Linear combinations

A linear combination consists of a series of variables multiplied by constants. For a series of k variables the linear combination could be expressed as follows.

Generic linear combination: $W = a_1W_1 + a_2W_2 + ... + a_kW_k$

The concern in Chapter 6 is for linear combinations of group means. This is expressed as follows.

For parameters, $\lambda = c_1\mu_{\bar{Y}_1} + c_2\mu_{\bar{Y}_2} + c_3\mu_{\bar{Y}_3} + ... + c_k\mu_{\bar{Y}_k}$

For statistics, $g = c_1\bar{Y}_1 + c_2\bar{Y}_2 + c_3\bar{Y}_3 + ... + c_k\bar{Y}_k$

The variance for a linear combination is given as the sum of the variances of the variables in the linear combination plus twice the covariances.

$$Var(W) = a_1^2Var(W_1) + a_2^2Var(W_2) + ... + a_k^2Var(W_k) +$$
$$2(a_1a_2Cov(W_1,W_2) + a_1a_3Cov(W_1,W_3) + a_2a_3Cov(W_2,W_3) + ...a_1a_kCov(W_1,W_k))$$

An important aspect of this calculation is that if the variables are independent, which is usually the case in Analysis of Variance, the covariances can be assumed to be zero. The variance of the linear combination is then the sum of the variances multiplied by the squares of the coefficients.

We have already seen one example of a linear combination in the t-test. The linear combination estimated for the t-test is $\hat{\delta} = c_1\bar{Y}_1 + c_2\bar{Y}_2$, where $c_1 = 1$ and $c_2 = -1$. Usually we are testing the null hypothesis $H_0 : \bar{Y}_1 - \bar{Y}_2 = 0$ or $H_0 : \delta = 0$. The variance for this linear combination is the sum of the variances of the means. We will assume the two groups are sampled independently and have no covariance. The variance for a mean is s^2/n, so the variance of this linear combination is the variance of the difference $\bar{Y}_1 - \bar{Y}_2$ which is $S^2_{\bar{Y}_1-\bar{Y}_2} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$. The t-test is done as the difference divided by the standard error which is $t = \frac{\left(\bar{Y}_1 - \bar{Y}_2\right) - \delta}{S_{\bar{Y}_1-\bar{Y}_2}}$, where δ is zero for $H_0 : \delta = 0$.

In Analysis of Variance there are more means than for the t-test. Assuming the "t" treatment levels in ANOVA are independent, the linear combination and variance is as follows.

$$g = c_1\bar{Y}_1 + c_2\bar{Y}_2 + c_3\bar{Y}_3 + ... + c_t\bar{Y}_t$$

$$Var(g) = \frac{c_1^2s_1^2}{n_1} + \frac{c_2^2s_2^2}{n_2} + \frac{c_3^2s_3^2}{n_3} + ... + \frac{c_t^2s_t^2}{n_t}$$

There are a few variations on this formula which are used when the variances are equal and can be pooled, or when the analysis is "balanced" and the n_i are equal. If the variances are equal they can be pooled into an estimate of a single variance given by

$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2 + ... + (n_t-1)s_t^2}{(n_1-1) + (n_2-1) + (n_3-1) + ... + (n_t-1)}$. The variance is then calculated as

$Var(g) = S_p^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3} + ... + \frac{c_t^2}{n_t}\right)$ and if $n_1=n_2=\ldots=n_t = n$, then

$Var(g) = \frac{S_p^2}{n}\left(c_1^2 + c_2^2 + c_3^2 + ... + c_t^2\right)$.

Before doing the linear contrasts, we should determine if the variances can be pooled or not. Although PROC GLM has some HOV (homogeneity of variance) tests, it has no capability of dealing with non-homogeneous variance if it is detected. PROC MIXED has a method of testing for HOV and can deal with non-homogeneous variance if detected. The original PROC MIXED model we ran was:

```
proc mixed data=MiceDiet cl covtest;
   Title2 'Analysis of Variance with PROC MIXED';
   class diet;
   model lifetime = diet;
run;
```

To test for homogeneity of variance we add the statement "repeated / group=diet;". This analysis is not a "repeated measures" analysis, but the "group=diet" option on this statement will cause PROC MIXED to fit separate variances to the variables specified in the group= option. In this case the variable is our treatment variable "diet".

```
proc mixed data=MiceDiet cl covtest;
   Title2 'Testing homogeniety of Variance with PROC MIXED';
   class diet;
   model lifetime = diet / ddfm=kr;
   repeated / group=diet;
run;
```

I have added one other modification to the program, the option "/ ddfm=kr" requests the Kenward-Rogers approach for handling denominator degrees of freedom. Other options are available, including the Satterthwaithe options used in the two sample t-test. However, this one appears to be the best for ANOVA situations.

The results of this test of homogeneity are given in the printed output. Note in particular that each diet now has it's own estimate of variance.

Covariance Parameter Estimates

Cov Parm	Group	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
Residual	DIET N/N85	26.2687	4.9643	5.29	<.0001	0.05	18.7234	39.5319
Residual	DIET N/R40	44.9356	8.2733	5.43	<.0001	0.05	32.2855	66.8452
Residual	DIET N/R50	60.3448	10.2001	5.92	<.0001	0.05	44.4538	86.6356
Residual	DIET NP	37.6223	7.6796	4.90	<.0001	0.05	26.1635	58.7189
Residual	DIET R/R50	44.6645	8.5172	5.24	<.0001	0.05	31.7464	67.4911
Residual	DIET lopro	48.8838	9.3218	5.24	<.0001	0.05	34.7453	73.8667

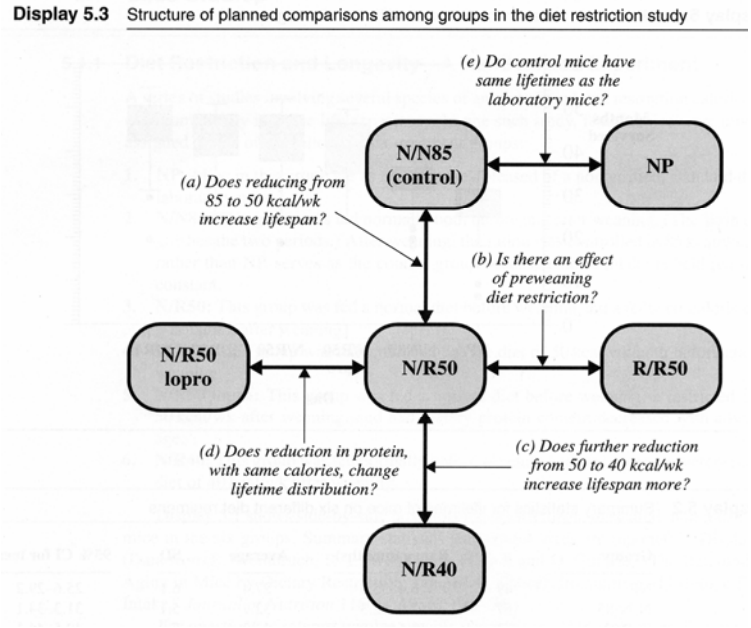
Also note that there is a new, separate test called the "Null Model Likelihood Ratio Test". This sections tests the model with separate variance as a full model (6 variances estimated at 1 d.f. each) against a reduced model with homogeneous variance (a single variance requiring only 1 d.f.). The test is for the 5 d.f. difference in the two models.

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
5	11.07	0.0500

In this case we would conclude ?a difference?. Very close, but we would conclude a difference exists among the variances. Recall that when we did the 5 t-tests we detected 4 with the same variance and one with different variances. Overall there appears to be a difference, marginal maybe, but a difference.

Linear combinations can be used to test specific hypotheses about the treatments using what are called "contrasts". Contrasts are easily set up in SAS. For example, in out first example (mouse diets) we tested 5 specific hypotheses: N/N85 versus N/R50, N/R50 versus R/R50, N/R40 versus N/R50, N/R50 versus lopro and N/N85 versus NP. These could have been done as contrasts in an ANOVA.



First the ANOVA, done in PROC GLM or PROC MIXED. We will use MIXED because it will also allow us to test for equal variances. Once we determine if we are justified in pooling the variances we proceed with our contrasts. Note that in PROC MIXED we can work with unequal variances, in PROC GLM we cannot.

Writing contrasts in SAS:

We want to test the linear contrast $g = c_1\overline{Y}_1 + c_2\overline{Y}_2 + c_3\overline{Y}_3 + ... + c_t\overline{Y}_t$ using either the variance $Var(g) = \frac{c_1^2 s_1^2}{n_1} + \frac{c_2^2 s_2^2}{n_2} + \frac{c_3^2 s_3^2}{n_3} + ... + \frac{c_t^2 s_t^2}{n_t}$ or $Var(g) = S_p^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \frac{c_3^2}{n_3} + ... + \frac{c_t^2}{n_t}\right)$. The SAS program keeps track of the S^2 values and the n_i values. All we need to do is determine what contrasts we want to do, and specify the values of c_i. This would be done as follows:

From the SAS section titled "Class Level Information" given in both GLM and MIXED we determine the order of the variables stored by SAS. This order will usually be alphanumeric, and for out example is [N/N85 N/R40 N/R50 NP R/R50 lopro].

```
Class Level Information
Class    Levels    Values
DIET          6    N/N85 N/R40 N/R50 NP R/R50 lopro
```

Now we need to determine what contrasts will test the hypotheses of interest. The hypotheses of interest were [N/N85 versus N/R50, N/R50 versus R/R50, N/R40 versus N/R50, N/R50 versus lopro and N/N85 versus NP]. In order to do the first test in our linear contrast we would need the following coefficients or "multipliers". $g = c_1\overline{Y}_{\text{N/N85}} + c_2\overline{Y}_{\text{N/R40}} + c_3\overline{Y}_{\text{N/R50}} + c_4\overline{Y}_{\text{NP}} + c_5\overline{Y}_{\text{R/R50}} + c_6\overline{Y}_{\text{lopro}}$. In order to

test $H_o: \bar{Y}_{\text{N/N85}} - \bar{Y}_{\text{N/R40}} = 0$ we need multiplier 1 to equal 1, multiplier 2 to equal –1, and all other multipliers to equal 0.

Test	Hypothesis $\mu_1 = \mu_2$	N/N85	N/R40	N/R50	NP	R/R50	lopro
A	N/N85 = N/R50	**1**	**0**	**–1**	**0**	**0**	**0**
B	N/R50 = R/R50	**0**	**0**	**1**	**0**	**–1**	**0**
C	N/R40 = N/R50	**0**	**1**	**–1**	**0**	**0**	**0**
D	N/R50 = lopro	**0**	**0**	**1**	**0**	**0**	**–1**
E	N/N85 = NP	**1**	**0**	**0**	**–1**	**0**	**0**

In practice, for testing purposes, it is not important if we calculate $\bar{Y}_{\text{N/N85}} - \bar{Y}_{\text{N/R40}}$ or $\bar{Y}_{\text{N/R40}} - \bar{Y}_{\text{N/N85}}$, the result is the same. Therefore, which treatment gets the "1" and which gets "–**1**" is not important.

These contrasts are set up is SAS in "contrast statements". The first statement below is a comment I created to keep track of the treatments and the order in which they occur. The other 5 statements are contrasts. A contrast starts with the word "contrast" followed by a description of up to 16 characters in single quotes. After the description comes the name of the group or treatment variable followed by the contrast. When writing contrasts I usually put the negative first, but this is not important.

```
*** diet treatments                N/N85 N/R40 N/R50 NP R/R50 lopro;
contrast 'A: N/N85 N/R50' diet      -1    0     1    0    0     0;
contrast 'B: N/R50 R/R50' diet       0    0    -1    0    1     0;
contrast 'C: N/R40 N/R50' diet       0   -1     1    0    0     0;
contrast 'D: N/R50 lopro' diet       0    0    -1    0    0     1;
contrast 'E: N/N85 NP'    diet      -1    0     0    1    0     0;
```

The results are given in the output.

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
A: N/N85 N/R50	1	122	70.40	<.0001
B: N/R50 R/R50	1	124	0.21	0.6474
C: N/R40 N/R50	1	129	4.97	0.0275
D: N/R50 lopro	1	123	3.96	0.0489
E: N/N85 NP	1	93.9	22.77	<.0001

Comparison of the t-test results with the ANOVA results.

Test	Hypothesis $\mu_1 = \mu_2$	Means	P > |t| (unpooled)	P > |t| (pooled)	P > |t| ANOVA
A	N/N85 = N/R50	32.691, 42.297	**<0.0001**	<0.0001	<0.0001
B	N/R50 = R/R50	42.297, 42.886	0.6474	**0.6532**	0.6472
C	N/R40 = N/R50	45.117, 42.297	0.0275	**0.0294**	0.0275
D	N/R50 = lopro	42.297, 39.686	0.0489	**0.0516**	0.0489
E	N/N85 = NP	32.691, 27.402	<0.0001	**<0.0001**	<0.0001

Which is better, t-tests of ANOVA? One big advantage to ANOVA is that if the variances can be pooled the combined variance is better and has more d.f. for testing. In this case we could not pool,

but the ANOVA in PROC MIXED can still handle this situation. In those cases where pooling is justified the tests have increased power.

In the mouse diet example the contrasts were clear. The contrasts were relatively simple "pairwise" contrasts. However, other contrasts are possible.

If we wanted to contrast N/N85 to the average of both N/R40 and N/R50 we would need to calculate $H_0 : \bar{Y}_{\text{N/N85}} - \frac{\bar{Y}_{\text{N/R40}} + \bar{Y}_{\text{N/R50}}}{2} = 0$. This can be simplified by multiplying through by 2, giving $H_0 : 2\bar{Y}_{\text{N/N85}} - (\bar{Y}_{\text{N/R40}} + \bar{Y}_{\text{N/R50}}) = 0$.

Similarly, if we wanted to compare the treatment lopro to the mean of all other treatments we could calculate $H_0 : \frac{\bar{Y}_{\text{N/N85}} + \bar{Y}_{\text{N/R40}} + \bar{Y}_{\text{N/R50}} + \bar{Y}_{\text{NP}} + \bar{Y}_{\text{R/R50}}}{5} - \bar{Y}_{\text{lopro}} = 0$ or multiply through by 5 to get $H_0 : \bar{Y}_{\text{N/N85}} + \bar{Y}_{\text{N/R40}} + \bar{Y}_{\text{N/R50}} + \bar{Y}_{\text{NP}} + \bar{Y}_{\text{R/R50}} - 5\bar{Y}_{\text{lopro}} = 0$.

NP and N/N85 versus all other treatments would test the hypothesis $H_0 : \frac{\bar{Y}_{\text{N/N85}} + \bar{Y}_{\text{NP}}}{2} - \frac{\bar{Y}_{\text{N/R40}} + \bar{Y}_{\text{N/R50}} + \bar{Y}_{\text{R/R50}} + \bar{Y}_{\text{lopro}}}{4} = 0$ and would simplify to $H_0 : 2\bar{Y}_{\text{N/N85}} + 2\bar{Y}_{\text{NP}} - (\bar{Y}_{\text{N/R40}} + \bar{Y}_{\text{N/R50}} + \bar{Y}_{\text{R/R50}} + \bar{Y}_{\text{lopro}}) = 0$

and the hypothesis NP and N/N85 versus N/R40, N/R50 and R/R50 can be expressed as either $H_0 : \frac{\bar{Y}_{\text{N/N85}} + \bar{Y}_{\text{NP}}}{2} - \frac{\bar{Y}_{\text{N/R40}} + \bar{Y}_{\text{N/R50}} + \bar{Y}_{\text{R/R50}}}{3} = 0$ and would simplify to $H_0 : 3\bar{Y}_{\text{N/N85}} + 3\bar{Y}_{\text{NP}} - (2\bar{Y}_{\text{N/R40}} + 2\bar{Y}_{\text{N/R50}} + 2\bar{Y}_{\text{R/R50}}) = 0$.

These contrasts are expressed below and done with PROC MIXED.

Alternative hypotheses	N/N85	N/R40	N/R50	NP	R/R50	lopro	SUM
1) N/N85 versus N/R40 and N/R50	**–1**	**0.5**	**0.5**	**0**	**0**	**0**	**0**
2) N/N85 versus N/R40 and N/R50	**–2**	**1**	**1**	**0**	**0**	**0**	**0**
3) lopro versus all others	**0.2**	**0.2**	**0.2**	**0.2**	**0.2**	**–1**	**0**
4) lopro versus all others	**1**	**1**	**1**	**1**	**1**	**–5**	**0**
5) NP and N/N85 versus all other treatments	**–0.5**	**0.25**	**0.25**	**–0.5**	**0.25**	**0.25**	**0**
6) NP and N/N85 versus all other treatments	**–2**	**1**	**1**	**–2**	**1**	**1**	**0**
7) NP and N/N85 versus N/R40, N/R50 and R/R50	**–0.5**	**0.333**	**0.333**	**–0.5**	**0.333**	**0**	**0**
8) NP and N/N85 versus N/R40, N/R50 and R/R50	**–3**	**2**	**2**	**–3**	**2**	**0**	**0**

Note that all contrasts must sum to one (1). The contrasts above all appear to sum to 1. When run we achieve the following results.

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
1	1	147	141.01	<.0001
2	1	147	141.01	<.0001
3	1	74.4	2.54	0.1154
4	1	74.4	2.54	0.1154
5	1	221	303.02	<.0001
6	1	221	303.02	<.0001
7	.	.	.	.
8	1	236	312.50	<.0001

Additional notes on contrasts:

1) Note that the results are identical with either the fractions or the integers (except for number 7). Also note that number 7 failed. SAS PROC MIXED produced nothing the the log indicating an error. PROC GLM produces the message "NOTE: CONTRAST 7 is not estimable.". Why can't contrast 7 be estimated?

```
contrast '7' diet  -0.5   0.333  0.333  -0.5  0.333     0;
```

The reason is that contrast 7 does not actually sum to zero. It sums to -0.001. SAS check this sum to 6 decimal places. Generally, using integers is more precise and less prone to errors.

2) Contrasts each have an α probability of error. It seems that we started using ANOVA to avoid doing 6 t-tests, each with an α probability of error. So what have we gained? We have gained the overall guarantee from the ANOVA that indeed something is significant. Once we know this then we can use contrasts and other techniques. If the ANOVA indicates that there are no significant treatment effects then we should not use the contrasts.

As an additional guarantee we should keep two restrictions: first we should not do more contrasts than we have degrees of freedom in the model and second we should only test *a priori* contrasts, not contrasts determined by examining the data *a posteriori*.

We will soon see some other techniques more appropriate *a posteriori* to testing.

Other examples

Recall the Spock trial judge example. Contrasts can be written to test the Spock judge from other judges. It is also possible to write joint contrasts to test multiple degree of freedom problems, such as the test among other judges in the Spock trial. These can be done in both PROC MIXED and PROC GLM. However, PROC MIXED does not produce SS for comparison to our EXTRA SS calculations. PROC GLM is used below.

```
proc glm data=Jury;
   class judge;
   model percent = judge;
*** judges in alphabetical order =>     A  B  C  D  E  F SPOCK;
   contrast 'Spock vrs others'   judge  1  1  1  1  1  1   -6;
   contrast 'Among other judges' judge -1  1  0  0  0  0    0,
                                 judge  0 -1  1  0  0  0    0,
                                 judge  0  0 -1  1  0  0    0,
                                 judge  0  0  0 -1  1  0    0,
                                 judge  0  0  0  0 -1  1    0;
run;
```

The results were:

```
Class Level Information

Class         Levels    Values
Judge              7    A B C D E F SPOCK

Number of Observations Read          46
Number of Observations Used          46

Dependent Variable: Percent
                                      Sum of
Source               DF         Squares     Mean Square    F Value    Pr > F
Model                 6     1927.080772      321.180129       6.72    <.0001
Error                39     1864.445255       47.806289
Corrected Total      45     3791.526027

R-Square     Coeff Var      Root MSE    Percent Mean
0.508260      26.01027      6.914209        26.58261

Source               DF       Type I SS     Mean Square    F Value    Pr > F
Judge                 6     1927.080772      321.180129       6.72    <.0001

Source               DF     Type III SS     Mean Square    F Value    Pr > F
Judge                 6     1927.080772      321.180129       6.72    <.0001

Contrast             DF     Contrast SS     Mean Square    F Value    Pr > F
Spock vrs others      1     1536.776942     1536.776942      32.15    <.0001
Among other judges    5      326.457869       65.291574       1.37    0.2582
```

General Linear Hypothesis test

Source	d.f.	SS	MS	F
Reduced model error	44	2190.903123		
Full model error	39	1864.445222		
Difference	5	326.457901	65.291580	1.365753
Full model error	39	1864.445222	47.806288	

The contrasts can do the same tests. The contrasts here separate the Model SS (1927.080772) into two components, Spock versus others (1536.776942) and among other judges (326.457869). The results also clearly show that the Spock judge differed from other judges and the other judges did not differ from one another.