A Brief and Friendly(?) Introduction to hierarchical (mixed-effects, multi-level) regression

ESSLLI 2012

Cluster-specific parameters
(“random effects”)

Parameters governing inter-cluster variability

Shared parameters
(“fixed effects”)

Roger Levy

UC San Diego
Department of Linguistics

8 Aug 2012
Goals of this talk

- Briefly review generalized linear models and how to use them
- Give a precise description of hierarchical (multi-level, mixed-effects) models
- Show how to draw inferences using a hierarchical model (*fitting* the model)
- Discuss how to interpret model parameter estimates
  - Fixed effects
  - Random effects
- Briefly discuss hierarchical logit models
- Discuss ongoing work on approaching standards for how to use multi-level models
Goal: model the effects of predictors (independent variables) $X$ on a response (dependent variable) $Y$. 
Reviewing generalized linear models I

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

The picture:
Reviewing generalized linear models I

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

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

The picture:
Assumptions of the generalized linear model (GLM):

1. Predictors \( \{X_i\} \) influence \( Y \) through the mediation of a linear predictor \( \eta \);
Assumptions of the generalized linear model (GLM):

1. Predictors \( \{X_i\} \) influence \( Y \) through the mediation of a linear predictor \( \eta \);
2. \( \eta \) is a linear combination of the \( \{X_i\} \):
Assumptions of the generalized linear model (GLM):

1. Predictors $\{X_i\}$ influence $Y$ through the mediation of a linear predictor $\eta$;

2. $\eta$ is a linear combination of the $\{X_i\}$:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N$$  (linear predictor)
Assumptions of the generalized linear model (GLM):

1. Predictors \( \{X_i\} \) influence \( Y \) through the mediation of a linear predictor \( \eta \);
2. \( \eta \) is a linear combination of the \( \{X_i\} \):
   \[
   \eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \quad \text{(linear predictor)}
   \]
3. \( \eta \) determines the predicted mean \( \mu \) of \( Y \)
   \[
   \eta = l(\mu) \quad \text{(link function)}
   \]
Assumptions of the generalized linear model (GLM):

1. Predictors \( \{X_i\} \) influence \( Y \) through the mediation of a linear predictor \( \eta \);

2. \( \eta \) is a linear combination of the \( \{X_i\} \):

\[
\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \quad \text{(linear predictor)}
\]

3. \( \eta \) determines the predicted mean \( \mu \) of \( Y \)

\[
\eta = l(\mu) \quad \text{(link function)}
\]

4. There is some noise distribution of \( Y \) around the predicted mean \( \mu \) of \( Y \):

\[
P(Y = y; \mu)
\]
Linear regression, which underlies ANOVA, is a kind of generalized linear model.
Linear regression, which underlies ANOVA, is a kind of generalized linear model.

- The predicted mean is just the linear predictor:

\[ \eta = l(\mu) = \mu \]
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 \( \sigma \):
  \[ \epsilon \sim N(0, \sigma) \]
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 (i.e., Gaussian) distributed around 0 with standard deviation \( \sigma \):
  \[ \epsilon \sim N(0, \sigma) \]

- This gives us the traditional linear regression equation:
  \[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \]
How do we fit the parameters $\beta_i$ and $\sigma$ (choose *model coefficients*)?

There are two major approaches (deeply related, yet different) in widespread use:
Reviewing GLMs IV

- How do we fit the parameters $\beta_i$ and $\sigma$ (choose model coefficients)?
- 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 $Y$
    
    $$\text{choose } \{\beta_i\} \text{ and } \sigma \text{ that make the likelihood } P(Y|\{\beta_i\}, \sigma) \text{ as large as possible}$$

\[
Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \\
\epsilon \sim N(0, \sigma)
\]
How do we fit the parameters $\beta_i$ and $\sigma$ (choose model coefficients)?

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

- Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data
How do we fit the parameters $\beta_i$ and $\sigma$ (choose model coefficients)?

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 $Y$

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

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

$$P(\{\beta_i\}, \sigma|Y) = \frac{P(Y|\{\beta_i\}, \sigma)P(\{\beta_i\}, \sigma)}{P(Y)}$$
How do we fit the parameters $\beta_i$ and $\sigma$ (choose model coefficients)?

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

- Bayesian inference: put a probability distribution on the model parameters and update it on the basis of what parameters best explain the data
  
  \[
  P(\{\beta_i\}, \sigma|Y) = \frac{\underbrace{P(Y|\{\beta_i\}, \sigma)P(\{\beta_i\}, \sigma)}_{\text{Likelihood}}}{P(Y)} \bigg(\underbrace{\text{Prior}}_{\text{Prior}}\bigg)
  \]
Reviewing GLMs V: a simple example

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

- Non-words with different neighborhood densities should have different average RT

- A simple model: assume that neighborhood density has a linear effect on average RT, and trial-level noise is normally distributed

\[ RT_i = \alpha + \beta x_i + \epsilon_i \sim N(0, \sigma) \]

- We need to draw inferences about \( \alpha \), \( \beta \), and \( \sigma \)

- e.g., "Does neighborhood density affect RT?" \( \rightarrow \) is \( \beta \) reliably non-zero?
You are studying non-word RTs in a lexical-decision task:

Word or non-word?

Non-words with different neighborhood densities should have different average RTs (number of neighbors of edit-distance 1).

A simple model: assume that neighborhood density has a linear effect on average RT, and trial-level noise is normally distributed. (n.b. wrong–RTs are skewed—but not horrible.)

\[ \text{If } x_i \text{ is neighborhood density, our simple model is } \text{RT}_i = \alpha + \beta x_i + \epsilon_i \sim N(0, \sigma) \]

We need to draw inferences about \( \alpha, \beta, \) and \( \sigma \). EXAMPLE: "Does neighborhood density affects RT?" → is \( \beta \) reliably non-zero?
Reviewing GLMs V: a simple example

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

  tpozt  Word or non-word?
  houze  Word or non-word?

- Non-words with different neighborhood densities should have different average RT

- A simple model: assume that neighborhood density has a linear effect on average RT, and trial-level noise is normally distributed

  \[ RT_i = \alpha + \beta x_i + \epsilon_i \]

  \( \epsilon_i \sim N(0, \sigma) \)

- We need to draw inferences about \( \alpha \), \( \beta \), and \( \sigma \)

- e.g., "Does neighborhood density affect RT?" → is \( \beta \) reliably non-zero?
You are studying non-word RTs in a lexical-decision task

tpozt  Word or non-word?
houze  Word or non-word?

Non-words with different *neighborhood densities* should have different average RT *(= number of neighbors of edit-distance 1)*
Reviewing GLMs V: a simple example

▶ You are studying non-word RTs in a lexical-decision task
  
  tpozt  Word or non-word?
  houze  Word or non-word?

▶ Non-words with different *neighborhood densities* should have different average RT *(= number of neighbors of edit-distance 1)*

▶ A simple model: assume that neighborhood density has a linear effect on average RT, and trial-level noise is normally distributed *(n.b. wrong–RTs are skewed—but not horrible.)*
You are studying non-word RTs in a lexical-decision task:

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

Non-words with different *neighborhood densities* should have different average RT *(= number of neighbors of edit-distance 1)*

A simple model: assume that neighborhood density has a *linear* effect on average RT, and trial-level noise is *normally distributed* *(n.b. wrong–RTs are skewed—but not horrible.)*

If $x_i$ is neighborhood density, our simple model is

$$RT_i = \alpha + \beta x_i + \epsilon_i \sim N(0, \sigma)$$

We need to draw inferences about $\alpha$, $\beta$, and $\sigma$.

*For example, does neighborhood density affect RT?* Is $\beta$ reliably non-zero?
You are studying non-word RTs in a lexical-decision task

tpozt  Word or non-word?
houze  Word or non-word?

Non-words with different neighborhood densities* should have different average RT *(= number of neighbors of edit-distance 1)

A simple model: assume that neighborhood density has a linear effect on average RT, and trial-level noise is normally distributed* *(n.b. wrong–RTs are skewed—but not horrible.)

If $x_i$ is neighborhood density, our simple model is

$$RT_i = \alpha + \beta x_i + \epsilon_i \sim N(0,\sigma)$$

We need to draw inferences about $\alpha$, $\beta$, and $\sigma$
You are studying non-word RTs in a lexical-decision task:

tpozt  Word or non-word?
houze  Word or non-word?

Non-words with different *neighborhood densities* should have different average RT *(= number of neighbors of edit-distance 1)*

A simple model: assume that neighborhood density has a *linear* effect on average RT, and trial-level noise is *normally distributed* *(n.b. wrong–RTs are skewed—but not horrible.)*

If $x_i$ is neighborhood density, our simple model is

$$RT_i = \alpha + \beta x_i + \epsilon_i \sim N(0,\sigma)$$

We need to draw inferences about $\alpha$, $\beta$, and $\sigma$

e.g., “Does neighborhood density affects RT?” → is $\beta$ reliably non-zero?
We’ll use length-4 nonword data from (Bicknell et al., 2010) (thanks!), such as:

Few neighbors

- gaty
- peme
- rixy

Many neighbors

- lish
- pait
- yine
We’ll use length-4 nonword data from (Bicknell et al., 2010) (thanks!), such as:

**Few neighbors**
- gaty
- peme
- rixy

**Many neighbors**
- lish
- pait
- yine

There’s a wide range of neighborhood density:
Here’s a translation of our simple model into R:

\[
RT \sim 1 + x
\]
Here’s a translation of our simple model into R:

\[ RT \sim 1 + x \]

The noise is implicit in asking R to fit a *linear* model.
Reviewing GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0,\sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim 1 + x \]
- The noise is implicit in asking R to fit a *linear* model
- (We can omit the 1; R assumes it unless otherwise directed)
Here’s a translation of our simple model into R:

\[ RT \sim x \]

The noise is implicit in asking R to fit a *linear* model.

(We can omit the 1; R assumes it unless otherwise directed.)
Here’s a translation of our simple model into R:

\[ RT \sim x \]

The noise is implicit in asking R to fit a \textit{linear} model

(We can omit the 1; R assumes it unless otherwise directed)

Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```r
> m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)

Gaussian noise, implicit intercept

[...]

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 382.997 | 26.837 | 14.271 | <2e-16 *** |
| neighbors | 4.828 | 6.553 | 0.737 | 0.466 |

> sqrt(summary(m)[["dispersion"]])

[1] 107.2248
Reviewing GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim x \]
- The noise is implicit in asking R to fit a *linear* model
- (We can omit the 1; R assumes it unless otherwise directed)
- Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```R
> m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)

[...]

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 382.997 | 26.837 | 14.271 | <2e-16 *** |
| neighbors | 4.828 | 6.553 | 0.737 | 0.466 |

> sqrt(summary(m)[["dispersion"]])
```

[1] 107.2248
Reviewing GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  \[ RT \sim x \]
  
  - The noise is implicit in asking R to fit a *linear* model
  
  - (We can omit the 1; R assumes it unless otherwise directed)

- Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

  ```r
  > m <- glm(RT ~ neighbors, d, family="gaussian")
  > summary(m)
  
  Call:
  glm(formula = RT ~ neighbors, family = "gaussian", data = d)
  
  Deviance Residuals:
  Min       1Q   Median       3Q      Max
  -11.824  -3.811  -0.119   3.709   16.075
  
  Coefficients:
  Estimate Std. Error t value Pr(>|t|)
  (Intercept) 382.997  26.837  14.271  <2e-16 ***
  neighbors   4.828   6.553   0.737   0.466
  
  (Dispersion parameter for gaussian family taken to be 107.225)
  
  Null deviance: 6327.13  on 499  degrees of freedom
  Residual deviance: 5466.06  on 498  degrees of freedom
  
  AIC: 3702.9
  
  Number of Fisher Scoring iterations: 2
  
  > sqrt(summary(m)[["dispersion"]])
  [1] 107.2248
  ```
Reviewing GLMs VII: maximum-likelihood model fitting

\[ RT_i = \alpha + \beta X_i + \epsilon_i \sim N(0, \sigma) \]

- Here’s a translation of our simple model into R:
  
  \[
  RT \sim x
  \]

- The noise is implicit in asking R to fit a linear model

- (We can omit the 1; R assumes it unless otherwise directed)

- Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```r
> m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)

[...]

  Estimate Std. Error t value Pr(>|t|)
(Intercept) 382.997  26.837  14.271  <2e-16 ***
neighbors  4.828   6.553   0.737   0.466

> sqrt(summary(m)[["dispersion"]])

[1] 107.2248
```

\hat{\alpha}, \hat{\beta}
Here’s a translation of our simple model into R:

\[ RT \sim x \]

- The noise is implicit in asking R to fit a linear model
- (We can omit the 1; R assumes it unless otherwise directed)

Example of fitting via maximum likelihood: one subject from Bicknell et al. (2010)

```R
> m <- glm(RT ~ neighbors, d, family="gaussian")
> summary(m)

Estimate Std. Error t value Pr(>|t|)
(Intercept) 382.997  26.837  14.271  <2e-16 ***
neighbors   4.828   6.553   0.737   0.466
```

\[ \hat{\alpha} \]

\[ \hat{\beta} \]

\[ \hat{\sigma} \]
Estimated coefficients are what underlies "best linear fit" plots.

<table>
<thead>
<tr>
<th>Intercept</th>
<th>383.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>neighbors</td>
<td>4.83</td>
</tr>
<tr>
<td>$\hat{\sigma}$</td>
<td>107.22</td>
</tr>
</tbody>
</table>
Estimated coefficients are what underlies “best linear fit” plots
Estimated coefficients are what underlies “best linear fit” plots.
Reviewing GLMs IX: Bayesian model fitting

\[
P(\{\beta_i\}, \sigma | Y) = \frac{\text{Likelihood}}{P(Y)} \cdot \frac{\text{Prior}}{P(\{\beta_i\}, \sigma)}
\]

- Alternative to maximum-likelihood:
  Bayesian model fitting

\[p_{MCMC} = 1 - \text{largest possible symmetric confidence interval wholly on one side of 0}\]
Reviewing GLMs IX: Bayesian model fitting

\[ P(\{\beta_i\}, \sigma | Y) = \frac{P(Y|\{\beta_i\}, \sigma) P(\{\beta_i\}, \sigma)}{P(Y)} \]

- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable

\[ p_{\text{MCMC}} = 0.46 \]

The HPD confidence region is 1 minus the largest possible symmetric confidence interval wholly on one side of 0.
Reviewing GLMs IX: Bayesian model fitting

\[ P(\{\beta_i\}, \sigma | Y) = \frac{P(Y|\{\beta_i\}, \sigma) P(\{\beta_i\}, \sigma)}{P(Y)} \]

- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable
- Multiply by likelihood \(\rightarrow\) posterior probability distribution over \((\alpha, \beta, \sigma)\)

\[ p_{MCMC}(\text{Baayen et al., 2008}) \text{ is 1 minus the largest possible symmetric confidence interval wholly on one side of 0} \]
Reviewing GLMs IX: Bayesian model fitting

Alternative to maximum-likelihood: Bayesian model fitting

Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable

Multiply by likelihood → posterior probability distribution over \((\alpha, \beta, \sigma)\)
Reviewing GLMs IX: Bayesian model fitting

\[ P(\{\beta_i\}, \sigma | Y) = \frac{\text{Likelihood}}{P(Y)} \cdot \frac{\text{Prior}}{P(\{\beta_i\}, \sigma)} \]

- **Alternative to maximum-likelihood:** Bayesian model fitting
- **Simple** (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable
- Multiply by likelihood \(\rightarrow\) posterior probability distribution over \((\alpha, \beta, \sigma)\)
- Bound the region of highest posterior probability containing 95% of probability density \(\rightarrow\) HPD confidence region
Reviewing GLMs IX: Bayesian model fitting

\[ P(\{\beta_i\}, \sigma|Y) = \frac{P(Y|\{\beta_i\}, \sigma) P(\{\beta_i\}, \sigma)}{P(Y)} \]

- Alternative to maximum-likelihood: Bayesian model fitting
- Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable
- Multiply by likelihood \(\rightarrow\) posterior probability distribution over \((\alpha, \beta, \sigma)\)
- Bound the region of highest posterior probability containing 95% of probability density \(\rightarrow\) HPD confidence region
Reviewing GLMs IX: Bayesian model fitting

Alternative to maximum-likelihood: Bayesian model fitting

Simple (uniform, non-informative) prior: all combinations of \((\alpha, \beta, \sigma)\) equally probable

Multiply by likelihood \(\rightarrow\) posterior probability distribution over \((\alpha, \beta, \sigma)\)

Bound the region of highest posterior probability containing 95% of probability density \(\rightarrow\) HPD confidence region

\[ P(\{\beta_i\}, \sigma | Y) = \frac{P(Y | \{\beta_i\}, \sigma) P(\{\beta_i\}, \sigma)}{P(Y)} \]

\(p_{MCMC}\) (Baayen et al., 2008) is 1 minus the largest possible symmetric confidence interval wholly on one side of 0
But of course experiments don't have just one participant

Different participants may have different idiosyncratic behavior

And items may have idiosyncratic properties too

We’d like to take these into account, and perhaps investigate them directly too.

This is what multi-level (hierarchical, mixed-effects) models are for!
Recap of the graphical picture of a single-level model:
Multi-level Models III: the new graphical picture

Cluster-specific parameters ("random effects")

Shared parameters ("fixed effects")

Parameters governing inter-cluster variability
Multi-level Models III: the new graphical picture

Cluster-specific parameters ("random effects")

Shared parameters ("fixed effects")

Parameters governing inter-cluster variability
Multi-level Models III: the new graphical picture

Cluster-specific parameters ("random effects")

Shared parameters ("fixed effects")

Parameters governing inter-cluster variability
Multi-level Models III: the new graphical picture

Cluster-specific parameters ("random effects")

Shared parameters ("fixed effects")

Cluster-specific parameters ("random effects")

Shared parameters ("fixed effects")
Multi-level Models III: the new graphical picture

Cluster-specific parameters ("random effects")

Shared parameters ("fixed effects")

Parameters governing inter-cluster variability
An example of a multi-level model:

- Back to your lexical-decision experiment
  - tpozt  Word or non-word?
  - houze  Word or non-word?

- Non-words with different *neighborhood densities* should have different average decision time
An example of a multi-level model:

- Back to your lexical-decision experiment
  
  - tpozt  
  - houze  
  - *Word or non-word?*

- Non-words with different *neighborhood densities* should have different average decision time

- Additionally, different participants in your study may also have:
  
  - different overall decision speeds
  - differing sensitivity to neighborhood density
An example of a multi-level model:

- Back to your lexical-decision experiment
  
  tpozt  \textit{Word or non-word?}
  
  houze \textit{Word or non-word?}

- Non-words with different \textit{neighborhood densities} should have different average decision time

- \textbf{Additionally}, different participants in your study may also have:
  
  - different overall decision speeds
  
  - differing sensitivity to neighborhood density

- You want to draw inferences about all these things at the same time
Once again we’ll assume for simplicity that the number of word neighbors $x$ has a linear effect on mean reading time, and that trial-level noise is normally distributed*
Once again we’ll assume for simplicity that the number of word neighbors $x$ has a linear effect on mean reading time, and that trial-level noise is normally distributed.*

Random effects, starting simple: let each participant $i$ have idiosyncratic differences in reading speed

$$RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \sim N(0, \sigma_b) \quad \text{Noise} \sim N(0, \sigma_\epsilon)$$
Once again we’ll assume for simplicity that the number of word neighbors $x$ has a linear effect on mean reading time, and that trial-level noise is normally distributed*

Random effects, starting simple: let each participant $i$ have idiosyncratic differences in reading speed

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

In R, we’d write this relationship as

$$RT \sim 1 + x + (1 | \text{participant})$$
Once again we’ll assume for simplicity that the number of word neighbors $x$ has a linear effect on mean reading time, and that trial-level noise is normally distributed.*

Random effects, starting simple: let each participant $i$ have idiosyncratic differences in reading speed

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

In R, we’d write this relationship as

$$RT \sim 1 + x + (1 | \text{participant})$$

Once again we can leave off the 1, and the noise term $\epsilon_{ij}$ is implicit
Once again we’ll assume for simplicity that the number of word neighbors $x$ has a linear effect on mean reading time, and that trial-level noise is normally distributed.

Random effects, starting simple: let each participant $i$ have idiosyncratic differences in reading speed

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

In R, we’d write this relationship as

$$RT \sim x + (1 \mid \text{participant})$$

Once again we can leave off the 1, and the noise term $\epsilon_{ij}$ is implicit.
One beauty of multi-level models is that you can simulate trial-level data
This is invaluable for achieving deeper understanding of both your analysis and your data
Multi-level Models VI: simulating data

\[ RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \]

\( \sim \mathcal{N}(0,\sigma_b) \)

Noise \( \sim \mathcal{N}(0,\sigma_e) \)

- One beauty of multi-level models is that you can simulate trial-level data
- This is invaluable for achieving deeper understanding of both your analysis and your data

```r
## simulate some data
> sigma.b <- 125 # inter-subject variation larger than
> sigma.e <- 40 # intra-subject, inter-trial variation
> alpha <- 500
> beta <- 12
> M <- 6 # number of participants
> n <- 50 # trials per participant
> b <- rnorm(M, 0, sigma.b) # individual differences
> nneighbors <- rpois(M*n,3) + 1 # generate num. neighbors
> subj <- rep(1:M,n)
> RT <- alpha + beta * nneighbors + b[subj] + rnorm(M*n,0,sigma.e) # simulate RTs!
```
Participant-level clustering is easily visible
Participant-level clustering is easily visible
Participant-level clustering is easily visible.

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

\[ RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \]

\( \sim N(0, \sigma_b) \)

\( \text{Noise} \sim N(0, \sigma_e) \)

Thus far, we’ve just defined a model and used it to generate data.
Statistical inference with multi-level models

\[ RT_{ij} = \alpha + \beta x_{ij} + b_i + \varepsilon_{ij} \sim N(0, \sigma_b) \quad \text{Noise} \sim N(0, \sigma_e) \]

- Thus far, we've just defined a model and used it to generate data
- We psycholinguists are usually in the opposite situation...
Statistical inference with multi-level models

\[ RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \]

\( \sim N(0, \sigma_b) \) \hspace{1cm} \text{Noise} \sim N(0, \sigma_e) 

▶ Thus far, we’ve just defined a model and used it to generate data
▶ We psycholinguists are usually in the opposite situation . . .
▶ We have data and we need to infer a model
  ▶ Specifically, the “fixed-effect” parameters \( \alpha, \beta, \) and \( \sigma_e \), plus the parameter governing inter-subject variation, \( \sigma_b \)
  ▶ e.g., hypothesis tests about effects of neighborhood density: can we reliably infer that \( \beta \) is \{non-zero, positive, . . .\}?
Statistical inference with multi-level models

\[ RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \]

\[ \sim N(0, \sigma_b) \quad \text{Noise} \sim N(0, \sigma_e) \]

- Thus far, we've just defined a model and used it to generate data.
- We psycholinguists are usually in the opposite situation... We have data and we need to infer a model.
  - Specifically, the “fixed-effect” parameters \( \alpha, \beta, \) and \( \sigma_e, \) plus the parameter governing inter-subject variation, \( \sigma_b \)
  - e.g., hypothesis tests about effects of neighborhood density: can we reliably infer that \( \beta \) is \{non-zero, positive, \ldots\}?
- Fortunately, we can use the same principles as before to do this:
  - The principle of maximum likelihood
  - Or Bayesian inference
Fitting a multi-level model using maximum likelihood

\[ RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \]

\[ \sim \mathcal{N}(0, \sigma_b) \quad \text{Noise} \sim \mathcal{N}(0, \sigma_\epsilon) \]

```r
> m <- lmer(time ~ neighbors.centered + (1 | participant), dat, REML=F)
> print(m, corr=F)

[...]
Random effects:
  Groups Name Variance Std.Dev.
  participant (Intercept) 4924.9  70.177
  Residual 19240.5  138.710

Number of obs: 1760, groups: participant, 44

Fixed effects:
  Estimate Std. Error  t value
(Intercept) 583.787     11.082   52.68
neighbors.centered 8.986      1.278    7.03
```
Fitting a multi-level model using maximum likelihood

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

\[
\begin{align*}
\hat{\alpha} & \quad \hat{\beta} \\
(\text{Intercept}) & \quad 583.787 \\
\text{neighbors.centered} & \quad 8.986
\end{align*}
\]

\[
\begin{align*}
\text{Estimate} & \quad \text{Std. Error} & \quad \text{t value} & \quad \text{t value} \\
(\text{Intercept}) & \quad 11.082 & \quad 52.68 \\
\text{neighbors.centered} & \quad 1.278 & \quad 7.03
\end{align*}
\]
Fitting a multi-level model using maximum likelihood

\[ \begin{align*}
RT_{ij} &= \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \\
&\sim \mathcal{N}(0, \sigma_b) \quad \text{Noise} \sim \mathcal{N}(0, \sigma_c)
\end{align*} \]

```r
> m <- lmer(time ~ neighbors.centered + (1 | participant), dat, REML=F)
> print(m, corr=F)

[...]
Random effects:
Groups     Name        Variance Std.Dev.
participant (Intercept) 4924.9    70.177
Residual             19240.5   138.710
Number of obs: 1760, groups: participant, 44

Fixed effects:
\[ \begin{align*}
\hat{\alpha} & \quad \text{Estimate} \quad \text{Std. Error} \quad t \quad \text{value} \\
(\text{Intercept}) & \boxed{583.787} \quad 11.082 \quad 52.68 \\
\text{neighbors.centered} & \boxed{8.986} \quad 1.278 \quad 7.03
\end{align*} \]
Fitting a multi-level model using maximum likelihood

\[ RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \]

\[ \sim N(0,\sigma_b) \quad \text{Noise} \sim N(0,\sigma_e) \]

```r
> m <- lmer(time ~ neighbors.centered + (1 | participant), dat, REML=F)
> print(m, corr=F)
```

[...]

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant</td>
<td>(Intercept)</td>
<td>4924.9</td>
<td>70.177</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>19240.5</td>
<td>138.710</td>
</tr>
</tbody>
</table>

Number of obs: 1760, groups: participant, 44

Fixed effects:

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>neighbors.centered</td>
<td>8.986</td>
<td>1.278</td>
<td>7.03</td>
</tr>
</tbody>
</table>

\[ \hat{\alpha} \]

\[ \hat{\beta} \]

\[ \hat{\sigma}_b \]
Fitting a multi-level model using maximum likelihood

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

```r
> m <- lmer(time ~ neighbors.centered + (1 | participant), dat, REML=F)
> print(m, corr=F)

[...]
Random effects:
Groups Name Variance Std.Dev.
participant (Intercept) 4924.9 70.177
Residual 19240.5 138.710
Number of obs: 1760, groups: participant, 44

Fixed effects:
\( \hat{\alpha} \) Estimate Std. Error t value
(Intercept) 583.787 11.082 52.68
neighbors.centered 8.986 1.278 7.03
\( \hat{\beta} \)
Interpreting parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>583.79</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.99</td>
</tr>
<tr>
<td>$\hat{\sigma}_b$</td>
<td>70.18</td>
</tr>
<tr>
<td>$\hat{\sigma}_\epsilon$</td>
<td>138.7</td>
</tr>
</tbody>
</table>

The fixed effects are interpreted just as in a traditional single-level model:

- The "average" RT for a non-word in this study is 583.79ms.
- Every extra neighbor increases "average" RT by 8.99ms.

Inter-trial variability $\sigma_\epsilon$ also has the same interpretation:

- Inter-trial variability for a given participant is Gaussian, centered around the participant+word-specific mean with standard deviation 138.7ms.

Inter-participant variability $\sigma_b$ is what's new:

- Variability in average RT in the population from which the participants were drawn has standard deviation 70.18ms.
Interpreting parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>583.79</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.99</td>
</tr>
<tr>
<td>$\sigma_b$</td>
<td>70.18</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>138.7</td>
</tr>
</tbody>
</table>

- The **fixed effects** are interpreted just as in a traditional single-level model:
Interpreting parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>583.79</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.99</td>
</tr>
<tr>
<td>$\hat{\sigma}_b$</td>
<td>70.18</td>
</tr>
<tr>
<td>$\hat{\sigma}_\epsilon$</td>
<td>138.7</td>
</tr>
</tbody>
</table>

- The fixed effects are interpreted just as in a traditional single-level model:
  - The “average” RT for a non-word in this study is 583.79 ms
The fixed effects are interpreted just as in a traditional single-level model:

- The “average” RT for a non-word in this study is 583.79ms.
- Every extra neighbor increases “average” RT by 8.99ms.
Interpreting parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>583.79</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.99</td>
</tr>
<tr>
<td>$\hat{\sigma}_b$</td>
<td>70.18</td>
</tr>
<tr>
<td>$\hat{\sigma}_\epsilon$</td>
<td>138.7</td>
</tr>
</tbody>
</table>

- The *fixed effects* are interpreted just as in a traditional single-level model:
  - The “average” RT for a non-word in this study is 583.79ms
  - Every extra neighbor increases “average” RT by 8.99ms
- Inter-trial variability $\sigma_\epsilon$ also has the same interpretation
Interpreting parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>583.79</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.99</td>
</tr>
<tr>
<td>$\hat{\sigma}_b$</td>
<td>70.18</td>
</tr>
<tr>
<td>$\hat{\sigma}_\epsilon$</td>
<td>138.7</td>
</tr>
</tbody>
</table>

- The **fixed effects** are interpreted just as in a traditional single-level model:
  - The “average” RT for a non-word in this study is 583.79ms
  - Every extra neighbor increases “average” RT by 8.99ms
- Inter-trial variability $\sigma_\epsilon$ also has the same interpretation
  - Inter-trial variability for a given participant is Gaussian, centered around the participant+word-specific mean with standard deviation 138.7ms
Interpreting parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>583.79</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.99</td>
</tr>
<tr>
<td>$\hat{\sigma}_b$</td>
<td>70.18</td>
</tr>
<tr>
<td>$\hat{\sigma}_\epsilon$</td>
<td>138.7</td>
</tr>
</tbody>
</table>

- The **fixed effects** are interpreted just as in a traditional single-level model:
  - The “average” RT for a non-word in this study is 583.79ms
  - Every extra neighbor increases “average” RT by 8.99ms
- Inter-trial variability $\sigma_\epsilon$ also has the same interpretation:
  - Inter-trial variability for a given participant is Gaussian, centered around the participant+word-specific mean with standard deviation 138.7ms
- Inter-participant variability $\sigma_b$ is what’s new:
Interpreting parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>583.79</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.99</td>
</tr>
<tr>
<td>$\hat{\sigma}_b$</td>
<td>70.18</td>
</tr>
<tr>
<td>$\hat{\sigma}_\epsilon$</td>
<td>138.7</td>
</tr>
</tbody>
</table>

- The **fixed effects** are interpreted just as in a traditional single-level model:
  - The “average” RT for a non-word in this study is 583.79ms
  - Every extra neighbor increases “average” RT by 8.99ms
- Inter-trial variability $\sigma_\epsilon$ also has the same interpretation
  - Inter-trial variability for a given participant is Gaussian, centered around the participant+word-specific mean with standard deviation 138.7ms
- Inter-participant variability $\sigma_b$ is what’s new:
  - Variability in average RT in the population from which the participants were drawn has standard deviation 70.18ms
Inferences about cluster-level parameters

\[
RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij}
\]

What about the participants’ idiosyncracies themselves—the \( b_i \)?
Inferences about cluster-level parameters

\[ RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \]

What about the participants’ idiosyncrasies themselves—the \( b_i \)?

We can also draw inferences about these—you may have heard about BLUPs
Inferences about cluster-level parameters

\[
RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij}
\]

\[\sim N(0,\sigma_b) \quad \text{Noise} \sim N(0,\sigma_\epsilon)\]

▶ What about the participants’ idiosyncracies themselves—the \(b_i\)?

▶ We can also draw inferences about these—you may have heard about **BLUPs**

▶ To understand these: committing to fixed-effect and random-effect parameter estimates determines a conditional probability distribution on participant-specific effects:

\[
P(b_i | \hat{\alpha}, \hat{\beta}, \hat{\sigma}_b, \hat{\sigma}_\epsilon)
\]
Inferences about cluster-level parameters

\[
RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij}
\]

\(\sim N(0, \sigma_b)\) \hspace{1cm} \text{Noise} \sim N(0, \sigma_c)

What about the participants’ idiosyncracies themselves—the \(b_i\)?

We can also draw inferences about these—you may have heard about BLUPs.

To understand these: committing to fixed-effect and random-effect parameter estimates determines a conditional probability distribution on participant-specific effects:

\[
P(b_i | \hat{\alpha}, \hat{\beta}, \hat{\sigma}_b, \hat{\sigma}_c)
\]

The BLUPS are the conditional modes of \(b_i\)—the choices that maximize the above probability.
Inferences about cluster-level parameters II

- The BLUP participant-specific “average” RTs for this dataset are black lines on the base of this graph.

- The solid line is a guess at their distribution.

- The dotted line is the distribution predicted by the model for the population from which the participants are drawn.

- Reasonably close correspondence.
The BLUP participant-specific “average” RTs for this dataset are black lines on the base of this graph.

- The solid line is a guess at their distribution.
- The dotted line is the distribution predicted by the model for the population from which the participants are drawn.
The BLUP participant-specific “average” RTs for this dataset are black lines on the base of this graph.

The solid line is a guess at their distribution.
The dotted line is the distribution predicted by the model for the population from which the participants are drawn.
Reasonably close correspondence.
Participants may also have idiosyncratic sensitivities to *neighborhood density*

R^T_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij}

\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})

\begin{align*}
\text{Random effects:} \\
\text{participant (Intercept): } & 4928.625, 70.2042 \\
\text{neighbors.centered: } & 19.421, 4.4069, -0.307 \\
\text{Residual: } & 19107.143, 138.2286
\end{align*}

These three numbers jointly characterize \( \hat{\Sigma}_b \).
Participants may also have idiosyncratic sensitivities to neighborhood density.

Incorporate by adding cluster-level slopes into the model:

\[
RT_{ij} = \alpha + \beta x_{ij} + b_1i + b_2i x_{ij} + \epsilon_{ij}
\]

where

\[\sim N(0, \Sigma_b)\]

\[\text{Noise} \sim N(0, \sigma_\epsilon)\]
Participants may also have idiosyncratic sensitivities to *neighborhood density*

Incorporate by adding cluster-level slopes into the model:

\[
RT_{ij} = \alpha + \beta x_{ij} + b_{1i} + b_{2i} x_{ij} + \epsilon_{ij}
\]

\[\sim N(0, \Sigma_b)\]

Noise \[\sim N(0, \sigma_\epsilon)\]

In R (once again we can omit the 1’s):

\[
RT \sim 1 + x + (1 + x \mid \text{participant})
\]
Participants may also have idiosyncratic sensitivities to \textit{neighborhood density}.

Incorporate by adding cluster-level slopes into the model:

\[
RT_{ij} = \alpha + \beta x_{ij} + b_{1i} + b_{2i} x_{ij} + \epsilon_{ij}
\]

\[
\sim N(0,\Sigma_b) \quad \text{Noise} \sim N(0,\sigma_{\epsilon})
\]

In R (once again we can omit the 1’s):

\[
RT \sim 1 + x + (1 + x | \text{participant})
\]

```r
> lmer(RT ~ neighbors.centered +
> (neighbors.centered | participant), dat,REML=F)
```

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant</td>
<td>(Intercept)</td>
<td>4928.625</td>
<td>70.2042</td>
<td></td>
</tr>
<tr>
<td></td>
<td>neighbors.centered</td>
<td>19.421</td>
<td>4.4069</td>
<td>-0.307</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>19107.143</td>
<td>138.2286</td>
<td></td>
</tr>
</tbody>
</table>
Participants may also have idiosyncratic sensitivities to neighborhood density.

Incorporate by adding cluster-level slopes into the model:

\[
RT_{ij} = \alpha + \beta x_{ij} + b_{1i} + b_{2i} x_{ij} + \epsilon_{ij}
\]

\[
\sim N(0, \Sigma_b) \quad \text{Noise} \sim N(0, \sigma_{\epsilon})
\]

In R (once again we can omit the 1’s):

\[
RT \sim 1 + x + (1 + x \mid \text{participant})
\]

```r
> lmer(RT ~ neighbors.centered +
     (neighbors.centered | participant), dat,REML=F)
```

These three numbers jointly characterize \(\hat{\Sigma}_b\):
Let’s talk a little more about cluster-level slopes

\[ RT_{ij} = \alpha + \beta x_{ij} + b_1 i + b_2 i x_{ij} + \epsilon_{ij} \]

\[ \sim N(0, \Sigma_b) \]

Noise \[ \sim N(0, \sigma_\epsilon) \]

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant</td>
<td>(Intercept)</td>
<td>4928.625</td>
<td>70.2042</td>
<td></td>
</tr>
<tr>
<td>neighbors</td>
<td>centered</td>
<td>19.421</td>
<td>4.4069</td>
<td>-0.307</td>
</tr>
</tbody>
</table>

The results of the `lmer()` fit are saying that the maximum-likelihood estimate of the covariance matrix \( \Sigma_b \) governing participant-specific variability is

\[ \hat{\Sigma}_b = \begin{pmatrix} 70 & -0.3097 \\ -0.3097 & 4.41 \end{pmatrix} \]
Inferences about cluster-level parameters IV

Let’s talk a little more about cluster-level slopes

\[ RT_{ij} = \alpha + \beta x_{ij} + b_{1i} + b_{2i} x_{ij} + \epsilon_{ij} \]

\( \sim N(0, \Sigma_b) \)  \( \sim N(0, \sigma_{\epsilon}) \)

We’ve said that participant-specific idiosyncracies are multivariate normally distributed around the origin with covariance matrix \( \Sigma_b \)

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant</td>
<td>(Intercept)</td>
<td>4928.625</td>
<td>70.2042</td>
<td></td>
</tr>
<tr>
<td></td>
<td>neighbors.centered</td>
<td>19.421</td>
<td>4.4069</td>
<td>-0.307</td>
</tr>
</tbody>
</table>
Let’s talk a little more about cluster-level slopes

\[ RT_{ij} = \alpha + \beta x_{ij} + b_{1i} + b_{2i} x_{ij} + \epsilon_{ij} \]

We’ve said that participant-specific idiosyncracies are multivariate normally distributed around the origin with covariance matrix \( \Sigma_b \)

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant</td>
<td>(Intercept)</td>
<td>4928.625</td>
<td>70.2042</td>
<td></td>
</tr>
<tr>
<td></td>
<td>neighbors.centered</td>
<td>19.421</td>
<td>4.4069</td>
<td>-0.307</td>
</tr>
</tbody>
</table>

The results of the `lmer()` fit are saying that the maximum-likelihood estimate of the covariance matrix \( \hat{\Sigma}_b \) governing participant-specific variability is

\[
\hat{\Sigma}_b = \begin{pmatrix}
70.20 & -0.3097 \\
-0.3097 & 4.41
\end{pmatrix}
\]
Inference about cluster-level parameters V

Visualizing some multivariate normal distributions:

Covariance matrix
\[ \Sigma_b = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 4 \end{pmatrix} \]

Perspective plot

Contour plot

Covariance matrix
\[ \Sigma_b = \begin{pmatrix} 2.5 & -0.13 \\ -0.13 & 2 \end{pmatrix} \]
Inference about cluster-level parameters VI

- In 2D we often visually summarize a multivariate normal distribution with a characteristic ellipse

\[
\Sigma_b = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 4 \end{pmatrix}
\]
In 2D we often visually summarize a multivariate normal distribution with a characteristic ellipse. This ellipse contains a certain proportion (here & conventionally, 95%) of the probability mass for the distribution in question.
Correlation visible in participant-specific BLUPs

Participants who were faster overall also tend to be more affected by neighborhood density.

\[
\hat{\Sigma}_b = \begin{pmatrix}
70.20 & -0.3097 & -0.3097 & 4.41 \\
-0.3097 & 1 & -1 & 1 \\
-0.3097 & -1 & 1 & -1 \\
4.41 & 1 & -1 & 1
\end{pmatrix}
\]
Participants

Participants who were faster overall also tend to be more affected by neighborhood density.
Participants who were faster overall also tend to be more affected by neighborhood density.

\[ \hat{\Sigma}_b = (70.20 - 0.3097 - 0.3097 4.41) \]
Correlation visible in participant-specific BLUPs
Correlation visible in participant-specific BLUPs
Participants who were faster overall also tend to be more affected by neighborhood density

\[ \hat{\Sigma}_b = \begin{pmatrix} 70.20 & -0.3097 \\ -0.3097 & 4.41 \end{pmatrix} \]
We can also use Bayes’ rule to draw inferences about fixed effects.
Bayesian inference for multilevel models

\[
P(\{\beta_i\}, \sigma_b, \sigma_\epsilon | Y) = \frac{\text{Likelihood}}{\text{Prior}} = \frac{P(Y | \{\beta_i\}, \sigma_b, \sigma_\epsilon)}{P(Y | \{\beta_i\}, \sigma_b, \sigma_\epsilon)}
\]

- We can also use Bayes’ rule to draw inferences about fixed effects.
- Computationally more challenging than with single-level regression; Markov-chain Monte Carlo (MCMC) sampling techniques allow us to approximate it.
Bayesian inference for multilevel models

\[ P(\{\beta_i\}, \sigma_b, \sigma_\epsilon | Y) = \frac{P(Y|\{\beta_i\}, \sigma_b, \sigma_\epsilon) P(\{\beta_i\}, \sigma_b, \sigma_\epsilon)}{P(Y)} \]

- We can also use Bayes’ rule to draw inferences about fixed effects
- Computationally more challenging than with single-level regression; Markov-chain Monte Carlo (MCMC) sampling techniques allow us to approximate it
If you have had any training in psychology, you be asking yourself:

*Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.*
Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.
Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.

- Yes, but there are several respects in which multi-level models go beyond AN(C)OVA:
Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.

- Yes, but there are several respects in which multi-level models go beyond AN(C)OVA:
  1. They handle *imbalanced datasets* just as well as balanced datasets
Why do you care??? II

Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.

▶ Yes, but there are several respects in which multi-level models go beyond AN(C)OVA:

1. They handle *imbalanced datasets* just as well as balanced datasets
2. You can use non-linear linking functions (e.g., *logit models* for binary-choice data)
Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.

▶ Yes, but there are several respects in which multi-level models go beyond AN(C)OVA:

1. They handle *imbalanced datasets* just as well as balanced datasets
2. You can use non-linear linking functions (e.g., logit models for binary-choice data)
3. You can cross cluster-level effects
   ▶ Every trial belongs to both a participant cluster and an item cluster
Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.

▶ Yes, but there are several respects in which multi-level models go beyond AN(C)OVA:

1. They handle *imbalanced datasets* just as well as balanced datasets
2. You can use non-linear linking functions (e.g., logit models for binary-choice data)
3. You can cross cluster-level effects
   - Every trial belongs to both a participant cluster and an item cluster
   - You can build a single unified model for inferences from your data
Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.

Yes, but there are several respects in which multi-level models go beyond AN(C)OVA:

1. They handle *imbalanced datasets* just as well as balanced datasets
2. You can use non-linear linking functions (e.g., logit models for binary-choice data)
3. You can cross cluster-level effects
   - Every trial belongs to both a participant cluster and an item cluster
   - You can build a single unified model for inferences from your data
   - ANOVA requires separate by-participants and by-items analyses (min-$F'$ is quite conservative)
Do I really care about these models? For hypothesis testing I could do everything you just did with an ANCOVA, treating participant as a random factor, or by looking at participant means.

- Yes, but there are several respects in which multi-level models go beyond AN(C)OVA:
  1. They handle *imbalanced datasets* just as well as balanced datasets
  2. You can use non-linear linking functions (e.g., logit models for binary-choice data)
  3. You can cross cluster-level effects
     - Every trial belongs to both a participant cluster and an item cluster
     - You can build a single unified model for inferences from your data
     - ANOVA requires separate by-participants and by-items analyses (min-$F'$ is quite conservative)
The logit link function for categorical data

- Much psycholinguistic data is *categorical* rather than *continuous*:
  - Yes/no answers to alternations questions
  - Speaker choice: *(realized *(that) her goals were unattainable)*
  - Cloze continuations, and so forth...
The logit link function for categorical data

- Much psycholinguistic data is *categorical* rather than *continuous*:
  - Yes/no answers to alternations questions
  - Speaker choice: *(realized (that) her goals were unattainable)*
  - Cloze continuations, and so forth...
- Linear models inappropriate; they predict continuous values
The logit link function for categorical data

- Much psycholinguistic data is *categorical* rather than *continuous*:
  - Yes/no answers to alternations questions
  - Speaker choice: *(realized (that) her goals were unattainable)*
  - Cloze continuations, and so forth...

- Linear models inappropriate; they predict continuous values
- We can stay within the multi-level generalized linear models framework but use different link functions and noise distributions to analyze categorical data
The logit link function for categorical data

- Much psycholinguistic data is *categorical* rather than *continuous*:
  - Yes/no answers to alternations questions
  - Speaker choice: *(realized (that) her goals were unattainable)*
  - Cloze continuations, and so forth...

- Linear models inappropriate; they predict continuous values
- We can stay within the multi-level generalized linear models framework but use different link functions and noise distributions to analyze categorical data
- e.g., the logit model (Agresti, 2002; Jaeger, 2008)
The logit link function for categorical data

- Much psycholinguistic data is *categorical* rather than *continuous*:
  - Yes/no answers to alternations questions
  - Speaker choice: *(realized* *(that)* *her goals were unattainable)*
  - Cloze continuations, and so forth...

- Linear models inappropriate; they predict continuous values

- We can stay within the multi-level generalized linear models framework but use different *link functions* and *noise distributions* to analyze categorical data

- e.g., the logit model *(Agresti, 2002; Jaeger, 2008)*

\[ \eta_{ij} = \alpha + \beta X_{ij} + b_i \]  

(linear predictor)
The logit link function for categorical data

- Much psycholinguistic data is *categorical* rather than *continuous*:
  - Yes/no answers to alternations questions
  - Speaker choice: *(realized *(that)* her goals were unattainable)*
  - Cloze continuations, and so forth...
- Linear models inappropriate; they predict continuous values
- We can stay within the multi-level generalized linear models framework but use different link functions and noise distributions to analyze categorical data
- e.g., the logit model (Agresti, 2002; Jaeger, 2008)

\[ \eta_{ij} = \alpha + \beta X_{ij} + b_i \]  
(linear predictor)

\[ \eta_{ij} = \log \frac{\mu_{ij}}{1 - \mu_{ij}} \]  
(link function)
The logit link function for categorical data

- Much psycholinguistic data is *categorical* rather than *continuous*:
  - Yes/no answers to alternations questions
  - Speaker choice: *(realized (that) her goals were unattainable)*
  - Cloze continuations, and so forth...

- Linear models inappropriate; they predict continuous values

- We can stay within the multi-level generalized linear models framework but use different link functions and noise distributions to analyze categorical data

- e.g., the logit model (Agresti, 2002; Jaeger, 2008)

\[
\eta_{ij} = \alpha + \beta X_{ij} + b_i \quad \text{(linear predictor)}
\]

\[
\eta_{ij} = \log \frac{\mu_{ij}}{1 - \mu_{ij}} \quad \text{(link function)}
\]

\[
P(Y = y; \mu_{ij}) = \mu_{ij} \quad \text{(binomial noise distribution)}
\]
We’ll look at the effect of neighborhood density on correct identification as non-word in Bicknell et al. (2010)
We’ll look at the effect of neighborhood density on correct identification as non-word in Bicknell et al. (2010)

Assuming that any effect is linear in log-odds space (see Jaeger (2008) for discussion)
We’ll look at the effect of neighborhood density on correct identification as non-word in Bicknell et al. (2010).

Assuming that any effect is linear in log-odds space (see Jaeger (2008) for discussion)

```r
> lmer(correct ~ neighbors.scaled + (neighbors.scaled
  / participant) + (1 | target), dat, family="binomial")
```

```
[...]
Random effects:
  Groups   Name          Variance  Std.Dev.  Corr
participant (Intercept)  1.139785  1.06761
               neighbors.scaled  0.030559  0.17481 -1.000
  target (Intercept)     0.213311  0.46186
Number of obs: 1760, groups: participant, 44; target, 40

Fixed effects:
        Estimate Std. Error   z value  Pr(>|z|)
(Intercept)   3.3593     0.2215    15.168 < 2e-16 ***
neighbors.scaled  -0.6360    0.1271   -5.005 5.59e-07 ***
```

*α*—participants usually right

*Σ*<sub>b</sub>

*S* (note there is no *σ*<sub>ϵ</sub> for logit models)

*β*—effect small compared with inter-subject variation
We’ll look at the effect of neighborhood density on correct identification as non-word in Bicknell et al. (2010).

Assuming that any effect is linear in log-odds space (see Jaeger (2008) for discussion)

```r
> lmer(correct ~ neighbors.scaled + (neighbors.scaled | participant) + (1 | target), dat, family="binomial")
```

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant</td>
<td>(Intercept)</td>
<td>1.139785</td>
<td>1.06761</td>
<td></td>
</tr>
<tr>
<td></td>
<td>neighbors.scaled</td>
<td>0.030559</td>
<td>0.17481</td>
<td>-1.000</td>
</tr>
<tr>
<td>target</td>
<td>(Intercept)</td>
<td>0.213311</td>
<td>0.46186</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs: 1760, groups: participant, 44; target, 40

α—participants usually right

Fixed effects:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|---------|
| (Intercept)    | 3.3593   | 0.2215     | 15.168  | < 2e-16 *** |
| neighbors.scaled | -0.6360 | 0.1271     | -5.005  | 5.59e-07 *** |
The logit link function for categorical data II

- We’ll look at the effect of neighborhood density on correct identification as non-word in Bicknell et al. (2010)
- Assuming that any effect is linear in log-odds space (see Jaeger (2008) for discussion)

```r
> lmer(correct ~ neighbors.scaled + (neighbors.scaled | participant) + (1 | target), dat, family="binomial")
```

 Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant</td>
<td>(Intercept)</td>
<td>1.139785</td>
<td>1.06761</td>
<td></td>
</tr>
<tr>
<td></td>
<td>neighbors.scaled</td>
<td>0.030559</td>
<td>0.17481</td>
<td>-1.000</td>
</tr>
<tr>
<td>target</td>
<td>(Intercept)</td>
<td>0.213311</td>
<td>0.46186</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs: 1760, groups: participant, 44; target, 40

Fixed effects:

| (Intercept)   | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------|----------|------------|---------|---------|
|               | 3.3593   | 0.2215     | 15.168  | < 2e-16 *** |
| neighbors.scaled | -0.6360 | 0.1271     | -5.005  | 5.59e-07 *** |
The logit link function for categorical data II

- We’ll look at the effect of neighborhood density on correct identification as non-word in Bicknell et al. (2010)
- Assuming that any effect is linear in log-odds space (see Jaeger (2008) for discussion)

```r
> lmer(correct ~ neighbors.scaled + (neighbors.scaled | participant) + (1 | target), dat, family="binomial")

 Fixed effects:  
 (Intercept) 3.3593 0.2215 15.168 < 2e-16 *** neighbors.scaled -0.6360 0.1271 -5.005 5.59e-07 ***  

Random effects:  
 Groups Name Variance Std.Dev. Corr
 participant (Intercept) 1.139785 1.06761
 neighbors.scaled 0.030559 0.17481 -1.000
 target (Intercept) 0.213311 0.46186

 Number of obs: 1760, groups: participant, 44; target, 40

 ^\alpha—participants usually right
 ^\beta—effect small compared with inter-subject variation
 ^\Sigma_{bs} (note there is no ^\sigma_{\epsilon} for logit models)
 ^\sigma_{bw}
Hierarchical (multi-level, mixed-effects) models may seem strange and foreign. But all you really need to understand them is three basic things:

- Generalized linear models
- The principle of maximum likelihood
- Bayesian inference

These models open up many new interesting doors!
The good, the bad, and the ugly

So now you have this new freedom in how to model your data!
The good, the bad, and the ugly

- So now you have this new freedom in how to model your data!
- But this opens up many new choices you didn’t have before
So now you have this new freedom in how to model your data!
But this opens up many new choices you didn’t have before
And if you make the wrong choices, you can draw the wrong inferences from your dataset
A note on $p$-values and philosophy of science

- Frequentist hypothesis testing means the Neyman-Pearson paradigm, with an asymmetry between null ($H_0$) and alternative ($H_1$) hypotheses.
Frequentist hypothesis testing means the Neyman-Pearson paradigm, with an asymmetry between null ($H_0$) and alternative ($H_1$) hypotheses.

A $p$-value from a dataset $D$ is how unlikely a given dataset was to be produced under $H_0$. 

Note that so-called "pMCMC" is NOT a $p$-value in the Neyman-Pearson sense!
Frequentist hypothesis testing means the Neyman-Pearson paradigm, with an asymmetry between null ($H_0$) and alternative ($H_1$) hypotheses.

A $p$-value from a dataset $D$ is how unlikely a given dataset was to be produced under $H_0$.

Note that so-called “$p_{MCMC}$” is NOT a $p$-value in the Neyman-Pearson sense!
A note on \( p \)-values and philosophy of science

- Frequentist hypothesis testing means the Neyman-Pearson paradigm, with an asymmetry between null (\( H_0 \)) and alternative (\( H_1 \)) hypotheses.
- A \( p \)-value from a dataset \( D \) is how unlikely a given dataset was to be produced under \( H_0 \).
- Note that so-called "\( p_{MCMC} \)" is NOT a \( p \)-value in the Neyman-Pearson sense!
- Weakness, both in practice and in principle: the alternative hypothesis is never actually used (except indirectly in determining optimal acceptance and rejection regions).
A note on \( p \)-values and philosophy of science

Alternative: Bayesian hypothesis testing, which is symmetric:

\[
\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0) P(H_0)}{P(D|H_1) P(H_1)}
\]

I am fundamentally Bayesian in my philosophy of science

But, weakness in practice: your likelihoods \( P(D|H_0) \) and \( P(D|H_1) \) can depend on fine details of your assumptions about \( H_0 \) and \( H_1 \)

I do not trust you to assess these likelihoods neutrally! (Nor should you trust me)

So for me, the \( p \)-value of your experiment serves as a rough indicator of how small \( P(D|H_0) \) may be

Technically, such a measure doesn't need to be a true Neyman-Pearson \( p \)-value (\( p \)-MCMC falls into this category)
A note on $p$-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0) P(H_0)}{P(D|H_1) P(H_1)}$$

- I am fundamentally Bayesian in my philosophy of science
A note on \( p \)-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

\[
\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0) P(H_0)}{P(D|H_1) P(H_1)}
\]

- I am fundamentally Bayesian in my philosophy of science

- But, weakness in practice: your likelihoods \( P(D|H_0) \) and \( P(D|H_1) \) can depend on fine details of your assumptions about \( H_0 \) and \( H_1 \)
A note on $p$-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0) P(H_0)}{P(D|H_1) P(H_1)}$$

- I am fundamentally Bayesian in my philosophy of science

- But, weakness in practice: your likelihoods $P(D|H_0)$ and $P(D|H_1)$ can depend on fine details of your assumptions about $H_0$ and $H_1$

- I do not trust you to assess these likelihoods neutrally! (Nor should you trust me)
A note on $p$-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

\[
\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0) P(H_0)}{P(D|H_1) P(H_1)}
\]

- I am fundamentally Bayesian in my philosophy of science

- But, weakness in practice: your likelihoods $P(D|H_0)$ and $P(D|H_1)$ can depend on fine details of your assumptions about $H_0$ and $H_1$

- I do not trust you to assess these likelihoods neutrally! (Nor should you trust me)

- So for me, the $p$-value of your experiment serves as a rough indicator of how small $P(D|H_0)$ may be
A note on $p$-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)P(H_0)}{P(D|H_1)P(H_1)}$$

- I am fundamentally Bayesian in my philosophy of science

- But, weakness in practice: your likelihoods $P(D|H_0)$ and $P(D|H_1)$ can depend on fine details of your assumptions about $H_0$ and $H_1$

- I do not trust you to assess these likelihoods neutrally! (Nor should you trust me)

- So for me, the $p$-value of your experiment serves as a rough indicator of how small $P(D|H_0)$ may be

- Technically, such a measure doesn’t need to be a true Neyman-Pearson $p$-value ($p_{MCMC}$ falls into this category)
I really care about fixed effects—what random effects do I use?

- Simplest possible example: look at naming times of a set of words in high-predictability versus low-predictability contexts.
I really care about fixed effects—what random effects do I use?

- Simplest possible example: look at naming times of a set of words in high-predictability versus low-predictability contexts
- Treat as a dichotomous factor
I really care about fixed effects—what random effects do I use?

- Simplest possible example: look at naming times of a set of words in high-predictability versus low-predictability contexts
- Treat as a dichotomous factor
- Quick review: what question are we asking when we do traditional by-participants and by-items ANOVAs (t-tests)?
I really care about fixed effects—what random effects do I use?

- Simplest possible example: look at naming times of a set of words in high-predictability versus low-predictability contexts
- Treat as a dichotomous factor
- Quick review: what question are we asking when we do traditional by-participants and by-items ANOVAs (t-tests)?
- Classic question: *above and beyond idiosyncratic sensitivities of different individuals to context-driven predictability*, are predictable words in general named faster than unpredictable words?

\[ \text{RT} \sim \text{Predictability} + \text{idiosyncratic sensitivities of different individuals (Predictability | Subj)} \]
I really care about fixed effects—what random effects do I use?

- Simplest possible example: look at naming times of a set of words in high-predictability versus low-predictability contexts
- Treat as a dichotomous factor
- Quick review: what question are we asking when we do traditional by-participants and by-items ANOVAs ($t$-tests)?
- Classic question: *above and beyond idiosyncratic sensitivities of different individuals to context-driven predictability*, are predictable words in general named faster than unpredictable words?
- In R:

$$\text{RT} \sim \underbrace{\text{Predictability}}_{\text{idiosyncratic sensitivities of different individuals}} + \underbrace{(\text{Predictability} | \text{Subj})}_{\text{above and beyond idiosyncratic sensitivities of different individuals to context-driven predictability}}$$
What random effects do I use?

Now consider a $2 \times 2$ design, where we want to assess interaction of factors A and B in face of idiosyncratic sensitivities of both individuals and linguistic items to experimental condition.
What random effects do I use?

Now consider a $2 \times 2$ design, where we want to assess interaction of factors A and B in face of idiosyncratic sensitivities of both individuals and linguistic items to experimental condition.

In R:

$$\text{Response} \sim A \times B + (A \times B|\text{Subj}) + (A \times B|\text{Item})$$
Now consider a $2 \times 2$ design, where we want to assess interaction of factors A and B in face of idiosyncratic sensitivities of both individuals and linguistic items to experimental condition.

In R:

$$\text{Response} \sim A \times B + (A \times B|\text{Subj}) + (A \times B|\text{Item})$$

Many of you may have had experience not being able to fit such a model to your data.
Example from Levy et al. (2012) self-paced reading:

1. The reporter interviewed the star about the movie *which was filmed in* the jungles of Vietnam. [VP-attached PP, RC adjacent]

2. The reporter interviewed the star about the movie *who was married to* the famous model. [VP-attached PP, RC distant]

3. The reporter interviewed the star of the movie *which was filmed in* the jungles of Vietnam. [NP-attached PP, RC adjacent]

4. The reporter interviewed the star of the movie *who was married to* the famous model. [NP-attached PP, RC distant]
Example from Levy et al. (2012) self-paced reading:

(1) The reporter interviewed the star about the movie *which was filmed in* the jungles of Vietnam. [VP-attached PP, RC adjacent]

(2) The reporter interviewed the star about the movie *who was married to* the famous model. [VP-attached PP, RC distant]

(3) The reporter interviewed the star of the movie *which was filmed in* the jungles of Vietnam. [NP-attached PP, RC adjacent]

(4) The reporter interviewed the star of the movie *who was married to* the famous model. [NP-attached PP, RC distant]
Example from Levy et al. (2012) self-paced reading:

```r
> lmer(rt ~ Cprep*Cloc + (cond - 1 | subj) + (cond - 1 | item), REML=F)

[...]

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>condabout local</td>
<td>9058.23</td>
<td>95.175</td>
<td></td>
</tr>
<tr>
<td></td>
<td>condabout nonlocal</td>
<td>29114.38</td>
<td>170.629</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>condof local</td>
<td>8661.62</td>
<td>93.068</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>condof nonlocal</td>
<td>10818.69</td>
<td>104.013</td>
<td>0.964</td>
</tr>
<tr>
<td>item</td>
<td>condabout local</td>
<td>0.00</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>condabout nonlocal</td>
<td>3662.90</td>
<td>60.522</td>
<td>NaN</td>
</tr>
<tr>
<td></td>
<td>condof local</td>
<td>399.24</td>
<td>19.981</td>
<td>NaN</td>
</tr>
<tr>
<td></td>
<td>condof nonlocal</td>
<td>316.46</td>
<td>17.789</td>
<td>NaN</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>21098.95</td>
<td>145.255</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs: 880, groups: subj, 44; item.factor, 20

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>424.823</td>
<td>18.193</td>
</tr>
<tr>
<td>Cprep</td>
<td>-15.308</td>
<td>6.690</td>
</tr>
<tr>
<td>Cloc</td>
<td>17.107</td>
<td>8.398</td>
</tr>
<tr>
<td>Cprep:Cloc</td>
<td>-22.935</td>
<td>5.670</td>
</tr>
</tbody>
</table>
```
Example from Levy et al. (2012) self-paced reading:

- I get concerned when I see NaN in my analysis! → perhaps consider tossing whatever source of variability led to that badness

\[
\text{lmer}(rt \sim \text{Cprep*Cloc} + (\text{cond} - 1 | \text{subj}) + (1 | \text{item}), \text{REML=F})
\]

[..]

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj.factor</td>
<td>condabout local</td>
<td>8874.35</td>
<td>94.204</td>
<td></td>
</tr>
<tr>
<td></td>
<td>condabout nonlocal</td>
<td>29637.98</td>
<td>172.157</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td>condof local</td>
<td>8707.14</td>
<td>93.312</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td>condof nonlocal</td>
<td>10994.36</td>
<td>104.854</td>
<td>0.969</td>
</tr>
<tr>
<td>item.factor</td>
<td>(Intercept)</td>
<td>247.19</td>
<td>15.722</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>21717.00</td>
<td>147.367</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs: 880, groups: subj.factor, 44; item.factor, 20

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>424.823</td>
<td>18.160</td>
<td>23.394</td>
</tr>
<tr>
<td>Cprep</td>
<td>-15.308</td>
<td>5.958</td>
<td>-2.569</td>
</tr>
<tr>
<td>Cloc</td>
<td>17.107</td>
<td>6.776</td>
<td>2.525</td>
</tr>
</tbody>
</table>
Example from Levy et al. (2012) self-paced reading:

- Comparing models with and without condition-specific sensitivity for items demonstrates that the data don’t incontrovertibly justify the full random effects structure for items.

```r
m1 <- lmer(rt ~ Cprep*Cloc + (cond - 1 | subj) + (1 | item))
m2 <- lmer(rt ~ Cprep*Cloc + (cond - 1 | subj) + (cond - 1 | item))
anova(m1, m2)
```

This is a likelihood-ratio test between models.

Determining which random effects structure to use is the problem of model selection.

The good news: model selection is extensively studied!

The bad news: there are many, many models to be selected from.
Example from Levy et al. (2012) self-paced reading:

- Comparing models with and without condition-specific sensitivity for items demonstrates that the data don’t incontrovertibly justify the full random effects structure for items

\[
\begin{align*}
> \ m1 \ &\leftarrow \ \text{lmer}(rt \sim \text{Cprep}\!\times\!\text{Cloc} + (\text{cond} - 1 \mid \text{subj}) + (1 \mid \text{item})) \\
> \ m2 \ &\leftarrow \ \text{lmer}(rt \sim \text{Cprep}\!\times\!\text{Cloc} + (\text{cond} - 1 \mid \text{subj}) + (\text{cond} - 1 \mid \text{item})) \\
> \ \text{anova}(m1,m2)
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>16</td>
<td>11461</td>
<td>11537</td>
<td>-5714.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m2</td>
<td>25</td>
<td>11465</td>
<td>11585</td>
<td>-5707.6</td>
<td>13.538</td>
<td>9</td>
<td>0.1397</td>
</tr>
</tbody>
</table>
Example from Levy et al. (2012) self-paced reading:

- Comparing models with and without condition-specific sensitivity for items demonstrates that the data don’t incontrovertibly justify the full random effects structure for items

```r
> m1 <- lmer(rt ~ Cprep*Cloc + (cond - 1 | subj) + (1 | item))
> m2 <- lmer(rt ~ Cprep*Cloc + (cond - 1 | subj) + (cond - 1 | item))
> anova(m1,m2)

Df  AIC  BIC logLik Chisq Chi Df Pr(>Chisq)
---
m1   16 11461 11537 -5714.3
m2   25 11465 11585 -5707.6  13.538  9       0.1397
```

- This is a likelihood-ratio test between models
- Determining which random effects structure to use is the problem of model selection
- The good news: model selection is extensively studied!
- The bad news: there are many, many models to be selected from
Models to be selected from

- Simple view for our $2 \times 2$ scenario (factors A and B)

  $\text{cond|subj, cond|item}$

  $\text{cond|subj, 1|item}$  $\text{1|subj, cond|item}$

  $\text{1|subj, 1|item}$

  $\text{1|subj}$  $\text{1|item}$

  no random effects
But wait! There are many other possible random-effects structures, e.g.

\[(A+B|\text{subj}) + (1|\text{item})\]
\[(1|\text{subj}) + (0+A|\text{subj}) + (1|\text{item})\]

...
But wait! There are many other possible random-effects structures, e.g.

\[(A+B|\text{subj}) + (1|\text{item})\]
\[(1|\text{subj}) + (0+A|\text{subj}) + (1|\text{item})\]
\[\ldots\]

What do these even mean?
Recall that a multivariate normal distribution is characterized by its covariance matrix.
Recall that a multivariate normal distribution is characterized by its covariance matrix.

For two conditions A1, A2:

\[
\begin{pmatrix}
A1 \\
A2
\end{pmatrix}
\begin{pmatrix}
\sigma & 1 \\
1 & \sigma
\end{pmatrix}
(1|\text{subj})
\]

\[
\begin{pmatrix}
A1 \\
A2
\end{pmatrix}
\begin{pmatrix}
\sigma_1 & 1 \\
1 & \sigma_2
\end{pmatrix}
\]

N/A (?)

\[
\begin{pmatrix}
A1 \\
A2
\end{pmatrix}
\begin{pmatrix}
\sigma_1 & 0 \\
0 & \sigma_2
\end{pmatrix}
(1|\text{subj}) + (0 + A|\text{subj})
\]

(but A must be recoded as numeric!)

\[
\begin{pmatrix}
A1 \\
A2
\end{pmatrix}
\begin{pmatrix}
\sigma_1 & \sigma_{12} \\
\sigma_{12} & \sigma_2
\end{pmatrix}
(A|\text{subj})
\]

If we add the possibility of no random effect, we get a hierarchy of 4 levels of richness.

With crossed participant and item random effects, 16 models in a lattice.

And the situation is much worse for more than two conditions.
Recall that a multivariate normal distribution is characterized by its covariance matrix.

For two conditions $A_1$, $A_2$:

\[
\begin{pmatrix}
A_1 & 1 \\
A_2 & 1
\end{pmatrix}
\begin{pmatrix}
\sigma \\
\sigma
\end{pmatrix} + (1|\text{subj})
\]

\[
\begin{pmatrix}
A_1 & 1 \\
A_2 & 1
\end{pmatrix}
\begin{pmatrix}
\sigma_1 \\
\sigma_2
\end{pmatrix} + N/A (?)
\]

\[
\begin{pmatrix}
A_1 & 0 \\
A_2 & 0
\end{pmatrix}
\begin{pmatrix}
\sigma_1 \\
\sigma_2
\end{pmatrix} + (1|\text{subj}) + (0 + A|\text{subj})
\]

(but $A$ must be recoded as numeric!)

If we add the possibility of no random effect, we get a hierarchy of 4 levels of richness.
Recall that a multivariate normal distribution is characterized by its covariance matrix.

For two conditions A1, A2:

\[
\begin{pmatrix}
A1 \\
A2
\end{pmatrix}
\begin{pmatrix}
\sigma & 1 \\
1 & \sigma
\end{pmatrix}
\]

\( (1|\text{subj}) \)

\[
\begin{pmatrix}
A1 \\
A2
\end{pmatrix}
\begin{pmatrix}
\sigma_1 & 1 \\
1 & \sigma_2
\end{pmatrix}
\]

N/A (?)

\[
\begin{pmatrix}
A1 \\
A2
\end{pmatrix}
\begin{pmatrix}
\sigma_1 & 0 \\
0 & \sigma_2
\end{pmatrix}
\]

\( (1|\text{subj}) + (0 + A|\text{subj}) \) (but A must be recoded as numeric!)

\[
\begin{pmatrix}
A1 \\
A2
\end{pmatrix}
\begin{pmatrix}
\sigma_1 & \sigma_{12} \\
\sigma_{12} & \sigma_2
\end{pmatrix}
\]

\( (A|\text{subj}) \)

If we add the possibility of no random effect, we get a hierarchy of 4 levels of richness.

With crossed participant and item random effects, 16 models in a lattice.
Recall that a multivariate normal distribution is characterized by its covariance matrix.

For two conditions A1, A2:

\[
\begin{pmatrix}
  A1 & \sigma \\
  A2 & 1
\end{pmatrix}
\begin{pmatrix}
  1 & \sigma \\
  1 & 1
\end{pmatrix}
\]

\((1|\text{subj})\)

\[
\begin{pmatrix}
  A1 & \sigma_1 \\
  A2 & 0
\end{pmatrix}
\begin{pmatrix}
  1 & \sigma_1 \\
  0 & \sigma_2
\end{pmatrix}
\]

\((1|\text{subj}) + (0 + A|\text{subj})\) (but A must be recoded as numeric!)

If we add the possibility of no random effect, we get a hierarchy of 4 levels of richness.

With crossed participant and item random effects, 16 models in a lattice.

And the situation is much worse for more than two conditions.
Which random effects do I use?

- I hope to have convinced you that you don’t want to do model selection by hand

- Seems to me there are two reasonable things to do:
  - Automate model selection
  - Use different techniques that allow model selection to be circumvented

- For the latter, we can use Bayesian inference to marginalize over uncertainty regarding random effects
- Can’t do this in lme4 (yet?), but in the meantime we can use JAGS

Ongoing joint work with Hal Tily: comparing these approaches
I hope to have convinced you that you don’t want to do model selection by hand

Seems to me there are two reasonable things to do:

- Automate model selection
- Use different techniques that allow model selection to be circumvented

For the latter, we can use Bayesian inference to marginalize over uncertainty regarding random effects

Can’t do this in \texttt{lme4} (yet?), but in the meantime we can use \texttt{JAGS}

Ongoing joint work with Hal Tily: comparing these approaches
Which random effects do I use?

- I hope to have convinced you that you don’t want to do model selection by hand
- Seems to me there are two reasonable things to do:
  - Automate model selection
Which random effects do I use?

I hope to have convinced you that you don’t want to do model selection by hand.

Seems to me there are two reasonable things to do:

- Automate model selection
- Use different techniques that allow model selection to be circumvented
Which random effects do I use?

- I hope to have convinced you that you don’t want to do model selection by hand
- Seems to me there are two reasonable things to do:
  - Automate model selection
  - Use different techniques that allow model selection to be circumvented
- For the latter, we can use Bayesian inference to marginalize over uncertainty regarding random effects
Which random effects do I use?

- I hope to have convinced you that you don’t want to do model selection by hand.
- Seems to me there are two reasonable things to do:
  - Automate model selection
  - Use different techniques that allow model selection to be circumvented
- For the latter, we can use Bayesian inference to marginalize over uncertainty regarding random effects.
- Can’t do this in \texttt{lme4} (yet?), but in the meantime we can use JAGS.
Which random effects do I use?

- I hope to have convinced you that you don’t want to do model selection by hand.
- Seems to me there are two reasonable things to do:
  - Automate model selection
  - Use different techniques that allow model selection to be circumvented
- For the latter, we can use Bayesian inference to marginalize over uncertainty regarding random effects.
- Can’t do this in lme4 (yet?), but in the meantime we can use JAGS.
- Ongoing joint work with Hal Tily: comparing these approaches.
Rubric for comparing techniques

Types of datasets to look at:

- Balanced datasets closest to meeting the standards for ANOVAs (=typical controlled psycholinguistic experiments)
- Balanced datasets where there are unbalanced control variables you’d like to incorporate into the analysis to improve signal-to-noise ratio
- Imbalanced datasets that otherwise look like the above two cases;
- Imbalanced datasets with lots of potential predictors and/or controls
Rubric for comparing techniques

Types of datasets to look at:

- Balanced datasets closest to meeting the standards for ANOVAs (=typical controlled psycholinguistic experiments)
- Balanced datasets where there are unbalanced control variables you’d like to incorporate into the analysis to improve signal-to-noise ratio
Rubric for comparing techniques

Types of datasets to look at:

- Balanced datasets closest to meeting the standards for ANOVAs (=typical controlled psycholinguistic experiments)
- Balanced datasets where there are unbalanced control variables you’d like to incorporate into the analysis to improve signal-to-noise ratio
- Imbalanced datasets that otherwise look like the above two cases;
Types of datasets to look at:

- Balanced datasets closest to meeting the standards for ANOVAs (=typical controlled psycholinguistic experiments)
- Balanced datasets where there are unbalanced control variables you’d like to incorporate into the analysis to improve signal-to-noise ratio
- Imbalanced datasets that otherwise look like the above two cases;
- Imbalanced datasets with lots of potential predictors and/or controls
# Some simple simulations (Barr et al., sion)

24-subject, 12- or 24-item balanced “experiment”, between- or within-items, a single 2-level experimental manipulation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>grand-average intercept</td>
<td>$\sim U(-3, 3)$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>grand-average slope</td>
<td>0 (H$_0$ true) or .8 (H$_1$ true)</td>
</tr>
<tr>
<td>$\tau_{00}$</td>
<td>by-subject variance of $S_{0s}$</td>
<td>$\sim U(0, 3)$</td>
</tr>
<tr>
<td>$\tau_{11}$</td>
<td>by-subject variance of $S_{1s}$</td>
<td>$\sim U(0, 3)$</td>
</tr>
<tr>
<td>$\rho_S$</td>
<td>correlation between ($S_{0s}, S_{1s}$) pairs</td>
<td>$\sim U(-.8,.8)$</td>
</tr>
<tr>
<td>$\omega_{00}$</td>
<td>by-item variance of $I_{0i}$</td>
<td>$\sim U(0, 3)$</td>
</tr>
<tr>
<td>$\omega_{11}$</td>
<td>by-item variance of $I_{1i}$</td>
<td>$\sim U(0, 3)$</td>
</tr>
<tr>
<td>$\rho_I$</td>
<td>correlation between ($I_{0i}, I_{1i}$) pairs</td>
<td>$\sim U(-.8,.8)$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>residual error</td>
<td>$\sim U(0, 3)$</td>
</tr>
<tr>
<td>$p_{\text{missing}}$</td>
<td>proportion of missing observations</td>
<td>$\sim U(.00,.05)$</td>
</tr>
</tbody>
</table>
### Super-brief summary of results

**Between-items design:**

<table>
<thead>
<tr>
<th></th>
<th>$N_{\text{items}}$</th>
<th>12</th>
<th>24</th>
<th>12</th>
<th>24</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type I Error</strong></td>
<td>at or near $\alpha = .05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>min-$F'$</td>
<td></td>
<td>.044</td>
<td>.045</td>
<td>.210</td>
<td>.328</td>
<td>.210</td>
</tr>
<tr>
<td>LMEM, Maximal, $\chi^2_{LR}$</td>
<td></td>
<td>.070</td>
<td>.058</td>
<td>.267</td>
<td>.364</td>
<td>.223</td>
</tr>
<tr>
<td>LMEM, No Random Correlations, $\chi^2_{LR}$</td>
<td></td>
<td>.069</td>
<td>.057</td>
<td>.267</td>
<td>.363</td>
<td>.223</td>
</tr>
<tr>
<td>LMEM, No Within-Unit Intercepts, $\chi^2_{LR}$</td>
<td></td>
<td>.081</td>
<td>.065</td>
<td>.288</td>
<td>.380</td>
<td>.223</td>
</tr>
<tr>
<td>LMEM, Maximal, $t$</td>
<td></td>
<td>.086</td>
<td>.065</td>
<td>.300</td>
<td>.382</td>
<td>.222</td>
</tr>
<tr>
<td>LMEM, No Random Correlations, $t$</td>
<td></td>
<td>.086</td>
<td>.064</td>
<td>.300</td>
<td>.382</td>
<td>.223</td>
</tr>
<tr>
<td>LMEM, No Within-Unit Intercepts, $t$</td>
<td></td>
<td>.100</td>
<td>.073</td>
<td>.323</td>
<td>.401</td>
<td>.222</td>
</tr>
<tr>
<td>$F_1 \times F_2$</td>
<td></td>
<td>.063</td>
<td>.077</td>
<td>.252</td>
<td>.403</td>
<td>.224</td>
</tr>
<tr>
<td><strong>Type I Error</strong></td>
<td>far exceeding $\alpha = .05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LMEM, Random Intercepts Only, $\chi^2_{LR}$</td>
<td></td>
<td>.102</td>
<td>.111</td>
<td>.319</td>
<td>.449</td>
<td>.216</td>
</tr>
<tr>
<td>LMEM, Random Intercepts Only, $t$</td>
<td></td>
<td>.128</td>
<td>.124</td>
<td>.360</td>
<td>.472</td>
<td>.217</td>
</tr>
<tr>
<td>LMEM, No Random Correlations, MCMC</td>
<td></td>
<td>.172</td>
<td>.192</td>
<td>.426</td>
<td>.582</td>
<td></td>
</tr>
<tr>
<td>LMEM, Random Intercepts Only, MCMC</td>
<td></td>
<td>.173</td>
<td>.211</td>
<td>.428</td>
<td>.601</td>
<td></td>
</tr>
<tr>
<td>$F_1$</td>
<td></td>
<td>.421</td>
<td>.339</td>
<td>.671</td>
<td>.706</td>
<td>.134</td>
</tr>
</tbody>
</table>

*Performance is sensitive to coding of the predictor (see appendix); simulations use deviation*
Within-items design:

<table>
<thead>
<tr>
<th></th>
<th>Type I Error at or near $\alpha = .05$</th>
<th>Type I Error exceeding $\alpha = .05$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N_{items}$ 12 24 12 24 12 12</td>
<td></td>
</tr>
<tr>
<td>min-$F'$</td>
<td>.027 .031 .327 .512 .327</td>
<td></td>
</tr>
<tr>
<td>LMEM, Maximal, $\chi^2_{LR}$</td>
<td>.059 .056 .460 .610 .433</td>
<td></td>
</tr>
<tr>
<td>LMEM, No Random Correlations, $\chi^2_{LR}$</td>
<td>.059 .056 .461 .610 .432</td>
<td></td>
</tr>
<tr>
<td>LMEM, No Within-Unit Intercepts, $\chi^2_{LR}$</td>
<td>.056 .055 .437 .596 .416</td>
<td></td>
</tr>
<tr>
<td>LMEM, Maximal, $t$</td>
<td>.072 .063 .496 .629 .434</td>
<td></td>
</tr>
<tr>
<td>LMEM, No Random Correlations, $t$</td>
<td>.072 .062 .497 .629 .432</td>
<td></td>
</tr>
<tr>
<td>LMEM, No Within-Unit Intercepts, $*$</td>
<td>.070 .064 .477 .620 .416</td>
<td></td>
</tr>
<tr>
<td>$F_1 \times F_2$</td>
<td>.057 .072 .440 .643 .416</td>
<td></td>
</tr>
<tr>
<td>$F_1$</td>
<td>.176 .139 .640 .724 .345</td>
<td></td>
</tr>
<tr>
<td>LMEM, No Random Correlations, MCMC</td>
<td>.187 .198 .682 .812</td>
<td></td>
</tr>
<tr>
<td>LMEM, Random Intercepts Only, MCMC</td>
<td>.415 .483 .844 .933</td>
<td></td>
</tr>
<tr>
<td>LMEM, Random Intercepts Only, $\chi^2_{LR}$</td>
<td>.440 .498 .853 .935</td>
<td></td>
</tr>
<tr>
<td>LMEM, Random Intercepts Only, $t$</td>
<td>.441 .499 .854 .935 .379</td>
<td></td>
</tr>
</tbody>
</table>

*Performance is sensitive to coding of the predictor (see appendix); simulations use deviation
Wisdom of the crowd: for traditional, perfectly balanced datasets, you lose relatively little with standard ANOVAs controlled studies
Tentative initial conclusions

- **Wisdom of the crowd**: for traditional, perfectly balanced datasets, *you lose relatively little with standard ANOVAs* controlled studies
  - In practice, I think most researchers informally pay attention to \( \max(F_1, F_2) \)
Tentative initial conclusions

- **Wisdom of the crowd**: for traditional, perfectly balanced datasets, *you lose relatively little with standard ANOVAs* controlled studies
  - In practice, I think most researchers informally pay attention to \( \max(F1,F2) \)
  - Clark (1973) warned that this is anticonservative, but much of this anticonservativity seems to be in principle only *(though see our results regarding dependence on sizes of random effects)*
Tentative initial conclusions

- **Wisdom of the crowd**: for traditional, perfectly balanced datasets, *you lose relatively little with standard ANOVAs controlled studies*
  - In practice, I think most researchers informally pay attention to $\max(F_1, F_2)$
  - Clark (1973) warned that this is anticonservative, but much of this anticonservativity seems to be in principle only (though see our results regarding dependence on sizes of random effects)
- Failing to include appropriate random-effects structure in your model is *horribly anti-conservative!*
Tentative initial conclusions

- **Wisdom of the crowd:** for traditional, perfectly balanced datasets, *you lose relatively little with standard ANOVAs* controlled studies
  - In practice, I think most researchers informally pay attention to \( \max(F1,F2) \)
  - Clark (1973) warned that this is anticonservative, but much of this anticonservativity seems to be in principle only (though see our results regarding dependence on sizes of random effects)
- Failing to include appropriate random-effects structure in your model is *horribly anti-conservative!*
- Both model selection and fully Bayesian analysis can address this problem
So why did I care about hierarchical models again?

- If all you ever work with is perfectly balanced factorial designs, then you’re probably OK with ANOVAs
So why did I care about hierarchical models again?

- If all you ever work with is perfectly balanced factorial designs, then you’re probably OK with ANOVAs
  - (On the other hand, when was the last time you tested for sphericity and considered using Hotelling’s $T^2$ test?)
So why did I care about hierarchical models again?

- *If all you ever work with is perfectly balanced factorial designs, then you’re probably OK with ANOVAs*
  - (On the other hand, when was the last time you tested for sphericity and considered using Hotelling’s $T^2$ test?)
- You may be able to eke out a bit more statistical power in these studies from the hierarchical analysis; but *be sure that you are not hallucinating that power!*
So why did I care about hierarchical models again?

- *If all you ever work with is perfectly balanced factorial designs, then you’re probably OK with ANOVAs*
  - (On the other hand, when was the last time you tested for sphericity and considered using Hotelling’s $T^2$ test?)
- You may be able to eke out a bit more statistical power in these studies from the hierarchical analysis; but *be sure that you are not hallucinating that power!*
- Hierarchical models really come into their own, however, with
So why did I care about hierarchical models again?

- *If all you ever work with is perfectly balanced factorial designs, then you’re probably OK with ANOVAs*
  - (On the other hand, when was the last time you tested for sphericity and considered using Hotelling’s $T^2$ test?)
- You may be able to eke out a bit more statistical power in these studies from the hierarchical analysis; but *be sure that you are not hallucinating that power!*
- Hierarchical models really come into their own, however, with
  - *Imbalanced datasets*
So why did I care about hierarchical models again?

- *If all you ever work with is perfectly balanced factorial designs,* then you’re probably OK with ANOVAs
  - (On the other hand, when was the last time you tested for sphericity and considered using Hotelling’s $T^2$ test?)
- You may be able to eke out a bit more statistical power in these studies from the hierarchical analysis; but *be sure that you are not hallucinating that power!*
- Hierarchical models really come into their own, however, with
  - Imbalanced datasets
  - *Appreciably non-linear responses*
So why did I care about hierarchical models again?

- *If all you ever work with is perfectly balanced factorial designs, then you’re probably OK with ANOVAs*
  - (On the other hand, when was the last time you tested for sphericity and considered using Hotelling’s $T^2$ test?)
- You may be able to eke out a bit more statistical power in these studies from the hierarchical analysis; but be sure that you are not hallucinating that power!
- Hierarchical models really come into their own, however, with
  - Imbalanced datasets
  - Appreciably non-linear responses
  - Injection of prior knowledge into (Bayesian) data analysis
So why did I care about hierarchical models again?

- *If all you ever work with is perfectly balanced factorial designs,* then you’re probably OK with ANOVAs
  - (On the other hand, when was the last time you tested for sphericity and considered using Hotelling’s $T^2$ test?)
- You may be able to eke out a bit more statistical power in these studies from the hierarchical analysis; but *be sure that you are not hallucinating that power!*
- Hierarchical models really come into their own, however, with
  - Imbalanced datasets
  - Appreciably non-linear responses
  - Injection of prior knowledge into (Bayesian) data analysis
  - Flexibility in specification of the model structure

More generally, establishing proper standards for use of hierarchical models on linguistic data opens a channel to a rich universe of data-analysis techniques.
So why did I care about hierarchical models again?

- *If all you ever work with is perfectly balanced factorial designs, then you’re probably OK with ANOVAs*
  - (On the other hand, when was the last time you tested for sphericity and considered using Hotelling’s $T^2$ test?)
- You may be able to eke out a bit more statistical power in these studies from the hierarchical analysis; but *be sure that you are not hallucinating that power!*
- Hierarchical models really come into their own, however, with
  - Imbalanced datasets
  - Appreciably non-linear responses
  - Injection of prior knowledge into (Bayesian) data analysis
  - Flexibility in specification of the model structure
- More generally, establishing proper standards for use of hierarchical models on linguistic data opens a channel to a rich universe of data-analysis techniques


