Lecture 2: The Generalized Linear Mixed Model

LSA 2013, LI539

Mixed Effect Models

T. Florian Jaeger

Brain and Cognitive Sciences
University of Rochester

July 12, 2013

Generalized Linear Mixed Models

- Experiments don't have just one participant / Corpora don't just have one speaker.
  - Different participants may have different idiosyncratic behavior.
  - And items may have idiosyncratic properties, too.

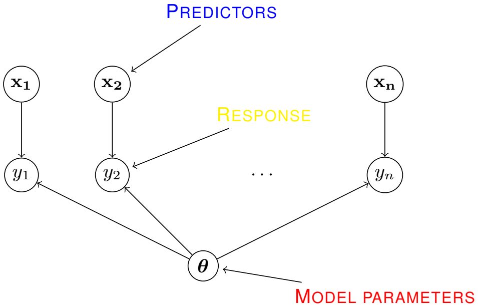
→ Violations of the assumption of independence!

NB: There may even be more clustered (repeated) properties and clusters may be nested (e.g. subjects ϵ dialects ϵ languages).

  - We'd like to take these into account, and perhaps investigate them.

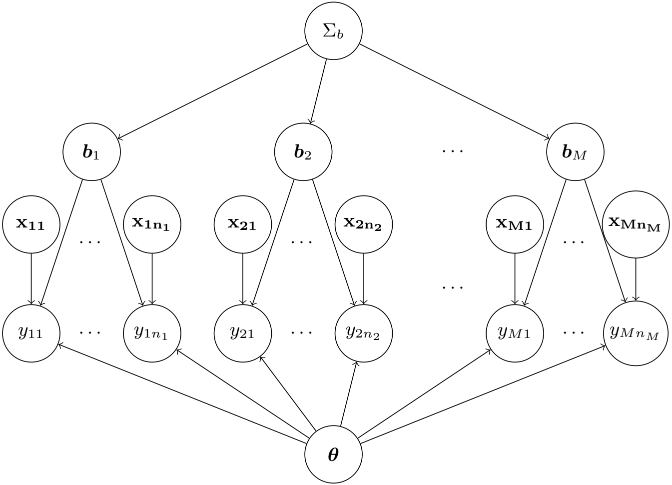
  → **Generalized Linear Mixed** or **Multilevel Models** (a.k.a. hierarchical, mixed-effects).

Generalized Linear Models



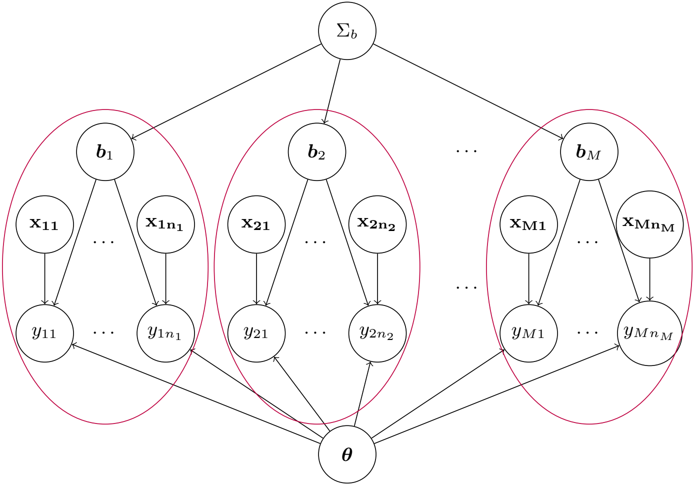
Generalized Linear Mixed Models

(provided by R. Levy)



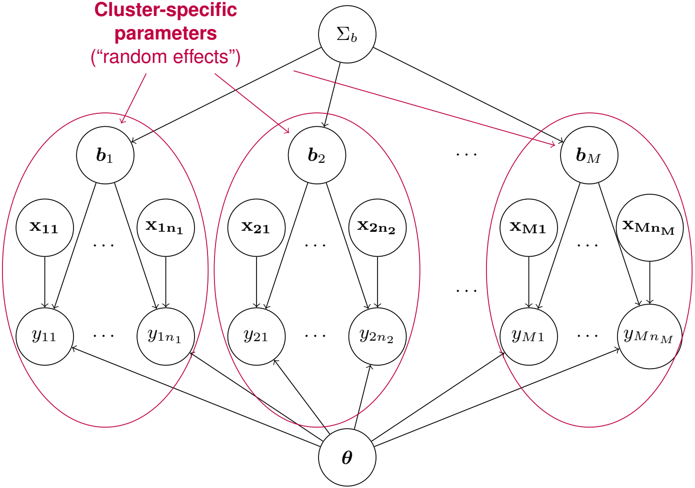
Generalized Linear Mixed Models

(provided by R. Levy)



Generalized Linear Mixed Models

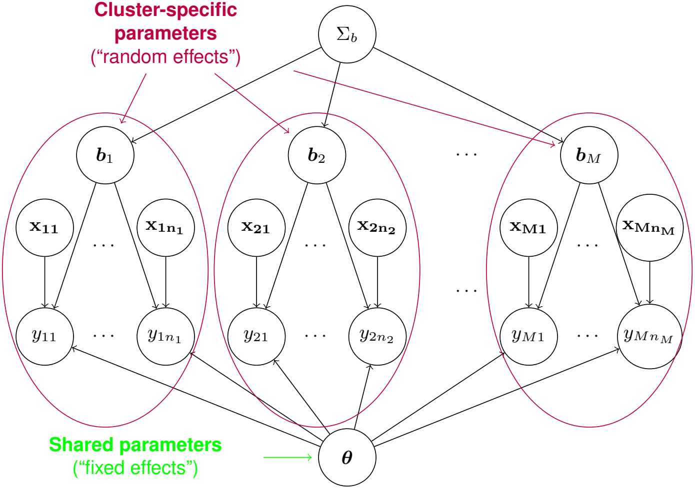
(provided by R. Levy)



Cluster-specific parameters
("random effects")

Generalized Linear Mixed Models

(provided by R. Levy)

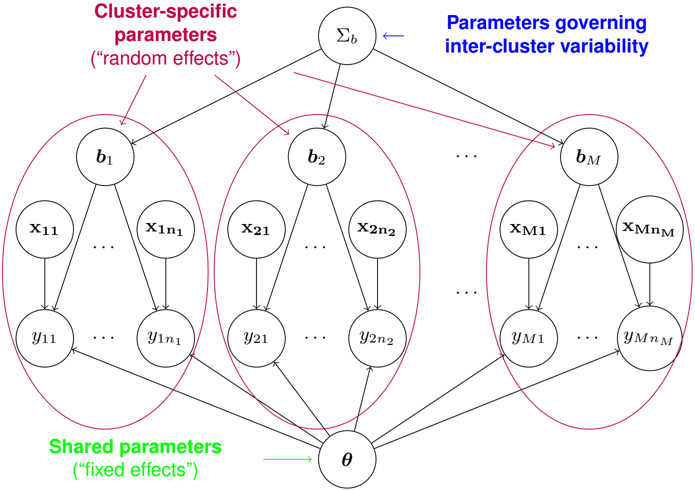


Cluster-specific parameters
(“random effects”)

Shared parameters
(“fixed effects”)

Generalized Linear Mixed Models

(provided by R. Levy)



Cluster-specific parameters
(“random effects”)

Parameters governing inter-cluster variability

Shared parameters
(“fixed effects”)

Introductions and Tutorials

- (Gelman & Hill, 2006): introduction to GLM and GLMM directed at *scientists in the social sciences*; Bayesian-inspired, but not fanatic; very intuitive; comes with R code, libraries, etc.; Gelman is also a contributor to `http://fivethirtyeight.blogs.nytimes.com/` and has a wonderful (and often funny) blog on multilevel/mixed modeling, `http://andrewgelman.com/`.
- (Baayen, Davidson, & Bates, 2008): introduction to linear mixed models directed at *psycholinguists*; one of the most cited JML articles of all times; parts are technical, but quite accessible; Douglas Bates, the main developer of `lme4`, is a co-author.
- (Jaeger, 2008): introduction to mixed logit models directed at *psycholinguists*; compares ANOVA and mixed logit models for the analysis of binary categorical data
- (Johnson, 2009): beautifully intuitive introduction to GLMMs directed at *sociolinguists*; compares GoldVarb with GLMM; great visualizations that showcase the differences; comes with `R` library for mixed effect GoldVarb, `Rbrul`
- (Jaeger, Graff, Croft, & Pontillo, 2011): introduction to LMMs directed at *typologists*; discusses fixed and random effects, challenges due to data sparsity, and simulation-based approaches to Type I&II error rate assessment (for mixed logit models, see also Dixon, 2008); explains how mixed models can be used to model genetic (family, subfamily, genera) relations as well as spational (language contact) effects.
- Roger Levy has lecture notes (in preparation for a book) on statistical models for linguists that also cover mixed models. He takes a graphical model approach.
- I don't do this often, but do not follow (Janssen, 2012) – it contains misleading advice about random effects (cf. Baayen et al., 2008; Barr, Levy, Scheepers, & Tily, 2013; Jaeger, 2008).

Linear Mixed Models: A simulated example

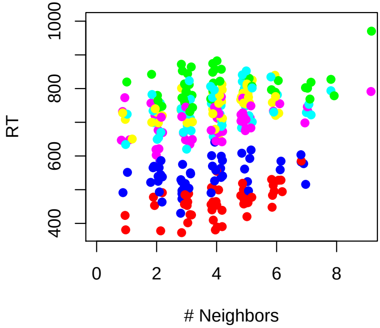
- Simulation of trial-level data can be invaluable for achieving deeper understanding of the data (based on example kindly provided by Roger Levy)

Linear Mixed Models: A simulated example

- Simulation of trial-level data can be invaluable for achieving deeper understanding of the data (based on example kindly provided by Roger Levy)

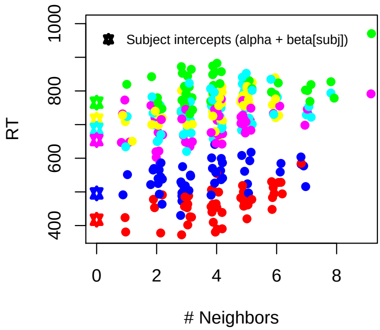
```
# parameters of simulation
sigma.b <- 125            # inter-subject variation larger than
sigma.e <- 40             # intra-subject, inter-trial variation
alpha <- 500              # intercept
beta <- 12                # slope (of neighborhood density, NHD)
M <- 6                    # number of participants
n <- 50                   # trials per participant
b <- rnorm(M, 0, sigma.b) # random by-subject intercept differences

# combine into data.frame
d = data.frame(
  Subject = rep(1:M,n),
  Neighbors = rpois(M*n,3) + 1,
  Noise = rnorm(M*n,0,sigma.e)
)
d$b = b[d$Subject]        # Subject intercept
d$RT =                    # simulate RTs!
  alpha +                 # add intercept
  beta * d$Neighbors +    # add effect of NHD
  d$b +                   # add by-subject intercept differences
  rnorm(M*n,0,sigma.e)    # add trial-level noise
```

A simulated example (cntd')



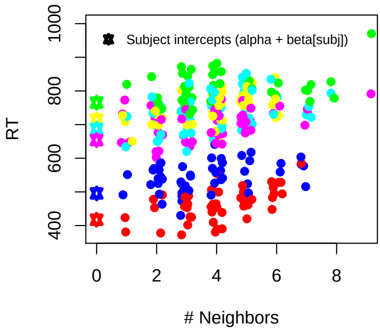
- Participant-level clustering is easily visible

A simulated example (cntd')



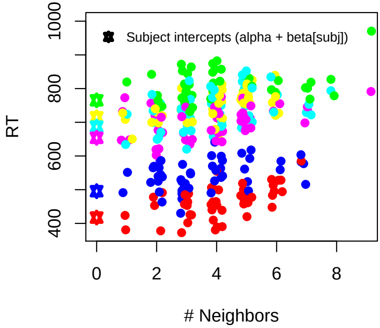
- Participant-level clustering is easily visible

A simulated example (cntd')



- Participant-level clustering is easily visible
- This reflects the fact that (simulated) inter-participant variation (125ms) is larger than (simulated) inter-trial variation (40ms)

A simulated example (cntd')



- Participant-level clustering is easily visible
- This reflects the fact that (simulated) inter-participant variation (125ms) is larger than (simulated) inter-trial variation (40ms)
- And the (simulated) effects of neighborhood density are also visible

Linear Mixed Models: A simulated example (cont'd)

```
# empirical SD of by-subject differences and noise
sd(b)

## [1] 92.94

sd(d$Noise)

## [1] 40.28

# analyze (simulated) data
library(lme4)
lmer(RT ~ Neighbors + (1 + Neighbors | Subject), d)

## Linear mixed model fit by REML
## Formula: RT ~ Neighbors + (1 + Neighbors | Subject)
##    Data: d
##   AIC  BIC logLik deviance REMLdev
##  3141 3163  -1565     3141    3129
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 7166.8   84.66
##           Neighbors      2.7    1.64    1.000
##  Residual             1862.7   43.16
## Number of obs: 300, groups: Subject, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   530.58      35.16   15.09
## Neighbors      10.21       1.67    6.13
```

Mixed Linear Model: Real data

- Back to our lexical-decision experiment:
- A variety of predictors seem to affect RTs, e.g.:
  - `Frequency`
  - `Trial`
  - `NativeLanguage`
  - Interactions

Mixed Linear Model: Real data

- Back to our lexical-decision experiment:
- A variety of predictors seem to affect RTs, e.g.:
  - `Frequency`
  - `Trial`
  - `NativeLanguage`
  - Interactions
- Additionally, different participants (a **grouping factor**) in your study may also have:
  - different overall decision speeds
  - differing sensitivity to e.g. Frequency.

Mixed Linear Model: Real data

- Back to our lexical-decision experiment:
- A variety of predictors seem to affect RTs, e.g.:
  - `Frequency`
  - `Trial`
  - `NativeLanguage`
  - Interactions
- Additionally, different participants (a **grouping factor**) in your study may also have:
  - different overall decision speeds
  - differing sensitivity to e.g. Frequency.
- We want to draw inferences about all these things at the same time or, about some of them, *conditional on* the others, including between-subject variability.

Mixed Linear Model

- Random effects, starting simple: let each participant j have idiosyncratic differences in RTs, $\mathbf{b_j}$.

Mixed Linear Model

- Random effects, starting simple: let each participant j have idiosyncratic differences in RTs, $\mathbf{b_j}$.
- Idea: Model distribution of subject differences as deviation from grand mean.
- Mixed models approximate these subject-specific deviations by assuming that these deviations are normally distributed.
- Grand mean reflected in ordinary intercept
  - → By-subject mean set to 0 (without loss of generality)
  - → Only additional parameter fit from data is standard deviatio.n

$$\boxed{RT_{ij} = \beta_0 + \beta_1 \texttt{Frequency}_{ij} + \overbrace{b_j}^{\text{By-subject differences}\sim N(0,\sigma_b)} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}}$$

Mixed linear model with one random intercept

$$\boxed{RT_{ij} = \beta_0 + \beta_1 \texttt{Frequency}_{ij} + \overbrace{b_j}^{\text{By-subject differences}\sim N(0,\sigma_b)} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}}$$

```
library(languageR)
library(lme4)
data(lexdec)
l1 = lmer(RT ~ Frequency + (1 | Subject), data=lexdec)
```

- ...where the term (...|Subject) can be read as '...by subjects'; this term specifies the by-subject **random effects** (in this case, only a by-subject **random intercept**)

NB: As was the case for `glm`, we can omit the `family` specification and the fixed effect intercept.

Interpretation of the output

$$RT_{ij} = \beta_0 + \beta_1 \texttt{Frequency}_{ij} + \overbrace{b_j}^{\text{By-subject differences}\sim N(0,\sigma_b)} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}$$

- Interpretation parallel to ordinary regression models:

```
## Linear mixed model fit by REML
## Formula: RT ~ Frequency + (1 | Subject)
##    Data: lexdec
##    AIC   BIC logLik deviance REMLdev
##   -858 -837    433     -881    -866
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 0.0237   0.154
##  Residual             0.0327   0.181
## Number of obs: 1659, groups: Subject, 21
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  6.58878    0.03773   174.6
## Frequency   -0.04287    0.00349   -12.3
##
## Correlation of Fixed Effects:
##           (Intr)
## Frequency -0.439
```

MCMC-sampling

- *t*-value anti-conservative

→ MCMC-sampling of coefficients to obtain non anti-conservative estimates

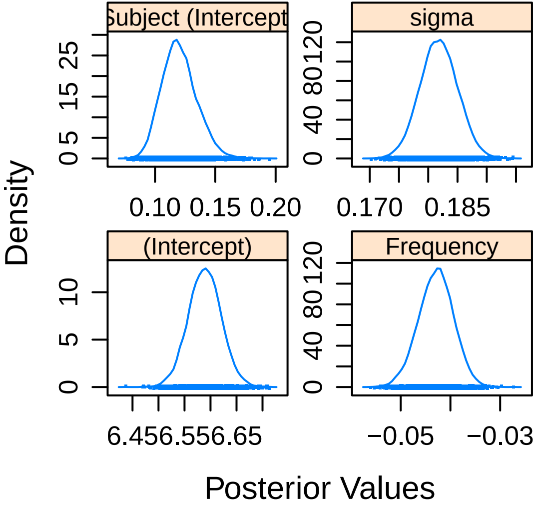
```
library(languageR)
p = pvals.fnc(l1, nsim = 10000, addPlot=F)
print(p)

## $fixed
##             Estimate MCMCmean HPD95lower HPD95upper  pMCMC Pr(>|t|)
## (Intercept)   6.5888   6.5891     6.5279     6.6495 0.0001        0
## Frequency    -0.0429  -0.0429    -0.0497    -0.0359 0.0001        0
##
## $random
##     Groups        Name Std.Dev. MCMCmedian MCMCmean HPD95lower HPD95upper
## 1  Subject (Intercept)   0.1541     0.1180   0.1197     0.0924     0.1494
## 2 Residual               0.1809     0.1817   0.1818     0.1755     0.1880
```

Interpretation of the output

- **estimates of coefficients** for fixed and random predictors.
- **predictions = fitted values**, just as for ordinary regression model.

```
cor(fitted(l1), lexdec$RT)^2

## [1] 0.4457
```

MCMC-sampling



MCMC-sampling: What does it do?

- `mcmcsamp()` (in the `lme4` package) uses Markov Chain Monte Carlo methods to generate samples from the **posterior distribution** of the parameters of a fitted (`lmer`) model

  NB: No normality is assumed during MCMC sampling
- The posterior distribution of parameter is their distribution conditional on the data and the assumptions of the (`lmer`) model.
- `pvals.fnc()` (in the `languageR` package) provides a summary (mean estimates, highest posterior density intervals) of the output of `mcmcsamp()`
- **Highest posterior intervals (HPDs)** are *one* common Bayesian equivalent of **confidence intervals**. HPDs are the shortest possible interval enclosing $(1-\alpha)\%$ of the posterior probability mass of a distribution.

  → HPDs are not necessarily symmetric around the mean.
- When applied to mixed models, whether we can reject the null hypothesis that a parameter is 0 (given the assumed significance level α)

`mcmcsamp()`: Technical details

- `mcmcsamp` uses **Gibbs sampling**
- In Gibbs sampling a (subset of) parameters is sampled conditional on assumed values for other (subsets of) parameters.
- When applied to mixed models, three subsets of parameters are independently sampled from:
  - the variance, σ_ϵ^2, of the per-observation noise
  - the parameters that determine the variance-covariance matrix, Σ_B, of the random effects (to be specific, sampling happens over the ST factor of the **Cholesky decomposition of the variance-covariance matrix**)
  - the random effects and the fixed effects (in the Bayesian formulation on the model the random effects are regarded as parameters).
- The starting values for mcmc-sampling from an `lmer` models are taken from the ML or REML fit of the model.
- For further detail, e.g., what we exactly sample from, see
  - `https://stat.ethz.ch/pipermail/r-help/2006-October/115585.html`
  - `https://stat.ethz.ch/pipermail/r-help/2006-August/110736.html`

MCMC-sampling: Example

```
m = mcmcsamp(l1, n=5, verbose=T)
str(m)

## Formal class 'merMCMC' [package "lme4"] with 9 slots
##   ..@ Gp      : int [1:2] 0 21
##   ..@ ST      : num [1, 1:5] 0.851 0.86 0.85 0.826 0.751
##   ..@ call    : language lmer(formula = RT ~ Frequency + (1 | Subject),
##   ..@ deviance: num [1:5] -881 -881 -881 -881 -881
##   ..@ dims    : Named int [1:18] 1 1659 2 21 1 1 1 1 2 5 ...
##   .. ..- attr(*, "names")= chr [1:18] "nt" "n" "p" "q" ...
##   ..@ fixef   : num [1:2, 1:5] 6.5888 -0.0429 6.5554 -0.0377 6.5631 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:2] "(Intercept)" "Frequency"
##   .. .. ..$ : NULL
##   ..@ nc      : int 1
##   ..@ ranef   : num[1:21, 0 ]
##   ..@ sigma   : num [1, 1:5] 0.181 0.179 0.181 0.186 0.186
```

MCMC-sampling: Example (cont'd)

```
m@ST

##        [,1]   [,2]   [,3]   [,4]   [,5]
## [1,] 0.8514 0.8604 0.8505 0.8263 0.7509

m@sigma

##        [,1]   [,2]   [,3]   [,4]   [,5]
## [1,] 0.1809 0.1793 0.1807 0.1864 0.1863

m@fixef

##                 [,1]    [,2]     [,3]     [,4]     [,5]
## (Intercept)  6.58878  6.5554  6.56309  6.68895  6.57046
## Frequency   -0.04287 -0.0377 -0.04182 -0.04715 -0.03655
```

- These values (samples of parameters drawn from the posterior distribution –as estimated by the MCMC method) are what `pvals.fnc()` bases the estimated HPD on that we use to test significance in the linear mixed model, thereby avoiding anti-conservative t-tests.

GLMM vs. GLM

- Mixed models (GLMMs) **inherit all advantages from Generalized Linear Models**, but are computationally and conceptually more complex.
- **From independence to conditional independence:** GLMM revises the assumption of independence and instead assumes *conditional independence*. For GLMs, we assumed that our observations $y_1, \ldots, y_n$ are independently sampled; for GLMMs, we assume that they are independent once the grouping factors are taken into account.

GLMM vs. GLM

- Mixed models (GLMMs) **inherit all advantages from Generalized Linear Models**, but are computationally and conceptually more complex.
- **From independence to conditional independence:** GLMM revises the assumption of independence and instead assumes *conditional independence*. For GLMs, we assumed that our observations $y_1, \ldots, y_n$ are independently sampled; for GLMMs, we assume that they are independent once the grouping factors are taken into account.

NB: GLMMs allow us to make assumptions about independence/conditional independence for each parameter in the model. That is powerful, but also a source of confusion (see below).

  - For example, in the example (linear) mixed model above we evaluated the intercept under the assumption of conditional independence, but we evaluated the `Frequency` effect under the stronger (less conservative) assumption of independence (no random by-subject slope for `Frequency` was included in the model).

Mixed vs. fixed effects for grouping factors

- The way GLMMs account for clusters in the data is efficient:
  - In the above example, 1 parameter was sufficient to model differences between k =21 subjects in their base RT.
  - Had we models these diffferences as fixed effects, we would have required 20 parameters ($k - 1$).

→ The assumption that between-cluster differences are normally distributed saves $k - 2$ degrees of freedom for individual differences in intercepts alone.

Mixed vs. separate by-cluster regressions

- An alternative way to account for individual differences are tests over separate by-cluster regressions (cf. Lorch & Myers, 1990):
  - Split the data by level of the grouping variable (e.g., subjects)
  - Step 1: Fit the same ordinary (GLM) regression model to each cluster
  - Step 2: Analyze the distribution of coefficient estimates from Step 1 to assess whether they are different from 0.
- GLMMs provide more power than by-cluster regression (e.g., Dixon, 2008)

  (in the alternative approach, information about the amount of data that went into each by-cluster estimate is not available at the final step of analysis: e.g., the difference between a model fit a subject with 10 data points compared to a subject with 100 data points)

GLMM vs. repeated-measures ANOVA

- Both repeated-measures ANOVA and GLMM are suited for the analysis of repeated-measures data (for which the assumption of independence is not warranted, cf. Clark (1973)).
- Some of the advantages of ANOVA include:
  - ANOVA is more commonly taught and more of your colleagues will be familiar with it (i.e., can provide better feedback)
  - Our fields have established standards on how to conduct and report ANOVA (though those can be a two-edged sword)
  - ANOVA is a robust tool that *typically* work even when some of its assumptions are violated (e.g., normality, sphericity; but see Dixon (2008); Jaeger (2008)).
  - ANOVA is computationally cheap (though this advantage is increasingly less important).

Some advantages of GLMM

- Like GLMs, appropriate models exists for a variety of types of dependent variables (unlike ANOVA, which is *intended* for continuous Gaussian data)
- Like for GLMs, the framework encourages us to think about the coding and conceptualization of our effects (though this advantage depends on the discipline of the user)
- GLMMs readily extend situations with multiple crossed grouping factors (e.g., subjects and items). ANOVA requires multiple separate analyses (F1, F2) that are known to be anti-conservative, but can be made conservative by combining them in an additional step (minF, cf. Clark (1973))
- GLMMs provide greater insight into the random effect structure:
  - We can specify random effect structures (this freedom comes with risks for the uninitiated)

    NB: For example, the linear mixed model shown above only contains a random intercept rather than random effects for *all* fixed effects.
  - We can look at the random difference between levels of a grouping variable.
  - We can explore the covariation structure of the random effects (within the random effects for a grouping factor)

So, GLMM or ANOVA?

- In all the current frenzy about mixed models, don't forget that the most important part is to understand your data.
- It's important that your familiar and comfortable with whatever analysis you choose.
- In case of doubt, perhaps risk loosing some power, but use what you understand (e.g., for balanced data the Type I and Type II error advantage of mixed models is negligible to non-existent).
- E.g., use ANOVA if you aren't comfortable with linear mixed models *unless* the assumptions of the ANOVA are *severly* violated.

→ GLMMs are a very powerful tool, but with the ability to do more also comes the risk of doing more wrongly.

What about all the things that could go wrong?

- Almost all common issues with GLMMs are also common issues with ordinary regression (GLMs).
  - Some of these issues are due to the fact that we can now analyze variables with ditributional properties that ANOVA isn't suited for (e.g., categorical data, count data) and those types of *variables* (not GLM or GLMM) come with their own challenges.
  - Other issues are either shared with ANOVA (overly influential cases, outliers) or are due to the relaxed constraints on the 'shape' of the data (e.g., unbalanced and sparse data or lack of full factorial balanced designs)
- Some issues are specific to –or at least more common in– mixed models:
  - Random effects
  - Increased computational complexity and issues with convergence.

Random effects: Examining individual differences

- Let's look at the by-subject adjustments to the intercept. These are called **B**est **U**nbiased **L**inear **P**redictors (**BLUPs**)
    - BLUPs are *not* fitted parameters. Only one degree of freedom was added to the model. The BLUPs are estimated posteriori based on the fitted model.

$$P(b_i|\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}_b, \widehat{\sigma}_\epsilon, \mathrm{X})$$

Random effects: Examining individual differences

- Let's look at the by-subject adjustments to the intercept. These are called **B**est **U**nbiased **L**inear **P**redictors (**BLUPs**)
  - BLUPs are *not* fitted parameters. Only one degree of freedom was added to the model. The BLUPs are estimated posteriori based on the fitted model.

$$P(b_i|\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}_b, \widehat{\sigma}_\epsilon, \mathrm{X})$$

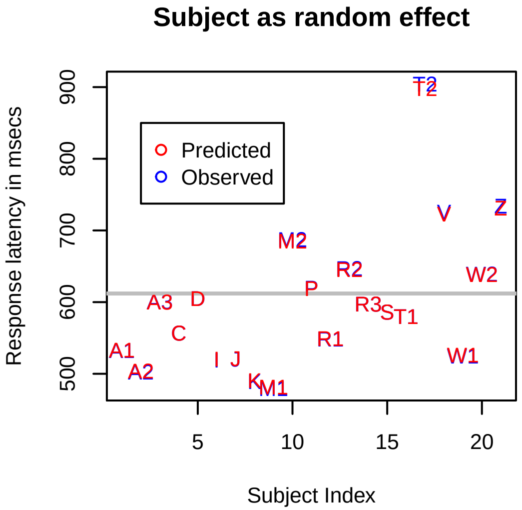
  - The BLUPs are the **conditional modes** of the $\mathbf{b_i}$s—the choices that maximize the above probability

Examining individual differences

NB: By-subjects adjustments are assumed to be centered around zero, but they don't necessarily do so:

```
mean(ranef(l1)$Subject[[1]])

## [1] -1.823e-13

head(ranef(l1)$Subject[1], 5)

##    (Intercept)
## A1    -0.10552
## A2    -0.16189
## A3     0.01243
## C     -0.06170
## D      0.02062
```

Shrinkage: Observed vs. fitted by-subject differences



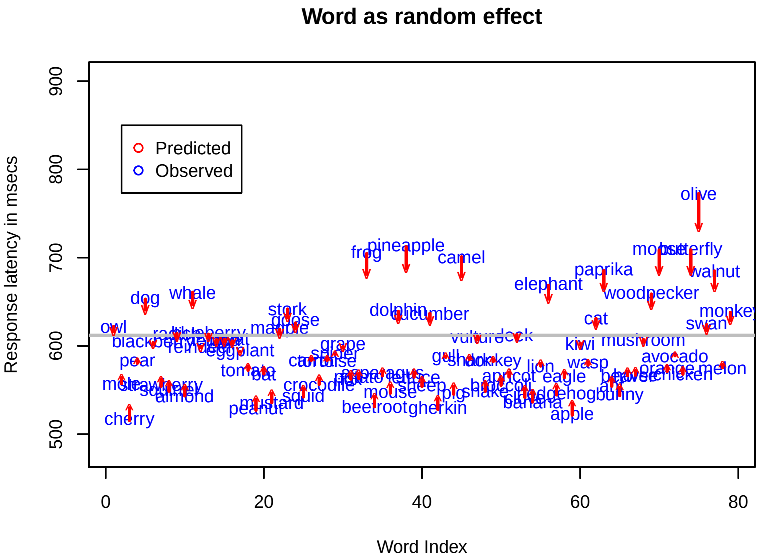
(for a nice introduction to shrinkage, see Kliegl, Masson, and Richter (2010))

Mixed models with several grouping factors

- Unlike with ANOVA, the linear mixed model can accommodate more than one random intercept
- These are *crossed* random effects.

```
l2 = lmer(RT ˜ 1 + (1 | Subject) + (1 | Word),
          data = lexdec)
head(ranef(l2)$Subject, 5)

##    (Intercept)
## A1    -0.10568
## A2    -0.16213
## A3     0.01245
## C     -0.06179
## D      0.02065

head(ranef(l2)$Word, 5)

##           (Intercept)
## almond        0.03859
## ant          -0.04993
## apple        -0.11279
## apricot      -0.01725
## asparagus     0.07771
```

Shrinkage: Observed vs. fitted by-subject differences



Mixed models with random slopes

- Not only the intercept, but any of the slopes (of the predictors) may differ between individuals.
- For example, subjects may show different sensitivity to `Frequency`:

```
l3 = lmer(RT ~ 1 + Frequency + (1 + Frequency | Subject) + (1 | Word),
          data = lexdec)
```

Mixed models with random slopes (cont'd)

```
print(l3, corr=F)

## Linear mixed model fit by REML
## Formula: RT ~ 1 + Frequency + (1 + Frequency | Subject) + (1 | Word)
##    Data: lexdec
##   AIC  BIC logLik deviance REMLdev
##  -943 -905    479     -970    -957
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Word     (Intercept) 0.00297  0.0545
##  Subject  (Intercept) 0.05647  0.2376
##           Frequency   0.00041  0.0202   -0.918
##  Residual             0.02917  0.1708
## Number of obs: 1659, groups: Word, 79; Subject, 21
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  6.58878    0.05925   111.2
## Frequency   -0.04287    0.00731    -5.9
```

NB: By default, R also fits a random effect for the covariance(s) between all random slopes within a grouping factor (shown above in the column `Corr` as the correlation between the variances). We'll return to that.

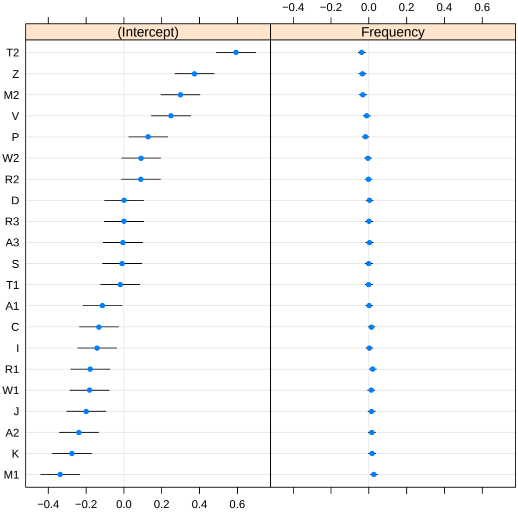
Mixed models with random slopes

- The BLUPs of the random slope reflect the by-subject adjustments to the overall `Frequency` effect.

```
head(ranef(l3)$Subject, 5)

##     (Intercept)  Frequency
## A1    -0.113082   0.002002
## A2    -0.237502   0.015898
## A3    -0.005239   0.003483
## C     -0.132056   0.014383
## D      0.001134   0.003810

head(ranef(l3)$Word, 5)

##              (Intercept)
## almond           0.01666
## ant             -0.02480
## apple           -0.04996
## apricot         -0.04152
## asparagus        0.03392
```

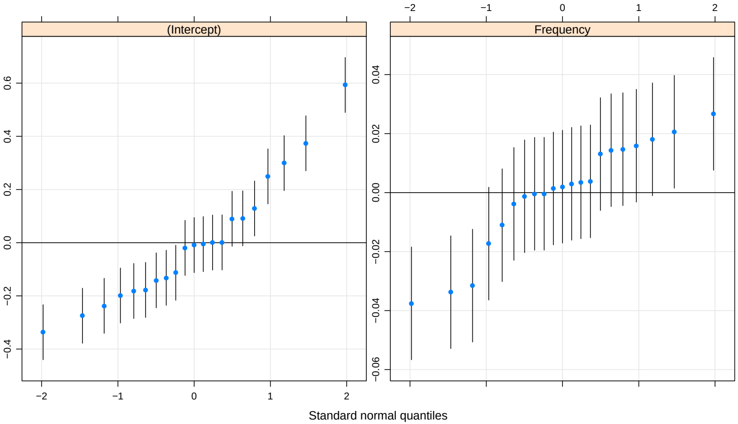
Visualizing the random effects

```
dotplot(ranef(l3))
```



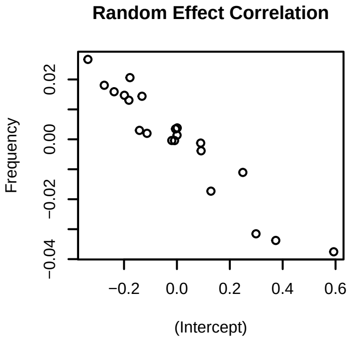
Visualizing the random effects

```
qqmath(ranef(l3))
```



Visualizing random effect correlations

```
par(cex=.5)
plot(ranef(l3)$Subject, main="Random Effect Correlation")
```



Question

- What does this correlation mean?

Visualizing random effect correlations

Answer

- The plot shows us that there is a negative correlation between the modes of the by-subject random intercepts and the modes of the by-subject random slopes for word frequency

  (this is reflected in the negative correlations of the two variances, -0.918, see above)

Visualizing random effect correlations

Answer

- The plot shows us that there is a negative correlation between the modes of the by-subject random intercepts and the modes of the by-subject random slopes for word frequency

  (this is reflected in the negative correlations of the two variances, -0.918, see above)

→ Subjects with a higher `intercept` tend to have smaller slopes for `Frequency`.

Question

- According to this model, do slower readers have larger or smaller effects of word frequency?

Visualizing random effect correlations

Answer

- Notice first that the axis of the plot are the by-subject BLUPs for the `intercept` and `Frequency`.

→ we should interpret them relative to the fixed coefficient estimates.

→ slower readers exhibit larger (more negative) effects of word frequency

Random effect correlations are sensitive to predictor transformation

NB: Random effect correlations can change (quantitatively or even qualitatively) when we transform a continuous predictor or change the coding of a categorical predictor.

- For example, here's the same model after centering `Frequency`:

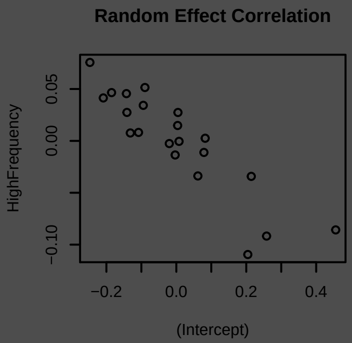
```
lexdec$cFrequency = lexdec$Frequency - mean(lexdec$Frequency)
l3b = lmer(RT ~ 1 + cFrequency + (1 + cFrequency | Subject),
          data= lexdec)
print(l3b, corr=F)

## Linear mixed model fit by REML
## Formula: RT ~ 1 + cFrequency + (1 + cFrequency | Subject)
##    Data: lexdec
##   AIC  BIC logLik deviance REMLdev
##  -876 -844    444     -902    -888
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 0.023740 0.1541
##           cFrequency  0.000387 0.0197   -0.814
##  Residual             0.032136 0.1793
## Number of obs: 1659, groups: Subject, 21
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  6.38509    0.03391   188.3
## cFrequency  -0.04287    0.00551    -7.8
```

Random effect correlations and predictor coding

- Here's an illustration that predictor coding, too, affects random effect correlations.

```
lexdec$HighFrequency =
    ifelse(lexdec$Frequency > median(lexdec$Frequency), 1, 0)
l4 = lmer(RT ~ 1 + HighFrequency + (1 + HighFrequency | Subject),
          data= lexdec)
print(l4, corr=F)

## Linear mixed model fit by REML
## Formula: RT ~ 1 + HighFrequency + (1 + HighFrequency | Subject)
##    Data: lexdec
##   AIC  BIC logLik deviance REMLdev
##  -912 -879    462     -935    -924
## Random effects:
##  Groups   Name          Variance Std.Dev. Corr
##  Subject  (Intercept)   0.03086  0.176
##           HighFrequency 0.00314  0.056    -0.811
##  Residual               0.03141  0.177
## Number of obs: 1659, groups: Subject, 21
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)     6.4440     0.0388   166.0
## HighFrequency  -0.1192     0.0150    -7.9
```

Visualization

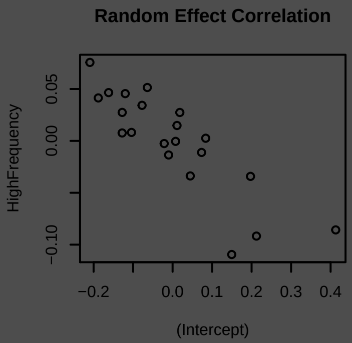


Random effect correlations and predictor coding

```
lexdec$HighFrequency = ifelse(lexdec$Frequency > median(lexdec$Frequency),
          .5, -.5)
l4b = lmer(RT ~ 1 + HighFrequency + (1 + HighFrequency | Subject),
          data= lexdec)
print(l4b, corr=F)

## Linear mixed model fit by REML
## Formula: RT ~ 1 + HighFrequency + (1 + HighFrequency | Subject)
##    Data: lexdec
##   AIC  BIC logLik deviance REMLdev
##  -912 -879    462     -935    -924
## Random effects:
##  Groups   Name          Variance Std.Dev. Corr
##  Subject  (Intercept)   0.02367  0.154
##           HighFrequency 0.00314  0.056    -0.744
##  Residual               0.03141  0.177
## Number of obs: 1659, groups: Subject, 21
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)     6.3843     0.0339   188.6
## HighFrequency  -0.1192     0.0150    -7.9

par(cex=.5)
plot(ranef(l4b)$Subject, main="Random Effect Correlation")
```

Random Effect Correlation

Visualization



Some useful commands

```
# use vcov(model) for variance-covariance matrix of fixed effects
# these determiner the standard errors of the fixed effects.
vcov(l4b)

## 2 x 2 Matrix of class "dpoMatrix"
##               [,1]       [,2]
## [1,]  0.0011460 -0.0003048
## [2,] -0.0003048  0.0002253


# get standard errors of fixed effects:
sqrt(diag(vcov(l4b)))

## [1] 0.03385 0.01501
```

Some useful commands (cont'd)

```
# use VarCorr(model) for variance-covariance matrix of random effects
# unlike the fixed effect variance-covariance matrix, only the *structure*
# of the VarCorr is known, the values are parameters that are estimated
# when we fit the model.
VarCorr(l4b)

## $Subject
##                (Intercept) HighFrequency
## (Intercept)        0.02367      -0.00641
## HighFrequency     -0.00641       0.00314
## attr(,"stddev")
##   (Intercept) HighFrequency
##       0.15384       0.05603
## attr(,"correlation")
##                (Intercept) HighFrequency
## (Intercept)         1.0000       -0.7436
## HighFrequency      -0.7436        1.0000
##
## attr(,"sc")
## [1] 0.1772


# for further details, see
# https://stat.ethz.ch/pipermail/r-help/2006-July/109308.html
```

Confidence intervals on random effects

- Sometimes you might want to get CIs for the random effects (e.g., are the variances significantly different from zero?)
- `lmer` does purposefully *not* provide standard error estimates for the estimates of the variance-covariance of the random effects

  (because the distribution of variance estimates is expected to be *not symmetric*, making the use of SE-based CIs problematic).
- Instead, it is recommended that MCMC-sampling is used:

```
s = mcmcsamp(l2, 50000)
HPDinterval(s)
```

Confidence intervals on random effects

- As a reminder, this was our l2 model:

```
## Linear mixed model fit by REML
## Formula: RT ˜ 1 + (1 | Subject) + (1 | Word)
##    Data: lexdec
##   AIC  BIC logLik deviance REMLdev
##  -891 -869    449     -904    -899
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Word     (Intercept) 0.00591  0.0768
##  Subject  (Intercept) 0.02377  0.1542
##  Residual             0.02984  0.1727
## Number of obs: 1659, groups: Word, 79; Subject, 21
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    6.385      0.035     182
```

Confidence intervals on random effects (cont'd)

```
## $fixef
##             lower upper
## (Intercept)  6.33 6.438
## attr(,"Probability")
## [1] 0.95
##
## $ST
##       lower  upper
## [1,] 0.3169 0.4647
## [2,] 0.5220 0.8349
## attr(,"Probability")
## [1] 0.95
##
## $sigma
##       lower upper
## [1,] 0.1686 0.181
## attr(,"Probability")
## [1] 0.95
```

Technical note: Cholesky decomposition

- Variance-covariance matrices are square, symmetrical matrices containing variances along the diagonal elements and covariances in the off-diagonal elements.
- The `ST` slot of a `mer-class` object (the output of `lmer()`) contains

  > *A list of S and T factors in the TSST' Cholesky factorization of the relative variance matrices of the random effects associated with each random-effects term.*

  . . .

  (see `help(‘‘mer-class’’)`)
- Cholesky decomposition deconstructs any n-by-n positive definite covariance matrix (all variance-covariance matrices are assumed to be positive definite) into an n-by-n triangular matrix (a matrix with zeros above the diagonal), postmultiplied by its transpose. (see also `http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf`, p. 22)

Technical note: Cholesky decomposition

- the `ST` values of an `lmer` object can be turned into variance-covariance matrices via `VarCorr` (see `?VarCorr`).

```
HPDinterval(VarCorr(s, type = "varcov"))

##          lower    upper
## [1,] 0.003052 0.006495
## [2,] 0.008181 0.020966
## [3,] 0.028418 0.032728
## attr(,"Probability")
## [1] 0.95


# (given in the order they appear in l2: variance for Word intercept,
# Subject intercept, Residual)
```

(cf. the estimates of the random effects in model l2 above, which this is based on)

What random slopes should I include?

- A mixed model with random slopes for all its predictors (incl. random intercept) is comparable in structure to an ANOVA

→ at least within the tradition of ANOVA designs, we should use mixed models with maximal random effect structures (Barr et al., 2013) (see also Jaeger, 2008; Baayen et al., 2008).

What random slopes should I include?

- A mixed model with random slopes for all its predictors (incl. random intercept) is comparable in structure to an ANOVA

→ at least within the tradition of ANOVA designs, we should use mixed models with maximal random effect structures (Barr et al., 2013) (see also Jaeger, 2008; Baayen et al., 2008).

- However, maximal random effect structures often don't even converge for balanced designs with lots of data. For more interesting (harder to analyze) data, models with *any* random slope often won't converge. Taken as an absolute standard, keeping it maximal is problematic (see also Jaeger et al., 2011).

Some suggestions

- Do *not* assume that only random intercepts are required (see Baayen et al., 2008; Jaeger, 2008). (actually, there seems to be only one paper that has ever suggested this, Janssen (2012))
- Feel free to explore the random effect structure. It's one of the appealing features of mixed models.
- Differences between **keeping it maximal** or using the **maximal random effect (RE) structure justified by model comparison** (the only two realistic options) are usually negligible (compared to, for example, common mistakes in statistical pratice; cf. Simmons, Nelson, and Simonsohn (2011)).
- For some types of data **keeping it maximal** seems to cause problems:
    - For highly heterogenous and sparse data (overfitting; lack of any insight)
    - For dichotomous data with many categorical cells

Maximal RE structure justified by model comparison

- One great feature of Mixed Models is that we can *assess* whether a certain random effect structure is actually warranted given the data.
- Just as nested ordinary regression models can be compared (cf. **stepwise regression**), we can compare models with nested random effect structures.
- Here, **model comparison** shows that the covariance parameter of `l3` significantly improves the model compared to `l2` with both the random intercept and slope for subjects, but no covariance parameter ($\chi^2(1) = 21.6$, $p < 0.0001$).
- The random slope overall is also justified ($\chi^2(2) = 24.1$, $p < 0.0001$).

Random vs. fixed effects

- A FAQ is when something should be considered a **random effect**, rather than a **fixed effect**

Random vs. fixed effects

- A FAQ is when something should be considered a **random effect**, rather than a **fixed effect**

- First off, not everybody likes this terminology since it can be misleading and since none of the frequently used criteria is absolute (Gelman & Hill, 2006)
- That said, here are a couple of considerations:
  - when sampling further data from our **population** would increase the number of levels of a factor, it is often considered a random effect
  - when we assume that differences in mean of an effect (i.e., the fixed effect **coefficients**) are randomly distributed with regard to another factor, that latter factor is commonly considered a random effect
  - Convention also often tells us what to do. For example, in psycholinguistics it is standard to consider both subjects and items as random effects.

Modeling a hierarchy of clusters (multilevel)

- Sometimes we want to recognize that there is a hierarchy of random effects.
- For example,
    - A sociolinguist analyzing behavioral repeated measures data from multiple speakers might consider random effects for subjects and (perhaps multiple levels of) dialect region
    - A typologist analyzing linguistic properties of languages in a sample might want to model random effects of language family, subfamily, genus, and alike (e.g., Atkinson, 2011; Jaeger et al., 2011).

An example

- Atkinson (2011) modeled phonological complexity of languages in such a model. He wanted to test the hypothesis that, during human migration, repeated founder events (groups splitting off, migrating, and creating new language communities) led to a shrinking of the phonological inventory. In his words,

  > *A series of founder events should produce a gradient of decreasing phonemic diversity with increasing distance from the origin [of language].*

- Atkinson obtain measures of phonological complexity (standardized number of consonants, vowels, and tones) from WALS (Haspelmath et al., 2005).
- He then predicted the complexity of a language's phonology as a function of its (log-transformed) population size, the distances of its population center from the origin of language, and the interaction of these two factors.

An example (cont'd)

- However, languages do not constitute independent samples. So, to capture that Atkinson inlcuded crossed random intercepts by language family, subfamily, and genus.

NB: when a 'lower' grouping factor (e.g., Subfamily) has no value (e.g., because only one sample from the higher grouping factor, e.g., Family, is available) we code that as 0.

Task

Let's think about why this way of specifying the random effects makes sense. Write down the linear predictor of such a model. Think about the BLUPs.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from africa. *Science*, *332*(6027), 346–349.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, *12*(4), 335–359.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434–446.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*(2), 281–320.

Janssen, D. P. (2012). Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants. *Behavior research methods*, *44*(1), 232–247.

Johnson, D. E. (2009). Getting off the goldvarb standard: Introducing rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, *3*(1), 359–383.

Kliegl, R., Masson, M. E., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition*, *18*(5), 655–681.

Lorch, R., & Myers, J. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Memory and Language*, *16*(1), 149–157.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.