means differs significantly from zero, with the tests for all possible contrasts simultaneously protected at the same significance level, usually 0.05.

When asked for Scheffé follow-ups to a one-way ANOVA, SAS tests all pairwise differences between means, but *there are infinitely many more contrasts in the same family that it does not do* — and they are all jointly protected against false significance at the 0.05 level.

It's a miracle. You can do infinitely many tests, all simultaneously protected. You do not have to know what they are in advance. It's an license for unlimited data fishing, at least within the class of contrasts of treatment means. And you can test up to $p-1$ contrasts simultaneously if you wish. They are all part of the same family.

Two more miracles:

- If the initial one-way ANOVA is not significant, it's *impossible* for any of the Scheffé follow-ups to be significant. This is not quite true of Bonferroni or Tukey.

- If the initial one-way ANOVA *is* significant, there *must* be a single contrast that is significantly different from zero. It may not be a pairwise difference, you may not think of it, and if you do find one it may not be easy to interpret, but there is at least one out there. Well, actually, there are infinitely many, but they may all be extremely similar to one another. Incidentally, if you test any *collection* of contrasts that includes a contrast that is significantly different from zero by a Scheffé test, then the Scheffé test for the collection will be significant too.

Given all this, clearly it is helpful to be able to test any set of contrast you wish. As you will see below, the `contrast` statement of `proc glm` lets you do it easily. For now, let's assume that you have done an initial F test for differences among p treatment means, it's statistically significant, and also you can get F tests for any contrast of collection of contrasts you specify.

As usual, the F tests for contrasts (which are sometimes optimistically called "planned comparisons") are designed in a vacuum, as if each one were the only test you would ever do on your data. But you can convert them into Scheffé follow-ups to the initial test by using a different critical value (Recall that if a test statistic is greater than the critical value, it's significant).

Suppose that the follow-up test you want to do involves s contrasts; for a test of a single difference between means or some other single contrast, $s = 1$. Compute the usual F statistic for testing the contrast, and compare it to a modified critical value that we will call F_{S-crit}; the S is for Scheffé. The formula for F_{S-crit} is

$$F_{S-crit} = \frac{p-1}{s} F_{crit}, \tag{6.1}$$

where F_{crit} is the critical value for the *initial* test — the one you are following up. You reject the null hypothesis and declare your Scheffé test significant if $F > F_{S-crit}$.

You can do as many of these tests as you want easily, using SAS and a small table of F_{S-crit} critical values. You can make the table you need with `proc iml`. This is illustrated in the example below; the code can easily be modified to suit any problem. Or, you can use a textbook table of the F distribution and a calculator.

Please take another look at Formula (6.1). Notice that multiplying by the number of means (minus one) is a kind of penalty for the richness of the infinite family of tests you could do, while dividing by the number of contrasts you're testing reduces the penalty because you're looking for something bigger. As soon as Mr. Scheffé discovered these tests, people started complaining that the penalty was very severe, and it was too hard to get significance. In my opinion, what's remarkable is not that a license for unlimited fishing is expensive, but that it's for sale at all. You can pay for it by increasing the sample size.

When sample sizes are unequal, SAS presents follow-up tests for pairwise differences between means in the form of confidence intervals. If the 95% confidence interval does not include zero, the test (Bonferroni, Tukey or Scheffé) is significant at 0.05. Since all three types of follow-up test point to exactly the same conclusions for these data, only the Scheffé will be reproduced here.

```
                General Linear Models Procedure

                  Scheffe's test for variable: SALES

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

           Alpha= 0.05  Confidence= 0.95  df= 15  MSE= 10.54667
                       Critical Value of F= 3.28738

      Comparisons significant at the 0.05 level are indicated by '***'.

                                   Simultaneous            Simultaneous
                                       Lower    Difference     Upper
           PACKAGE                  Confidence    Between   Confidence
          Comparison                   Limit       Means       Limit

   5Col No Cartoon - 5Colour Cartoon     7.700       0.859    14.541  ***
   5Col No Cartoon - 3Colour Cartoon    12.600       6.150    19.050  ***
   5Col No Cartoon - 3Col No Cartoon    13.800       7.350    20.250  ***
   5Colour Cartoon - 5Col No Cartoon    -7.700     -14.541    -0.859  ***
   5Colour Cartoon - 3Colour Cartoon     4.900      -1.941    11.741
   5Colour Cartoon - 3Col No Cartoon     6.100      -0.741    12.941
   3Colour Cartoon - 5Col No Cartoon   -12.600     -19.050    -6.150  ***
   3Colour Cartoon - 5Colour Cartoon    -4.900     -11.741     1.941
   3Colour Cartoon - 3Col No Cartoon     1.200      -5.250     7.650
   3Col No Cartoon - 5Col No Cartoon   -13.800     -20.250    -7.350  ***
   3Col No Cartoon - 5Colour Cartoon    -6.100     -12.941     0.741
   3Col No Cartoon - 3Colour Cartoon    -1.200      -7.650     5.250
```

Notice that the critical value for the initial test (F_{crit}, not F_{S-crit}) for performing more tests is conveniently provided.

This pairwise confidence interval format is not so easy to look at, even if the significant differences are indicated by "***." For one thing, each comparison is given twice, once in each direction. For another, the actual means are not printed, just the differences between means. It helps to re-arrange the means from lowest to highest. This next display is not part of the SAS output; it's SAS output edited with a word processor.

```
Level of              ------------SALES------------
PACKAGE           N       Mean              SD

5Col No Cartoon   5     27.2000000       3.96232255

5Colour Cartoon   4     19.5000000       2.64575131
3Colour Cartoon   5     14.6000000       2.30217289
3Col No Cartoon   5     13.4000000       3.64691651
```

Now we see that the 5-colour No Cartoon treatment is significantly different from each of the others, which are not significantly different from each other. That's the kind of package design they should use; from a marketing standpoint, we're done. But let's look at some more follow-up tests anyway.

Testing Contrasts The `proc glm` in `kenton.sas` continues

```
/* Test some custom contrasts */
contrast '3Colourvs5Colour' package 1  1 -1 -1;
contrast 'Cartoon'          package 1 -1  1 -1;
contrast 'CartoonDepends'   package 1 -1 -1  1;
/* Test a collection of contrasts */
contrast 'Overall F'        package 1 -1  0  0,
                            package 0  1 -1  0,
                            package 0  0  1 -1;
/* Test effects of Colour and Cartoons simultaneously, allowing for
   a possible interaction */
contrast 'ColorCartoon'     package 1  1 -1 -1,
                            package 1 -1  1 -1;
```

The syntax for specifying a contrast goes: The word `contrast`, a label for the test in single or double quotes (this will appear in the output), the name of the independent variable, the coefficients of the contrast (the a values), and a semicolon to end the statement. If you are testing more than one contrast simultaneously, put a comma after the first one, repeat the independent variable name, and give another set of coefficients. The last contrast ends with a semi-colon instead of a comma. As the example shows, you can do as many tests as you like.

6.2.4 Proper Follow-ups

We will describe a set of tests as *proper follow-ups* to to an initial test if

1. The null hypothesis of the initial test logically implies the null hypotheses of all the tests in the follow-up set.

2. All the tests are jointly protected against Type I error (false significance) at a known significance level, usually $\alpha = 0.05$.

The first property requires explanation. First, consider that the Tukey tests, which are limited to pairwise differences between means, automatically satisfy this, because if all the population means are equal, then each pair is equal to each other. But it's possible to make mistakes with Bonferroni and Scheffé if you're not careful.

Here's why the first property is important. Suppose the null hypothesis of a follow-up test *does* follow logically from the null hypothesis of the initial test. Then, if the null hypothesis of the follow-up is false (there's really something going on), then the null hypothesis of the initial test must be incorrect too, and this is one way in which the initial null hypothesis is false. Thus if we correctly reject the follow-up null hypothesis, we have uncovered one of the ways in which the initial null hypothesis is false. In other words, we have (partly, perhaps) identified where the initial effect comes from.

On the other hand, if the null hypothesis of a potential follow-up test is *not* implied by the null hypothesis of the initial test, then the truth or untruth of the follow-up null hypothesis does not tell us *anything* about the null hypothesis of the initial test. They are in different domains. For example, suppose we conclude $2\mu_1$ is different from $3\mu_2$. Great, but if we want to know how the statement $\mu_1 = \mu_2 = \mu_3$ might be wrong, it's irrelevant.

If you stick to testing contrasts as a follow-up to a one-way ANOVA, you're fine. This is because if a set of population means are all equal, then any contrast of those means is equal to zero. That is, the null hypothesis of the initial test automatically implies the null hypotheses of any potential follow-up test, and everything is okay. Furthermore, if you try to specify a linear combination that is not a contrast with the `contrast` statement of `proc glm`, SAS will just say something like `NOTE: CONTRAST SOandSO is not estimable` in the log file. There is no other error message or warning; the test just does not appear in your list file.

If you really want a linear combination that is not a contrast, use the `estimate` statement. It will give the sample value of any linear combination of treatment means, along with a t-test for whether the linear combination is significantly different from zero. Here's output from the `estimate` statement in `kenton.sas`:

```
                                                 Standard
Parameter                        Estimate           Error    t Value    Pr > |t|

3Colourvs5Colour              -9.35000000      1.49705266      -6.25      <.0001
```

Note $t^2 = F$ immediately below.

6.2.5 Converting Tests for Contrasts into Scheffé tests

Here is the output from the `contrast` statements.

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

3Colourvs5Colour             1     411.4000000     411.4000000      39.01    <.0001
Cartoon                      1      49.7058824      49.7058824       4.71    0.0464
CartoonDepends               1      93.1882353      93.1882353       8.84    0.0095
Overall F                    3     588.2210526     196.0736842      18.59    <.0001
ColorCartoon                 2     479.5888889     239.7944444      22.74    <.0001
```

By ordinary one-at-a-time F tests, all the tests are significant. But let's treat them as Scheffé tests. To do this, we need the F_{S-crit} critical values for $s = 1$, 2 and 3. Actually we don't need one for $s = 3$, because by (6.1), it's the same as the critical value of the initial test. And in fact, any test of $p-1$ non-redundant contrasts is equivalent to the initial one-way ANOVA, always.

It's easy to get the F_{S-crit} values from `proc iml`. The following code is written carefully so that you can use it for any problem by just modifying the vales of `numdf` and `dendf` (and maybe `alpha` if you don't want to use 0.05).

```
proc iml;
     title2 'Table of critical values for all possible Scheffe tests';
     numdf = 3;   /* Numerator degrees of freedom for initial test (p-1) */
     dendf = 15;  /* Denominator degrees of freedom for initial test (n-p) */
     alpha = 0.05;
     critval = finv(1-alpha,numdf,dendf);
     zero = {0 0}; S_table = repeat(zero,numdf,1);  /* Make empty matrix */
     /* Label the columns */
     namz = {"Number of Contrasts in followup test"
             "    Scheffe Critical Value"};
     mattrib S_table colname=namz;
     do i = 1 to numdf;
        s_table(|i,1|) = i;
        s_table(|i,2|) = numdf/i *  critval;
     end;
     reset noname; /* Makes output look nicer in this case */
     print "Initial test has"  numdf " and " dendf "degrees of freedom."
           "Using significance level alpha = " alpha;
     print s_table;
```

Here is the output.

```
                  Kenton Oneway Example From Neter et al.                   8
          Table of critical values for all possible Scheffe tests
                                                    13:35 Friday, March 16, 2007

     Initial test has         3  and         15 degrees of freedom.
               Using significance level alpha =       0.05


     Number of Contrasts in followup test     Scheffe Critical Value

                                        1                  9.8621463
                                        2                  4.9310732
                                        3                  3.2873821
```

For the one-degree-of-freedom tests (single contrasts) we need $F > 9.86$ for significance. This means `3Colourvs5Colour` is significant, but `Cartoon` and `CartoonDepends` are not, even though `CartoonDepends` has a p-value of 0.0095 by the one-at-a-time test. `ColourCartoon` is also significant, because $22.74 > 4.93$. And of course `Overall F` is significant; it's the initial test.

6.2.6 Extensions

This section provides a brief but very powerful extension of the Scheffé tests to multiple regression, and Scheffé-like tests for logistic regression.

Multiple Regression

Suppose the initial hypothesis is that d regression coefficients all are equal to zero. We will follow up the initial test by testing whether s linear combinations of these regression coefficients are different from zero; $s \leq d$. Notice that now we are testing *linear combinations*, not just contrasts. If a set of coefficients are all zero, then any linear combination (weighted sum) of the coefficients is also zero. Thus the null hypotheses of the follow-up tests are implied by the null hypotheses of the initial test. As in the case of Scheffé tests for contrasts in one-way ANOVA, using an adjusted critical value guarantees simultaneous protection for all the follow-up tests at the same significance level as the initial test. This means we have proper follow-ups.

The formula for the modified critical value is

$$F_{S-crit} = \frac{d}{s} F_{crit}, \tag{6.2}$$

where again, the null hypothesis of the initial test is that d regression coefficients are all zero, and the null hypothesis of the follow-up test is that s linear combinations of those coefficients are equal to zero.

For convenience, here is the `proc iml` code to produce a table of adjusted critical values.

```
proc iml;
     title2 'Scheffe tests for Regression: Critical values';
     numdf = 3;   /* Numerator degrees of freedom for initial test (d) */
     dendf = 15;  /* Denominator degrees of freedom for initial test (n-d-1) */
     alpha = 0.05;
     critval = finv(1-alpha,numdf,dendf);
     zero = {0 0}; S_table = repeat(zero,numdf,1);  /* Make empty matrix */
     /* Label the columns */
     namz = {"Number of linear combos in followup test"
             "    Scheffe Critical Value"};
     mattrib S_table colname=namz;
     do i = 1 to numdf;
        s_table(|i,1|) = i;
        s_table(|i,2|) = numdf/i *  critval;
     end;
     reset noname; /* Makes output look nicer in this case */
     print "Initial test has "  numdf " and " dendf "degrees of freedom."
           "Using significance level alpha = " alpha;
     print s_table;
```

The Scheffé tests for contrasts in a one-way ANOVA are special cases of this, because anything you can do with factorial analysis of variance, you can do with dummy variable regression. It's very convenient with `test` statements in `proc reg`.

Logistic Regression

For logistic regression, there are Scheffé-like followups called *union-intersection tests*. The true Scheffé tests are a special kind of union-intersection method that applies to the (multivariate) normal linear model. Scheffé tests have one property that is not true of union-intersection follow-ups in general: the guaranteed existence of a significant one-degree-of-freedom test. This is tied to geometric properties of the multivariate normal distribution.

Just as in normal regression, the suppose that the initial null hypothesis is that d coefficients in the logistic regression model are all equal to zero. Suppose the initial hypothesis is that d regression coefficients all are equal to zero. We will follow up by testing whether s linear combinations of these regression coefficients are different from zero; $s \leq d$. There is no adjustment.

The critical value for the follow-up tests is exactly that of the initial test: a chi-square with d degrees of freedom. This principle applies to both likelihood ratio and Wald tests. In fact, it is true of likelihood ratio and Wald tests in general, not just in logistic regression.

Bibliographic citation

If you want to report the use of union-intersection tests or Scheffé tests for regression, or even Scheffé tests for more than one contrast in a one-way design, you will have difficulty finding it in any published Statistics text. Like Scheffé's original 1953 article [13], they almost universally stick to single contrasts. And it's usually not too helpful to cite unpublished material like this document.

Hochberg and Tamhane's (1987) monograph *Multiple comparison procedures* [7] is a good source for the tests of multiple linear combinations in regression, of which the tests of contrasts presented here are a special case. It's not very readable to non-statisticians, though. The same can be said of Gabriel's (1969) article [6], which is the primary source for the union-intersection follow-ups. But you can just trust me and cite them anyway.

Bibliography

[1] Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, **187**, 398-403.

[2] Cody, R. P. and Smith, J. K. (1991). *Applied statistics and the SAS programming language.* (4th Edition) Upper Saddle River, New Jersey: Prentice-Hall.

[3] Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. New York: Rand McNally.

[4] Feinberg, S. (1977) *The analysis of cross-classified categorical data.* Cambridge, Massachusetts: MIT Press.

[5] Fisher, R. A. (1925) *Statistical methods for research workers.* London: Oliver and Boyd.

[6] Gabriel, K. R. (1969). "Simultaneous test procedures — some theory of multiple comparisons." *Ann. Math. Statist.*, 40, 224–250.

[7] Hochberg, Y., and Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

[8] Moore, D. S. and McCabe, G. P. (1993). *Introduction to the practice of statistics.* New York: W. H. Freeman.

[9] Neter, J., Kutner, M. H., Nachhtscheim, C. J. and Wasserman, W. (1996) *Applied linear statistical models.* (4th Edition) Toronto: Irwin.

[10] Roethlisberger, F. J. (1941). *Management and morale*. Cambridge, Mass.: Harvard University Press.

[11] Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Croft.

[12] Rosenthal, R. and Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectation and pupils' intellectual development.* New York: Holt, Rinehart and Winston.

[13] Scheffé, H. (1953). "A method for judging all contrasts in the analysis of variance." *Biometrika*, 40, 87–104.