*<u>Topics for these notes</u>*:

- *Linear mixed models with a random intercept, multi-sample data*
- *Repeated measures ANOVA*
- *Simple random effects and mixed effect models*
- *ICC and reliability*

*<u>Associated reading</u>: Sections 2.2 and 3 in 'LMM introduction…' course notes.*

## *2.2 Multiple-sample data*

- Here, we now consider data with two factors (e.g., group and time), which is the type that is probably more commonly analyzed and published. Here are some examples:

*Example 1*: Hypothetical myostatin application involving group (myostatin, control) and time (24, 48 and 72 hours), where experimental units are measured at each time point. In the 2×3 factorial experiment with 4 experimental units at each treatment combination, the sample size was 24. For this newly proposed experiment, each experimental unit would be measured 3 times so that 8 experimental units would be required.

*Example 2*: A clinical trial is conducted where subjects are randomized to treatment or placebo and monitored over several months.

# The model:

$$
Y_{hij} = \mu + \underbrace{\gamma_h}_{\text{Group}} + \overbrace{\tau_j}^{\text{Time}} + \underbrace{(\gamma\tau)_{hj}}_{\text{G}\times\text{T}} + \overbrace{b_{i(h)}}^{\text{Subject(group)}} + \underbrace{\varepsilon_{hij}}_{\text{error at individual time point}}
$$

where $b_{i(h)} \sim iid\ N(0, \sigma_b^2)$ independent of $\varepsilon_{hij} \sim iid\ N(0, \sigma_\varepsilon^2)$. Sum to 0 restrictions can be placed on G, T and G×T effects.

## *Example with SAS using the 'dog data'*

Reiczigel (Biometrics, 1999) describes an experiment of Sterczer et al. (1996) using two-dimensional ultrasonography to study the effect of cholagogues on changes in gallbladder volume (GBV) in three groups of healthy dogs. One group received cholechystokynin, another received clanobutin, and the third was a control group. GBV values were determined immediately before the administration of the test substance and at 30 minute intervals thereafter.

## 2.2.1 *The fit using LMMs Solution for Fixed Effects*

```
*for comparison;
proc mixed data=uni_dogs;
 class id group time;
 model y = group time group*time / solution;
 random intercept / subject=id(group); run;
```

| | |
|---|---|
| Covariance Structure | *Variance Componenets* |
| Subject Effect | *id(group)* |
| Estimation Method | *REML* |
| Residual Var. Method | *Profile* |
| Fixed Effects SE Method | *Model-Based* |
| DF Method | *Containment* |

Dimensions

| | |
|---|---|
| Covariance Parameters | 2 |
| Columns in X | 24 |
| Columns in Z Per Subject | 1 |
| Subjects | 18 |
| Max Obs Per Subject | 5 |
| Number of Observations Used | 90 |

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate |
|---|---|---|
| Intercept | id(group) | 46.6443 |
| Residual | | 0.6897 |

Fit Statistics

| | |
|---|---|
| -2 Res Log Likelihood | 299.3 |
| AIC (smaller is better) | 303.3 |
| AICC (smaller is better) | 303.4 |
| BIC (smaller is better) | 305.0 |

id(group) means 'ID within group'. It is used here since subject ID's are not unique across the experiment, just within groups. For RM ANOVA in PROC GLM (later), we need to include id(group) regardless of uniqueness of subject ID's.

| Effect | gp | time | Est. | SE | DF | t | Pr>\|t\| |
|---|---|---|---|---|---|---|---|
| Interc. | | | 16.75 | 2.81 | 15 | 5.96 | <.0001 |
| group | ch | | 2.40 | 3.97 | 15 | 0.60 | 0.5547 |
| group | cl | | -1.62 | 3.97 | 15 | -0.41 | 0.6892 |
| group | co | | 0 | . | . | . | . |
| time | | 0 | -0.08 | 0.48 | 60 | -0.16 | 0.8735 |
| time | | 30 | -0.45 | 0.48 | 60 | -0.93 | 0.3571 |
| time | | 60 | -0.04 | 0.48 | 60 | -0.08 | 0.9393 |
| time | | 90 | -0.23 | 0.48 | 60 | -0.49 | 0.6283 |
| time | | 120 | 0 | . | . | . | . |
| grp*tm | ch | 0 | 0.63 | 0.68 | 60 | 0.93 | 0.3541 |
| grp*tm | ch | 30 | -5.06 | 0.68 | 60 | -7.46 | <.0001 |
| grp*tm | ch | 60 | -3.87 | 0.68 | 60 | -5.71 | <.0001 |
| grp*tm | ch | 90 | -1.35 | 0.68 | 60 | -1.99 | 0.0508 |
| grp*tm | ch | 120 | 0 | . | . | . | . |
| grp*tm | cl | 0 | 0.78 | 0.68 | 60 | 1.13 | 0.2617 |
| grp*tm | cl | 30 | -2.02 | 0.68 | 60 | -2.98 | 0.0041 |
| grp*tm | cl | 60 | -1.20 | 0.68 | 60 | -1.77 | 0.0815 |
| grp*tm | cl | 90 | -0.03 | 0.68 | 60 | -0.04 | 0.9707 |
| grp*tm | cl | 120 | 0 | . | . | . | . |
| grp*tm | co | 0 | 0 | . | . | . | . |
| grp*tm | co | 30 | 0 | . | . | . | . |
| grp*tm | co | 60 | 0 | . | . | . | . |
| grp*tm | co | 90 | 0 | . | . | . | . |
| grp*tm | co | 120 | 0 | . | . | . | . |

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| group | 2 | 15 | 0.24 | 0.7869 |
| time | 4 | 60 | 44.43 | <.0001 |
| group*time | 8 | 60 | 13.38 | <.0001 |

## *2.2.2 RM ANOVA approach*

**<u>The ANOVA table, including expected mean squares.</u>** Note: $Q_T$ is a function of time effects; the greater the value, the more the difference between $\tau_j$ parameters. Similar for G, G×T. In the table above, $n_h$=# of subjects in group $h$, $n_{tot}$=total sample size.

| Source | DF | SS | MS | $E$(MS) |
|---|---|---|---|---|
| G | $s-1$ | $r\sum n_h(\bar{Y}_{h\bullet\bullet}-\bar{Y}_{\bullet\bullet\bullet})^2$ | $SS_G/(s-1)$ | $\sigma_\varepsilon^2+r\sigma_b^2+Q_G$ |
| T | $r-1$ | $n_{tot}\sum(\bar{Y}_{\bullet\bullet j}-\bar{Y}_{\bullet\bullet\bullet})^2$ | $SS_T/(r-1)$ | $\sigma_\varepsilon^2+Q_T$ |
| G×T | $(s-1)(r-1)$ | $\sum\sum n_h(\bar{Y}_{h\bullet j}-\bar{Y}_{h\bullet\bullet}-\bar{Y}_{\bullet\bullet j}+\bar{Y}_{\bullet\bullet\bullet})^2$ | $SS_{G*T}/[(s-1)(r-1)]$ | $\sigma_\varepsilon^2+Q_{GT}$ |
| Subject(Group) | $n_{tot}-s$ | $r\sum\sum(\bar{Y}_{hi\bullet}-\bar{Y}_{h\bullet\bullet})^2$ | $SS_{S(G)}/(n_{tot}-s)$ | $\sigma_\varepsilon^2+r\sigma_b^2$ |
| Residual | $(n_{tot}-s)(r-1)$ | $\sum\sum\sum(Y_{hij}-\bar{Y}_{h\bullet j}-\bar{Y}_{hi\bullet}+\bar{Y}_{h\bullet\bullet})^2$ | $SS_R/[(n_{tot}-s)(r-1)]$ | $\sigma_\varepsilon^2$ |
| Total | $n_{tot}r-1$ | | | |

# Tests (based on model with sum-to-0 restrictions)

**Group×Time**

$H_0$: All $(\gamma\tau)_{hj} = 0$ (or $Q_{GT}=0$)

Use $F = MS_{GT}/MS_R$

**Group**

$H_0$: All $\gamma_h = 0$ (or $Q_G=0$)

Use $F = MS_G/MS_{S(G)}$

**Subject (i.e., Subject(Group))**

$H_0$: $\sigma_b^2 = 0$

Use $F = MS_{S(G)}/MS_R$

Estimating the ICC may be more informative than running a test for subject variance.

**Time**

$H_0$: All $\tau_j = 0$ (or $Q_T=0$)

Use $F = MS_T/MS_R$

- It is important to use the correct MSE (denominator of the *F*-statistic). We cannot use the usual MSE for all tests.

- In the 'split-plot' model there are two random terms, the random subject effect that expresses variability between subjects (whole-plot error), and the usual error term (sub-plot error).

- It makes sense that the test for *group* should use the $MS_{subjects(group)}$ in the denominator, since it is the between-subject measure of error, and the *group* test is based on comparisons between subjects.

- The tests for *time* and *time*group* uses the usual MSE, which involves differences at the subject-time level.

- More formally, we can use the expected mean squares as a guide to indicate which terms to use in the *F*-statistic. We look for *E*(MS) quantities such that the difference is only a term that involves the parameter being tested.

    - For example, the *E*(MS) for *group* is $\sigma_\varepsilon^2 + r\sigma_b^2 + Q_{\mathrm{G}}$ and the *E*(MS) for *subjects(group)* is $\sigma_\varepsilon^2 + r\sigma_b^2$, where $Q_{\mathrm{G}}$ involves sum-of-squared *group* effects.

    - Under $H_0$, $Q_{\mathrm{G}}=0$ and the *E*(MS) quantities are the same; in this case the *F* statistic has a central *F* distribution. Increasing values of the test statistic will support $H_A$: $Q_{\mathrm{G}}>0$ (unequal group means). This makes sense since the *E*(MS) in the numerator increases (as $Q_{\mathrm{G}}$ increase) while the *E*(MS) in the denominator stays the same.

```
proc glm data=uni_dogs; class id group time;
 model y = group time group*time id(group);
 random id(group) / test;run;
```

## Output summary

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 29 | 3819.001669 | 131.689713 | 190.93 | <.0001 |
| Error | 60 | 41.384513 | 0.689742 | | |
| Corrected Total | 89 | 3860.386182 | | | |

| R-Square | Coeff Var | Root MSE | y Mean |
|---|---|---|---|
| 0.989280 | 5.177222 | 0.830507 | 16.04156 |

This test uses incorrect denominator MS.

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| group | 2 | 113.950016 | 56.975008 | 82.60 | <.0001 |
| time | 4 | 122.575149 | 30.643787 | 44.43 | <.0001 |
| group*time | 8 | 73.807018 | 9.225877 | 13.38 | <.0001 |
| id(group) | 15 | 3508.669487 | 233.911299 | 339.13 | <.0001 |

| Source | Type III Expected Mean Square |
|---|---|
| group | Var(Error) + 5 Var(id(group)) + Q(group,group*time) |
| time | Var(Error) + Q(time,group*time) |
| group*time | Var(Error) + Q(group*time) |
| id(group) | Var(Error) + 5 Var(id(group)) |

Tests of Hypotheses for Mixed Model Analysis of Variance

Group comparisons involve between-subject differences. Thus, the "MSE" for the group test is $MS_{id(group)}$.

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| *group | 2 | 113.950016 | 56.975008 | 0.24 | 0.7869 |
| Error: MS(id(group)) | 15 | 3508.669487 | 233.911299 | | |
| *time | 4 | 122.575149 | 30.643787 | 44.43 | <.0001 |
| group*time | 8 | 73.807018 | 9.225877 | 13.38 | <.0001 |
| id(group) | 15 | 3508.669487 | 233.911299 | 339.13 | <.0001 |
| Error: MS(Error) | 60 | 41.384513 | 0.689742 | | |

* This test assumes one or more other fixed effects are zero.

Time and group*time comparisons involve between-time differences; the MSE for these tests is the usual MSE.

# <u>Notes</u>:

- The partial (or likelihood ratio) $F$-tests for standard GLMs and repeated measures ANOVA involve the ratio of two mean square (MS) quantities.

- The numerator MS can be expressed as $MS_{source}$ ('source' including the particular item(s) you are testing). This can be computed as $(SSE_{red} - SSE_{full})/s$, where $s$ is the difference between rank $\mathbf{X}$ for full and reduced models (i.e., difference in degrees of freedom between the two models).

- The full model has the source item included, while the reduced model does not. The denominator is the 'error' mean square, but in repeated measures ANOVA we need to make sure we are using the correct one, it may not necessarily be the standard MSE.

- Using the expected mean square quantities as a guide, the only difference between E(MS) for the denominator and E(MS) for the numerator should involve a quantity that can directly used to test the ‘source’.

- While F-tests for RM ANOVA are conducted by using sums of squares quantities, those using LMM methods (with PROC MIXED) are not. For the latter, the *F*-statistic is calculated as a function of the estimated parameters in the model; the test (p-value) is calibrated by selection of the denominator degrees of freedom, which will be discussed in more detail later.

- We discussed how the implied covariance structure for repeated measures outcomes for the random intercept model (1) is *compound symmetric*.

  - In RM ANOVA, $F$-tests will be accurate in the slightly more general case that variances and covariances must meet *sphericity* conditions.

  - Sphericity holds when $Var(Y_{ij} - Y_{ij'})$ is the same value for all $j \neq j'$. The compound symmetric structure is one that satisfies sphericity.

  - Goodness-of-fit tests to assess sphericity are easily obtained from statistical software such as SAS. If sphericity does not appear to hold, adjusted $F$-test can be employed (G-G and H-F), but they tend to be overly conservative.

- Since LMM methods can do what RM ANOVA can, and more (i.e., provide inference for a broader array of models), in practice I generally do not use RM ANOVA. However there may be a few instances when using RM ANOVA instead of LMM methods may be advantageous, which are discussed in the Classical methods notes.

- For other models, RM ANOVA will still yield the same results as LMM methods, however the calculations are a bit more complex (see course notes for more detail).

- For other complex models or data (e.g. missing data), RM ANOVA may not yield exactly equivalent results as LMM methods, but often they will be very close.

## *2.2.3 Contrasts*

### *2.2.3.1 Contrasts for time and group*time*

Orthogonal polynomial contrasts

See the Appendix to get the form of polynomial contrasts. (These are actually orthonormal polynomial contrasts; they can be rescaled if desired – i.e. to get integers.)

Example – consider an experiment with 2 groups and 3 times, with the 'means model'

$$Y_{hij} = \mu_{hj} + b_{i(h)} + \varepsilon_{hij}$$

where $\mu_{hj} = \mu + \gamma_h + \tau_j + (\gamma\tau)_{hj}$ , and *h*, *i* and *j* index group, subject and time, respectively. We can write $\boldsymbol{\mu}^t = (\mu_{11}\ \mu_{12}\ \mu_{13}\ \mu_{21}\ \mu_{22}\ \mu_{23})$.

## Test for time effect

- There are 2 degrees of freedom associated with the time effect; remember the test of main effect for time can be written as: $H_0$: $\bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 2} = \bar{\mu}_{\bullet 3}$. We learned in the GLM review that one way to get a full-rank model is to impose some constraints on the parameters.

- When the constraints are: $\sum_i \gamma_i = 0$; $\sum_j \tau_j = 0$; $(\gamma\tau)_{i\bullet} = 0$ for each $i$, $(\gamma\tau)_{\bullet j} = 0$ for each $j$, the test above can be equivalently written as

  $H_0$: $\tau_1 = \tau_2 = \tau_3$

  or

  $H_0$: $\mathbf{C\tau}=\mathbf{0}$, where $\mathbf{C}=\begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$ and $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)^t$.

- The sum of squares for the test above can also be broken down based on a set of orthogonal polynomial contrasts. The **C** matrix is: $\mathbf{C}=\begin{pmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix}$ (first row is linear contrast, second is quadratic).
  - The sums of squares for this test will be the same as for the main effect test of time (if the rows of **C** are considered in the same test).
  - If you want rows of **C** to be considered in the same test in SAS, you separate the rows by commas. To get tests for the linear and quadratic components, we simply write two contrasts statements to carry this out. I.e., we test $H_0$: $\mathbf{c}_1^t\boldsymbol{\tau}=0$ for linear and $H_0$: $\mathbf{c}_2^t\boldsymbol{\tau}=0$ for quadratic, where $\mathbf{c}_i^t$, $i$=1,2 are rows of **C**.
  - If you want the rows of **C** (or groups of rows of **C**) to be considered in different tests, write different contrast statements and separate them by semicolons.

## Test for group×time effect

There are 2×1=2 degrees of freedom for the group×time effect. The test for group×time can be written as:

$$H_0: \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}$$

or

$$H_0: (\gamma\tau)_{11} - (\gamma\tau)_{21} = (\gamma\tau)_{12} - (\gamma\tau)_{22} = (\gamma\tau)_{13} - (\gamma\tau)_{23}$$

There are 2 degrees of freedom for this test; can you write the **C** matrix?

We can break the SS for *group*time* into pieces if we have orthogonal contrasts, similar to what we did for *time*.

A set of orthogonal polynomial contrasts for the group*time effect is:

$$\begin{pmatrix} -1 & 0 & 1 & 1 & 0 & -1 \\ 1 & -2 & 1 & -1 & 2 & -1 \end{pmatrix} = \begin{pmatrix} \mathbf{c}_1^t \\ \mathbf{c}_2^t \end{pmatrix}$$

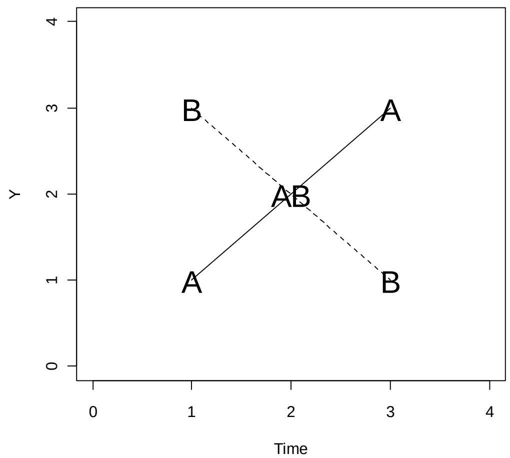
The test associated with the first row is

H0: $\mathbf{c}_1^t\boldsymbol{\mu} = 0$ or H0: $-\mu_{11} + \mu_{13} = -\mu_{21} + \mu_{23}$.

I.e., the test addresses the question: Is the linear trend for Group A different than the linear trend for Group B? Similar for the second row, but with quadratic…

# Illustration of linear-by-linear and quadratic-by-quadratic interaction.

Presence of linear (or ‘linear × linear’) interaction, but no quadratic interaction.



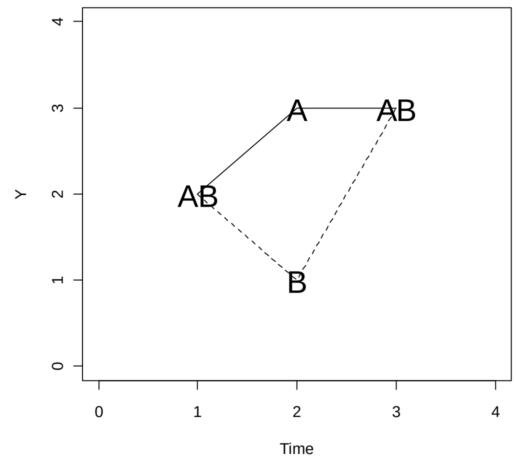

A L: -1(1) + 0(2) + 1(3) = 2

Q: 1(1) -2(2) + 1(3) = 0

B L: -1(3) + 0(2) + 1(1) = -2

Q: 1(3) -2(2) + 1(1) = 0

Presence of quadratic (or ‘quadratic × quadratic’) interaction, but no linear interaction.




A L: -1(2) + 0(3) + 1(3) = 1

Q: 1(2) – 2(3) +1(3) = -1

B L: -1(2) + 0(1) + 1(3) = 1

Q: 1(2) – 2(1) + 1(3) = 3

## Contrasts using PROC MIXED and the dog data

- There are different sets of contrasts that may be of interest. If we were interested in how effects over time compare with the baseline value, we could add the ‘contrast(1)’ option in the REPEATED statement to obtain the tests; these four contrasts are not orthogonal.

- Another possibility is to consider orthogonal polynomial contrasts. There are 3 groups and 5 times (the actual experiment involved more times). Thus the orthogonal polynomial contrasts can be written up to the 4 degree (quartic) term. Each polynomial component has $g$–1 = 3–1 = 2 degrees of freedom. You can verify that the sums of squares for the set of contrasts for *time* add up to $SS_{time}$, and similarly, that the sums of squares for the set of contrasts for *group*×*time* add up to $SS_{group \times time}$. SAS code and output for the repeated measures ANOVA follow.

```
proc mixed data=uni_dogs; class id group time;

 model y = group time group*time / solution;
 random intercept / subject=id(group);
 contrast 'linear' time -2 -1 0 1 2;
 contrast 'quadratic' time 2 -1 -2 -1 2;
 contrast 'cubic' time -1 2 0 -2 1;
 contrast 'quartic' time 1 -4 6 -4 1;
 contrast 'lxl' group*time -2 -1 0 1 2 2 1 0 -1 -2 0 0 0 0 0,
                 group*time -2 -1 0 1 2 0 0 0 0 0 2 1 0 -1 -2;
 contrast 'qxq' group*time 2 -1 -2 -1 2 -2 1 2 1 -2 0 0 0 0 0,
                 group*time 2 -1 -2 -1 2 0 0 0 0 0 -2 1 2 1 -2;
 contrast 'cxc' group*time -1 2 0 -2 1 1 -2 0 2 -1 0 0 0 0 0,
                 group*time -1 2 0 -2 1 0 0 0 0 0 1 -2 0 2 -1;
 contrast '4x4' group*time 1 -4 6 -4 1 -1 4 -6 4 -1 0 0 0 0 0,
                 group*time 1 -4 6 -4 1 0 0 0 0 0 -1 4 -6 4 -1;
 lsmeans group*time; run;
```

SAS output for contrasts:

| Label | Num DF | Den DF | F Value | Pr > F | Label | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|---|---|---|---|---|
| linear | 1 | 60 | 4.63 | 0.0355 | lxl | 2 | 60 | 1.46 | 0.2405 |
| quadratic | 1 | 60 | 111.46 | <.0001 | qxq | 2 | 60 | 37.52 | <.0001 |
| cubic | 1 | 60 | 55.61 | <.0001 | cxc | 2 | 60 | 14.23 | <.0001 |
| quartic | 1 | 60 | 6.01 | 0.0171 | 4x4 | 2 | 60 | 0.29 | 0.7489 |

# Comparison using PROC GLM and the dog data

```
proc glm data=uni_dogs; class id group time;
 model y = group time group*time id(group);
 random id(group) / test;
 contrast 'linear' time -2 -1 0 1 2;
 contrast 'quadratic' time 2 -1 -2 -1 2;
 contrast 'cubic' time -1 2 0 -2 1;
 contrast 'quartic' time 1 -4 6 -4 1;
  contrast 'lxl' group*time -2 -1 0 1 2 2 1 0 -1 -2 0 0 0 0 0,
                 group*time -2 -1 0 1 2 0 0 0 0 0 2 1 0 -1 -2;
  contrast 'qxq' group*time 2 -1 -2 -1 2 -2 1 2 1 -2 0 0 0 0 0,
                 group*time 2 -1 -2 -1 2 0 0 0 0 0 -2 1 2 1 -2;
  contrast 'cxc' group*time -1 2 0 -2 1 1 -2 0 2 -1 0 0 0 0 0,
                 group*time -1 2 0 -2 1 0 0 0 0 0 1 -2 0 2 -1;
  contrast '4x4' group*time 1 -4 6 -4 1 -1 4 -6 4 -1 0 0 0 0 0,
                 group*time 1 -4 6 -4 1 0 0 0 0 0 -1 4 -6 4 -1; run;
```

| Contrast | DF | Contrast SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| linear | 1 | 3.19200500 | 3.19200500 | 4.63 | 0.0355 |
| quadratic | 1 | 76.88038135 | 76.88038135 | 111.46 | <.0001 |
| cubic | 1 | 38.35526722 | 38.35526722 | 55.61 | <.0001 |
| quartic | 1 | 4.14749532 | 4.14749532 | 6.01 | 0.0171 |
| lxl | 2 | 2.01316333 | 1.00658167 | 1.46 | 0.2405 |
| qxq | 2 | 51.76348651 | 25.88174325 | 37.52 | <.0001 |
| cxc | 2 | 19.62958111 | 9.81479056 | 14.23 | <.0001 |
| 4x4 | 2 | 0.40078683 | 0.20039341 | 0.29 | 0.7489 |

Add up to $SS_{time}$.

Add up to $SS_{group*time}$.

- One of the differences between the contrast results here and those previous with the PROC MIXED output (contrasts will be shown ahead) is that here we have sums of squares quantities, whereas for MIXED, the F-statistic that is a function of the estimated parameters in the model is given, along with numerator and denominator DF, and the p-value; there are no sums of squares or ANOVA table with MIXED. Later we will discuss how the denominator DF is used to calibrate the correct p-value for the test statistic for contrasts in LMM methods.

### *2.2.3.2 Contrasts for group*

Although fairly straightforward for mixed models, these contrasts are a bit more complicated for RM ANOVA models. See the course notes for more detail.

## *3 Simple random effects and mixed effects models, ICC and reliability*

- There are different types of ICC that can be estimated. Previously, we considered an ICC that makes sense to use when repeated measures are taken over time (a fixed effect), and subject is modeled as a random effect. Here, we will compare two types of ICC.

- Consider a motivating example where raters (or judges) are asked to give scores for individuals. *From NYU website:* ***When judges subjectively evaluate phenomena, measurement error is often found in their assessment. The careful and responsible researcher will assess this error before applying their ratings to the study of any targeted phenomena.*** **The calculation of ICC's will give us insight into issues of rater reliability and measurement error for the given application. For more detail on the generic data set and measures of ICC, see** Shrout, P.E. & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability, *Psychological Bulletin*, Vol. 86, 2, 420-428.

The data appears as follows:

| | Rater | | | |
|---|---|---|---|---|
| **Subject** | **1** | **2** | **3** | **4** |
| **1** | 7 | 8 | 3 | 5 |
| **2** | 2 | 4 | 4 | 1 |
| **3** | 1 | 2 | 6 | 1 |
| **4** | 5 | 5 | 7 | 2 |
| **5** | 8 | 9 | 5 | 6 |
| **6** | 9 | 10 | 6 | 7 |

- ICC near 1 ⇒ stronger reliability; ICC near 0 ⇒ weak reliability. (Some have suggested that 0.75 is a reasonable cut-point for 'good' reliability, but certainly it should depend on the application at hand.)

- We can develop a mixed model that has a random term for subjects, and either a random or fixed term for judges. Whether we use a random or fixed term for raters depends on what type of inference we wish to make. If we are only interested in inference on the given raters, then we can treat judges as a fixed term. If the judges are drawn from a larger population of judges and we are interested in all judges, then we may treat judges as a random term.

```
data rater; input subject rater y @@; datalines;
1 1 7 1 2 8 1 3 3 1 4 5 2 1 2 2 2 4 2 3 4 2 4 1 3 1 1 3 2 2 3 3 6 3 4 1 4 1 5 4
2 5 4 3 7 4 4 2 5 1 8 5 2 9 5 3 5 5 4 6 6 1 9 6 2 10 6 3 6 6 4 7
;
```

```
*Approach 1 - random judges;
proc mixed data=rater;
 class subject rater; model y=;
 random subject rater;
 ods output covparms=cov1; run;
proc transpose data=cov1 out=cov_out1
 prefix=sigma2_; id CovParm; run;
data cov_out1; set cov_out1;
icc_app1=sigma2_subject/
  (sigma2_subject+sigma2_rater+
   sigma2_residual); run;
proc print data=cov_out1;
 var icc_app1; run;
```

Covariance Parameter Estimates

| Cov Parm | Estimate | |
|---|---|---|
| subject | 4.1444 | ($\hat{\sigma}_S^2$) |
| rater | 0.6611 | ($\hat{\sigma}_R^2$) |
| Residual | 3.2972 | ($\hat{\sigma}_\varepsilon^2$) |

| Obs | icc_case_2 | |
|---|---|---|
| 1 | 0.51148 | ($\hat{\sigma}_S^2/[\hat{\sigma}_S^2+\hat{\sigma}_R^2+\hat{\sigma}_\varepsilon^2]$) |

```
*Approach 2 - fixed judges;
proc mixed data=rater;
 class subject rater; model y=rater;
 random subject; ods output covparms=cov2;
run;
proc transpose data=cov2 out=cov_out2
 prefix=sigma2_; id CovParm; run;
data cov_out2; set cov_out2;
icc_app2=sigma2_subject/
 (sigma2_subject+sigma2_residual); run;
proc print data=cov_out2; var icc_app2;run;
```

Covariance Parameter Estimates

| Cov Parm | Estimate | |
|---|---|---|
| subject | 4.1444 | ($\hat{\sigma}_S^2$) |
| Residual | 3.2972 | ($\hat{\sigma}_\varepsilon^2$) |

Type 3 Tests of Fixed Effects

| Effect | NumDF | DenDF | F Value | Pr > F |
|---|---|---|---|---|
| rater | 3 | 15 | 2.20 | 0.1300 |

| Obs | icc_case_3 | |
|---|---|---|
| 1 | 0.55692 | ($\hat{\sigma}_S^2/[\hat{\sigma}_S^2+\hat{\sigma}_\varepsilon^2]$) |

## **Approach 1**:

$$Y_{ij} = \mu + b_{iS} + b_{jR} + \varepsilon_{ij} \,,$$

where $i$ denotes subject and $j$ denotes judge;
$b_{iS} \sim N(0, \sigma_S^2)$, $b_{jR} \sim N(0, \sigma_R^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, all independent.

This is random effects model, which is a special type of mixed model.

## **Approach 2**:

$$Y_{ij} = \mu + b_{iS} + \kappa_j + \varepsilon_{ij} \,,$$

where $i$ denotes subject and $j$ denotes judge;
$b_{iS} \sim N(0, \sigma_S^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, all independent;
$\kappa_j$ are fixed effects for judge.

## <u>Summary and notes</u>:

- The estimated proportion of variance in a score $Y_{ij}$ due to between-subject variance is 0.51 if we are making inferences for the larger population of judges; it is 0.56 is we are making inferences for one of the given judges. These values indicate weak to moderate reliability in the judges' ratings.
- Shrout and Fleiss also consider a random interaction term (rater*subject), however including this term in the models and ICC equations does not affect the ICC estimates for the specific analysis presented above.
- One can compute ICC estimates for single ratings or average ratings (see Shrout and Fleiss). The ICC calculations above are for single ratings.
- Note that we could have obtained similar estimates using RM ANOVA via PROC GLM. One advantage of using PROC MIXED to fit the random effects (upper left) or mixed (upper right) model is that you don't have to work with MS quantities, but you are directly given variance component estimates.
- The ICC estimates we calculated before were for mixed models, treating time as fixed and subject as random, like 'Approach 2' above.