

## Stat 414 - Day 13

### Correlated Observations

#### Last Time:

- Restricted (aka residual) maximum likelihood estimation (REML) addresses bias in the ML variance component estimation and is preferred when estimating or testing variance components.
- Estimation method needs to match across models for comparison

**Example 1:** Caffeine is widely used as a stimulant – but are there other ways to get the same effects, with little to no downside? To begin to answer this question, a study compared the effects of caffeine with theobromine, which is the active chemical naturally found in chocolate and is an alkaloid with a similar molecular structure and effects on people as caffeine (Scott & Chen, 1944). To measure the effects of these two different chemicals, the researchers trained subjects to tap their fingers in such a way that the rate could be measured. After learning/practicing this type of finger tapping, participants took either took a caffeine pill (200 mg), a theobromine pill (200 mg), or a placebo, and then their finger tapping rate was measured two hours later.

```
var(Taps)
```

Consider a one-way ANOVA on the stimulants:

```
summary(aov(Taps ~ Stimulant))
```

- (a) Is the difference among the stimulants statistically significant? (Be clear how you are deciding.)

One reason for the lack of significance is that there are at least *two* sources of variation in this study: the stimulants and the participants.

- (b) Is there significant person to person variation?

*#using effect coding*

```
summary(modelB <- lm(Taps ~ participant, contrasts = list(participant =
contr.sum)))
```

- (c) What is the  $R^2$  for this analysis (show computation details) and what does it tell you?

We note that each subject participated in all 3 treatments, in random order, giving us a randomized block design.

- (d) How many observations were taken in this study? How many participants were in this study?

- (e) Is it reasonable to consider the observations in this study independent from each other? What might the variance-covariance matrix of the residuals look like?

Correlated data is encountered in nearly every field. In education, student scores from a particular teacher are typically more similar than scores of other students who have had a different teacher. Here we expect repeated measurements on the same individual to be more similar than finger tap measurements from other participants. In political polling, opinions from members of the same household are usually more similar than opinions of members from other randomly selected households. Correlation of observations within the same teacher or patient or household is referred to as intraclass correlation.

One way to measure the intraclass correlation = between subject variance / total variance; can we “reliably” predict an observation based on which group (e.g., subject) it comes from. Can also be interpreted as the expected correlation between two randomly drawn observations that are in the same group. Rather than finding the correlation between all possible pairs of observations within each group, we will compute (from Day 3)

$$ICC = (MSGroups - MSEError) / (MSGroups + (k - 1)MSEError)$$

Sources	Df	MS	E(MS)
Groups	I - 1	MSGroups	$k\sigma_s^2 + \sigma_e^2$
Error	N - I	MS Error	$\sigma_e^2$

- (f) What should ICC be if the observations in a group are perfectly correlated? If they are independent?

```
anova(modelB)
```

- (g) Calculate and interpret the intraclass correlation for the subjects in the stimulant study.

So instead of thinking we have 4 x 3 independent observations, we can compute the “effective sample size” as  $\frac{I \times n}{(n-1) \times ICC}$  where  $n$  is the group size.

- (h) What is the effective sample size for this study?

So a more appropriate analysis (which was most interested in the stimulants) is to adjust for the participant effects, and explore the “within person” treatment effects (assuming they are the same as we don’t have enough replication to test for interactions).

```
library(ggplot2)
ggplot(data = fingertapstudy, aes(y=Taps, x = Stimulant, color=participant)) +
  geom_point() + theme_bw() +
  facet_wrap(~ participant)
summary(aov(Taps ~ participant + Stimulant))
```

- (i) Now what do you conclude about the Stimulants? (Use both the graph of the raw data and the anova output)

*To Think About:* Suppose we had a larger study with lots more participants. What would be a downside to including the participant variable in the model?