Notes on Maxwell & Delaney

PSYCH 710

November 28, 2023

11 single-factor within-subject designs

The ideas of blocking and the analysis of covariance were discussed in Chapter 9. By explicitly incorporating individual differences among subjects into the design and analysis of experiments, blocking and ancova can significantly reduce error variance and, consequently, increase the power of our statistical tests. One can think of **within-subjects designs** as the ultimate extension of the blocking approach: each block consists of a single subject. If certain assumptions are met, this type of design can significantly increase the sensitivity of an experiment.

Up to now we have discussed experimental designs in which different treatments have been administered to different groups of subjects. Such designs are called *between-subjects designs*. In a within-subject design, each subject receives *all* treatments. This type of design is sometimes referred to as a *repeated-measures design*, because multiple measures are obtained on each subject. However, some statisticians prefer to use repeated-measures to refer only to designs in which the same dependent measure is taken at multiple times.

There are, of course, a wide range of within-subjects experimental designs. I will focus on designs in which the multiple measures obtained from each subject are all of the same type. I will *not* address designs that obtain measures that differ qualitatively. For example, we could imagine a situation where each subject provides dependent measures of response time and response accuracy in a visual detection task. Such data should be analyzed using a multivariate approach described in chapters 13 and 14 in Maxwell and Delaney (2004), and will not be considered here.

The basic approach we will take to analyze within-subject data is to treat subjects as an experimental factor. However, there are several aspects of within-subjects designs that make the analysis more difficult. The first is that, in most within-subject designs, there is only one observation per cell (i.e., subject-treatment combination). That is to say, each subject is tested once, and only once, in each experimental condition. This means that we cannot measure "within-cell" error as we did in factorial experiments. The second distinguishing characteristic of this design is that the subjects factor is a **random factor**. Most experimental treatments are **fixed factors**: if the experiment was repeated, we would

use the same levels on the experimental variables. Subjects, however, is a random factor: If the experiment was repeated, we would use different "levels" on the subjects factor (i.e., we would get different subjects). A random factor adds variability to our measurements and, as we shall see, alters the analysis. Finally, unlike in previous designs, the observations in within-subjects designs are correlated, not independent. In other words, the residuals of our model are likely to have structure, and such structure will affect our analysis.

One more thing. Traditional methods for analyzing within-subject designs — the methods described here — are applicable only to *balanced* designs. In some ways, this is a major limitation of this approach. For example, the requirement of balanced data means that we must discard all of the data from a subject even if we lack a measurement in only one condition. Nevertheless, all of the following analyses assume that the design is balanced. Also, we will only consider designs with a single, fixed within-subject factor.

11.1 linear models

We start with the model

$$Y_{ij} = \mu + \alpha_j + \pi_i + (\pi\alpha)_{ij} + \epsilon_{ij} \tag{1}$$

where Y_{ij} is the score from subject i in condition j, μ is the intercept, α_j is the effect associated with condition j, π_i is the effect associated with subject i, $(\pi\alpha)_{ij}$ is the effect of the interaction between subject i and condition j, and ϵ_{ij} is the error for subject i in condition j.

We have a problem. The model in Equation 1 has too many parameters. Consider an experiment that has $n = 8$ subjects and $a = 4$ treatments for a total of 32 observations. Equation 1 includes an intercept (i.e., μ), $3 = a - 1$ treatment effects, $7 = n - 1$ subject effects, and $21 = (a-1)(n-1)$ interaction effects, which add up to a total of $1+3+7+21 = 32$ free parameters, which equals the number of observations. Therefore, the error terms in Equation 1 will be zero. The problem is that we have included an interaction term in our model, but there is no way to estimate that parameter from our data. Stated another way, there is no way to determine if the difference between the observation in each cell and the prediction based on the treatment and subject effects is due to an interaction or error. Therefore, I will simplify the model by dropping the interaction term

$$Y_{ij} = \mu + \alpha_j + \pi_i + \epsilon_{ij} \tag{2}$$

This interaction-less model is the full model for a design that consists of a single within-subjects factor and that has only a single measurement per cell. Note that the interaction effects do not disappear. Instead, they are incorporated into the error term (i.e., the residuals). By doing this, we are in some sense equating the *treatment* $\times$ *subjects* interaction and error. I will return to this point below.

The null hypothesis being tested is

$$H0 \ : \ \alpha_1 = \alpha_2 = \cdots = \alpha_j = 0 \tag{3}$$

so a restricted model is

$$Y_{ij} = \mu + \pi_i + \epsilon_{ij} \tag{4}$$

The F test is computed the usual way

$$F = \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F} \tag{5}$$

where E_F and E_R are $SS_{residuals}$ from the full and reduced models, respectively. The degrees of freedom for the residuals in the full model are worth examining. The degrees of freedom equal the number of observations minus the number of parameters in the model. You can verify that $df_F = (n-1)(a-1)$. Thus, $df_{residuals}$ for the model in Equation 2 equals the degrees of freedom for the (deleted) *treatment* $\times$ *subjects* interaction term. Once again, we see that there is a connection between the *treatment* $\times$ *subjects* interaction and error. The degrees of freedom for the reduced model is $n(a-1)$, so $df_R - df_F = a - 1$.

11.2 model coefficients

When the model in Equation 2 is fit to data, the best-fitting (least-squares) coefficients are:

$$\begin{aligned} \hat{\mu} &= \bar{Y}_{\cdot\cdot} \\ \hat{\alpha}_j &= \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \\ \hat{\pi}_i &= \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \end{aligned}$$

The intercept, μ, is the mean of all scores. The effect of treatment j (α_j) is the mean of the scores in treatment j, averaged across all subjects, minus the grand mean. The effect of subject i (π_i) is the mean score of subject i, averaged across treatments, minus the grand mean.

11.3 expected mean squares

Consider, again, Equation 2. Let's imagine that we fit this model to many sets of data and calculated the *average* parameter values. It can be shown[1] that the average, or expected, values of the parameters are the ones listed in Table 1. Notice that the expected value of $MS_{residuals}$ is the ths sum of the population error variance plus a term related to the *treatment* $\times$ *subjects* interaction. This result should not be surprising, because Equation 2 was created essentially by folding the interaction term into error. It is important for you to realize that the residuals and the interaction term are perfectly confounded in this

Table 1: Expected Mean Squares for a design that has one, fixed within-subjects factor (`treatment`).

Effect	Type	E(Mean Square)
treatment	fixed	$\sigma_e^2 + n\sum_{j=1}^{a}\alpha_j^2/(a-1)$
subjects	random	$\sigma_e^2 + a\sigma_\pi^2$
residuals		$\sigma_e^2 + \sigma_{\pi\alpha}^2$

design: The residual term *is* the interaction and *vice versa*. For this reason, some statistics packages label the residual term as *Treatment* × *Subjects*.

Table 1 shows that $MS_{treatment}$ also is influenced by the interaction term. Because the interaction contributes equally to $MS_{treatment}$ and $MS_{residuals}$, it is reasonable to conclude that $\sigma_\alpha^2 > 0$ when $MS_{treatment} > MS_{residuals}$. In other words, a comparison of the models shown in Equations 2 and 4 provides a reasonable test of the null hypothesis. Notice, however, it is *not* reasonable to evaluate the effect of subject by dividing $MS_{subjects}$ by $MS_{residuals}$. In fact, there is no unambiguous F test for the effect of subjects. This lack of a test is not really a problem, however, because rarely are we interested in showing that subjects differ beyond what is expected by chance.

11.4 sphericity

A one-way, between-subjects ANOVA assumes that error variance is constant across conditions. A similar assumption is made in the one-way within-subjects ANOVA: Specifically, the assumption is that the error variances for all of the dependent variables are equal. Another fundamental assumption in the between-subjects ANOVA is that the errors – i.e., the ϵ_{ijk}'s – are independent. This assumption is reasonable in between-subjects designs that randomly assign subjects to conditions, but it is less reasonable in within-subjects studies. Indeed, it is reasonable to expect that errors for a given subject will be correlated, to some degree, across conditions. Therefore, the independent-errors assumption needs to be relaxed if we are to conduct a reasonable analysis of data collected in within-subjects experiments. Instead of assuming independence, we will assume that the errors exhibit a specific form of dependency, or correlation. In particular, the assumption is that all of the covariances[2] between dependent variables are equal. This combination of assumptions – equal variances for all dependent variables, and equal covariances between each pair of dependent variables – is known as the assumption of **compound symmetry**. The F calculated in Equation 5 is distributed as an F statistic with $df = [(a-1), (n-1)(a-1)]$

[1] See Kirk (1995) for a derivation of the expected mean squares.

[2] Given random variables X and Y with expected values $E(X)$ and $E(Y)$, the covariance of X and Y is $E(XY) - E(X)E(Y)$.

if the dependent variables exhibit compound symmetry[3].

Although compound symmetry is sufficent for the F to be distributed correctly, it is not a necessary condition. Huynh and Feldt (1970) and Rouanet and Lépine (1970) showed that the F value is distributed as an F statistic if the variances of all *differences* among the dependent variables have the same variance:

$$\sigma^2_{Y_j - Y_k} = \sigma^2_j + \sigma^2_k - 2\sigma_{jk} \ (j \neq k) \tag{6}$$

where σ_{jk} is the covariance between dependent measures j and k. When this condition is met, the variance-covariance matrix of the dependent variables is said to be spherical, so this assumption is known as the **sphericity assumption**. It is important to remember that the sphericity assumption applies to all tests of a within-subject factor. Also, **the sphericity assumption is necessarily true whenever the F test for the within-subject factor has one degree of freedom in the numerator**.

Unfortunately, the sphericity assumption often is not valid. In such cases, the F value calculated in Equation 5 will not be distributed as F with the expected degrees of freedom. However, it will be distributed *approximately* as an F statistic with lower degrees of freedom (Box, 1954). Therefore, one strategy for dealing with violations of the sphericity assumption is to adjust our degrees of freedom before evaluating F. The modified degrees of freedom are $\epsilon(a-1)$ and $\epsilon(n-1)(a-1)$, where ϵ is a number indicating the degree to which the sphericity assumption is violated. When sphericity exists, $\epsilon = 1$; otherwise, ϵ is less than one with a minimum value of $1/(a-1)$.

In practice, of course, ϵ is not known and so must be estimated from the data. One strategy is to simply assume that ϵ is at its minimum value. When $\epsilon = 1/(a-1)$, $df_1 = 1$ and $df_2 = (n-1)$, and a within-subjects F test using these degrees of freedom is referred to as the Geisser-Greenhouse **conservative F test**, or as using the **lower-bound adjustment** of the degrees of freedom (Geisser and Greenhouse, 1958) . Alternatively, we can derive numerical estimates of ϵ from the variance-covariance matrix of our dependent measures. Geisser and Greenhouse proposed one such estimate, $\hat{\epsilon}$. The degrees of freedom for the **adjusted F test** are $df_1 = \hat{\epsilon}(a-1)$ and $df_2 = \hat{\epsilon}(n-1)(a-1)$. The Geisser-Greenhouse adjustment controls Type I error but is more powerful than the lower-bound adjustment. It is, however, slightly conservative as the true population ϵ approaches one, and therefore Huynh and Feldt (1976) proposed a different adjustment based on their estimate of epsilon denoted as $\tilde{\epsilon}$. The Huynh-Feldt procedure yields adjusted degrees of freedom $df_1 = \tilde{\epsilon}(a-1)$ and $df_2 = \tilde{\epsilon}(n-1)(a-1)$. It, too, is more powerful than the lower-bound adjustment, and it is slightly less conservative than the Geisser-Greenhouse adjustment. Both the Geisser-Greenhouse and Huynh-Feldt adjustments are acceptable procedures, although the former probably does a slightly better job at controlling Type I error rates. The conservative F test, as its name implies, is the most conservative test available: if the conservative F test is significant, than tests based on $\hat{\epsilon}$ and $\tilde{\epsilon}$ will be significant, too.

[3]Note that the same assumption is made for data collected in between-subjects designs, except that the standard assumption is that the covariances are all zero.

11.5 R example

In this section I will analyze the data presented in Table 11.5 in Maxwell and Delaney (2004). The data are from a fictitious experiment that measured cognitive ability in 12 children at 30, 36, 42, and 48 months of age. First, I read the data file, which contains the data frames `mw115` and `mw115L`.

```
load(url("http://pnb.mcmaster.ca/bennett/psy710/datasets/mw11_5.rda") )
summary(mw115)
```

```
##      age.30          age.36        age.42        age.48         subj
##  Min.   : 84.0   Min.   : 84   Min.   : 92   Min.   : 88   s1     :1
##  1st Qu.: 94.5   1st Qu.: 99   1st Qu.:101   1st Qu.:104   s2     :1
##  Median :104.0   Median :107   Median :106   Median :112   s3     :1
##  Mean   :103.0   Mean   :107   Mean   :110   Mean   :112   s4     :1
##  3rd Qu.:110.8   3rd Qu.:117   3rd Qu.:124   3rd Qu.:124   s5     :1
##  Max.   :129.0   Max.   :128   Max.   :132   Max.   :133   s6     :1
##                                                           (Other):6
```

```
summary(mw115L)
```

```
##       subj      age          score
##  s1     : 4   a30:12   Min.   : 84.0
##  s2     : 4   a36:12   1st Qu.: 98.2
##  s3     : 4   a42:12   Median :107.0
##  s4     : 4   a48:12   Mean   :108.0
##  s5     : 4            3rd Qu.:117.2
##  s6     : 4            Max.   :133.0
##  (Other):24
```

Each data frame contains four measures taken on each of 12 subjects: `mw11` is in the so-called wide format, and `mw11L` is in the so-called long format. Each format is useful in different contexts.

11.5.1 Anova (`car` package)

To analyze these data, we first use `lm()` to create a `multivariate model` of the between-subjects effects. The four dependent variables – `age.30`, `age.36`, `age.42` and `age.48` – are combined into a single matrix using the `cbind()` command. There is no between-subjects variable in this experiment, so the model contains only an intercept, which in R is represented as 1:

```
options(contrasts=c("contr.sum","contr.poly"))
# multivariate linear model:
# include only intercept in this case:
mw115.mlm <- lm(cbind(age.30,age.36,age.42,age.48) ~ 1, data=mw115)
```

Now I have to inform R about the nature of the within-subjects design. First, I create a factor that contains the levels within-subjects variable:

```
(age<-factor(x= c("a30","a36","a42","a48"),ordered=FALSE) )


## [1] a30 a36 a42 a48
## Levels: a30 a36 a42 a48
```

Note that the order of the levels is the same as the order of my model's multivariate dependent variable. Next, I have to specify the within-subjects design with a formula of the form $Y \sim myFactor$, or $Y \sim A+B+A:B$. There is only one within-subjects variable, *age*, so the appropriate formula for the current experiment is $Y \sim age$. Finally, I use the `Anova` command in the `car` package to convert our multivariate R object, `mw115.mlm`, into an anova object. In the next command, note how I pass information about the within-subjects factor in the `idata` parameter, and the within-subjects design in the `idesign` parameter:

```
#install.packages("car")
library(car)
mw115.aov <- Anova(mw115.mlm,idata=data.frame(age),idesign=~age,type="III")
```

Information about the within-subjects design was specified by a *one-sided* model formula, $\sim age$. The `type` parameter is used to tell `Anova` to calculate Type III sums of squares. The fact that our data are balanced (and that we do not have a between-subjects factor) means that Type II and III sums of squares will be equivalent, and therefore the `type` parameter is superfluous in this case. Finally, we can construct an anova table with the `summary` command. Setting `multivariate` to false suppresses the print out of multivariate tests:

```
summary(mw115.aov,multivariate=FALSE)


##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##             Sum Sq num Df Error SS den Df F value  Pr(>F)
## (Intercept) 559872      1     6624     11  929.74 5.6e-12 ***
## age            552      3     2006     33    3.03   0.043 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Mauchly Tests for Sphericity
##
##     Test statistic p-value
## age          0.243  0.0177
##
##
## Greenhouse-Geisser and Huynh-Feldt Corrections
##  for Departure from Sphericity
##
##     GG eps Pr(>F[GG])
## age   0.61      0.075 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     HF eps Pr(>F[HF])
## age 0.7249    0.06354
```

Several pieces of information are printed in the summary. The first part shows the standard ANOVA table: the p value for `age` is the one calculated if we assume that the sphericity assumption is valid. The second part shows the results of the Mauchly test for sphericity: the significant p value indicates that the deviation from sphericity is significant. The Mauchly test has been criticized for having low power, and therefore you might want to use a liberal Type I error rate (e.g., $\alpha = 0.1$) when using it to evaluate sphericity. The final part of the output show $\hat{\epsilon}$ and $\tilde{\epsilon}$, as well as the corrected p values for the test of a main effect of `age`. Note that neither p value is significant. In a paper, you could report this result as follows:

> [At the beginning of your Results section…] Statistical analyses were done with R (R Development Core Team, 2008). For within-subject tests of more than 1 degree-of-freedom, the Huynh-Feldt estimate of sphericity ($\tilde{\epsilon}$) was used to adjust p values of F tests conducted on within-subject variables (Maxwell and Delaney, 2004). [And later, when reporting the result of this analysis…] The effect of Age was not significant, $F(3, 33) = 3.027$, $\tilde{\epsilon} = 0.72$, $p = 0.063$.

11.5.2 `aov`, `aov_car`, & `Error`

The ANOVA table produced by `Anova` does not include information about subjects. Normally this is not a problem because there is no F test that can be done to evaluate the effect of subject. However, the complete table is useful on those occasions when we want to compute the variance component for subjects. The following code shows how to produce such a table using the long-format data frame and the `aov` command. Note how we

designate `subj` as a random factor and `age` as a within-subject factor with the `Error` term in our formula.

```
options(contrasts=c("contr.sum","contr.poly"))
mw115L.aov.01 <- aov(score~1+age+Error(subj/age),data=mw115L)
summary(mw115L.aov.01) # p values assume sphericity
```

```
##
## Error: subj
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 11   6624     602
##
## Error: subj:age
##           Df Sum Sq Mean Sq F value Pr(>F)
## age        3    552   184.0    3.03  0.043 *
## Residuals 33   2006    60.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p value for age assumes sphericity. Note that no F test is performed on `subj`.

Next, we use `aov_car` in the `afex` package. The model specification is the same as in `aov`; however, the summary table includes the Mauchly test and adjusted p values.

```
library(afex)
options("contrasts")
```

```
## $contrasts
## [1] "contr.sum"  "contr.poly"
```

```
mw115L.aov.car.01 <- aov_car(score~1+age+Error(subj/age),data=mw115L)
summary(mw115L.aov.car.01)
```

```
##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##             Sum Sq num Df Error SS den Df F value  Pr(>F)
## (Intercept) 559872      1     6624     11  929.74 5.6e-12 ***
## age            552      3     2006     33    3.03   0.043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Mauchly Tests for Sphericity
##
```

```
##     Test statistic p-value
## age          0.243  0.0177
##
##
## Greenhouse-Geisser and Huynh-Feldt Corrections
##  for Departure from Sphericity
##
##     GG eps Pr(>F[GG])
## age   0.61      0.075 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     HF eps Pr(>F[HF])
## age 0.7249    0.06354
```

11.5.3 `lmer` & `lme`

Finally, we can analyze the data using `lmer` and `lme` in the `lme4` and `nlme4` packages. These procedures compute best-fitting coefficients using maximum likelihood or restricted (or residual) maximum likelihood (REML), and differ in several fundamental ways from the ANOVA models described previously. As we will see, it is possible to construct mixed models that allow for by-subject adjustment of the intercept and fixed effect, and that include allow for general forms of the residual variance-covariance matrix. However, the mixed-effects model shown below includes a random effect that (like the ANOVA models) only allows for a by-subject adjustment of the intercept. This is the simplest mixed model for these data, and it assumes that the residuals on the various levels of the within-subjects factor are independent and distributed normally with a constant variance. Hence, the ANOVA table includes a p value for the effect of `age` that assumes sphericity. The random effect `subj` is evaluated with `ranova`.

```
library(lmerTest)
mw115L.lmer.01 <- lmer(score~age+(1|subj),data=mw115L)
anova(mw115L.lmer.01) # fixed effects


## Type III Analysis of Variance Table with Satterthwaite's method
##     Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
## age    552     184     3    33    3.03  0.043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


ranova(mw115L.lmer.01) # random effects
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## score ~ age + (1 | subj)
##            npar logLik AIC  LRT Df Pr(>Chisq)
## <none>        6   -172 356
## (1 | subj)    5   -185 380 26.3  1    2.9e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To evaluate the fixed effect without assuming sphericity or adjusting p values, we can compare our model to one that does not include `age`. The significant chi-square test suggests that `age` accounts for a significant portion of the variance and therefore we should not remove it from the model.

```
mw115L.lmer.00 <- lmer(score~1+(1|subj),data=mw115L) # intercept only
anova(mw115L.lmer.00,mw115L.lmer.01) # compare models


## Data: mw115L
## Models:
## mw115L.lmer.00: score ~ 1 + (1 | subj)
## mw115L.lmer.01: score ~ age + (1 | subj)
##                npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
## mw115L.lmer.00    3 371 377   -183      365
## mw115L.lmer.01    6 369 380   -178      357  8.75  3      0.033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The mixed model also can be fit to the data using `lme` in the `nlme` package. One advantage of using `lme` is that it allows you to define the structure of the within-subject residuals. The following code constructs three models that differ in terms of that structure. The first treats the residuals of the various levels of `age` as indendendent (i.e., the correlation between levels is zero). The second model allows the levels to be correlated, but the correlation is constant. The third model puts no constraints on the correlations between levels. All of the models assume that the residuals in each level are distributed normally with a constant variance. Note that the value of 0.5 in the following code is used to initialize the parameters and that other values yield the same results.

```
library(nlme)
# assume independence:
cog.nlme.00 <- lme(score~age,data=mw115L,
                random=~1|subj,
                correlation=NULL)
```

```
# assume compound symmetry:
cog.nlme.01 <- lme(score~age,data=mw115L,
                   random=~1|subj,
                   correlation=corCompSymm(value=0.5,form=~1|subj))
# no constraints on between-level correlations:
cog.nlme.02 <- lme(score~age,data=mw115L,
                   random=~1|subj,
                   correlation=corSymm(value=c(.5,.5,.5,.5,.5,.5),form=~1|subj))
```

The model's correlation structure is listed by the **summary** command, which is shown below only for the third model. You can see that the correlation is greatest between similar ages (e.g., 30 and 36 months) and declines as the ages become separated in time.

```
summary(cog.nlme.02) # no constraints
```

```
## Linear mixed-effects model fit by REML
##   Data: mw115L
##     AIC BIC logLik
##   352.6 374 -164.3
##
## Random effects:
##  Formula: ~1 | subj
##         (Intercept) Residual
## StdDev:       10.53    9.957
##
## Correlation Structure: General
##  Formula: ~1 | subj
##  Parameter estimate(s):
##  Correlation:
##   1     2     3
## 2 0.601
## 3 0.391 0.579
## 4 0.226 0.050 0.718
## Fixed effects:  score ~ age
##             Value Std.Error DF t-value p-value
## (Intercept)   108     3.736 33  28.910  0.0000
## age1           -5     1.953 33  -2.561  0.0152
## age2           -1     1.940 33  -0.515  0.6097
## age3            2     1.368 33   1.462  0.1531
##  Correlation:
##      (Intr) age1   age2
## age1 -0.018
## age2 -0.015  0.131
## age3  0.164 -0.820 -0.248
##
## Standardized Within-Group Residuals:
##     Min      Q1     Med      Q3     Max
## -1.5286 -0.5366 -0.1105  0.6271  1.5012
```

```
##
## Number of Observations: 48
## Number of Groups: 12
```

The next block of code uses ANOVA to evaluate the fixed effect. The p values for the `age` F test in the first two models are identical. This result makes sense because the assumption of independent within-subject errors made in the first model is, as far as the F test concerned, equivalent to the assumption of compound symmetry in the second model. In other words, forcing the correlation between levels to be zero is a *particular instance* of compound symmetry, whereas the second model allows the variance-covariance matrix to be possess *any* type of compound symmetry. The results of the F test on the fixed factor, `age`, differ slightly in the third model, which used a more complex variance-covariance matrix.

```
anova(cog.nlme.00) # independent
```

```
##             numDF denDF F-value p-value
## (Intercept)     1    33   929.7  <.0001
## age             3    33     3.0  0.0432
```

```
anova(cog.nlme.01) # compound symmetry
```

```
##             numDF denDF F-value p-value
## (Intercept)     1    33   929.7  <.0001
## age             3    33     3.0  0.0432
```

```
anova(cog.nlme.02) # no constraints
```

```
##             numDF denDF F-value p-value
## (Intercept)     1    33   916.4  <.0001
## age             3    33     2.7  0.0633
```

Finally, we can compare the overall fit of the three models with the `anova` command, which confusingly evaluates the goodness-of-fit with chi-square tests. Note that the first two models provide the same fit, although the second model uses one more degree of freedom (to estimate the correlation between the residuals). The third model uses six more degrees of freedom than the first model to estimate the six correlations between residuals on the four levels of `age`. The third model provides a significantly better fit than the other two.

```
anova(cog.nlme.00,cog.nlme.01,cog.nlme.02) # 3rd model fits better
```

```
##             Model df   AIC   BIC logLik   Test L.Ratio p-value
## cog.nlme.00     1  6 355.5 366.2 -171.8
## cog.nlme.01     2  7 357.5 370.0 -171.8 1 vs 2    0.00  1.0000
## cog.nlme.02     3 12 352.6 374.0 -164.3 2 vs 3   14.91  0.0108
```

11.6 variance components & intra-class correlation

The estimated values of the mean squares for our design are shown in Table 1. If we assume that the subject × treatment interaction is zero, then the expected values for subject and the error differ only by $a\sigma_\pi^2$, and therefore the variance component for subject could be estimated with the formula

$$\hat{\sigma}_\pi^2 = \frac{\text{MS}_{subjects} - \text{MS}_{error}}{a}$$

which for our data yields $\hat{\sigma}_\pi^2 = 135.4$. Association strength for the random factor, as indexed by the intra-class correlation is

$$\hat{\rho}_\pi = \frac{\hat{\sigma}_\pi^2}{\hat{\sigma}_\pi^2 + \hat{\sigma}_e^2}$$

or $\hat{\rho}_\pi = 135.4/(135.4 + 60.79) = 0.69$. Of course these estimates depend on the assumption that $\sigma_{\pi\alpha}^2 = 0$, which may not be reasonable.

ANOVA estimates of the variance components also can be calculated using `anovaMM` in the `VCA` package. Random effects are specified in the model formula by enclosing the terms in parentheses. The values of the variance components are the same as those calcuated from the mean squares with the assumption that $\sigma_{\pi\alpha}^2 = 0$. In addition, the table lists the intra-class correlations in the `%Total` column.

```
library(VCA)
aov.vca <- anovaMM(score~age+(subj),Data=mw115L) # note capital D in Data
print(aov.vca,digits=3)
```

```
##
##
## ANOVA-Type Estimation of Mixed Model:
## ------------------------------------
##
##      [Fixed Effects]
##
```

```
##    int agea30 agea36 agea42 agea48
##    112     -9     -5     -2      0
##
##
##      [Variance Components]
##
##   Name  DF     SS    MS      VC      %Total SD     CV[%]
## 1 total 18.117              196.136 100    14.005 12.967
## 2 subj  11     6624 602.182 135.348 69.007 11.634 10.772
## 3 error 33     2006 60.788  60.788  30.993 7.797  7.219
##
## Mean: 108 (N = 48)
##
## Experimental Design: balanced  |  Method: ANOVA
```

Finally, we can extract variance components from `lmer` and `lme` objects using the `VarCorr` command. For this simple, *balanced* design, the variance components and intra-class correlations from the `lmer` object are equal to the ANOVA estimates

```
# lmer
lmer.vca <- VarCorr(mw115L.lmer.01) # independence
print(lmer.vca,comp=c("Variance","Std.Dev."))
```

```
##  Groups   Name        Variance Std.Dev.
##  subj     (Intercept) 135.3    11.6
##  Residual              60.8     7.8
```

```
library(performance)
icc(mw115L.lmer.01,by_group=T)
```

```
## # ICC by Group
##
## Group |   ICC
## -------------
## subj  | 0.690
```

The variance components differ across the models fit with `lme`. The variance components from the first model, which assumed that the residuals were independent across levels of `age`, are the same as the anova estimates. The variance component for `subj` is slightly smaller in the other two models. Using the third model, our estimate of the intra-class correlation for subject is 0.62.

```r
# lme
VarCorr(cog.nlme.00) # independence

## subj = pdLogChol(1)
##             Variance StdDev
## (Intercept) 135.35   11.634
## Residual     60.79    7.797

VarCorr(cog.nlme.01) # compound symmetry

## subj = pdLogChol(1)
##             Variance StdDev
## (Intercept)  52.1     7.218
## Residual    144.0    12.002

VarCorr(cog.nlme.02) # no constraints

## subj = pdLogChol(1)
##             Variance StdDev
## (Intercept) 110.89   10.530
## Residual     99.14    9.957

( icc.subj <- 131.54 / (131.54 + 78.49) )

## [1] 0.6263
```

11.7 association strength & effect size for the fixed effect

Maxwell and Delaney (2004) present an equation for ω^2, which represents the proportion of the variance of the treatment effects relative to the sum of all of the effects in the model:

$$\omega_A^2 = \frac{\sigma_\alpha^2}{\sigma_e^2 + \sigma_\pi^2 + \sigma_\alpha^2}$$

It can be estimated from the anova table with the formula:

$$\hat{\omega}_A^2 = \frac{(a-1)(\text{MS}_A - \text{MS}_{AxS})}{\text{SS}_{Total} + \text{MS}_S} \tag{7}$$

a is the number of levels on the fixed factor, A. For the data analyzed in the previous section, omega squared for `age` is

$$\frac{3(184 - 60.8)}{6624 + 552 + 2006 + 602} = 0.04$$

Note that omega-squared is less than zero when $F_A < 1$. In such cases it is standard practice to set omega-squared to zero (Kirk, 1995). Omega squared can be used to derive an estimate of effect size, Cohen's f (Kirk, 1995):

$$\hat{f}_A = \sqrt{\frac{\hat{\omega}_A^2}{1 - \hat{\omega}_A^2}} \tag{8}$$

For our example, $\hat{f} = 0.204$.

Measures of effect size and association strength for the fixed factor also can be calculated using commands in the **effectsize** package. Note that the value returned by **cohens_f** is calculated by substituting η^2 for ω^2 in Eq. 8.

```
library(effectsize)
omega_squared(mw115L.aov.car.01)

## # Effect Size for ANOVA (Type III)
##
## Parameter | Omega2 (partial) |       95% CI
## -------------------------------------------
## age       |             0.04 | [0.00, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].

cohens_f(mw115L.aov.car.01) # # uses eta-squared = 0.22 in Eq 8

## # Effect Size for ANOVA (Type III)
##
## Parameter | Cohen's f (partial) |      95% CI
## ------------------------------------------
## age       |                0.52 | [0.07, Inf]
##
## - One-sided CIs: upper bound fixed at [Inf].
```

Finally, we calculate association strength and effect size using an **lmer** object.

```
omega_squared(mw115L.lmer.01) # estimated from F value of 3.03
```

```
## # Effect Size for ANOVA (Type III)
##
## Parameter | Omega2 (partial) |       95% CI
## ------------------------------------------
## age       |             0.14 | [0.00, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```

```
cohens_f(mw115L.lmer.01) # uses eta-squared = 0.22 in Eq 8
```

```
## # Effect Size for ANOVA (Type III)
##
## Parameter | Cohen's f (partial) |      95% CI
## ---------------------------------------------
## age       |                0.52 | [0.07, Inf]
##
## - One-sided CIs: upper bound fixed at [Inf].
```

These values differ from previous ones because `effectsize` estimates them from the F statistic rather than the SS and MS values in the ANOVA table:. See the vignette, *Effect Size from Test Statistics* in the `effectsize` package, for more details.

```
anova(mw115L.lmer.01)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##     Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
## age    552     184     3    33    3.03  0.043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
F_to_eta2(3.03,3,33)
```

```
## Eta2 (partial) |       95% CI
## ------------------------------
## 0.22           | [0.00, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```

```
F_to_omega2(3.03,3,33)
```

```
## Omega2 (partial) |       95% CI
## --------------------------------
## 0.14             | [0.00, 1.00]
##
## - One-sided CIs: upper bound fixed at [1.00].
```

```
F_to_f(3.03,3,33)
```

```
## Cohen's f (partial) |      95% CI
## ---------------------------------
## 0.52                | [0.07, Inf]
##
## - One-sided CIs: upper bound fixed at [Inf].
```

11.8 linear comparisons among treatments

The strategy for conducting comparisons among treatments is similar to the one used previously for between-subjects designs. We start by creating a set of contrast weights that represent the comparison of interest: there should be one weight for each dependent measure, and all of the weights must sum to zero. Next, we create a composite score, ψ_i, for each subject that is simply the sum weighted dependent variables:

$$\psi_i = \sum_{j=1}^{a} c_j Y_{ij}$$

Finally, the values of the composite scores are evaluated using a `t.test`. If the null hypothesis being evaluated by our comparison is non-directional, then we use `t.test` to test the null hypothesis that the composite scores were drawn from a zero-mean population. Consider, again, the data analyzed in section 11.5. Let us use a linear trend analysis to evaluate the hypothesis that `scores` followed a linear trend across `age`. The first step is to transform our data frame into a matrix of numbers:

```
dat.mat <- with(mw115,cbind(age.30,age.36,age.42,age.48) )
dat.mat
```

```
##       age.30 age.36 age.42 age.48
##  [1,]    108     96    110    122
##  [2,]    103    117    127    133
##  [3,]     96    107    106    107
##  [4,]     84     85     92     99
##  [5,]    118    125    125    116
##  [6,]    110    107     96     91
##  [7,]    129    128    123    128
##  [8,]     90     84    101    113
##  [9,]     84    104    100     88
## [10,]     96    100    103    105
## [11,]    105    114    105    112
## [12,]    113    117    132    130
```

The variable `dat.mat` represents the data as a four-column matrix: each row contains the data from one subject, and each column contains the data from a single age. Next, we create the contrast weights and then the composite scores using the matrix multiplication operator `%*%`. Note that the order of the terms is important here, so pay attention!

```
lin.trend<-c(-1.5,-0.5,0.5,1.5);
lin.scores<-dat.mat %*% lin.trend;
lin.scores
```

```
##       [,1]
##  [1,]   28
##  [2,]   50
##  [3,]   16
##  [4,]   26
##  [5,]   -3
##  [6,]  -34
##  [7,]   -4
##  [8,]   43
##  [9,]    4
## [10,]   15
## [11,]    6
## [12,]   33
```

Finally, we use `t.test` to evaluate the null hypothesis that the linear trend scores are drawn from a distribution with a mean of zero.

```
t.test(lin.scores)
```

```
##
##      One Sample t-test
##
## data:  lin.scores
## t = 2.2, df = 11, p-value = 0.05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   0.2702 29.7298
## sample estimates:
## mean of x
##        15
```

The t test is significant ($t = 2.24$, $df = 11$, $p = 0.046$) so we reject the null hypothesis that our composite scores, which represent the linear trend across age, are zero. Note that in this case the sphericity assumption *must* be valid because our comparison is a single degree of freedom test. Hence, there is a significant linear association between `score` and `age`.

It is also possible to do directional tests of our hypothesis. For example, here is how we could test the hypothesis that there is an *increasing* linear trend with age.

```
t.test(lin.scores,alternative="greater")
```

```
##
##      One Sample t-test
##
## data:  lin.scores
## t = 2.2, df = 11, p-value = 0.02
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  2.981    Inf
## sample estimates:
## mean of x
##        15
```

The null hypothesis is that the composite scores are drawn from a population with a mean $\mu_{composite} \leq 0$; the alternative hypothesis is $\mu_{composite} > 0$. The test is significant, so we reject the null hypothesis in favor of the alternative. Note that the results of our directional test depends critically on the sign of the contrast scores:

```
lin.trend<-c(1.5,0.5,-0.5,-1.5);
lin.scores<-dat.mat%*%lin.trend;
lin.scores
```

```
##       [,1]
##  [1,]  -28
##  [2,]  -50
##  [3,]  -16
##  [4,]  -26
##  [5,]    3
##  [6,]   34
##  [7,]    4
##  [8,]  -43
##  [9,]   -4
## [10,]  -15
## [11,]   -6
## [12,]  -33
```

```
t.test(lin.scores,alternative="greater")
```

```
##
##      One Sample t-test
```

```
##
## data:  lin.scores
## t = -2.2, df = 11, p-value = 1
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -27.02    Inf
## sample estimates:
## mean of x
##       -15
```

The bottom line is that you have to have a very clear understanding of how your contrast weights are related to your `t.test`.

11.8.1 comparisions using `emmeans`

In the previous section, the t test that evaluated the linear trend used a so-called local error term that was calculated using the linear trend scores and had 11 degrees of freedom. Alternatively, we could have used the so-called global error term from the original ANOVA, which was based on all of the mean square residuals (within-subjects) and had 33 degrees of freedom. To highlight the difference between the two error terms, the t test for linear trend is recast here as an ANOVA that evaluates the null hypothesis that the grand mean (i.e., the intercept) is zero. To do this correctly, it is important to use the same weights for the linear trend that we used for the t test.

```
wLinear <- c(-1.5,-0.5,0.5,1.5);
lin.scores<-dat.mat %*% wLinear;
summary(aov(lin.scores~1),intercept=T) # using local error term
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## (Intercept)  1   2700    2700    5.02  0.047 *
## Residuals   11   5912     537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, we use the error term from the original ANOVA to recalculate the F and p values. However, before recalculating the F, we normalize the mean square by the sum of the squared contrasts weights.

```
w2 <- sum(wLinear^2) # sum of squared contrast weights
MS.lin.trend <- 2700/w2 # normalized mean square
summary(mw115L.aov.01)[[2]]
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## age        3    552   184.0    3.03  0.043 *
## Residuals 33   2006    60.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MS.err <- 60.8
df.err <- 33
F.lin.trend.global <- MS.lin.trend /MS.err
p.new <- 1-pf(F.lin.trend.global,1,df.err)
c(F.lin.trend.global,p.new) # list F and p values
```

```
## [1] 8.881579 0.005373
```

The **emmeans** package can perform contrasts on within-subjects factors. Interestingly, it sometimes sometimes uses the local estimate error and other times uses the global error. For example, it uses the local estimate of error (df=11) when the input is an **aov_car** object:

```
library(emmeans)
# aov_car object:
# mw115L.aov.car.01 <- aov_car(score~1+age+Error(subj/age),data=mw115L)
mw115L.aov.car.emm <- emmeans(mw115L.aov.car.01,specs="age")
# uses local estimate of error:
contrast(mw115L.aov.car.emm,method=list(linear=wLinear))
```

```
##  contrast estimate   SE df t.ratio p.value
##  linear         15 6.69 11   2.241  0.0466
```

However, the global error estimate (df=33) is used when the input is an **aov** or **lmer** object

```
# aov object:
# mw115L.aov.01 <- aov(score~1+age+Error(subj/age),data=mw115L)
mw115L.aov.emm <- emmeans(mw115L.aov.01,specs="age")
# lmer object:
# mw115L.lmer.01 <- lmer(score~age+(1|subj),data=mw115L)
mw115L.lmer.emm <- emmeans(mw115L.lmer.01,specs="age")
# contrast uses the uses global estimate of error:
contrast(mw115L.aov.emm,method=list(linear=wLinear))
```

```
##  contrast estimate   SE df t.ratio p.value
##  linear         15 5.03 33   2.981  0.0054
```

```
contrast(mw115L.lmer.emm,method=list(linear=wLinear))
```

```
##  contrast estimate   SE df t.ratio p.value
##  linear         15 5.03 33   2.981  0.0054
##
## Degrees-of-freedom method: kenward-roger
```

```
2.981^2 # square t to get value of F statistic
```

```
## [1] 8.886
```

References

Box, G. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, i: Effect of inequality of variance and of correlation of errors in the two-way classification. *Annals of Mathematical Statistics*, 25(290-302).

Geisser, S. and Greenhouse, S. (1958). An extensioin of box's rersults on the use of the f distribution in multivariate analysis. *Annals of Mathematical Statistics*, 29:885–91.

Huynh, H. and Feldt, L. (1970). Conditions under which mean square ratios in repeated measures designs have exact f-distributions. *Journal of the American Statistical Association*, 65:1582–89.

Huynh, H. and Feldt, L. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1(69-82).

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Brooks/-Cole, 3rd edition.

Maxwell, S. E. and Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective*. Lawrence Erlbaum Associates, Mahwah, N.J., 2nd ed edition.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rouanet, H. and Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: Anova and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, 23:147–63.