Outline for the rest of the course

For lectures:

- This week and next, I'll motivate and exemplify *mixed effects models*.
- In the final lecture, I'll talk about making publication-quality visualizations using `ggplot2`, R's implementation of Wilkinson's *grammar of graphics*.

And for assignments:

- Homework 7 will extend our work on model interpretation.
- Homework 8 will cover mixed effects regression modeling.
- I'll produce a brief study guide for the final for the final practicum.
- You'll have one week to complete the take-home final.

Outline today

- Review of homework 5 (up to 20-30 minutes; more in the practicum)
- Discussion of homework 7 (up to 20-30 minutes; more in the practicum)
- Introducing *mixed effects models*:
  - The *language-as-fixed-effects* fallacy
  - Subjects aren't fixed effects either…
  - Random slopes and intercepts
  - Random effects syntax
- In practicum:
  - More homework

Homework 5 review

Part 1: multiple regression

Loading the data

```
> library(lmtest)
> d <- read.table("ANAE-O.tsv", header = TRUE)
```

Computing Euclidean distance

```
> euclidean.distance <- function(x1, y1, x2, y2) {
+     x.delta <- x1 - x2
+     y.delta <- y1 - y2
+     return(sqrt(x.delta * x.delta + y.delta * y.delta))
+ }
> d$distance <- with(d, euclidean.distance(o.F1, o.F2, oh.F1,
oh.F2))
```

Fitting the model

```
> d$age <- d$age - mean(d$age, na.rm = TRUE)
> contrasts(d$gender) <- contr.sum
> contrasts(d$dialect) <- contr.sum
> r <- lm(distance ~ age + gender + dialect, data = d)
```

Testing for an effect of age

```
> lrtest(lm(distance ~ gender + dialect, data = d), r)

Likelihood ratio test
Model 1: distance ~ gender + dialect
Model 2: distance ~ age + gender + dialect
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  25 -2502.6
2  26 -2499.0  1 7.2191   0.007213 **
```

Testing for an effect of gender

```
> lrtest(lm(distance ~ age + dialect, data = d), r)

Likelihood ratio test
Model 1: distance ~ age + dialect
Model 2: distance ~ age + gender + dialect
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  25 -2499.5
2  26 -2499.0  1 1.0834     0.2979
```

Testing for an effect of dialect

```
> lrtest(lm(distance ~ age + gender, data = d), r)

Likelihood ratio test
Model 1: distance ~ age + gender
Model 2: distance ~ age + gender + dialect
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   4 -2688.5
2  26 -2499.0 22 378.93  < 2.2e-16 ***
```

Post-hoc tests on dialect

```
> TukeyHSD(aov(r), which = c("dialect"))
...
```

The two least merged dialects are Providence and New York City; NYC is significantly different (at $\alpha = .05$) than every other dialect except Providence.

Part 2: multicollinearity

Measuring multicollinearity

```
> n <- ncol(d)
> for (n1 in 1:(n - 1)) {
+     x <- d[, n1]
+     x.name <- names(d)[n1]
+     for (n2 in (n1 + 1):n) {
+         y <- d[, n2]
+         y.name <- names(d)[n2]
+         cat(sprintf("r(%s, %s) = %+.3f\n",
+                     x.name, y.name, cor(x, y)))
+     }
+ }
```

Measuring multicollinearity

```
r(occ, res) = +0.575
r(occ, sc1) = +0.677
r(occ, sc2) = +0.536
r(res, sc1) = +0.343
r(res, sc2) = +0.252
r(sc1, sc2) = +0.754
```

Of course you don't have to use loops to do this:

```
> with(d, cor(occ, res))
> ...
```

Standardizing the independent variables

```
> for (name in names(d)) {
+       x <- d[[name]]
+       x <- (x - mean(x)) / sd(x)
+       d[[name]] <- x # Writes it back into the data frame.
+       cat(sprintf("%s: Xbar = %+.3f, s = %.3f\n",
+                   name, mean(x), sd(x)))
+ }
occ: Xbar = +0.000, s = 1.000
res: Xbar = -0.000, s = 1.000
sc1: Xbar = +0.000, s = 1.000
sc2: Xbar = -0.000, s = 1.000
```

Residualizing the independent variables

```
> d$res <- residuals(lm(res ~ occ, data = d))
> d$sc1 <- residuals(lm(sc1 ~ occ + res, data = d))
> d$sc2 <- residuals(lm(sc2 ~ occ + res + sc1, data = d))
```

Testing the residualization

```
> n <- ncol(d)
> for (n1 in 1:(n - 1)) {
+     x <- d[, n1]
+     x.name <- names(d)[n1]
+     for (n2 in (n1 + 1):n) {
+         y <- d[, n2]
+         y.name <- names(d)[n2]
+         cat(sprintf("r(%s, %s) = %+.3f\n",
+                     x.name, y.name, cor(x, y)))
+     }
+ }
```

Testing the residualization

```
r(occ, res) = +0.000
r(occ, sc1) = +0.000
r(occ, sc2) = -0.000
r(res, sc1) = -0.000
r(res, sc2) = +0.000
r(sc1, sc2) = -0.000
```

Of course you don't have to use loops to do this:

```
> with(d, cor(occ, res))
> ...
```

Questions?

Homework 7

Mixed effects regression I

Repeatability and randomness

- *Fixed* (or *repeatable*) effects:
  - values drawn from a small number of finite values
  - data in sample usually encompasses all possible values
  - hypothesized to correlate with the DV
  - E.g., gender, L1, experimental condition
- *Random* effects:
  - large (or infinite) number of possible values
  - data found in sample contains a mere subset of the possible values; "sample, not population"
  - E.g., subject identity, stimulus identity

This isn't always a hard distinction: e.g., Labov's 23 North American English dialects.

Nearly all continuous-valued IVs are *fixed* in the relevant sense.

Nestedness and crossedness

A categorical independent variable *A* is said to be *nested in* variable *B* if a value for a level of *A* implies a value for *B*.

E.g., in a between-subjects study, subject identity (or L1, etc.) is nested in the condition or treatment; each subject has only one value for condition.

A categorical independent variable *A* is said to be *crossed with* variable *B* if every level of *B* is found at every level of *A*.

E.g., in a within-subjects study, subject identity (or L1, etc.) is crossed with condition or treatment; each subject has trials in all conditions.

For quasi-experimental and observational studies these definitions can be relaxed.

Example: Columbus [ʃ]tr

Independent variables include:

- Subject
- Emphasis ("less" or "more")
- Gender ("male" or "female")
- Age (dichotomized as "young", "middle", and "old")
- Class ("working class", "lower middle class", "upper middle class")

Which variables are fixed? Which are random?

Which are nested? Which are crossed?

Example: lexical decision (after Baayen 2008)

Independent variables include:

- Subject
- Stimulus
- Native language ("English", "other")
- Trial number
- Word frequency
- Word length

Which variables are fixed? Which are random?

Which are nested? Which are crossed?

The nested, between-subjects design

Subject 1: A1 B1 C1 D1 E1
Subject 2: A1 B1 C1 D1 E1
Subject 3: A2 B2 C2 D2 E2
Subject 4: A2 B2 C2 D2 E2

(Alphabetic characters: stimuli; numbers: condition.)

The crossed, random, within-subjects design

Subject 1: A1 B1 C1 D1 A2 B2 C2 D2
Subject 2: D2 B1 D1 C1 A2 B2 C2 A1
Subject 3: C1 D1 C2 A2 B1 D2 B2 A1
Subject 4: D1 A2 A1 B1 C1 B2 C2 D2

(Alphabetic characters: stimuli; numbers: condition.)

Subject and item

The identity of a speaker/signer/participant/informant is a *subject* random variable.

And the identity of a stimulus/word/example/etc. is an *item* random variable.

(There are of course other logical possibilities.)

Why we care

Random effects can require many, many degrees of freedom.

Yet, we're not interested in interpreting their coefficients or using them for null hypothesis testing: they are simply *nuisance variables*.

But if we didn't include the random effects in our estimation procedure, we are violating independence assumptions (e.g., that subjects are more self-similar than they are similar to each other).

Nesting produces multicollinearity: if *A* is nested by *B*, there can be perfect correlations (e.g., $R = 1$) between several of their dummy-coding columns!

Subjects as random effects

In many behavioral studies, the largest source of variance is subject.

Some subjects may have had their Wheaties and a cup of coffee that morning; some subjects (who after all, are quite likely to be college freshmen) may have been up all night...

The language-as-fixed-effects fallacy

Clark (1973), in a famous paper, argues that researchers ignore the variability associated with items (e.g., and their intrinsic difficulty) in experimental work. In other words, we must also treat items as *random effects*.

Such variability can only be safely ignored if it turns out to be statistically trivial, or if we have used all possible items that meet a given selection criteria.

Clark proposes a statistic (min F') intended to address this problem; see Raaijmakers et al. 1999 and Baayen 2007 for critique.

Against aggregation

Aggregating by subject artificially deflates variance.

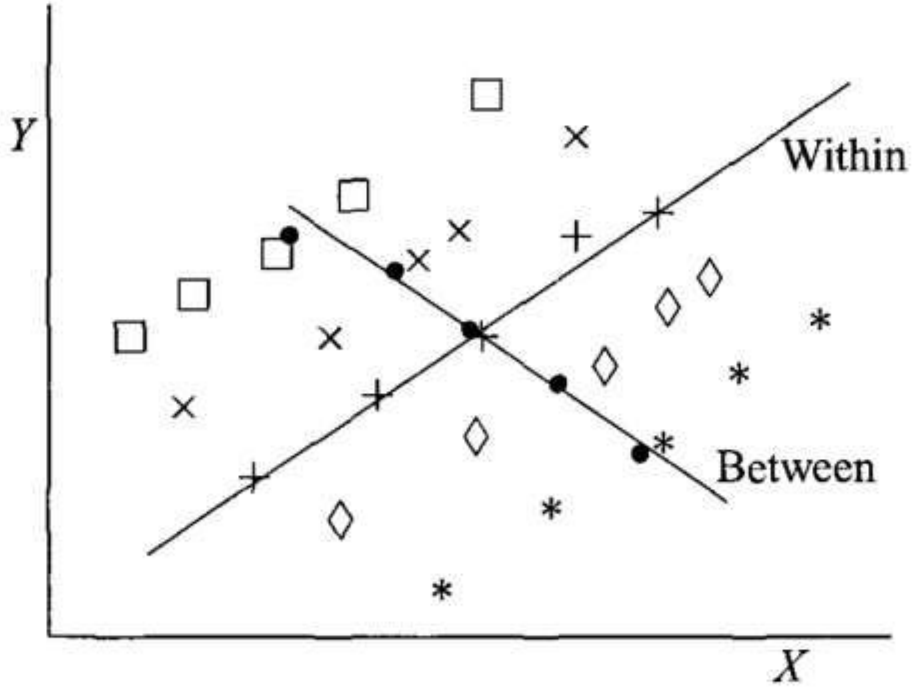
Aggregating by item does the same.

It is not easy to combine both types of aggregation, or to use aggregation in the presence of nesting.

Simpson's paradox

Subjects: $*, \Diamond, +, \times, \Box$

Subject Means:



The mixed effects solution

Mixed effects models are an extension of linear models that allow us to compute the coefficients for fixed effects while also specifying random effects such as subject or item (or both at the same time).

The random effect coefficients (and standard errors) are, strictly speaking, not parameters or coefficients of the model, nor are they estimated the same way.

Compared to linear regression, these models are much more complex to estimate. Key details were worked out by Pinheiro & Bates (2000) and became widely available a few years later as part of R's `lme4` packages; but it is still quite possible to specify a mixed-effects model that is simply "too big".

Random intercepts

A random intercept is characterized by:

- a many-leveled categorical variable defining a random effect,
- a normal distribution with a mean $\bar{X} = 0$ and a standard deviation s estimated from the data, and
- a mapping from each level of the categorical variable to a value on that normal distribution.

Note that the estimates are *shrunk* or *regressed to the mean* to prevent overfitting according to somewhat arcane criteria.

Example: lexical decision study (1/)

Consider a lexical decision study in which the dependent variable is reaction time.

(Usually we ignore non-word trials.)

The primary fixed effect is native language: "English" or "other".

We also have an important covariate: word frequency (estimated using some large reference corpus).

Example: lexical decision study (2/)

Subject ID "nests" with native language: each participant is either a native English speaker or not.

Item "nests" with frequency: each item has a given corpus frequency.

Item (the word or non-word presented to the participant) is between subjects: no participant sees the same item twice, but items do re-occur across the study.

Example: lexical decision study (3/)

This calls for a random effects of subject and item; we expect that *most* of the variance in RT will be associated with subjects' ability, and with items.

Assuming the model has a (fixed, non-random) intercept, the subject random effect adds or subtracts a subject-specific quantity to the predicted RT.

Similarly, the item random effect adds or subtracts a item-specific quantity to the predicted RT.

Naturally, the random effects do not generalize to new subjects or items (i.e., ones not present in the original study): there is no way to estimate their random intercept adjustments.

R formula syntax for random intercepts

In R we write the random intercept term as `(1 | X)` where `X` is the random effect variable. We can have multiple random intercepts:

```
RT ~ (1 | Subj) + (1 | Item)
```

Lexical decision model (after Baayen 2008)

```
> library(languageR)  # Install it if you haven't.
> library(lme4)       # Ditto.
> library(lmtest)
> print(names(lexdec))
 [1] "Subject"        "RT"          "Trial"        ...
 [5] "NativeLanguage" "Correct"     "PrevType"
 [9] "Word"           "Frequency"   "FamilySize"   ...
[13] "Length"         "Class"       "FreqSingular" ...
[17] "DerivEntropy"   "Complex"     "rInfl"        ...
[21] "SubjFreq"       "meanSize"    "meanWeight"   ...
[25] "BNCc"           "BNCd"        "BNCcRatio"
```

Simple linear regression

```
> d <- subset(lexdec, Correct == "correct")
> r <- lm(RT ~ NativeLanguage + Frequency + Trial, data = d)
> summary(r)
...

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.5483795  0.0248372 263.652   <2e-16 ***
NativeLanguageOther  0.1474721  0.0109455  13.473   <2e-16 ***
Frequency           -0.0426626  0.0042878  -9.950   <2e-16 ***
Trial               -0.0002163  0.0001144  -1.891   0.0588 .
...
```

Mixed linear regression (1/)

```
> r.mixed <- lmer(RT ~ NativeLanguage + Frequency + Trial +
+                   (1 | Subject) + (1 | Word), data = d)
```

Mixed linear regression (2/)

```
> summary(r.mixed)
...
Random effects:
 Groups   Name        Variance Std.Dev.
 Word     (Intercept) 0.003376 0.05811
 Subject  (Intercept) 0.016729 0.12934
 Residual             0.028633 0.16921
...
```

Mixed linear regression (3/)

```
> summary(r.mixed)
...
Fixed effects:
                       Estimate Std. Error t value
(Intercept)           6.551e+00  4.916e-02 133.250
NativeLanguageOther   1.521e-01  5.768e-02   2.637
Frequency            -4.378e-02  6.140e-03  -7.130
Trial                -1.930e-04  9.117e-05  -2.117
...
```

Hypothesis testing

It is difficult-to-impossible to compare fixed and mixed effects models using the likelihood ratio test because it is difficult to estimate just how many degrees of freedom are associated with a given random effect.

However, it is straightforward to compare two mixed effects models *so long as*:

- they have the same random effects structure, and
- one "nests" the other (in the likelihood ratio test sense),

because then we just need to know how they *differ* in degrees of freedom.

Testing for an effect of native language

```
> r.mixed.nativelanguage <- lmer(RT ~ Frequency + Trial +
+                              (1 | Subject) + (1 | Word),
+                              data = d)
> lrtest(r.mixed.nativelanguage, r.mixed)
...
Model 1: RT ~ Frequency + Trial + (1 | Subject) + (1 | Word)
Model 2: RT ~ NativeLanguage + Frequency + Trial + (1 |
Subject) + (1 | Word)
  #Df LogLik Df  Chisq Pr(>Chisq)
1   6 466.33
2   7 467.50  1 2.3453     0.1257
```

For next week

- *Random slopes*, interactions between fixed and mixed effects (e.g., maybe some subjects show less of a trial-fatigue effect than others?)
- *generalized* linear mixed effects models such as *logistic mixed effects regression*
- One last round of practice with model interpretation and reporting

Recommended reading on mixed effects models

Baayen (§7) does a good job both motivating and exemplifying mixed effects models. (It's a bit overlong because of all the case studies.)

Johnson (§4.4, §7.3-7.4) also is useful.

If you're interested in learning more about *ordinal regression*, read

- Baayen (§6.3.2) and
- Gorman & Johnson (2014: 226-229).

(This is all indicated on the course website.)

Questions? Please take them to email, or Slack.