Lecture 11 continued

Everything in this section really belongs in Lecture 11.

1.1 Testing for interactions

The *english* dataset includes average naming latencies not only for college-age speakers but also for speakers age 60 and over. This degree of age difference turns out to have a huge effect on naming latency (Figure 1):

```
histogram(~ RTnaming | AgeSubject, english)
```

Clearly, college-age speakers are faster at naming words than speakers over age 60. We may be interested in including this information in our model. In Lecture 10 we already saw how to include both variables in a multiple regression model. Here we will investigate an additional possibility: that different levels of frication may have different effects on mean naming latency depending on speaker age. For example, we might think that fricatives, which our linear model above indicates are the hardest class of word onsets, might be even harder for elderly speakers than they are for the young. When these types of inter-predictor contingencies are included in a statistical model they are called *interactions*.

It is instructive to look explicitly at the linear model that results from introducing interactions between multiple categorical predictors. We will take *old* as the baseline value of speaker age, and leave *burst* as the baseline value of frication. This means that the “baseline” predictor set involves an
old-group speaker naming a burst-initial word, and the intercept $\alpha$ will express the predicted mean latency for this combination. There are seven other logically possible combinations of age and frication; thus our full model will have to have seven dummy indicator variables, each with its own parameter. There are many ways to set up these dummy variables; we'll cover perhaps the most straightforward way. In addition to $X_{\{1,2,3\}}$ for the non-baseline levels of frication, we add a new variable $X_4$ for the non-baseline levels of speaker age (young). This set of dummy variables allows us to encode all eight possible groups, but it doesn’t allow us to estimate separate parameters for all these groups. To do this, we need to add three more dummy variables, one for each of the non-baseline frication levels when coupled with the non-baseline age level. This gives us the following complete set of codings:

<table>
<thead>
<tr>
<th>Frication</th>
<th>Age</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
</tr>
</thead>
<tbody>
<tr>
<td>burst</td>
<td>old</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>frication</td>
<td>old</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>long</td>
<td>old</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>short</td>
<td>old</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>burst</td>
<td>young</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>frication</td>
<td>young</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>long</td>
<td>young</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>short</td>
<td>young</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 1: Histogram of naming latencies for young (ages $\sim 22.6$) versus old (ages $> 60$ speakers)
We can test this full model against a strictly additive that allows for effects of both age and initial phoneme class, but not for interactions—that is, one with only \(X_{1,2,3,4}\). In R, the formula syntax \texttt{Frication*AgeSpeaker} indicates that an interaction between the two variables should be included in the model.

```r
> m.0 <- lm(exp(RTnaming) ~ Frication + AgeSubject, english)
> m.A <- lm(exp(RTnaming) ~ Frication * AgeSubject, english)
> anova(m.0, m.A)
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4563</td>
<td></td>
<td>3977231</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4560</td>
<td>3</td>
<td>7019</td>
<td>2.6871</td>
<td>0.0449</td>
</tr>
</tbody>
</table>

Note that there are three more parameters in the model with interactions than in the additive model, which fits degrees of freedom listed in the analysis of variance table. As we can see, there is some evidence that frication interacts with speaker age. We get the same result with \texttt{aov()}:

```r
> summary(aov(exp(RTnaming) ~ Frication * AgeSubject, english))
```

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frication</td>
<td>3</td>
<td>341494</td>
<td>113831</td>
<td>130.7414 &lt; 2e-16 ***</td>
</tr>
<tr>
<td>AgeSubject</td>
<td>1</td>
<td>42099187</td>
<td>42099187</td>
<td>48353.1525 &lt; 2e-16 ***</td>
</tr>
<tr>
<td>Frication:AgeSubject</td>
<td>3</td>
<td>7019</td>
<td>2340</td>
<td>2.6871 0.04489 *</td>
</tr>
</tbody>
</table>

1.2 ANOVA in its more general form

Although the picture in Figure ?? is the way that linear model comparison is classically done, and is appropriate for the ANOVA comparisons that we
have looked at so far, the partitioning of the variance in ANOVA can get more complicated. The call to \texttt{aov()} we just made in (1) for the interaction between \texttt{Frication} and \texttt{AgeSubject} actually partitioned the variance as shown in Figure 2. In each line of the summary for (1), the variance inside the box corresponding to the predictor of interest is being compared with the Residual Error box in Figure 2. The ratio of mean squares is \( F \)-distributed in all these cases. One of the somewhat counterintuitive consequences of this approach is that you can test for main effects of one predictor (say \texttt{Frication}) while accounting for idiosyncratic interactions of that predictor with another variable.

2 A bit more on the \( F \) distribution

By popular demand, here’s a bit more about the \( F \) distribution. There’s really not much to say about this distribution except that, crucially, it is the distribution of the ratio of two \( \chi^2 \) random variables. Because the variance of a sample is distributed as a \( \chi^2 \) random variable, the ratio of variances in linear models can be compared to the \( F \) distribution.

More formally, if \( U \sim \chi^2_m \) and \( V \sim \chi^2_n \), we have
$F_{m,n} \sim \frac{U/m}{V/n}$

It is useful to play a bit with the $F$ distribution to see what it looks like. In general, the cumulative distribution is more interesting and pertinent than the probability density function (unless you have an anomalously low $F$ statistic).

## 3 A case study

This is the outcome of a self-paced reading experiment conducted by Hannah Rohde, in collaboration with me and Andy Kehler.

The question under investigation is whether certain kinds of verbs (implicit causality (IC) verbs) such as “detest”, which intuitively demand some sort of explanation, can affect readers’ online syntactic attachment preferences.

(2) a. John detests the children of the musician who is generally arrogant and rude (IC,LOW)
b. John detests the children of the musician who are generally arrogant and rude (IC,HIGH)
c. John babysits the children of the musician who is generally arrogant and rude (nonIC,LOW)
d. John babysits the children of the musician who are generally arrogant and rude (nonIC,HIGH)

Hannah hypothesized that the use of an IC verb should facilitate reading of high-attached RCs, which are generally found in English to be harder to read than low-attached RCs (Cuetos and Mitchell, 1988). The reasoning here is that the IC verbs demand an explanation, and one way of encoding that explanation linguistically is through a relative clause. In these cases, the most plausible type of explanation will involve a clause in which the object of the IC verb plays a role, so an RC modifying the IC verb’s object should become more expected. This stronger expectation may facilitate processing when such an RC is seen (Levy, 2007).

The stimuli for the experiment consist of 20 quadruplets of sentences of the sort above. Such a quadruplet is called an EXPERIMENTAL ITEM in the
language of experimental psychology. The four different variants of each item are called the conditions. Since a participant who sees one of the sentences in a given item is liable to be strongly influenced in her reading of another sentence in the item, the convention is only to show each item once to a given participant. To achieve balance, each participant will be shown five items in each condition.

<table>
<thead>
<tr>
<th>Participant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ic,high</td>
<td>NONic,high</td>
<td>ic,low</td>
<td>NONic,low</td>
<td>ic,high</td>
</tr>
<tr>
<td>2</td>
<td>NONic,low</td>
<td>ic,high</td>
<td>NONic,high</td>
<td>ic,low</td>
<td>NONic,low</td>
</tr>
<tr>
<td>3</td>
<td>ic,low</td>
<td>NONic,low</td>
<td>ic,high</td>
<td>NONic,high</td>
<td>ic,low</td>
</tr>
<tr>
<td>4</td>
<td>NONic,high</td>
<td>ic,low</td>
<td>NONic,low</td>
<td>ic,high</td>
<td>NONic,high</td>
</tr>
<tr>
<td>5</td>
<td>ic,high</td>
<td>NONic,high</td>
<td>ic,low</td>
<td>NONic,low</td>
<td>ic,high</td>
</tr>
</tbody>
</table>

The experimental data will be analyzed for effects of verb type and attachment level, and more crucially for an interaction between these two effects. For this reason, we plan to conduct a two-way ANOVA.

In self-paced reading, the observable effect of difficulty at a given word often shows up a word or two downstream, so in this case we will focus on the first word after the disambiguator—i.e., “generally”. This is called the first spillover region. First we read in the complete dataset, zoom in on the results at this region, and look at the distribution of reading times for each condition. A boxplot (also known as a box-and-whiskers diagram) is a good tool for this kind of visualization, and for identifying outliers.¹

```r
dat1 <- read.table("results.final.txt", quote="",sep="\t",header=T)
dat1 <- subset(dat1,subj != "subj2" & subj != "subj10" & subj != "subj50") # these subjects answered questions at chance
dat1$subj <- factor(dat1$subj) # eliminate these levels from the factor
spillover.1 <- subset(dat1,expt==1 & correct ==1 & crit=="RC_VERB+1") # focus on first spillover region,
                   # only correctly-answered sentences
```

¹In a boxplot, the upper and lower bounds of the box are the first and third quartile of the data; the length of this box is called the inter-quartile range, or IQR. The solid-line “whiskers” are placed at the farthest points that lie no more than 1.5 × IQR from the edges of the box. Any points that lie beyond these whiskers are considered “outliers” and plotted individually.
As you can see in Figure 3, there are lots of outliers – though the situation looks better in log-space. Histograms reveal similar results (Figure 4):

```r
> spillover.1$verb <- factor(spillover.1$verb) # remove "NONE"
  # levels of factor
> spillover.1$attachment <- factor(spillover.1$attachment)
> boxplot(rt ~ verb*attachment,spillover.1)
> boxplot(log(rt) ~ verb*attachment,spillover.1)
```

It is by no means obvious what to do in this kind of situation where there are so many outliers, and the raw response departs so severely from normality. In the case of self-paced reading, the dominant convention is to perform OUTLIER REMOVAL: use some relatively standardized criterion for identifying outliers, and deal with those outliers in some way.

There are several different ways outlier removal is handled in the literature; in this situation we shall apply one of the more common procedures. For each condition, we calculate for each point a Z-SCORE, which is a measure of how many sample standard deviations the point lies away from the sample mean. Points with a Z-score of magnitude above 4 are simply thrown away:
Figure 4: Histograms for raw and log reading times at the first spillover region for the IC-RC experiment

```r
> cbind.list <- function(l) {
    result <- c()
    for(i in 1:length(l)) {
        result <- cbind(result,l[[i]])
    }
    result
}

> get.z.score <- function(response,conds.list) {
    means <- tapply(response,conds.list,mean)
    sds <- tapply(response,conds.list,sd)
    (response - means[cbind.list(conds.list)]) /
     sds[cbind.list(conds.list)]
}

> z <- with(spillover.1,get.z.score(rt,list(verb,attachment)))
> sum(abs(z) > 4) # 14 points are flagged this way as outliers
[1] 14
> length(z) # 14 out of 933 is a lot of outliers at 4sd!!!!
    # But this is typical for self-paced reading experiments
[1] 933
> spillover.1.to.analyze <- subset(spillover.1,abs(z) <= 4)
```

We take a quick look at what our resulting data look like (some fields of
the data frame have been omitted):

```r
> head(spillover.1.to.analyze)
   Subj Item Verb Attachment Crit   RT
         9     1   IC     high RC_VERB+1 365.27
        22    10  nonIC    low RC_VERB+1 616.43
        35     1   IC     low RC_VERB+1 255.56
        48     1  nonIC     high RC_VERB+1 626.26
        61     1   IC     high RC_VERB+1 330.45
        74     1  nonIC    low RC_VERB+1 434.66
```

and now we are ready to conduct our two-way ANOVA.

## 4 The comparisons to make

In this experiment, four factors characterize each stimulus: a particular subject reads a particular item that appears with particular values of the verb and attachment manipulations. Verb and attachment have two levels each, so if we had \( m \) subjects and \( n \) items we would in principle need at least \( 2 \times 2 \times m \times n \) observations to consider a full classic linear model with interactions of all possible types. However, because each subject saw each item only once, we only have \( m \times n \) observations. Therefore it is not possible to construct the full model.

For many years dating back to ?, the gold standard in this situation has been to construct two separate analyses: one for subjects, and one for items. In the analysis over subjects, we take as our individual data points the mean value of all the observations in each cell of Subject × Verb × Attachment—that is, we AGGREGATE, or average, across items. Correspondingly, in the analysis over items, we aggregate across subjects. We can use the function `aggregate()` to perform this averaging:

```r
sp.1.subj <- with(spillover.1.to.analyze,aggregate(list(rt=rt),
             list(subj=subj,verb=verb,attachment=attachment),mean))
sp.1.item <- with(spillover.1.to.analyze,aggregate(list(rt=rt),
             list(item=item,verb=verb,attachment=attachment),mean))
```

The view of the resulting data for the analysis over subjects can be seen in Table 1. This setup is called a WITHIN-SUBJECTS or REPEATED-MEASURES
design because each subject participates in each condition—or, in another
manner of speaking, we take multiple measurements for each subject. Designs
in which, for some predictor factor, each subject participates in only one
condition are called BETWEEN-SUBJECTS designs.

The way we partition the variance for this type of analysis can be seen in
Figure 5. Because we have averaged things out so we only have one obser-
vation per Subject/Verb/Attachment combination, there will be no variation in
the Residual Error box. Each test for an effect of a predictor sets of interest
(verb, attachment, and verb:attachment) is performed by comparing the
variance explained by the predictor set \( P \) with the variance associated with
arbitrary random interactions between the subject and \( P \). This is equivalent
to performing a model comparison between the following two linear models,
where \( i \) range over the subjects and \( j \) over the conditions in \( P \):

\[
rt_{ij} = \alpha + B_i \text{Subj}_i + \epsilon_{ij} \quad \text{(null hypothesis)} \tag{2}
\]

\[
rt_{ij} = \alpha + B_i \text{Subj}_i + \beta_j P_j + \epsilon_{ij} \quad \text{(alternative hypothesis)} \tag{3}
\]

There is an added wrinkle here, which is that the \( B_i \) are not technically free
parameters but rather are themselves assumed to be random and normally
distributed. However, this difference does not really affect the picture here.
(In a couple of weeks, when we get to mixed-effects models, this difference
will become more prominent and we’ll learn how to handle it in a cleaner
and more unified way.)

Fortunately, \texttt{aov()} is smart enough to know to perform all these model
comparisons in the appropriate way, by use of the \texttt{Error()} specification in
your model formula. This is done as follows, for subjects:

| Verb | Attachment | 1   | 2   | 3   | 4   | 5   | ...
|------|------------|-----|-----|-----|-----|-----|-----
| IC   | High       | 280.7 | 396.1 | 561.2 | 339.8 | 546.1 | ...
|      | Low        | 256.3 | 457.8 | 547.3 | 408.9 | 594.1 | ...
| nonIC| High       | 340.9 | 507.8 | 786.7 | 369.8 | 453.0 | ...
|      | Low        | 823.7 | 311.4 | 590.4 | 838.3 | 298.9 | ...

Table 1: Repeated-measures (within-subjects) view of item-aggregated data
for subjects ANOVA
Figure 5: The picture for this 2 × 2 ANOVA, where Verb and Attachment are the fixed effects of interest, and subjects are a random factor.

> summary(aov(rt ~ verb * attachment
   + Error(subj/(verb *attachment)), sp.1.subj))

Error: subj
   Df Sum Sq Mean Sq F value Pr(>F)
Residuals 54 4063007  75241

Error: subj:verb
   Df Sum Sq  Mean Sq F value   Pr(>F)
verb   1  48720 48720.0  7.0754 0.01027 *
Residuals 54 371834  6886 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Error: subj:attachment
   Df Sum Sq  Mean Sq   F value  Pr(>F)
attachment 1  327   327.0  0.04060   0.841
Residuals 54 434232  8041 

Linguistics 251 lecture 12 notes, page 11     Roger Levy, Fall 2007
Error: subj:verb:attachment

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb:attachment</td>
<td>1</td>
<td>93759</td>
<td>93759</td>
<td>6.853</td>
</tr>
<tr>
<td>Residuals</td>
<td>54</td>
<td>738819</td>
<td>13682</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

and for items:

```r
> summary(aov(rt ~ verb * attachment + Error(item/(verb *attachment)), sp.1.item))
```

Error: item

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals</td>
<td>19</td>
<td>203631</td>
<td>10717</td>
<td></td>
</tr>
</tbody>
</table>

Error: item:verb

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb</td>
<td>1</td>
<td>21181</td>
<td>21181</td>
<td>3.5482</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>113419</td>
<td>5969</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Error: item:attachment

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>attachment</td>
<td>1</td>
<td>721</td>
<td>721</td>
<td>0.093</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>147299</td>
<td>7753</td>
<td></td>
</tr>
</tbody>
</table>

Error: item:verb:attachment

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb:attachment</td>
<td>1</td>
<td>38211</td>
<td>38211</td>
<td>5.4335</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>133615</td>
<td>7032</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Fortunately, the by-subjects and by-items analysis yield largely similar results: they both point towards (a) a significant main effect of verb type;
and (b) more interestingly, a significant interaction between verb type and attachment level. To interpret these, we need to look at the means of each condition. It is conventional in psychological experimentation to show the condition means from the aggregated data for the by-subjects analysis:

```r
> with(sp.1.subj, tapply(rt, list(verb), mean))
  IC  nonIC
 452.2940 482.0567

> with(sp.1.subj, tapply(rt, list(verb, attachment), mean))

<table>
<thead>
<tr>
<th></th>
<th>high</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC</td>
<td>430.4316</td>
<td>474.1565</td>
</tr>
<tr>
<td>nonIC</td>
<td>501.4824</td>
<td>462.6309</td>
</tr>
</tbody>
</table>
```

The first spillover region was read more quickly in the implicit-causality verb condition than in the non-IC verb condition. The interaction was a CROSSTERE INTERACTION: in the high attachment conditions, the first spillover region was read more quickly for IC verbs than for non-IC verbs; but for the low attachment conditions, reading was faster for non-IC verbs than for IC verbs.

We interpreted this result to indicate that IC verbs do indeed facilitate processing of high-attaching RCs, to the extent that this becomes the preferred attachment level.

**References**
