An Introduction to Linear and Logit Multilevel Models

Day 1

Florian Jaeger

M. Gillespie & P. Graff

May 3, 2010

Overview - Day 1

- **Lecture 1:**
  - (Re-)Introducing Ordinary Regression
  - Comparison to ANOVA
  - General*ized* Linear Models
  - Generalized Linear Mixed Models (Multilevel Models)
  - Trade-offs
- **Talk(s):**
  - Efficiency in Production
  - Syntax in Flux
- **Tutorial 1: Contrast Coding (M. Gillespie)**
  - Implementing specific hypotheses
  - Coding types: treatment, effect (sum), Helmert, and polynomial coding
  - Interactions: centering

Overview - Day 2

- **Tutorial 2: Interactions, Centering, and more (M. Gillespie)**
- **Lecture 2:**
  - Common Issues and Solutions in Regression Modeling (Mixed or not)
    - outliers
    - collinearity
    - model evaluation
- **Tutorial 3: Testing Linguistic Theories with Logistic Regression (P. Graff)**
  - Nested and non-nested model comparison: AIC, BIC, etc.
- **Tutorial 4: BYOD - Group Therapy** (M. Gillespie, P. Graff, F. Jaeger)
- **Please ask/add to the discussion any time!**

Acknowledgments

- I've incorporated (and modified) a couple of slides prepared by:
  - Victor Kuperman (Stanford)
  - Roger Levy (UCSD)

  ... with their permission (naturalmente!)
- I am also grateful for feedback from:
  - Austin Frank (Rochester)
  - Previous audiences to similar workshops at CUNY, Haskins, Rochester, Buffalo, UCSD, MIT.
- For more materials, check out:
  - http://www.hlp.rochester.edu/
  - http://wiki.bcs.rochester.edu:2525/HlpLab/StatsCourses
  - http://hlplab.wordpress.com/ (e.g. multinomial mixed models code)

Generalized Linear Models

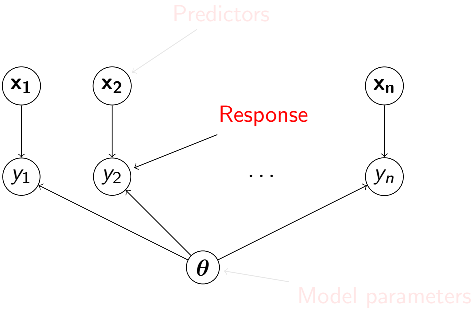
Goal: model the effects of predictors (independent variables) $\mathbf{X}$ on a response (dependent variable) Y.

The picture:

Generalized Linear Models

Goal: model the effects of predictors (independent variables) $\mathbf{X}$ on a response (dependent variable) Y.

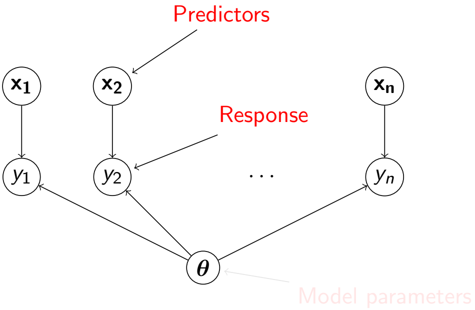
The picture:



Generalized Linear Models

Goal: model the effects of predictors (independent variables) $\mathbf{X}$ on a response (dependent variable) Y.

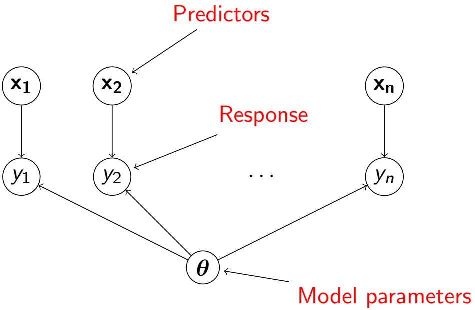
The picture:



Generalized Linear Models

Goal: model the effects of predictors (independent variables) $\mathbf{X}$ on a response (dependent variable) Y.

The picture:



Reviewing GLMs

Assumptions of the generalized linear model (GLM):

- Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
- η is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \qquad \text{(linear predictor)}$$

- η determines the predicted mean μ of Y

$$\eta = g(\mu) \qquad \text{(link function)}$$

- There is some noise distribution of Y around the predicted mean μ of Y:

$$P(Y = y; \mu)$$

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

- Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
- η is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \qquad \text{(linear predictor)}$$

- η determines the predicted mean μ of Y

$$\eta = g(\mu) \qquad \text{(link function)}$$

- There is some noise distribution of Y around the predicted mean μ of Y:

$$P(Y = y; \mu)$$

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

- Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
- η is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \qquad \text{(linear predictor)}$$

- η determines the predicted mean μ of Y

$$\eta = g(\mu) \qquad \text{(link function)}$$

- There is some noise distribution of Y around the predicted mean μ of Y:

$$P(Y = y; \mu)$$

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

- Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
- η is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \qquad \text{(linear predictor)}$$

- η determines the predicted mean μ of Y

$$\eta = g(\mu) \qquad \text{(link function)}$$

- There is some noise distribution of Y around the predicted mean μ of Y:

$$P(Y = y; \mu)$$

Reviewing GLMs

Assumptions of the generalized linear model (GLM):

- Predictors $\{X_i\}$ influence Y through the mediation of a linear predictor η;
- η is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \qquad \text{(linear predictor)}$$

- η determines the predicted mean μ of Y

$$\eta = g(\mu) \qquad \text{(link function)}$$

- There is some noise distribution of Y around the predicted mean μ of Y:

$$P(Y = y; \mu)$$

Reviewing Linear Regression

Linear regression, which underlies ANOVA, is a kind of generalized linear model.

- The predicted mean is just the linear predictor:

$$\eta = l(\mu) = \mu$$

- Noise is normally (=Gaussian) distributed around 0 with standard deviation σ:

$$\epsilon \sim N(0, \sigma)$$

- This gives us the traditional linear regression equation:

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean } \mu = \eta} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

Reviewing Linear Regression

Linear regression, which underlies ANOVA, is a kind of generalized linear model.

- The predicted mean is just the linear predictor:

$$\eta = l(\mu) = \mu$$

- Noise is normally (=Gaussian) distributed around 0 with standard deviation σ:

$$\epsilon \sim N(0, \sigma)$$

- This gives us the traditional linear regression equation:

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean } \mu = \eta} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

Reviewing Linear Regression

Linear regression, which underlies ANOVA, is a kind of generalized linear model.

- The predicted mean is just the linear predictor:

$$\eta = l(\mu) = \mu$$

- Noise is normally (=Gaussian) distributed around 0 with standard deviation σ:

$$\epsilon \sim N(0, \sigma)$$

- This gives us the traditional linear regression equation:

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean } \mu = \eta} + \overbrace{\epsilon}^{\text{Noise} \sim N(0,\sigma)}$$

Reviewing Linear Regression

Linear regression, which underlies ANOVA, is a kind of generalized linear model.

- The predicted mean is just the linear predictor:

$$\eta = l(\mu) = \mu$$

- Noise is normally (=Gaussian) distributed around 0 with standard deviation σ:

$$\epsilon \sim N(0, \sigma)$$

- This gives us the traditional linear regression equation:

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean } \mu = \eta} + \overbrace{\epsilon}^{\text{Noise}\sim N(0,\sigma)}$$

Reviewing Logistic Regression

Logistic regression, too, is a kind of generalized linear model.

- The linear predictor:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$$

- The link function g is the logit transform:

$$\mathrm{E(Y)} = p = \mathrm{g}^{-1}(\eta) \Leftrightarrow$$
$$\mathrm{g}(p) = \ln \frac{p}{1-p} = \eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n \quad (1)$$

- The distribution around the mean is taken to be binomial.

Reviewing Logistic Regression

Logistic regression, too, is a kind of generalized linear model.

- The linear predictor:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$$

- The link function g is the logit transform:

$$\mathrm{E}(\mathrm{Y}) = p = \mathrm{g}^{-1}(\eta) \Leftrightarrow$$
$$\mathrm{g}(p) = \ln \frac{p}{1-p} = \eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n \quad (1)$$

- The distribution around the mean is taken to be binomial.

Reviewing Logistic Regression

Logistic regression, too, is a kind of generalized linear model.

- The linear predictor:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$$

- The link function g is the logit transform:

$$\mathrm{E}(\mathrm{Y}) = p = \mathrm{g}^{-1}(\eta) \Leftrightarrow$$
$$\mathrm{g}(p) = \ln \frac{p}{1-p} = \eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n \quad (1)$$

- The distribution around the mean is taken to be binomial.

Reviewing Logistic Regression

Logistic regression, too, is a kind of generalized linear model.

- The linear predictor:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$$

- The link function g is the logit transform:

$$\begin{aligned} \mathrm{E}(\mathrm{Y}) = p = \mathrm{g}^{-1}(\eta) \Leftrightarrow \\ \mathrm{g}(p) = \ln \frac{p}{1-p} = \eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n \end{aligned} \quad (1)$$

- The distribution around the mean is taken to be binomial.

Reviewing GLM

- Poisson regression
- Beta-binomial model (for low count data, for example)
- Ordered and unordered multinomial regression.
- ...

The Linear Model

- Let's start with the Linear Model (linear regression, multiple linear regression)

A simple example

- You are studying word RTs in a lexical-decision task

  tpozt *Word or non-word?*
  house *Word or non-word?*

A simple example

- You are studying word RTs in a lexical-decision task

  `tpozt` *Word or non-word?*

  `house` *Word or non-word?*

A simple example

- You are studying word RTs in a lexical-decision task

`tpozt`	*Word or non-word?*
`house`	*Word or non-word?*

Data: Lexical decision RTs

- Data set based on Baayen et al. (2006; available through `languageR` library in the free statistics program R)

Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Journal of Memory and Language 55 (2006) 290–313
Journal of Memory and Language
www.elsevier.com/locate/jml

Morphological influences on the recognition of monosyllabic monomorphemic words

R.H. Baayen [a,*], L.B. Feldman [b], R. Schreuder [c]

[a] Radboud University Nijmegen and Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands
[b] State University of New York at Albany, Department of Psychology, SS112 Albany, NY 12222, USA
[c] Radboud University Nijmegen, P.O. Box 310, 6500 AH Nijmegen, The Netherlands

Received 15 July 2005; revision received 28 March 2006

Data: Lexical decision RTs

- Lexical Decisions from 79 concrete nouns each seen by 21 subjects (1,659 observation).
- **Outcome:** log lexical decision latency `RT`
- **Inputs:**
  - factor (e.g. `NativeLanguage`: *English* or *Other*)
  - continuous predictors (e.g. `Frequency`).

```
> library(languageR)
> head(lexdec[,c(1,2,5,10,11)])
  Subject       RT NativeLanguage Frequency FamilySize
1      A1 6.340359        English  4.859812  1.3862944
2      A1 6.308098        English  4.605170  1.0986123
3      A1 6.349139        English  4.997212  0.6931472
4      A1 6.186209        English  4.727388  0.0000000
5      A1 6.025866        English  7.667626  3.1354942
6      A1 6.180017        English  4.060443  0.6931472
```

A simple example

- A simple model: assume that `Frequency` has a *linear* effect on average (log-transformed) RT, and trial-level noise is *normally distributed*
- If x_i is `Frequency`, our simple model is

$$\boxed{RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}}$$

- We need to draw inferences about α, β, and σ
- e.g., "Does `Frequency` affects RT?" $\rightarrow$ is β reliably non-zero?

A simple example

- A simple model: assume that `Frequency` has a *linear* effect on average (log-transformed) RT, and trial-level noise is *normally distributed*
- If x_i is `Frequency`, our simple model is

$$\boxed{RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}}$$

- We need to draw inferences about α, β, and σ
- e.g., "Does `Frequency` affects RT?"$\longrightarrow$ is β reliably non-zero?

A simple example

- A simple model: assume that `Frequency` has a *linear* effect on average (log-transformed) RT, and trial-level noise is *normally distributed*
- If x_i is `Frequency`, our simple model is

$$\boxed{RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}}$$

- We need to draw inferences about α, β, and σ
- e.g., "Does `Frequency` affects RT?"$\rightarrow$ is β reliably non-zero?

A simple example

- A simple model: assume that `Frequency` has a *linear* effect on average (log-transformed) RT, and trial-level noise is *normally distributed*
- If x_i is `Frequency`, our simple model is

$$\boxed{RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}}$$

- We need to draw inferences about α, β, and σ
- e.g., "Does `Frequency` affects RT?"$\rightarrow$ is β reliably non-zero?

Reviewing GLMs: A simple example

$$RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}$$

- Here's a translation of our simple model into R:

```
> glm(RT ~ 1 + Frequency, data=lexdec,
+ family="gaussian")
[...]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5887   0.022296 295.515   <2e-16 ***
Frequency    -0.0428   0.004533  -9.459   <2e-16 ***
> sqrt(summary(l)[["dispersion"]])
[1] 0.2353127
```

$\hat{\alpha}$ $\hat{\beta}$ $\hat{\sigma}$

Reviewing GLMs: A simple example

$$RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}$$

- Here's a translation of our simple model into R:

```
> glm(RT ~ 1 + Frequency, data=lexdec,
+ family="gaussian")
[...]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5887   0.022296 295.515   <2e-16 ***
Frequency    -0.0428   0.004533  -9.459   <2e-16 ***
> sqrt(summary(l)[["dispersion"]])
[1] 0.2353127
```

$\widehat{\alpha}$ $\widehat{\beta}$ $\widehat{\sigma}$

Reviewing GLMs: A simple example

$$RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}$$

- Here's a translation of our simple model into R:

```
> glm(RT ~ 1 + Frequency, data=lexdec,
+ family="gaussian")
[...]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5887   0.022296 295.515   <2e-16 ***
Frequency    -0.0428   0.004533  -9.459   <2e-16 ***
> sqrt(summary(l)[["dispersion"]])
[1] 0.2353127
```

$\hat{\alpha}$ $\hat{\beta}$ $\hat{\sigma}$

Reviewing GLMs: A simple example

$$RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}$$

- Here's a translation of our simple model into R:

```
> glm(RT ~ 1 + Frequency, data=lexdec,
+ family="gaussian")
[...]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5887   0.022296 295.515   <2e-16 ***
Frequency    -0.0428   0.004533  -9.459   <2e-16 ***
> sqrt(summary(l)[["dispersion"]])
[1] 0.2353127
```

$\widehat{\alpha}$ $\widehat{\beta}$ $\widehat{\sigma}$

Linear Model with just an intercept

- The intercept is a predictor in the model (usually one we don't care about).

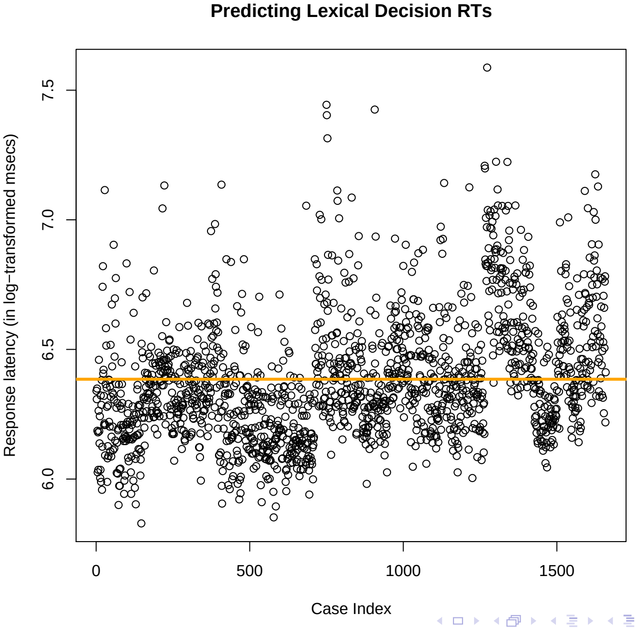
$\longrightarrow$ A significant intercept indicates that it is different from zero.

```
> l.lexdec0 = lm(RT ~ 1, data=lexdec)
> summary(l.lexdec0)
[...]
Residuals:
     Min        1Q    Median        3Q       Max
-0.55614 -0.17048 -0.03945  0.11695  1.20222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.385090   0.005929    1077   <2e-16 ***
[...]
```

NB: Here, intercept encodes overall mean.

Visualization of Intercept Model

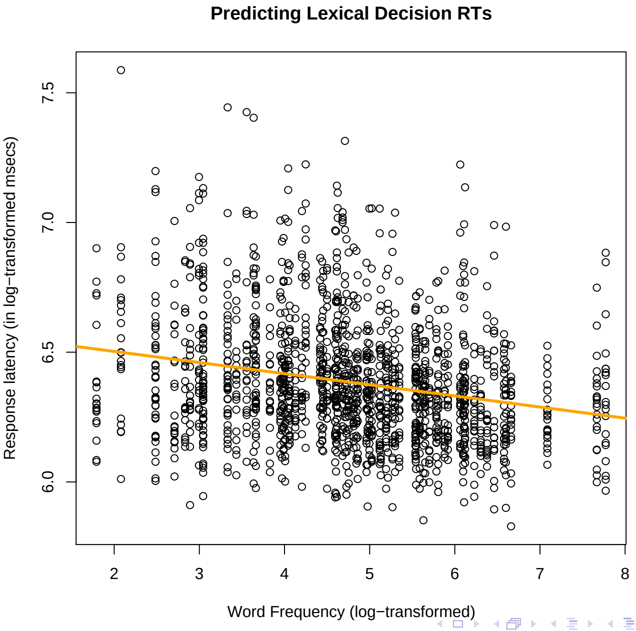


Linear Model with one predictor

```
> l.lexdec1 = lm(RT ~ 1 + Frequency, data=lexdec)
```

- Classic geometrical interpretation: Finding slope for the predictor that minimized the squared error.
  - **NB:** Never forget the directionality in this statement (the error in predicting the outcome is minimized, not the distance from the line).
  - **NB:** Maximum likelihood (ML) fitting is the more general approach as it extends to other types of Generalized Linear Models. ML is identical to least-squared error for Gaussian errors.

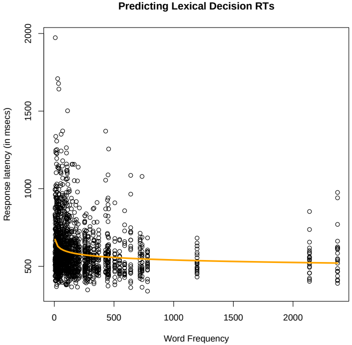
Frequency effect on RT



Linearity Assumption

NB: Like AN(C)OVA, the linear model assumes that the outcome is linear *in the coefficients* (**linearity assumption**).

- This does not mean that the outcome and the **input variable** have to be linearly related (cf. previous page).
- To illustrate this, consider that we can back-transform the log-transformed `Frequency` (→ **transformations** may be necessary).



Adding further predictors

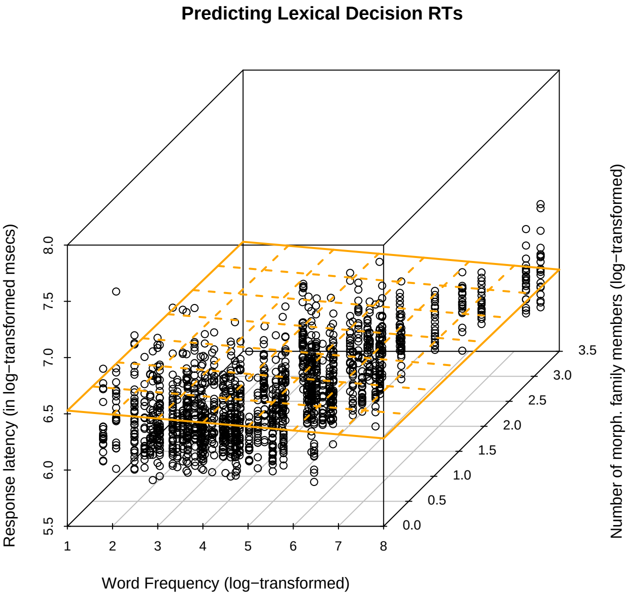
- `FamilySize` is the number of words in the morphological family of the target word.
- For now, we are assuming two independent effects.

```
> l.lexdec1 = lm(RT ~ 1 + Frequency + FamilySize,
+ data=lexdec)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.563853   0.026826 244.685  < 2e-16 ***
Frequency   -0.035310   0.006407  -5.511 4.13e-08 ***
FamilySize  -0.015655   0.009380  -1.669   0.0953 .
```

Question

- On the previous slide, is the interpretation of the output clear?
- What is the interpretation of the intercept?
- How much faster is the most frequent word expected to be read compared to the least frequent word?

Frequency and Morph. Family Size



Continuous and categorical predictors

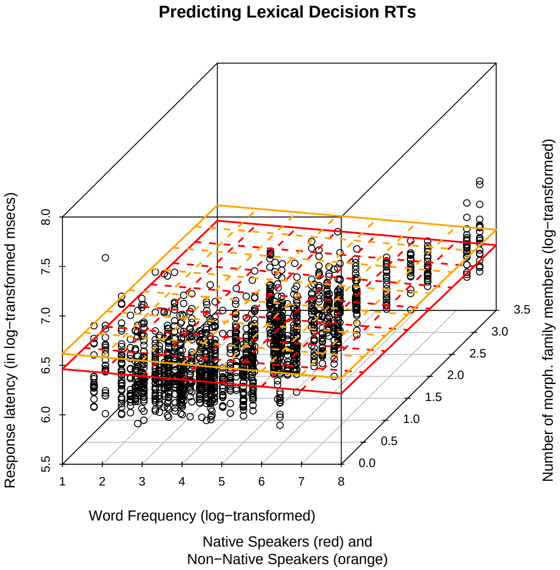
```
> l.lexdec1 = lm(RT ~ 1 + Frequency + FamilySize +
+ NativeLanguage, data=lexdec)
```

- Recall that we're describing the output as a linear combination of the predictors.

$\longrightarrow$ Categorical predictors need to be coded numerically.

  - The default is dummy/treatment coding for regression (cf. **sum/contrast coding** for ANOVA).

Adding Native Language



Question

- Remember that a Generalized Linear Model predicts the mean of the outcome as a linear combination.
- In the previous figure, what does 'mean' mean here?

Interactions

- Interactions are products of predictors.
- Significant interactions tell us that the slope of a predictor differs for different values of the other predictor.

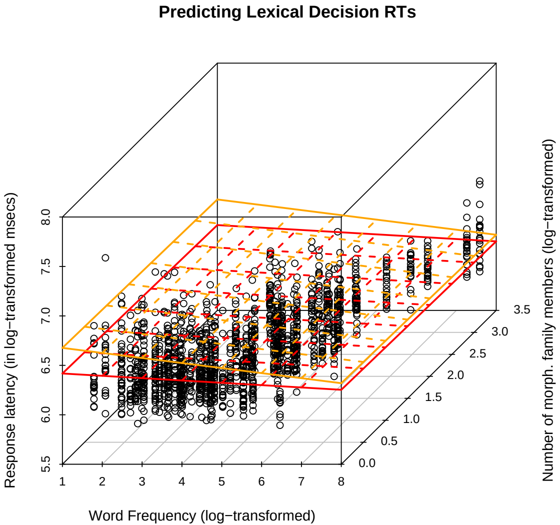
```
> l.lexdec1 = lm(RT ~ 1 + Frequency + FamilySize +
+ NativeLanguage + Frequency:NativeLanguage,
+ data=lexdec)
Residuals:
     Min       1Q   Median       3Q      Max
-0.66925 -0.14917 -0.02800  0.11626  1.06790

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.441135   0.031140 206.847  < 2e-16
Frequency              -0.023536   0.007079  -3.325 0.000905
FamilySize             -0.015655   0.008839  -1.771 0.076726
NativeLanguageOther     0.286343   0.042432   6.748 2.06e-11
Frequency:NatLangOther -0.027472   0.008626  -3.185 0.001475
```

Question

- On the previous slide, how should we interpret the interaction?

Interaction: Frequency & Native Language



Interaction with Native Speakers (red) and
Non−Native Speakers (orange)

Linear Model vs. ANOVA

- Shared with ANOVA:
  - Linearity assumption (though many types of non-linearity can be investigated)
  - Assumption of normality, but part of a more general framework that extends to other distribution in a conceptually straightforward way.
  - Assumption of independence

  NB: ANOVA is linear model with categorical predictors.
- Differences:
  - Generalized Linear Model
  - Consistent and transparent way of treating continuous and categorical predictors.
  - Regression encourages a priori explicit coding of hypothesis $\longrightarrow$ reduction of post-hoc tests $\longrightarrow$ decrease of family-wise error rate.

Hypothesis testing in psycholinguistic research

- Typically, we make predictions not just about the existence, but also the *direction* of effects.
- Sometimes, we're also interested in effect *shapes* (non-linearities, etc.)
- Unlike in ANOVA, regression analyses reliably test hypotheses about **effect direction**, **effect shape**, and **effect size** without requiring post-hoc analyses if (a) *the predictors in the model are coded appropriately* (cf. M. Gillespie's tutorial later today) and (b) *the model can be trusted* (cf. tomorrow).

Determining the parameters

- How do we choose parameters (model coefficients) β_i and σ?
- We find the *best* ones.
- There are two major approaches (deeply related, yet different) in widespread use:
  - The principle of maximum likelihood: pick parameter values that maximize the probability of your data Y

    choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y|\{\beta_i\}, \sigma)$ *as large as possible*

  - Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data

Determining the parameters

- How do we choose parameters (model coefficients) β_i and σ?
- We find the *best* ones.
- There are two major approaches (deeply related, yet different) in widespread use:
  - The principle of maximum likelihood: pick parameter values that maximize the probability of your data Y

    choose $\{\beta_i\}$ and σ that make the likelihood $P(Y|\{\beta_i\}, \sigma)$ as large as possible

  - Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data

Determining the parameters

- How do we choose parameters (model coefficients) β_i and σ?
- We find the *best* ones.
- There are two major approaches (deeply related, yet different) in widespread use:
  - The principle of maximum likelihood: pick parameter values that maximize the probability of your data Y

    choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y|\{\beta_i\}, \sigma)$ *as large as possible*

  - Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data

Determining the parameters

- How do we choose parameters (model coefficients) β_i and σ?
- We find the *best* ones.
- There are two major approaches (deeply related, yet different) in widespread use:
  - The principle of maximum likelihood: pick parameter values that maximize the probability of your data Y

    choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y|\{\beta_i\}, \sigma)$ *as large as possible*

  - Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data

$$P(\{\beta_i\}, \sigma|Y) = \frac{P(Y|\{\beta_i\}, \sigma) \overbrace{P(\{\beta_i\}, \sigma)}^{\text{Prior}}}{P(Y)}$$

Determining the parameters

- How do we choose parameters (model coefficients) β_i and σ?
- We find the *best* ones.
- There are two major approaches (deeply related, yet different) in widespread use:
  - The principle of maximum likelihood: pick parameter values that maximize the probability of your data Y

    choose $\{\beta_i\}$ and σ that make the likelihood $P(Y|\{\beta_i\},\sigma)$ as large as possible

  - Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data

$$P(\{\beta_i\},\sigma|Y) = \frac{\overbrace{P(Y|\{\beta_i\},\sigma)}^{\text{Likelihood}}\overbrace{P(\{\beta_i\},\sigma)}^{\text{Prior}}}{P(Y)}$$

Penalization, Regularization, etc.

- Modern moderns are often fit using maximization of likelihood combined with some sort of **penalization**, a term that 'punished' high model complexity (high values of the coefficients).
- cf. Baayen, Davidson, and Bates (2008) for a nice description.

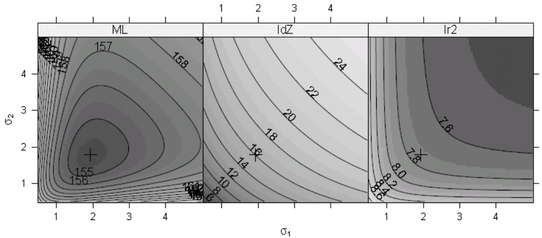


Figure 2. Contours of the profiled deviance as a function of the relative standard deviations of the item random effects and the subject random effects. The leftmost panel shows the deviance, the function that is minimized at the maximum likelihood estimates, the middle panel shows the component of the deviance that measures model complexity and the rightmost panel shows the component of the deviance that measures fidelity of the fitted values to the observed data.

Generalized Linear Mixed Models

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

$\longrightarrow$ Violations of the assumption of independence!

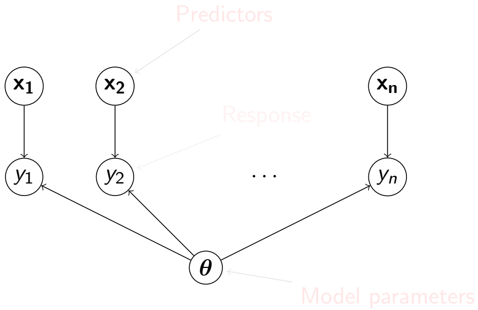
NB: There may even be more clustered (repeated) properties and clusters may be nested (e.g. subjects ϵ dialects ϵ languages).

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

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

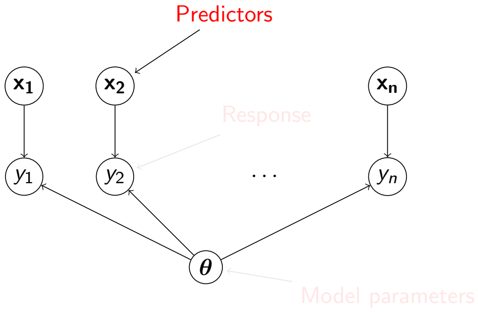
Recall: Generalized Linear Models

The picture:



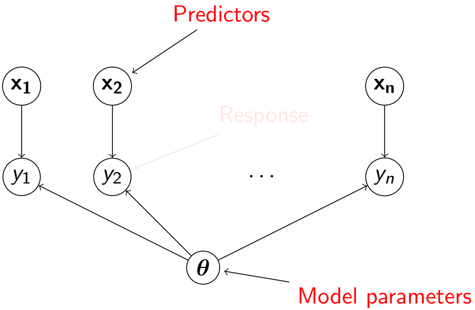
Recall: Generalized Linear Models

The picture:



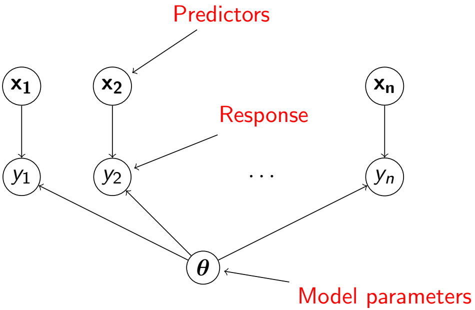
Recall: Generalized Linear Models

The picture:

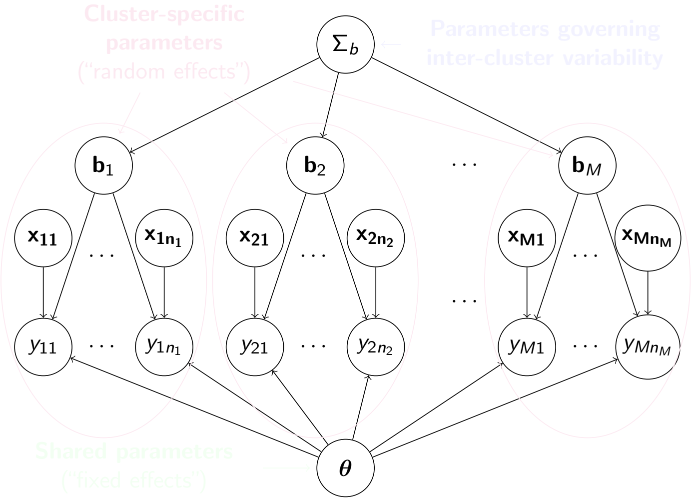


Recall: Generalized Linear Models

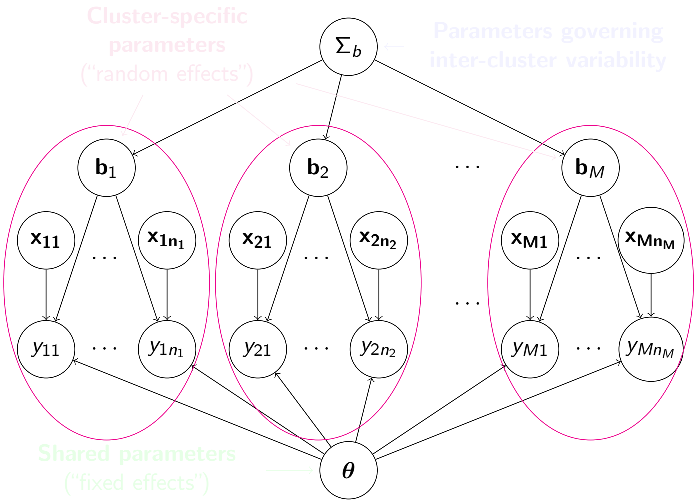
The picture:



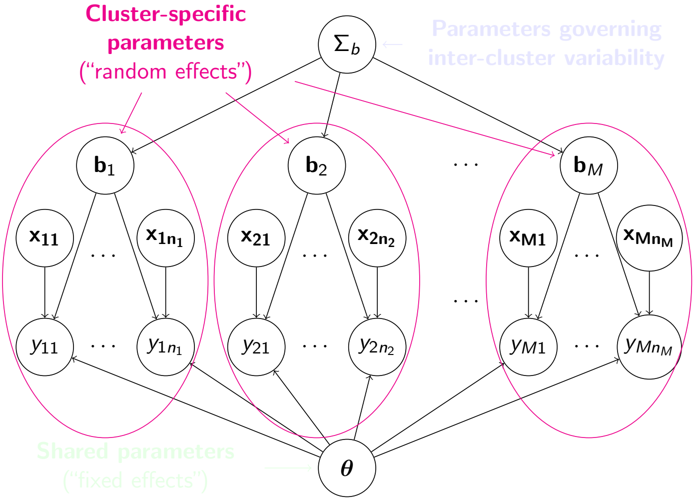
Generalized Linear Mixed Models



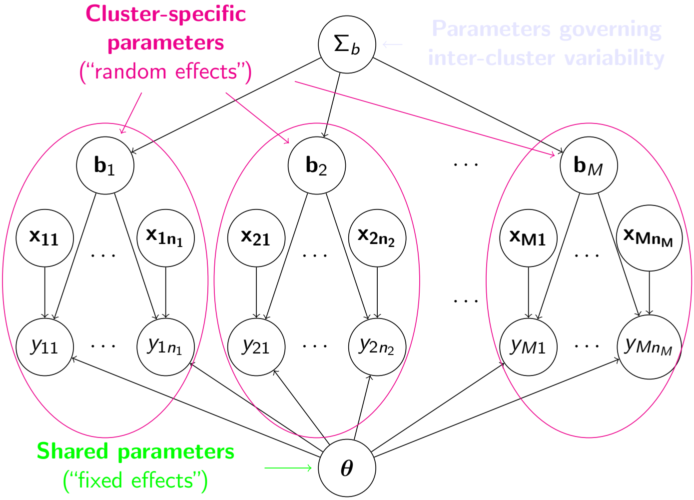
Generalized Linear Mixed Models



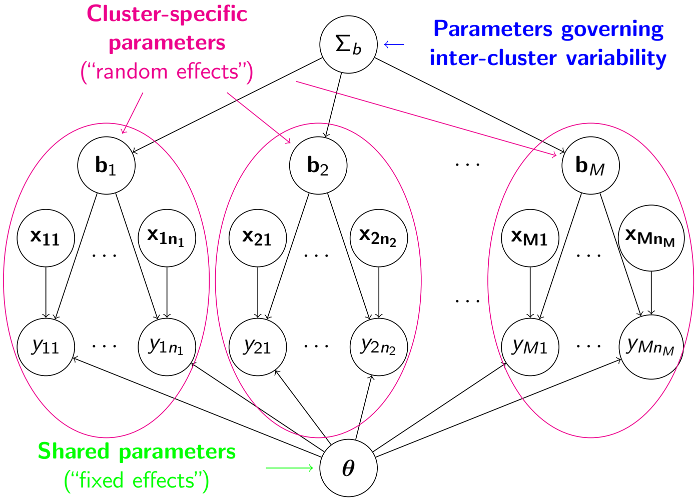
Generalized Linear Mixed Models



Generalized Linear Mixed Models



Generalized Linear Mixed Models



Mixed Linear Model

- Back to our lexical-decision experiment:
- A variety of predictors seem to affect RTs, e.g.:
  - `Frequency`
  - `FamilySize`
  - `NativeLanguage`
  - Interactions
- Additionally, different participants in your study may also have:
  - different overall decision speeds
  - differing sensitivity to e.g. Frequency.
- You want to draw inferences about all these things at the same time

Mixed Linear Model

- Back to our lexical-decision experiment:
- A variety of predictors seem to affect RTs, e.g.:
  - `Frequency`
  - `FamilySize`
  - `NativeLanguage`
  - Interactions
- Additionally, different participants in your study may also have:
  - different overall decision speeds
  - differing sensitivity to e.g. Frequency.
- You want to draw inferences about all these things at the same time

Mixed Linear Model

- Back to our lexical-decision experiment:
- A variety of predictors seem to affect RTs, e.g.:
  - `Frequency`
  - `FamilySize`
  - `NativeLanguage`
  - Interactions
- Additionally, different participants in your study may also have:
  - different overall decision speeds
  - differing sensitivity to e.g. Frequency.
- You want to draw inferences about all these things at the same time

Mixed Linear Model

- Random effects, starting simple: let each participant i have idiosyncratic differences in reaction times (RTs)

$$\boxed{RT_{ij} = \alpha + \beta x_{ij} + \overbrace{b_i}^{\sim N(0,\sigma_b)} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}}$$

Mixed linear model with one random intercept

- Idea: Model distribution of subject differences as deviation from grand mean.
- Mixed models approximate deviation by fitting a normal distribution.
- Grand mean reflected in ordinary intercept
  - $\longrightarrow$ By-subject mean can be set to 0
  - $\longrightarrow$ Only additional parameter fit from data is variance.

```
> lmer.lexdec0 = lmer(RT ~ 1 + Frequency +
+ (1 | Subject), data=lexdec)
```

Interpretation of the output

$$RT_{ij} = \alpha + \beta x_{ij} + \overbrace{b_i}^{\sim N(0,\sigma_b)} + \overbrace{\epsilon_{ij}}^{\text{Noise}\sim N(0,\sigma_\epsilon)}$$

- Interpretation parallel to ordinary regression models:

```
Formula: RT ~ 1 + Frequency + (1 | Subject)
   Data: lexdec
    AIC   BIC logLik deviance REMLdev
 -844.6 -823  426.3     -868  -852.6
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 0.024693 0.15714
 Residual             0.034068 0.18457
Number of obs: 1659, groups: Subject, 21

Fixed effects:
             Estimate Std. Error t value
(Intercept)  6.588778   0.026981  244.20
Frequency   -0.042872   0.003555  -12.06
```

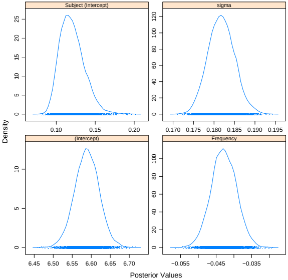
MCMC-sampling

- *t*-value anti-conservative

$\longrightarrow$ MCMC-sampling of coefficients to obtain non anti-conservative estimates

```
> pvals.fnc(lmer.lexdec0, nsim = 10000)
$fixed
            Estimate MCMCmean HPD95lower HPD95upper  pMCMC Pr(>|t|)
(Intercept)   6.5888   6.5886     6.5255     6.6516 0.0001        0
Frequency    -0.0429  -0.0428    -0.0498    -0.0359 0.0001        0

$random
    Groups        Name Std.Dev. MCMCmedian MCMCmean HPD95lower HPD95upper
1  Subject (Intercept)   0.1541     0.1188   0.1205     0.0927     0.1516
2 Residual               0.1809     0.1817   0.1818     0.1753     0.1879
```



Interpretation of the output

- So many new things! What is the output of the linear mixed model?
- **estimates of coefficients** for fixed and random predictors.
- **predictions = fitted values**, just as for ordinary regression model.

```
> cor(fitted(lmer.lexdec0), lexdec$RT)^2
[1] 0.4357668
```

Interpretation of the output

- So many new things! What is the output of the linear mixed model?
- **estimates of coefficients** for fixed and random predictors.
- **predictions = fitted values**, just as for ordinary regression model.

```
> cor(fitted(lmer.lexdec0), lexdec$RT)^2
[1] 0.4357668
```

Mixed models vs. ANOVA

- Mixed models **inherit all advantages from Generalized Linear Models**.
- Unlike the ordinary linear model, the linear mixed model now acknowledges that there are slower and faster subjects.
- This is done without wasting $k - 1$ degrees of freedom on k subjects. We only need one parameter!
- Unlike with ANOVA, we can actually look at the random differences ($\rightarrow$ individual differences).

Mixed models with one random intercept

- 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 b_is—the choices that maximize the above probability

Mixed models with one random intercept

- 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 b_is—the choices that maximize the above probability

Mixed models with one random intercept

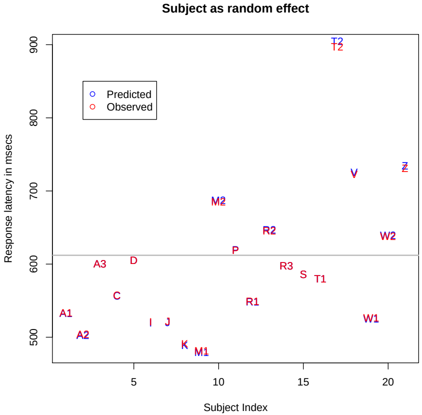
NB: By-subjects adjustments are assumed to be centered around zero, but they don't necessarily do so (here: $-2.3\text{E-}12$).

```
head(ranef(lexdec.lmer0))
$Subject
    (Intercept)
A1 -0.082668694
A2 -0.137236138
A3  0.009609997
C  -0.064365560
D   0.022963863
...
```

Mixed models with one random intercept

- Observed and fitted values of by-subject means.

```
> p = exp(as.vector(unlist(coef(lmer.lexdec0)$Subject)))
> text(p, as.character(unique(lexdec$Subject)), col = "red")
> legend(x=2, y=850, legend=c("Predicted", "Observed"),
+ col=c("blue", "red"), pch=1)
```



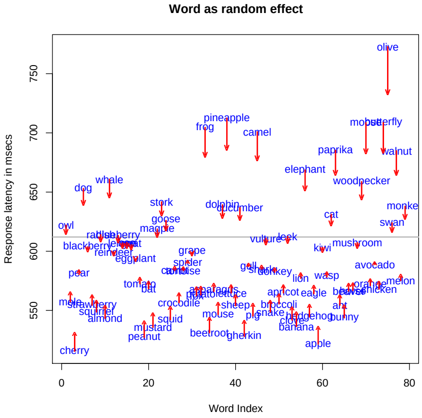
Mixed models with more random intercepts

- Unlike with ANOVA, the linear mixed model can accommodate more than one random intercept, if we think this is necessary/adequate.
- These are *crossed* random effects.

```
> lexdec.lmer1 = lmer(RT ~ 1 + (1 | Subject) + (1 | Word),
+ data = lexdec)
> ranef(lmer.lexdec1)
$Word
              (Intercept)
almond         0.0164795993
ant           -0.0245297186
apple         -0.0494242968
apricot       -0.0410707531
...
$Subject
    (Intercept)
A1 -0.082668694
A2 -0.137236138
A3  0.009609997
```

Mixed models with more random intercepts

- Shrinkage becomes even more visible for fitted by-word means



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`:

```
> lmer.lexdec2 = lmer(RT ~ 1 + Frequency +
+ (1 | Subject) + (0 + Frequency | Subject) +
+ (1 | Word),
+ data=lexdec)
Random effects:
 Groups   Name        Variance   Std.Dev.
 Word     (Intercept) 0.00295937 0.054400
 Subject  Frequency   0.00018681 0.013668
 Subject  (Intercept) 0.03489572 0.186804
 Residual             0.02937016 0.171377
Number of obs: 1659, groups: Word, 79; Subject, 21

Fixed effects:
             Estimate Std. Error t value
(Intercept)  6.588778   0.049830  132.22
Frequency   -0.042872   0.006546   -6.55
```

Mixed models with random slopes

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

```
> ranef(lmer.lexdec2)
$Word
             (Intercept)
almond      0.0164795993
ant        -0.0245297186
...
$Subject
     (Intercept)      Frequency
A1 -0.1130825633  0.0020016500
A2 -0.2375062644  0.0158978707
A3 -0.0052393295  0.0034830009
C  -0.1320599587  0.0143830430
D   0.0011335764  0.0038101993
I  -0.1416446479  0.0029889156
...
```

Mixed model vs. ANOVA

- A mixed model with random slopes for all its predictors (incl. random intercept) is comparable in structure to an ANOVA
- Unlike ANOVA, random effects can be fit for several grouping variables in one single model.
  - $\longrightarrow$ More power (e.g. Baayen 2004; Dixon, 2008).
- No nesting assumptions *need* to be made (for examples of nesting in mixed models, see Barr, 2008 and his blog). As in the examples, so far, random effects can be crossed.
- Assumptions about variance-covariance matrix can be tested
  - No need to rely on assumptions (e.g. sphericity).
  - Can test whether specific random effect is needed (**model comparison**).

Random Intercept, Slope, and Covariance

- Random effects (e.g. intercepts and slopes) may be correlated.
  - By default, R fits these covariances, introducing additional degrees of freedom (parameters).
  - Note the simpler syntax.

```
> lmer.lexdec2 = lmer(RT ~ 1 + Frequency +
+ (1 + Frequency | Subject) +
+ (1 | Word),
+ data=lexdec)
```

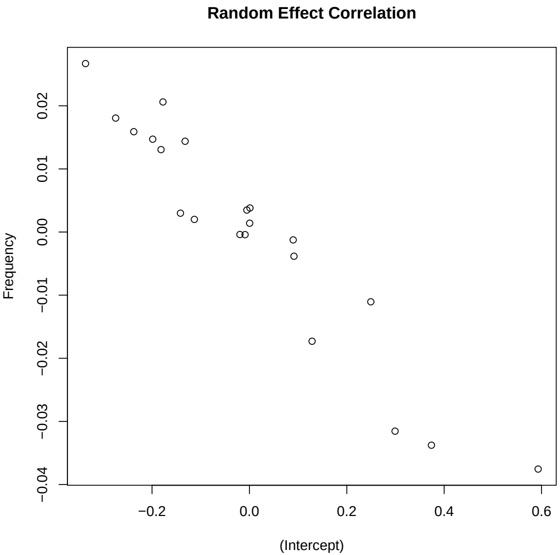
Random Intercept, Slope, and Covariance

```
Random effects:
 Groups   Name        Variance   Std.Dev. Corr
 Word     (Intercept) 0.00296905 0.054489
 Subject  (Intercept) 0.05647247 0.237639
          Frequency   0.00040981 0.020244 -0.918
 Residual             0.02916697 0.170783
Number of obs: 1659, groups: Word, 79; Subject, 21

Fixed effects:
             Estimate Std. Error t value
(Intercept)  6.588778   0.059252  111.20
Frequency   -0.042872   0.007312   -5.86
```

- What do such covariance parameters mean?

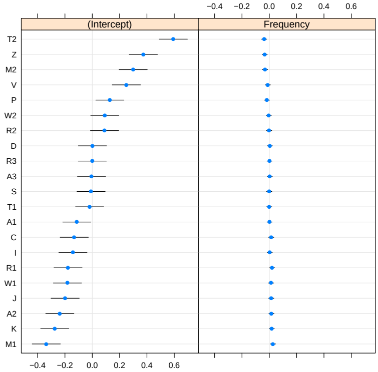
Covariance of random effects: An example



Plotting Random Effects: Example

- Plotting random effects sorted by magnitude of first BLUP (here: intercept) and with posterior variance-covariance of random effects conditional on the estimates of the model parameters and on the data.

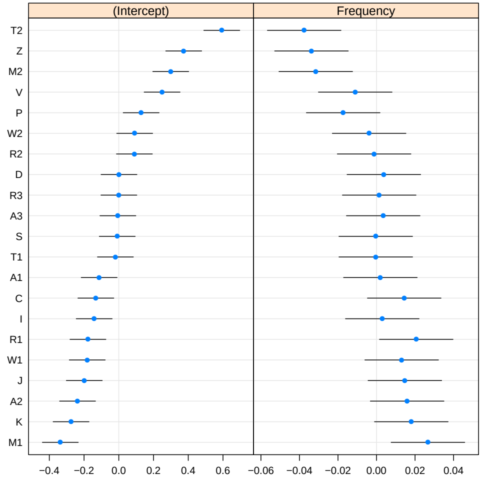
```
> dotplot(ranef(lmer.lexdec3, postVar=TRUE))
```



Plotting Random Effects: Example

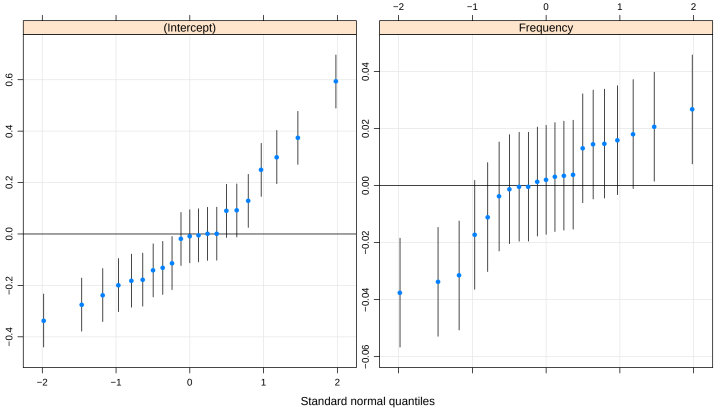
- Plotted without forcing scales to be identical:

```
> dotplot(ranef(lmer.lexdec3, postVar=TRUE),
+ scales = list(x =
+ list(relation = 'free')))[["Subject"]]
```



Plotting Random Effects: Example

- Plotting observed against theoretical quantiles:



Is the Random Slope Justified?

- 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 `lmer.lexdec3` significantly improves the model compared to `lmer.lexdec2` 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 ($\chi^2(2) = 24.1$, $p < 0.0001$).

$\longrightarrow$ Despite the strong correlation, the two random effects for subjects are needed (given the fixed effect predictors in the model).

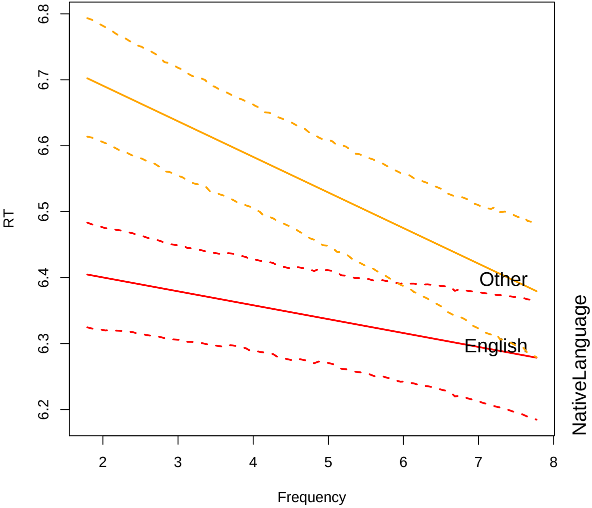
Interactions

```
> lmer.lexdec4b = lmer(RT ~ 1 + NativeLanguage * (
+ Frequency + FamilySize + SynsetCount +
+ Class) +
+ (1 + Frequency | Subject) + (1 | Word),
+ data=lexdec)
[...]
Fixed effects:
                             Estimate Std. Error t value
(Intercept)                  6.385090   0.030425  209.86
cNativeEnglish              -0.155821   0.060533   -2.57
cFrequency                  -0.035180   0.008388   -4.19
cFamilySize                 -0.019757   0.012401   -1.59
cSynsetCount                -0.030484   0.021046   -1.45
cPlant                      -0.050907   0.015609   -3.26
cNativeEnglish:cFrequency    0.032893   0.011764    2.80
cNativeEnglish:cFamilySize   0.018424   0.015459    1.19
cNativeEnglish:cSynsetCount -0.022869   0.026235   -0.87
cNativeEnglish:cPlant        0.082219   0.019457    4.23
```

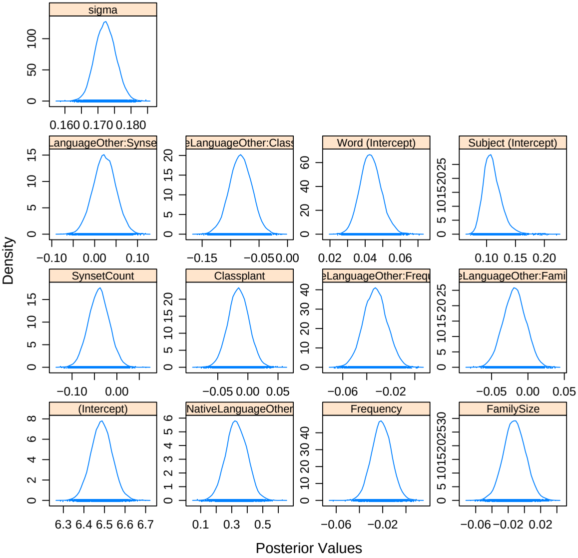
Interactions

```
> p.lmer.lexdec4b = pvals.fnc(lmer.lexdec4b,
nsim=10000, withMCMC=T)
> p.lmer.lexdec$fixed
                       Estimate MCMCmean HPD95lower HPD95upper  pMCMC Pr(>|t|)
(Intercept)              6.4867   6.4860     6.3839     6.5848 0.0001   0.0000
NativeLanguageOther      0.3314   0.3312     0.1990     0.4615 0.0001   0.0000
Frequency               -0.0211  -0.0210    -0.0377    -0.0048 0.0142   0.0156
FamilySize              -0.0119  -0.0120    -0.0386     0.0143 0.3708   0.3997
SynsetCount             -0.0403  -0.0401    -0.0852     0.0050 0.0882   0.0920
Classplant              -0.0157  -0.0155    -0.0484     0.0181 0.3624   0.3767
NatLang:Frequency       -0.0329  -0.0329    -0.0515    -0.0136 0.0010   0.0006
NatLang:FamilySize      -0.0184  -0.0184    -0.0496     0.0109 0.2416   0.2366
NatLang:SynsetCount      0.0229   0.0230    -0.0297     0.0734 0.3810   0.3866
NatLang:Classplant      -0.0822  -0.0825    -0.1232    -0.0453 0.0001   0.0000
```

Visualizing an Interactions



MCMC


sigma
LanguageOther:Synse
eLanguageOther:Class
Word (Intercept)
Subject (Intercept)
SynsetCount
Classplant
eLanguageOther:Frequ
eLanguageOther:Fami
(Intercept)
NativeLanguageOther
Frequency
FamilySize
Density
Posterior Values

Mixed Logit Model

- So, what do we need to change if we want to investigate, e.g. a binary (categorical) outcome?

Recall that ...

logistic regression is a kind of generalized linear model.

- The linear predictor:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$$

- The link function g is the logit transform:

$$\mathrm{E}(\mathrm{Y}) = p = \mathrm{g}^{-1}(\eta) \Leftrightarrow$$
$$\mathrm{g}(p) = \ln \frac{p}{1-p} = \eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n \quad (2)$$

- The distribution around the mean is taken to be binomial.

Recall that ...

logistic regression is a kind of generalized linear model.

- The linear predictor:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$$

- The link function g is the logit transform:

$$\begin{aligned} \mathrm{E}(\mathrm{Y}) = p = \mathrm{g}^{-1}(\eta) \Leftrightarrow \\ \mathrm{g}(p) = \ln \frac{p}{1-p} = \eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n \end{aligned} \tag{2}$$

- The distribution around the mean is taken to be binomial.

Mixed Logit Models

- Mixed Logit Models are a type of Generalized Linear *Mixed* Model.
- More generally, one advantage of the mixed model approach is its flexibility. Everything we learned about mixed *linear* models extends to other types of distributions within the exponential family (binomial, multinomial, poisson, beta-binomial, ...)

Caveat There are some implementational details (depending on your stats program, too) that may differ.

An example

- The same model as above, but now we predict whether participants' answer to the lexical decision task was correct.
- **Outcome:** Correct vs. incorrect answer (binomial outcome)
- **Predictors:** same as above

```
> lmer.lexdec.answer4 = lmer(Correct == "correct" ~ 1 +
+   NativeLanguage * (
+ Frequency + FamilySize + SynsetCount +
+ Class) +
+ (1 + Frequency | Subject) + (1 | Word),
+ data=lexdec, family="binomial")
```

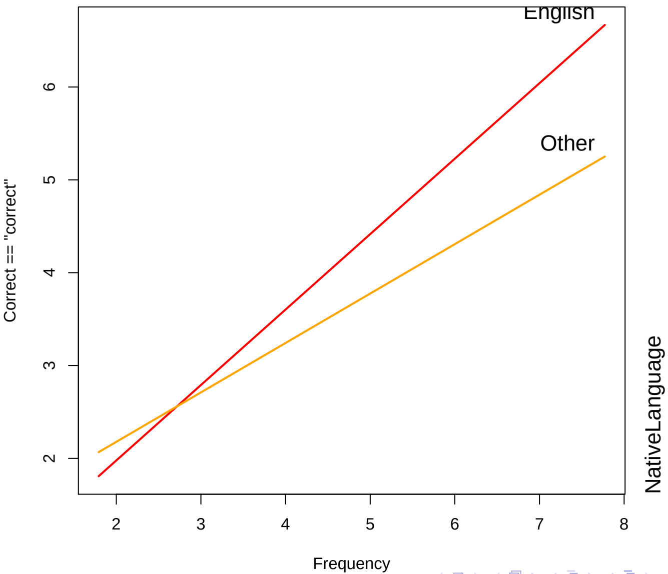
NB: The only difference is the outcome variable and the family (assumed noise distribution) now is binomial (we didn't specify it before because "gaussian" is the default).

Mixed Logit Output

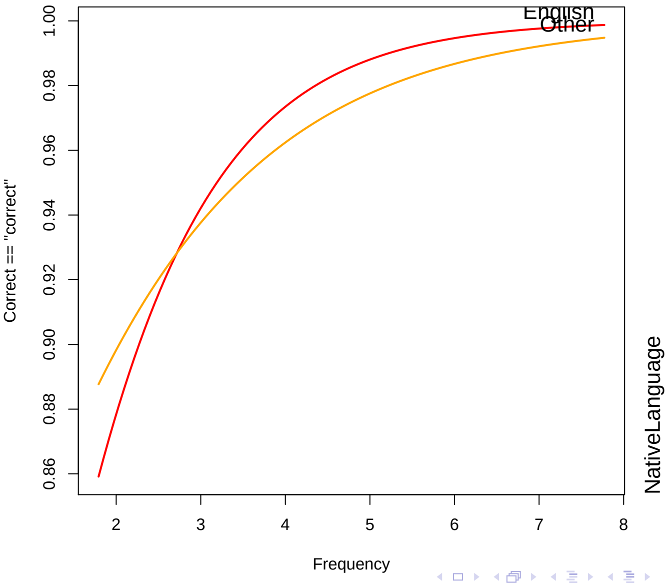
```
[...]
 AIC   BIC logLik deviance
 495 570.8 -233.5      467
Random effects:
 Groups  Name        Variance Std.Dev. Corr
 Word    (Intercept) 0.78368  0.88526
 Subject (Intercept) 2.92886  1.71139
         Frequency   0.11244  0.33532  -0.884
Number of obs: 1659, groups: Word, 79; Subject, 21

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                   4.3612     0.3022  14.433  < 2e-16 ***
cNativeEnglish                0.2828     0.5698   0.496  0.61960
cFrequency                    0.6925     0.2417   2.865  0.00417 **
cFamilySize                  -0.2250     0.3713  -0.606  0.54457
cSynsetCount                  0.8152     0.6598   1.235  0.21665
cPlant                        0.8441     0.4778   1.767  0.07729 .
cNativeEnglish:cFrequency     0.2803     0.3840   0.730  0.46546
cNativeEnglish:cFamilySize   -0.2746     0.5997  -0.458  0.64710
cNativeEnglish:cSynsetCount  -2.6063     1.1772  -2.214  0.02683 *
cNativeEnglish:cPlant         1.0615     0.7561   1.404  0.16035
```

Interaction in logit space



Interaction in probability space



Why not ANOVA?

- ANOVA over proportion has several problems (cf. Jaeger, 2008 for a summary)
  - Hard to interpret output
  - Violated assumption of homogeneity of variances

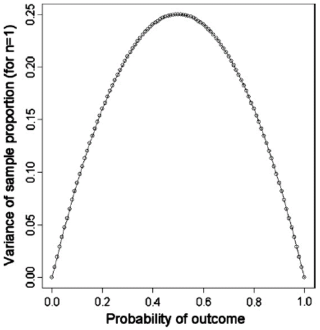
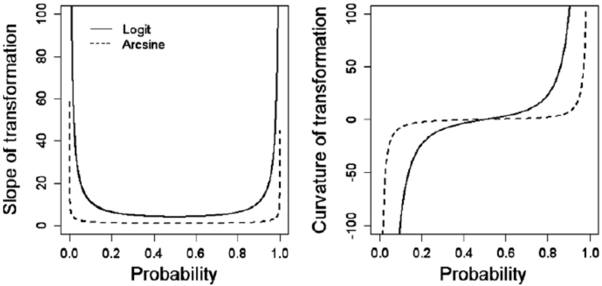


Fig. 1. Variance of sample proportion depending on p (for $n = 1$).

Why not ANOVA?

- These problems can be address via transformations, weighted regression, etc., But why should we do this is if there is an adequate approach that does not need fudging and has more power?



Summary

- There are a lot of issues, we have not covered today (by far most of these are not particular to mixed models, but apply equally to ANOVA).
- The mixed model approach has many advantages:
  - Power (especially on unbalanced data)
  - No assumption of homogeneity of variances
  - Random effect structure can be explored, understood.
  - Extendability to a variety of distributional families
  - Conceptual transparency
  - Effect direction, shape, size can be easily understood and investigated.

  → You end up getting another perspective on your data

Summary

- There are a lot of issues, we have not covered today (by far most of these are not particular to mixed models, but apply equally to ANOVA).
- The mixed model approach has many advantages:
    - Power (especially on unbalanced data)
    - No assumption of homogeneity of variances
    - Random effect structure can be explored, understood.
    - Extendability to a variety of distributional families
    - Conceptual transparency
    - Effect direction, shape, size can be easily understood and investigated.

  → You end up getting another perspective on your data

Summary

- There are a lot of issues, we have not covered today (by far most of these are not particular to mixed models, but apply equally to ANOVA).
- The mixed model approach has many advantages:
  - Power (especially on unbalanced data)
  - No assumption of homogeneity of variances
  - Random effect structure can be explored, understood.
  - Extendability to a variety of distributional families
  - Conceptual transparency
  - Effect direction, shape, size can be easily understood and investigated.

  $\longrightarrow$ You end up getting another perspective on your data

Modeling schema