Advanced Probabilistic Modeling in R

Day 2: hypothesis testing and mixed-effects regression

Roger Levy
UC San Diego
LSA 2015 Summer Institute
23 July 2015
Model likelihood
Model likelihood

- You’ve seen likelihood before with Bayes Rule in action
Model likelihood

- You’ve seen LIKELIHOOD before with Bayes Rule in action
- In statistical inference, you want to think of likelihood as a function of model parameters $\theta$, holding data $y$ constant

$$\text{Lik}(y|\theta) = \prod_{i} P(y_i|\theta)$$
Model likelihood

• You’ve seen LIKELIHOOD before with Bayes Rule in action
• In statistical inference, you want to think of likelihood as a function of model parameters $\theta$, holding data $y$ constant

$$\text{Lik}(y|\theta) = \prod_i P(y_i|\theta)$$

• It’s also often written

$$\text{Lik}(y; \theta)$$
Model likelihood

- You’ve seen likelihood before with Bayes Rule in action.
- In statistical inference, you want to think of likelihood as a function of model parameters $\theta$, holding data $y$ constant.

\[ \text{Lik}(y|\theta) = \prod_i P(y_i|\theta) \]

- It’s also often written as:

\[ \text{Lik}(y; \theta) \]

- The maximum likelihood for a dataset of a model class $M$ is the highest possible likelihood of the data under any choice of parameters for $M$.

\[ \max \text{Lik}_M(y) \]
The likelihood ratio test
The likelihood ratio test

- You have two possible probabilistic models for data $y$:
The likelihood ratio test

• You have two possible probabilistic models for data $y$:
  • The $k_1$-parameter NULL-HYPOTHESIS model $H_0$; and
The likelihood ratio test

- You have two possible probabilistic models for data $y$:
  - The $k_1$-parameter NULL-HYPOTHESIS model $H_0$; and
  - The $k_2$-parameter ALTERNATIVE-HYPOTHESIS model $H_A$, inside which $H_0$ is NESTED
The likelihood ratio test

- You have two possible probabilistic models for data $\mathbf{y}$:
  - The $k_1$-parameter NULL-HYPOTHESIS model $H_0$; and
  - The $k_2$-parameter ALTERNATIVE-HYPOTHESIS model $H_A$, inside which $H_0$ is NESTED
    - NESTING means that any distribution over your data representable by $H_0$ can be exactly mimicked by some parameter setting $H_A$ of
The likelihood ratio test

- You have two possible probabilistic models for data $y$:
  - The $k_1$-parameter NULL-HYPOTHESIS model $H_0$; and
  - The $k_2$-parameter ALTERNATIVE-HYPOTHESIS model $H_A$, inside which $H_0$ is NESTED
    - NESTING means that any distribution over your data representable by $H_0$ can be exactly mimicked by some parameter setting $H_A$ of
  - Then if $H_0$ is true, the statistic
    
    $\begin{align*}
    \Delta L &= -2 \log \frac{\max \text{Lik}_{H_0}(y)}{\max \text{Lik}_{H_A}(y)} \\
    &= -2 \log \max \text{Lik}_{H_A}(y) - 2 \log \max \text{Lik}_{H_A}(y)
    \end{align*}$
The likelihood ratio test

- You have two possible probabilistic models for data $y$:
  - The $k_1$-parameter NULL-HYPOTHESIS model $H_0$; and
  - The $k_2$-parameter ALTERNATIVE-HYPOTHESIS model $H_A$, inside which $H_0$ is NESTED
    - NESTING means that any distribution over your data representable by $H_0$ can be exactly mimicked by some parameter setting $H_A$ of
  - Then if $H_0$ is true, the statistic
    \[
    -2 \log \frac{\max \text{Lik}_{H_0}(y)}{\max \text{Lik}_{H_A}(y)}
    \]
    \[
    = 2 \log \max \text{Lik}_{H_A}(y) - 2 \log \max \text{Lik}_{H_A}(y)
    \]
  - is $\chi^2$-distributed with $k_2-k_1$ degrees of freedom.
The $\chi^2$ distribution

Probability density function

Cumulative distribution function

$p(x)$  

$P(X < x)$

1 d.f.  
2 d.f.  
3 d.f.  
6 d.f.

1 d.f.  
2 d.f.  
3 d.f.  
6 d.f.
An example
An example

- Possessors in English can surface as postnominal of genitives, or as prenominal ’s genitives:
  - *the performance of the specialists*
  - *the specialists’ performance*
An example

- Possessors in English can surface as postnominal of genitives, or as prenominal ’s genitives:
  - the performance of the specialists
  - the specialists’ performance

- In 2-conjunct coordinated NPs where both conjuncts have possessors, all four combinations of possessor surfacing are possible:
  - the stock of the company and the performance of the specialists
  - the stock of the company and the specialists’ performance
  - the company’s stock and the performance of the specialists
  - the company’s stock and the specialists’ performance
An example

- Possessors in English can surface as postnominal of genitives, or as prenominal ‘s genitives:
  - *the performance of the specialists*
  - *the specialists’ performance*

- In 2-conjunct coordinated NPs where both conjuncts have possessors, all four combinations of possessor surfacing are possible:
  - *the stock of the company and the performance of the specialists*
  - *the stock of the company and the specialists’ performance*
  - *the company’s stock and the performance of the specialists*
  - *the company’s stock and the specialists’ performance*

- But are like-constituent combinations more **probable**? (Linguistically: is there a **parallelism effect**?)
Contingency table for coordinate-NP possessives

<table>
<thead>
<tr>
<th></th>
<th>Prenom</th>
<th>Postnom</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Left Conjunct</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prenom</td>
<td>77</td>
<td>15</td>
</tr>
<tr>
<td>Postnom</td>
<td>20</td>
<td>39</td>
</tr>
</tbody>
</table>

*Parsed Wall Street Journal section of the Penn Treebank*
### Contingency table for coordinate-NP possessives

<table>
<thead>
<tr>
<th>Left Conjunct</th>
<th>Prenom</th>
<th>Postnom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right conjunct</td>
<td>Prenom</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Postnom</td>
<td>20</td>
</tr>
</tbody>
</table>

- Is there evidence here for a parallelism effect?
Contingency table for coordinate-NP possessives

<table>
<thead>
<tr>
<th>Left Conjunct</th>
<th>Right conjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prenom Postnom</td>
</tr>
<tr>
<td>Prenom</td>
<td>77 15</td>
</tr>
<tr>
<td>Postnom</td>
<td>20 39</td>
</tr>
</tbody>
</table>

- Is there evidence here for a parallelism effect?
- Probabilistically, is it the case that

\[ \text{Conj}_1 \text{ possessor } \perp \text{Conj}_2 \text{ possessor} \mid \{\text{both conjuncts have possessors}\} \]
### Fisher’s exact test? Chi-squared test?

<table>
<thead>
<tr>
<th>Left Conjunct</th>
<th>Prenom</th>
<th>Postnom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenom</td>
<td>77</td>
<td>15</td>
</tr>
<tr>
<td>Postnom</td>
<td>20</td>
<td>39</td>
</tr>
</tbody>
</table>
Fisher’s exact test? Chi-squared test?

<table>
<thead>
<tr>
<th>Left Conjunct</th>
<th>Right conjunct</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prenom</td>
<td>77</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Postnom</td>
<td>20</td>
<td>39</td>
</tr>
</tbody>
</table>

• You may be familiar with:
Fisher’s exact test? Chi-squared test?

- You may be familiar with:
  - Fisher’s exact test for 2x2 contingency tables
Fisher’s exact test? Chi-squared test?

You may be familiar with:
- Fisher’s exact test for 2x2 contingency tables
- The chi-squared test for generalized contingency tables
You may be familiar with:

- Fisher’s exact test for 2x2 contingency tables
- The chi-squared test for generalized contingency tables

Both of these results yield the inference of highly significant non-independence of conjuncts ($p<10^{-9}$)
Fisher’s exact test? Chi-squared test?

You may be familiar with:

- Fisher’s exact test for 2x2 contingency tables
- The chi-squared test for generalized contingency tables

- Both of these results yield the inference of highly significant non-independence of conjuncts ($p<10^{-9}$)
- …but there is a subtle problem with this analysis!
<table>
<thead>
<tr>
<th>Left Conjunct</th>
<th>Right conjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenom</td>
<td>Postnom</td>
</tr>
<tr>
<td></td>
<td>Prenom</td>
</tr>
<tr>
<td></td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Left Conjunct</td>
<td>Prenom</td>
</tr>
<tr>
<td>--------------</td>
<td>--------</td>
</tr>
<tr>
<td>Prenom</td>
<td>77</td>
</tr>
<tr>
<td>Postnom</td>
<td>20</td>
</tr>
</tbody>
</table>

$\text{Conj}_1 \text{ possessor} \perp \text{Conj}_2 \text{ possessor} \mid \{\text{both conjuncts have possessors}\}$
The null-hypothesis model:

\[
\text{Conj}_1 \text{ possessor } \perp \text{ Conj}_2 \text{ possessor } \mid \{\text{both conjuncts have possessors}\}
\]

- **The null-hypothesis model:**

  ![Diagram](image)
Conj₁ possessor ⊥ Conj₂ possessor | \{both conjuncts have possessors\}

- The null-hypothesis model:

- The information used to construct the contingency table:
Conj₁ possessor ⊥ Conj₂ possessor | \{both conjuncts have possessors\}

- The null-hypothesis model:

- The information used to construct the contingency table:

*Breaks d-separation and thus conditional independence!*
An example of “false parallelism”

\[ P(\text{NP}_1 = \text{Post}) = 0.5 \quad \text{(independent of NP}_2) \]
\[ P(\text{NP}_2 = \text{Post}) = 0.5 \quad \text{(independent of NP}_1) \]
\[ P(\text{Pre precedes Post}) = 0.9 \]
An example of “false parallelism”

- Hypothetical data:

\[
\begin{align*}
P(NP_1 = \text{Post}) &= 0.5 \\
P(NP_2 = \text{Post}) &= 0.5 \\
P(\text{Pre precedes Post}) &= 0.9
\end{align*}
\]

<table>
<thead>
<tr>
<th>Right conjunct</th>
<th>Prenom</th>
<th>Postnom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenom</td>
<td>50</td>
<td>90</td>
</tr>
<tr>
<td>Postnom</td>
<td>10</td>
<td>50</td>
</tr>
</tbody>
</table>
An example of “false parallelism”

Hypothetical data:

- $P(\text{NP}_1 = \text{Post}) = 0.5$ (independent of NP$_2$)
- $P(\text{NP}_2 = \text{Post}) = 0.5$ (independent of NP$_1$)
- $P(\text{Pre precedes Post}) = 0.9$

<table>
<thead>
<tr>
<th>Right conjunct</th>
<th>Prenom</th>
<th>Postnom</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Left Conjunct</strong></td>
<td>Prenom</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Postnom</td>
<td>10</td>
</tr>
</tbody>
</table>

- This contingency table passes Fisher’s exact test at $p<0.01$, but the dependency is due only to ordering!
Testing parallelism without ordering

Conj_1 possessor ⊥ Conj_2 possessor | {both conjuncts have possessors}
We can test

\[ \text{Conj}_1 \text{ possessor} \perp \text{Conj}_2 \text{ possessor} \mid \{\text{both conjuncts have possessors}\} \]
Testing parallelism without ordering

• We can test

\[ \text{Conj}_1 \ \text{possessor} \perp \text{Conj}_2 \ \text{possessor} \mid \{\text{both conjuncts have possessors}\} \]

by “unobserving” the conjunct order
Testing parallelism without ordering

- We can test

\[ \text{Conj}_1 \text{ possessor } \perp \text{Conj}_2 \text{ possessor} \mid \{ \text{both conjuncts have possessors} \} \]

by “unobserving” the conjunct order

\[ \text{NP}_1 \, \text{NP}_2 \]

\[ \text{Order} \]

<table>
<thead>
<tr>
<th>Left Conjunct</th>
<th>Prenom</th>
<th>Postnom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenom</td>
<td>77</td>
<td>15</td>
</tr>
<tr>
<td>Postnom</td>
<td>20</td>
<td>39</td>
</tr>
</tbody>
</table>
Testing parallelism without ordering

• We can test

\[ \text{Conj}_1 \text{ possessor} \perp \text{Conj}_2 \text{ possessor} \mid \{ \text{both conjuncts have possessors} \} \]

by “unobserving” the conjunct order

• Instead of counting left/right conjunct status we count # of conjuncts
Testing parallelism without ordering

- We can test
  \[ \text{Conj}_1 \text{ possessor} \perp \text{Conj}_2 \text{ possessor} \mid \{\text{both conjuncts have possessors}\} \]

  by “unobserving” the conjunct order

- Instead of counting left/right conjunct status we count # of conjuncts

<table>
<thead>
<tr>
<th>Left Conjunct</th>
<th>Right conjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenom Postnom</td>
<td>Prenom Postnom</td>
</tr>
<tr>
<td>77 15</td>
<td>20 39</td>
</tr>
</tbody>
</table>

  \[\begin{align*}
    &2 \text{ Prenom}, &1 \text{ Prenom}, &0 \text{ Postnom}, \\
    &0 \text{ Postnom} &1 \text{ Postnom} &2 \text{ Prenom} \\
    &77 &35 &39
  \end{align*}\]
Formulating the likelihood-ratio test
Formulating the likelihood-ratio test

- $H_0$: the probability of a conjunct having a postnominal possessor (given that it has a possessor) is always $p$
Formulating the likelihood-ratio test

- $H_0$: the probability of a conjunct having a postnominal possessor (given that it has a possessor) is always $p$
- $H_1$: possessor realization across conjuncts is dependent even ignoring ordering effects
Formulating the likelihood-ratio test

- $H_0$: the probability of a conjunct having a postnominal possessor (given that it has a possessor) is always $p$
- $H_1$: possessor realization across conjuncts is dependent even ignoring ordering effects

<table>
<thead>
<tr>
<th></th>
<th>2 Prenom,</th>
<th>1 Prenom,</th>
<th>0 Postnom,</th>
<th>0 Postnom</th>
<th>1 Postnom</th>
<th>2 Prenom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>77</td>
<td>35</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$</td>
<td>$(1-p)^2$</td>
<td>$2p(1-p)$</td>
<td>$p^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_A$</td>
<td>$p_1$</td>
<td>$p_2$</td>
<td>$1-p_1-p_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Formulating the likelihood-ratio test

- $H_0$: the probability of a conjunct having a postnominal possessor (given that it has a possessor) is always $p$
- $H_1$: possessor realization across conjuncts is dependent even ignoring ordering effects

<table>
<thead>
<tr>
<th></th>
<th>2 Prenom</th>
<th>1 Prenom</th>
<th>0 Postnom</th>
<th>0 Postnom</th>
<th>1 Postnom</th>
<th>2 Prenom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>77</td>
<td>35</td>
<td>39</td>
<td>35</td>
<td>39</td>
<td>35</td>
</tr>
<tr>
<td>$H_0$</td>
<td>$(1-p)^2$</td>
<td>$2p(1-p)$</td>
<td>$p^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_A$</td>
<td>$p_1$</td>
<td>$p_2$</td>
<td>$1-p_1-p_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$H_0 \quad p = \frac{35 + 2 \times 39}{2 \times 77 + 2 \times 35 + 2 \times 39} = 0.37$$

$$H_1 \quad p_1 = \frac{77}{77 + 35 + 39} = 0.51, \quad p_2 = \frac{35}{77 + 35 + 39} = 0.23$$
<table>
<thead>
<tr>
<th>Observed</th>
<th>77</th>
<th>35</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>$(1-p)^2$</td>
<td>$2p(1-p)$</td>
<td>$p^2$</td>
</tr>
<tr>
<td>$H_A$</td>
<td>$p_1$</td>
<td>$p_2$</td>
<td>$1-p_1p_2$</td>
</tr>
</tbody>
</table>
\[ H_0 \quad p = \frac{35 + 2 \times 39}{2 \times 77 + 2 \times 35 + 2 \times 39} = 0.37 \]

\[ H_1 \quad p_1 = \frac{77}{77 + 35 + 39} = 0.51, \quad p_2 = \frac{35}{77 + 35 + 39} = 0.23 \]
\[ H_0 \quad p = \frac{35 + 2 \times 39}{2 \times 77 + 2 \times 35 + 2 \times 39} = 0.37 \]

\[ H_1 \quad p_1 = \frac{77}{77 + 35 + 39} = 0.51, \quad p_2 = \frac{35}{77 + 35 + 39} = 0.23 \]

\[
\log \text{max Lik}_{H_0}(y) = 77 \log (1 - 0.37)^2 + 35 \log 2 \times 0.37 \times (1 - 0.37) + 39 \log 0.37^2 \\
= -175.4
\]

\[
\log \text{max Lik}_{H_A}(y) = 77 \log 0.51 + 35 \log 0.23 + 39 \log (1 - 0.51 - 0.23) \\
= -155.8
\]
\[
\begin{array}{|c|c|c|c|}
\hline
\text{2 Prenom, 1 Prenom, 0 Postnom, 0 Postnom, 1 Postnom, 2 Prenom} & 77 & 35 & 39 \\
\hline
\text{\textit{H}}_0 & (1-p)^2 & 2p(1-p) & p^2 \\
\text{\textit{H}}_A & p_1 & p_2 & 1-p_1p_2 \\
\hline
\end{array}
\]

\[
H_0 \quad p = \frac{35 + 2 \times 39}{2 \times 77 + 2 \times 35 + 2 \times 39} = 0.37
\]

\[
H_1 \quad p_1 = \frac{77}{77 + 35 + 39} = 0.51, \quad p_2 = \frac{35}{77 + 35 + 39} = 0.23
\]

\[
\log \max \text{Lik}_{H_0}(y) = 77 \log(1 - 0.37)^2 + 35 \log 2 \times 0.37 \times (1 - 0.37) + 39 \log 0.37^2 \\
= -175.4
\]

\[
\log \max \text{Lik}_{H_A}(y) = 77 \log 0.51 + 35 \log 0.23 + 39 \log(1 - 0.51 - 0.23) \\
= -155.8
\]

\[
-2 \log \frac{\max \text{Lik}_{H_0}(y)}{\max \text{Lik}_{H_A}(y)} = 39.2
\]
\[
\begin{array}{c|ccc}
\text{2 Prenom, 1 Prenom, 0 Postnom,} & 77 & 35 & 39 \\
\text{0 Postnom 1 Postnom 2 Prenom} & (1-p)^2 & 2p(1-p) & p^2 \\
\hline
H_0 & p = \frac{35 + 2 \times 39}{2 \times 77 + 2 \times 35 + 2 \times 39} = 0.37 \\
H_A & p_1 = \frac{77}{77 + 35 + 39} = 0.51, \quad p_2 = \frac{35}{77 + 35 + 39} = 0.23 \\
\end{array}
\]

\[
\log \max \text{Lik}_{H_0}(y) = 77 \log (1 - 0.37)^2 + 35 \log 2 \times 0.37 \times (1 - 0.37) + 39 \log 0.37^2 \\
= -175.4 \\
\log \max \text{Lik}_{H_A}(y) = 77 \log 0.51 + 35 \log 0.23 + 39 \log (1 - 0.51 - 0.23) \\
= -155.8 \\
\]

\[
-2 \log \frac{\max \text{Lik}_{H_0}(y)}{\max \text{Lik}_{H_A}(y)} = 39.2 \quad \text{Chi-squared test with 1 d.f. shows that this is highly significant — p << 0.001}
\]
Bayesian hypothesis testing
Bayesian hypothesis testing

- For a collection of hypotheses \( \{H_i\} \), we simply use Bayes’ Rule to compute posterior hypothesis probabilities:

\[
P(H_i | y) = \frac{P(y | H_i) P(H_i)}{P(y)}
\]
Bayesian hypothesis testing

• For a collection of hypotheses \( \{H_i\} \), we simply use Bayes’ Rule to compute posterior hypothesis probabilities:

\[
P(H_i | y) = \frac{P(y | H_i) P(H_i)}{P(y)}
\]

• The BAYES FACTOR quantifies the evidence presented by data \( y \) in favor of one hypothesis over another

\[
\left( \frac{P(H | y)}{P(H' | y)} \right) = \frac{P(y | H)}{P(y | H')} \cdot \frac{P(H)}{P(H')}
\]
Bayesian hypothesis testing

• For a collection of hypotheses \( \{H_i\} \), we simply use Bayes’ Rule to compute posterior hypothesis probabilities:

\[
P(H_i|y) = \frac{P(y|H_i)P(H_i)}{P(y)}
\]

• The BAYES FACTOR quantifies the evidence presented by data \( y \) in favor of one hypothesis over another:

\[
\frac{P(H|y)}{P(H'|y)} = \frac{P(y|H)}{P(y|H')} \frac{P(H)}{P(H')}
\]

\[\text{Bayes Factor}\]
Bayesian hypothesis testing

For a collection of hypotheses \( \{H_i\} \), we simply use Bayes’ Rule to compute posterior hypothesis probabilities:

\[
P(H_i | y) = \frac{P(y | H_i) P(H_i)}{P(y)}
\]

The BAYES FACTOR quantifies the evidence presented by data \( y \) in favor of one hypothesis over another

\[
\frac{P(H | y)}{P(H' | y)} = \frac{P(y | H)}{P(y | H')} \frac{P(H)}{P(H')}
\]

Bayes Factor

Note that Bayes Factor eliminates the effect of the prior (which isn’t taking into account the data)
Sequential dependencies in language

da ta da ta da da da da ta ta ta ta ta ta da da da da
Sequential dependencies in language

- Consider the syllable sequence:
  da ta da ta da da da da ta ta ta ta ta ta ta ta da da da da
Sequential dependencies in language

- Consider the syllable sequence:
  da ta da ta da da da da ta ta ta da ta ta da da da da
- Does the preceding syllable affect choice of the next?
Sequential dependencies in language

• Consider the syllable sequence:
  da ta da ta da da da da da ta ta ta ta ta ta da da da da da

• Does the preceding syllable affect choice of the next?

\[ H_1: \text{independence/unigram model} \]
\[ P(da) = \pi \]
\[ P(da|\emptyset) = \pi_\emptyset \]
\[ P(da|da) = \pi_{da} \]
\[ P(da|ta) = \pi_{ta} \]

\[ H_2: \text{non-independence/bigram model} \]
Sequential dependencies in language

• Consider the syllable sequence:

• Does the preceding syllable affect choice of the next?

\[
\begin{align*}
H_1 &: \text{ independence/unigram model} \\
H_2 &: \text{ non-independence/bigram model}
\end{align*}
\]

\[
\begin{align*}
P(da) &= \pi \\
P(da|\emptyset) &= \pi_\emptyset \\
P(da|da) &= \pi_{da} \\
P(da|ta) &= \pi_{ta}
\end{align*}
\]

• We want to compute Bayes Factor:

\[
\frac{P(y|H_1)}{P(y|H_2)}
\]
Sequential dependencies in language

• Consider the syllable sequence:

• Does the preceding syllable affect choice of the next?

\[ H_1: \text{independence/unigram model} \quad \quad H_2: \text{non-independence/bigram model} \]

\[
P(da) = \pi \\
P(da|\emptyset) = \pi_\emptyset \\
P(da|da) = \pi_{da} \\
P(da|ta) = \pi_{ta}
\]

• We want to compute Bayes Factor:

\[
\frac{P(y|H_1)}{P(y|H_2)}
\]

• To do this we must marginalize over each models’s parameters \( \theta \):

\[
P(y|H) = \int_\theta P(y|\theta, H) \, d\theta
\]
### Marginal data probability

<table>
<thead>
<tr>
<th>$H_1$: independence/unigram model</th>
<th>$P(da) = \pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_2$: non-independence/bigram model</td>
<td>$P(da</td>
</tr>
<tr>
<td></td>
<td>$P(da</td>
</tr>
<tr>
<td></td>
<td>$P(da</td>
</tr>
</tbody>
</table>
Marginal data probability

\[ H_1: \text{independence/unigram model} \]
\[ P(da) = \pi \]
\[ P(da|\emptyset) = \pi_{\emptyset} \]
\[ P(da|da) = \pi_{da} \]
\[ P(da|ta) = \pi_{ta} \]

\[ H_2: \text{non-independence/bigram model} \]

- Put 1,1 Beta priors (uniform) on all params in both models
Marginal data probability

- **$H_1$: independence/unigram model**
  - $P(da) = \pi$

- **$H_2$: non-independence/bigram model**
  - $P(da|\emptyset) = \pi_\emptyset$
  - $P(da|da) = \pi_{da}$
  - $P(da|ta) = \pi_{ta}$

- Put 1,1 Beta priors (uniform) on all params in both models

- **Trick:**

  \[
p(\pi|\pi \text{ is beta-}(\alpha_1, \alpha_2) \text{ distributed}) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1}
  \]
Marginal data probability

\[
\begin{align*}
H_1: \text{independence/unigram model} & \quad P(da) = \pi \\
H_2: \text{non-independence/bigram model} & \quad P(da|\emptyset) = \pi_\emptyset \\
& \quad P(da|da) = \pi_{da} \\
& \quad P(da|ta) = \pi_{ta}
\end{align*}
\]

- Put 1,1 Beta priors (uniform) on all params in both models

- **Trick:**
  \[
p(\pi|\pi \text{ is beta-}(\alpha_1, \alpha_2) \text{ distributed}) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1}
\]

SO
\[
\int_0^1 \pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1} d\pi = B(\alpha_1, \alpha_2)
\]
Marginal data probability

<table>
<thead>
<tr>
<th>$H_1$: independence/unigram model</th>
<th>$P(da) = \pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_2$: non-independence/bigram model</td>
<td>$P(da</td>
</tr>
<tr>
<td></td>
<td>$P(da</td>
</tr>
<tr>
<td></td>
<td>$P(da</td>
</tr>
</tbody>
</table>

- Put 1,1 Beta priors (uniform) on all params in both models

- **Trick:**

  $$p(\pi|\pi \text{ is beta-}(\alpha_1, \alpha_2) \text{ distributed}) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}$$

  so

  $$\int_0^1 \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1} d\pi = B(\alpha_1, \alpha_2)$$

  due to properness.
Marginal data probability

<table>
<thead>
<tr>
<th>$H_1$: independence/unigram model</th>
<th>$P(da) = \pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_2$: non-independence/bigram model</td>
<td>$P(da</td>
</tr>
<tr>
<td></td>
<td>$P(da</td>
</tr>
<tr>
<td></td>
<td>$P(da</td>
</tr>
</tbody>
</table>

- Put 1,1 Beta priors (uniform) on all params in both models
- **Trick:**

$$p(\pi|\pi \text{ is beta-}(\alpha_1, \alpha_2) \text{ distributed}) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1} (1-\pi)^{\alpha_2-1}$$

so

$$\int_0^1 \pi^{\alpha_1-1} (1-\pi)^{\alpha_2-1} d\pi = B(\alpha_1, \alpha_2)$$

due to properness.


12 da’s, 9 ta’s
Marginal data probability

\[ H_1: \text{independence/unigram model} \quad P(da) = \pi \]
\[ H_2: \text{non-independence/bigram model} \quad P(da|\emptyset) = \pi_\emptyset \]
\[ P(da|da) = \pi_{da} \]
\[ P(da|ta) = \pi_{ta} \]

- Put 1,1 Beta priors (uniform) on all params in both models

- **Trick:**

\[
p(\pi|\pi \text{ is beta-}(\alpha_1, \alpha_2) \text{ distributed}) = \frac{1}{B(\alpha_1,\alpha_2)} \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}
\]

so

\[
\int_{0}^{1} \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1} d\pi = B(\alpha_1,\alpha_2)
\]
due to properness.

da ta da ta da da da ta ta ta da ta ta ta da da da da

12 da’s, 9 ta’s

\[
P(y|H_1) = \int_{0}^{1} \pi^{12}(1-\pi)^9 d\pi = B(13,10)
\]

\[= 1.55 \times 10^7 \]
Marginal data probability of bigram model

\[ H_1: \text{independence/unigram model} \]
\[ P(da) = \pi \]
\[ P(da|\emptyset) = \pi_{\emptyset} \]
\[ P(da|da) = \pi_{da} \]
\[ P(da|ta) = \pi_{ta} \]

\[ H_2: \text{non-independence/bigram model} \]

\[ da \, ta \, da \, ta \, da \, da \, da \, da \, da \, ta \, ta \, da \, ta \, ta \, ta \, da \, da \, da \, da \, da \]
Marginal data probability of bigram model

\[ H_1: \text{independence/unigram model} \quad P(da) = \pi \]
\[ H_2: \text{non-independence/bigram model} \quad P(da|\emptyset) = \pi_\emptyset \]
\[ P(da|da) = \pi_{da} \]
\[ P(da|ta) = \pi_{ta} \]

\[ \begin{array}{c|cc}
\text{Context} & \text{da} & \text{ta} \\
\hline
\emptyset & 1 & 0 \\
da & 7 & 4 \\
ta & 4 & 5 \\
\end{array} \]
Marginal data probability of bigram model

\[ H_1: \text{independence/unigram model} \]
\[ P(da) = \pi \]
\[ H_2: \text{non-independence/bigram model} \]
\[ P(da) = \pi_\emptyset \]
\[ P(da|da) = \pi_{da} \]
\[ P(da|ta) = \pi_{ta} \]

\[
d a \ t a \ d a \ d a \ t a \ d a \ d a \ d a \ t a \ d a \ t a \ d a \ t a \ d a \ d a \ d a
\]

<table>
<thead>
<tr>
<th>Context</th>
<th>da</th>
<th>ta</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\emptyset)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>da</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>ta</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

\[
P(y|H_2) = \int_0^1 \pi_{\emptyset}^1 d\pi_{\emptyset} \int_0^1 \pi_{da}^7 (1 - \pi_{da})^4 d\pi_{da} \int_0^1 \pi_{ta}^4 (1 - \pi_{ta})^5 d\pi_{ta}
\]
\[
= B(2, 1) B(8, 5) B(5, 6)
\]
\[
= 1 \times 10^7
\]
Marginal data probability of bigram model

\( H_1: \) independence/unigram model

\[
\begin{align*}
P(da) &= \pi \\
P(da|\emptyset) &= \pi_\emptyset \\
P(da|da) &= \pi_{da} \\
P(da|ta) &= \pi_{ta}
\end{align*}
\]

\( H_2: \) non-independence/bigram model

\[
\begin{align*}
\text{Outcome} & \quad \text{da} \quad \text{ta} \\
\text{Context} & \quad \emptyset \quad \text{da} \quad \text{ta} \\
\emptyset & \quad 1 \quad 0 \\
da & \quad 7 \quad 4 \\
ta & \quad 4 \quad 5 \\
\end{align*}
\]

\[
P(y|H_2) = \int_0^1 \pi_\emptyset d\pi_\emptyset \int_0^1 \pi_{da} (1 - \pi_{da})^4 d\pi_{da} \int_0^1 \pi_{ta} (1 - \pi_{ta})^5 d\pi_{ta}
\]

\[
= B(2, 1)B(8, 5)B(5, 6)
\]

\[
= 1 \times 10^7
\]

\[
\text{Bayes Factor:} \quad \frac{P(y|H_1)}{P(y|H_2)} = \frac{1.55 \times 10^{-7}}{1 \times 10^{-7}}
\]

\[
= 1.55
\]
A guide to interpreting Bayes Factors

\[
\frac{P(y|H_1)}{P(y|H_2)} \quad \text{Strength of evidence}
\]

<table>
<thead>
<tr>
<th>Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–3</td>
<td>barely worth mentioning</td>
</tr>
<tr>
<td>3–10</td>
<td>substantial</td>
</tr>
<tr>
<td>10–30</td>
<td>strong</td>
</tr>
<tr>
<td>30–100</td>
<td>very strong</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>decisive</td>
</tr>
</tbody>
</table>

(Jeffreys, 1961)
Background & gameplan

- We use generalized linear models (GLMs) to determine the structure of influence of predictors on a response
Background & gameplan

- We use generalized linear models (GLMs) to determine the structure of influence of predictors on a response
- Now we’ll bring together the GLM with the idea of hierarchical models
We use generalized linear models (GLMs) to determine the structure of influence of predictors on a response.

Now we’ll bring together the GLM with the idea of hierarchical models.

This model will accommodate inter-cluster variability in both “average” cluster response (variability in the intercept parameter) and cluster-specific sensitivity to predictor variables (variability in slope/weight parameters).
We use generalized linear models (GLMs) to determine the structure of influence of predictors on a response.

Now we’ll bring together the GLM with the idea of hierarchical models.

This model will accommodate inter-cluster variability in both “average” cluster response (variability in the intercept parameter) and cluster-specific sensitivity to predictor variables (variability in slope/weight parameters).

Result: hierarchical, or mixed-effect, generalized linear model.
Goal: model the effects of predictors (independent variables) $X$ on a response (dependent variable) $Y$. 
Review: Generalized linear models I

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

The picture:
Review: 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) $X$ 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 \text{linear predictor} \( \eta \);
2. \( \eta \) is a linear combination of the \( \{X_i\} \):

\[
\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \quad \text{(linear predictor)}
\]
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 \( \{X_i\} \):
   \[
   \eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N \tag{linear predictor}
   \]
3. \( \eta \) determines the predicted mean \( \mu \) of \( Y \)
   \[
   \eta = l(\mu) \tag{link function}
   \]
4. There is some noise distribution of \( Y \) around the predicted mean \( \mu \) of \( Y \):
   \[
   P(Y = y; \mu)
   \]
GLMs III

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

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

- The predicted mean is just the linear predictor:

\[
\eta = l(\mu) = \mu
\]
GLMs III

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 (\(\equiv\)Gaussian) distributed around 0 with standard deviation \(\sigma\):

\[ \epsilon \sim N(0, \sigma) \]
GLMs III

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 \( \sigma \):

\[ \epsilon \sim N(0, \sigma) \]

- This gives us the traditional linear regression equation:

\[ Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \sim N(0, \sigma) \]
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:
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 $Y$
  
  choose $\{\beta_i\}$ and $\sigma$ that make the likelihood $P(Y|\{\beta_i\}, \sigma)$ as large as possible
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 $Y$
  
  $\text{choose } \{\beta_i\} \text{ and } \sigma \text{ that make the likelihood } P(Y|\{\beta_i\},\sigma) \text{ as large as possible}$

  - 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

$$P(\{\beta_i\}, \sigma|Y) = \frac{\underbrace{P(Y|\{\beta_i\}, \sigma)}_{\text{Likelihood}} \cdot \underbrace{P(\{\beta_i\}, \sigma)}_{\text{Prior}}}{P(Y)}$$
GLMs V: a simple example

► You are studying non-word RTs in a lexical-decision task
GLMs V: a simple example

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

Word or non-word?
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  
  tpozt \textit{Word or non-word?}
  
  houze \textit{Word or non-word?}
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  
  tpozt \textit{Word or non-word?}
  houze \textit{Word or non-word?}

- Non-words with different \textit{neighborhood densities} should have different average RT \((= \text{number of neighbors of edit-distance 1})\)
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  - tpozt \textit{Word or non-word?}
  - houze \textit{Word or non-word?}

- Non-words with different \textit{neighborhood densities} should have different average RT \((=\text{number of neighbors of edit-distance 1})\)

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

- You are studying non-word RTs in a lexical-decision task
  
  \[
  \text{tpozt} \quad \text{Word or non-word?} \\
  \text{houze} \quad \text{Word or non-word?}
  \]

- 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

  \[
  \begin{align*}
  RT_i &= \alpha + \beta x_i + \epsilon_i \\
  \epsilon_i &\sim N(0, \sigma)
  \end{align*}
  \]
GLMs V: a simple example

- You are studying non-word RTs in a lexical-decision task
  
  tpozt  \textit{Word or non-word?}
  
  houze  \textit{Word or non-word?}

- Non-words with different \textit{neighborhood densities}\footnote{should have different average RT \((\equiv \text{number of neighbors of edit-distance 1})\)} should have different average RT

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

- 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 \(\alpha, \beta, \text{ and } \sigma\)
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.)*

- 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?
GLMs VI

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

<table>
<thead>
<tr>
<th>Few neighbors</th>
<th>Many neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>gaty peme rixy</td>
<td>lish pait yine</td>
</tr>
</tbody>
</table>
We’ll use length-4 nonword data from (Bicknell et al., 2010) (thanks!), such as:

\[\begin{align*}
\text{Few neighbors} & \quad \text{Many neighbors} \\
gaty & \quad \text{peme} & \quad \text{rixy} & \quad \text{lish} & \quad \text{pait} & \quad \text{yine}
\end{align*}\]

There’s a wide range of neighborhood density:
GLMs VII: maximum-likelihood model fitting

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.
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.)
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)
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)
```

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
```
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 \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)

[...]

| 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
```
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
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)

[...]

\( \hat{\alpha} \) 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
\( \hat{\beta} \)
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)

[...]  
\[ \hat{\alpha} \]
\[
\begin{array}{lrrrr}
\text{(Intercept)} & 382.997 & 26.837 & 14.271 & <2e-16 \text{ ***} \\
\text{neighbors} & 4.828 & 6.553 & 0.737 & 0.466
\end{array}
\]

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

[1] 107.2248
```

\( \hat{\sigma} \)

\( \hat{\beta} \)
### GLMs: maximum-likelihood fitting VIII

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>383.00</td>
</tr>
<tr>
<td>neighbors</td>
<td>4.83</td>
</tr>
<tr>
<td>$\hat{\sigma}$</td>
<td>107.22</td>
</tr>
</tbody>
</table>
GLMs: maximum-likelihood fitting VIII

<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
GLMs: maximum-likelihood fitting VIII

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>383.00</td>
</tr>
<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
GLMs IX: Bayesian model fitting

Alternative to maximum-likelihood: Bayesian model fitting
GLMs IX: Bayesian model fitting

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

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

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

- \(p_{\text{MCMC}}\) (Baayen et al., 2008) is 1 minus the largest possible symmetric confidence interval wholly on one side of 0
Mixed-effects/hierarchical GLMs

The non-hierarchical GLM picture:

\[ \theta \]
\[ x_1 \]
\[ y_1 \]
\[ x_2 \]
\[ y_2 \]
\[ \cdots \]
\[ x_n \]
\[ y_n \]

Predictors

Response

Model parameters
Mixed-effects/hierarchical GLMs
Mixed-effects/hierarchical GLMs
Mixed-effects/hierarchical GLMs

Cluster-specific parameters ("random effects")

\[ \theta \]

\[ \Sigma_b \]

\[ b_1 \]

\[ x_{11} \quad \ldots \quad x_{1n_1} \]

\[ y_{11} \quad \ldots \quad y_{1n_1} \]

\[ b_2 \]

\[ x_{21} \quad \ldots \quad x_{2n_2} \]

\[ y_{21} \quad \ldots \quad y_{2n_2} \]

\[ \ldots \]

\[ b_M \]

\[ x_{M1} \quad \ldots \quad x_{Mn_M} \]

\[ y_{M1} \quad \ldots \quad y_{Mn_M} \]
Mixed-effects/hierarchical GLMs

Cluster-specific parameters ("random effects")

Shared parameters ("fixed effects")

\[ 
\begin{align*}
  \theta & \rightarrow \Sigma_b \\
  \Sigma_b & \rightarrow b_1 \\
  b_1 & \rightarrow x_{11} \rightarrow y_{11} \\
  \vdots & \vdots \vdots \vdots \vdots \vdots \vdots \\
  b_M & \rightarrow x_{M1} \rightarrow y_{M1} \\
  \vdots & \vdots \vdots \vdots \vdots \vdots \vdots \\
  \end{align*} 
\]
Mixed-effects/hierarchical GLMs

Cluster-specific parameters ("random effects")

Shared parameters ("fixed effects")

Parameters governing inter-cluster variability
Multi-level Models IX

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
Multi-level Models IX

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

- Additionally, different participants in your study may also have:
  - different overall decision speeds
  - differing sensitivity to neighborhood density
Multi-level Models IX

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

- 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
Multi-level Models IX: Model construction

- 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

$$RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij} \sim N(0,\sigma_b) \quad \text{Noise} \sim N(0,\sigma_\epsilon)$$
Multi-level Models IX: Model construction

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

$$RT_{ij} = \alpha + \beta x_{ij} + b_i + \varepsilon_{ij}$$

In R, we’d write this relationship as

$$RT \sim 1 + x + (1 \mid \text{participant})$$
Multi-level Models IX: Model construction

- 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

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

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

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

- In \( \mathbb{R} \), we’d write this relationship as

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

- 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

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

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.
Multi-level Models X: simulating data

<table>
<thead>
<tr>
<th>$RT_{ij} = \alpha + \beta x_{ij} + b_i + \epsilon_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sim N(0, \sigma_b)$</td>
</tr>
<tr>
<td>Noise $\sim N(0, \sigma_e)$</td>
</tr>
</tbody>
</table>

- 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 X: simulating data

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

- 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) #
```
Participant-level clustering is easily visible
Multi-level Models XI: simulating data

- Participant-level clustering is easily visible
Participant-level clustering is easily visible

This reflects the fact that inter-participant variation (125ms) is larger than inter-trial variation (40ms)
Multi-level Models XI: simulating data

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

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

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

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

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

- 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

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

- 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, . . . \}?
Statistical inference with multi-level models

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

\( \sim N(0, \sigma_b) \) 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, ...\}?
- 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

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

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

\[
\begin{align*}
m & \leftarrow \text{lmer}(\text{time} \sim \text{neighbors.centered} + \\
& \quad \quad \quad \quad \quad (1 \mid \text{participant}), \text{dat}, \text{REML=F}) \\
\end{align*}
\]

\[
\begin{align*}
\text{print}(m, \text{corr=F})
\end{align*}
\]

[...]

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></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>583.787</td>
<td>11.082</td>
<td>52.68</td>
<td></td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.986</td>
<td>1.278</td>
<td>7.03</td>
<td></td>
</tr>
</tbody>
</table>
Fitting a multi-level model using maximum likelihood

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

\[
> m <- \text{lmer}\text{()}\text{time ~ neighbors.centered + (1 | participant), dat, REML=F)}
> \text{print}\text{(m, corr=F)}
\]

[...]
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>( \hat{\alpha} )</td>
<td>583.787</td>
<td>11.082</td>
<td>52.68</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.986</td>
<td>1.278</td>
<td>7.03</td>
</tr>
</tbody>
</table>
Fitting a multi-level model using maximum likelihood

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

\[
\begin{align*}
> m <- \text{lmer(time} \sim \text{neighbors.centered} + \\
\quad (1 \mid \text{participant}), \text{dat}, \text{REML=F})
\end{align*}
\]

\[
> \text{print(m, corr=F)}
\]

[...]

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>(\hat{\alpha})</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>
<td></td>
</tr>
</tbody>
</table>
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:

<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:

|             | 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

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

\[
\begin{align*}
> m & \leftarrow \text{lmer(time} \sim \text{neighbors.centered} + (1 \mid \text{participant}), \text{dat}, \text{REML}=\text{F}) \\
> \text{print}(m, \text{corr}=\text{F})
\end{align*}
\]

[...]  
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></td>
<td>Residual</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>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>583.787</td>
<td>11.082</td>
</tr>
<tr>
<td>neighbors.centered</td>
<td>8.986</td>
<td>1.278</td>
</tr>
</tbody>
</table>
## 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>
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:
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
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
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
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

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

- What about the participants’ idiosyncracies themselves—the \( b_i \)?
- We can also draw inferences about these—once again, a common estimate of them is known as the **BLUP**
Inferences about cluster-level parameters

\[ RT_{ij} = \alpha + \beta x_{ij} + \hat{b}_i + \hat{\epsilon}_{ij} \sim N(0,\sigma_b) \text{ Noise}\sim N(0,\sigma_\epsilon) \]

- What about the participants’ idiosyncracies themselves—the \( b_i \)?
- We can also draw inferences about these—once again, a common estimate of them is known as the BLUP
- 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

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

- What about the participants’ idiosyncracies themselves—the \( b_i \)?
- We can also draw inferences about these—once again, a common estimate of them is known as the BLUP
- To understand these: committing to fixed-effect and random-effect parameter estimates determines a conditional probability distribution on participant-specific effects:

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

- 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 solid line is a guess at their distribution.
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.
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*.
Inference about cluster-level parameters III

- Participants may also have idiosyncratic sensitivities to *neighborhood density*
- Incorporate by adding cluster-level slopes into the model:

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

\[
\sim N(0, \Sigma_b) \quad \text{Noise} \sim N(0, \sigma_{\epsilon})
\]
Inference about cluster-level parameters III

- Participants may also have idiosyncratic sensitivities to *neighborhood density*
- Incorporate by adding cluster-level slopes into the model:

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

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

\[\text{Noise} \sim N(0, \sigma_\varepsilon)\]

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

\[RT \sim 1 + x + (1 + x \mid \text{participant})\]
Inference about cluster-level parameters III

- Participants may also have idiosyncratic sensitivities to *neighborhood density*
- Incorporate by adding cluster-level slopes into the model:

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

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

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

> `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>
Inference about cluster-level parameters III

- Participants may also have idiosyncratic sensitivities to *neighborhood density*
- Incorporate by adding cluster-level slopes into the model:

\[
RT_{ij} = \alpha + \beta x_{ij} + b_1i + b_2ix_{ij} + \varepsilon_{ij}
\]

- 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)
```

[...]

**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>

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

\[ RT_{ij} = \alpha + \beta x_{ij} + b_{1i} + b_{2i} x_{ij} + \epsilon_{ij} \]
Inferences about cluster-level parameters IV

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

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

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

- 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>neighbors.centered</td>
<td></td>
<td>19.421</td>
<td>4.4069</td>
<td>-0.307</td>
</tr>
</tbody>
</table>


Inferences about cluster-level parameters IV

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

\[
RT_{ij} = \alpha + \beta x_{ij} + b_1 i + b_2 i x_{ij} + \epsilon_{ij} \sim N(0, \Sigma_b) \quad \text{Noise} \sim N(0, \sigma_\epsilon)
\]

- We’ve said that participant-specific idiosyncrasies 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 \texttt{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

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

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

\[
\Sigma_b = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 4 \end{pmatrix}
\]
Inference about cluster-level parameters VI

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

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

- This ellipse contains a certain proportion (here & conventionally, 95%) of the probability mass for the distribution in question
Inference about cluster-level parameters VII
Inference about cluster-level parameters VII

Participants

Intercept

neighbors

0  5  10  15

1  400  500  600  700  800
Inference about cluster-level parameters VII

Participants

Intercept

neighbors

Participants
Inference about cluster-level parameters VII

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

\[
\hat{\Sigma}_b = \begin{pmatrix} 70.20 & -0.3097 \\ -0.3097 & 4.41 \end{pmatrix}
\]
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
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
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
What random effects structure to use in drawing inferences about fixed effects?

- We looked at models that both *included* and *didn’t include* random slopes for the #-neighbors effect in this dataset.
What random effects structure to use in drawing inferences about fixed effects?

- We looked at models that both *included* and *didn’t include* random slopes for the #-neighbors effect in this dataset
- Which one is the right model?
What random effects structure to use in drawing inferences about fixed effects?

- We looked at models that both included and didn’t include random slopes for the #-neighbors effect in this dataset
- Which one is the right model?
- This is a very general problem with no one solution, but I’ll describe what I think are good answers for the situation where one ultimately wants to make inferences about the importance of a “fixed-effect” (shared) model parameter
What random effects structure to use in drawing inferences about fixed effects?

- We looked at models that both included and didn’t include random slopes for the #-neighbors effect in this dataset.
- Which one is the right model?
- This is a very general problem with no one solution, but I’ll describe what I think are good answers for the situation where one ultimately wants to make inferences about the importance of a “fixed-effect” (shared) model parameter.
- There are two situations:
What random effects structure to use in drawing inferences about fixed effects?

- We looked at models that both included and didn’t include random slopes for the #-neighbors effect in this dataset
- Which one is the right model?
- This is a very general problem with no one solution, but I’ll describe what I think are good answers for the situation where one ultimately wants to make inferences about the importance of a “fixed-effect” (shared) model parameter
- There are two situations:
  1. When the (average) value that a fixed effect takes varies across clusters
What random effects structure to use in drawing inferences about fixed effects?

- We looked at models that both included and didn’t include random slopes for the #-neighbors effect in this dataset.
- Which one is the right model?
- This is a very general problem with no one solution, but I’ll describe what I think are good answers for the situation where one ultimately wants to make inferences about the importance of a “fixed-effect” (shared) model parameter.
- There are two situations:
  1. When the (average) value that a fixed effect takes varies across clusters.
  2. When the value that a fixed effect takes varies within some or all clusters.
Predictors varying between clusters

Hypothetical relationship observed for three words:
Predictors varying between clusters

If we were to ignore the potential cross-cluster variability here, it would look like we have good evidence for a word frequency effect.
Predictors varying between clusters

- If we were to ignore the potential cross-cluster variability here, it would look like we have good evidence for a word frequency effect.
- But we have measurements for only three words!
Predictors varying between clusters

▶ If we were to ignore the potential cross-cluster variability here, it would look like we have good evidence for a word frequency effect
▶ But we have measurements for only three words!
▶ Suppose that there were no effect of word frequency, but words themselves varied idiosyncratically in their ease of recognition
Predictors varying between clusters

- If we were to ignore the potential cross-cluster variability here, it would look like we have good evidence for a word frequency effect.
- But we have measurements for only three words!
- Suppose that there were no effect of word frequency, but words themselves varied idiosyncratically in their ease of recognition.
- But the probability that the observed means would have this monotonicity would still be $\frac{1}{6}$.
Predictors varying between clusters

- If we were to ignore the potential cross-cluster variability here, it would look like we have good evidence for a word frequency effect.
- But we have measurements for only three words!
- Suppose that there were no effect of word frequency, but words themselves varied idiosyncratically in their ease of recognition.
- But the probability that the observed means would have this monotonicity would still be $\frac{1}{6}$.
- To address this issue we need a random intercept.
Predictors varying between clusters

- If we were to ignore the potential cross-cluster variability here, it would look like we have good evidence for a word frequency effect.
- But we have measurements for only three words!
- Suppose that there were no effect of word frequency, but words themselves varied idiosyncratically in their ease of recognition.
- But the probability that the observed means would have this monotonicity would still be $\frac{1}{6}$.
- To address this issue we need a random intercept.
- Our model will wind up answering the question of whether there is a systematic trend across words for frequency sensitivity, *above and beyond idiosyncratic variation among words*. 
Predictors varying within clusters

Hypothetical frequency-based responses for five different individual participants:
It looks like we have good evidence for frequency-sensitivity of the response
Predictors varying within clusters II

- Classic question: *above and beyond idiosyncratic sensitivities of different individuals to context-driven predictability*, are predictable words in general named faster than unpredictable words?
The nonwords experiment

- The Bicknell et al. (2010) experiment had many different participants and many different nonwords

\[ \text{response} \sim X + (1 \mid \text{Word}) + (1 + X \mid \text{Participant}) \]
The nonwords experiment

- The Bicknell et al. (2010) experiment had many different participants and many different nonwords
- Each nonword has only one different number of neighbors, of course

\[
\text{response} \sim X + (1 \mid \text{Word}) + (1 + X \mid \text{Participant})
\]
The nonwords experiment

- The Bicknell et al. (2010) experiment had many different participants and many different nonwords.
- Each nonword has only one different number of neighbors, of course.
- Each participant is exposed to nonwords with many different numbers of neighbors.

\[
\text{response} \sim X + (1 \mid \text{Word}) + (1 + X \mid \text{Participant})
\]
The nonwords experiment

- The Bicknell et al. (2010) experiment had many different participants and many different nonwords
- Each nonword has only one different number of neighbors, of course
- Each participant is exposed to nonwords with many different numbers of neighbors
- Hence, variation in neighborhood density is between-words but within-participant

\[
\text{response} \sim X + (1 \mid \text{Word}) + (1 + X \mid \text{Participant})
\]
The nonwords experiment

- The Bicknell et al. (2010) experiment had many different participants and many different nonwords
- Each nonword has only one different number of neighbors, of course
- Each participant is exposed to nonwords with many different numbers of neighbors
- Hence, variation in neighborhood density is between-words but within-participant
- In the formula syntax of R’s lme4 package:

\[
\text{response} \sim X + (1 \mid \text{Word}) + (1 + X \mid \text{Participant})
\]
Hypothesis testing for LMEMs

- Exactly as for GLMs, the variance-covariance matrix of the fixed-effects covariance matrix contains a lot of information about confidence level in the parameters.
Hypothesis testing for LMEMs

- Exactly as for GLMs, the variance-covariance matrix of the fixed-effects covariance matrix contains a lot of information about confidence level in the parameters.
- It can be used to determine a t-statistic for each parameter.
Hypothesis testing for LMEMs

- Exactly as for GLMs, the variance-covariance matrix of the fixed-effects covariance matrix contains a lot of information about confidence level in the parameters.
- It can be used to determine a t-statistic for each parameter.
- As with GLMs (but not as with LMs), the properties of this statistic are asymptotic—it is asymptotically normal.
Hypothesis testing for LMEMs

- Exactly as for GLMs, the variance-covariance matrix of the fixed-effects covariance matrix contains a lot of information about confidence level in the parameters.
- It can be used to determine a t-statistic for each parameter.
- As with GLMs (but not as with LMs), the properties of this statistic are asymptotic—it is asymptotically normal.
- Likewise, the likelihood-ratio test can be used to compare models differing in fixed effects structure alone.
Hypothesis testing for LMEMs

- Exactly as for GLMs, the variance-covariance matrix of the fixed-effects covariance matrix contains a lot of information about confidence level in the parameters.
- It can be used to determine a t-statistic for each parameter.
- As with GLMs (but not as with LMs), the properties of this statistic are asymptotic—it is asymptotically normal.
- Likewise, the likelihood-ratio test can be used to compare models differing in fixed effects structure alone.
- It’s slightly anticonservative, but not too bad in practice.
Hypothesis testing for LMEMs

- Exactly as for GLMs, the variance-covariance matrix of the fixed-effects covariance matrix contains a lot of information about confidence level in the parameters.
- It can be used to determine a t-statistic for each parameter.
- As with GLMs (but not as with LMs), the properties of this statistic are asymptotic—it is asymptotically normal.
- Likewise, the likelihood-ratio test can be used to compare models differing in fixed effects structure alone.
- It’s slightly anticonservative, but not too bad in practice.
- Finally, models differing in random effects structure alone can in principle be compared with likelihood-ratio tests.
Hypothesis testing for LMEMs

- Exactly as for GLMs, the variance-covariance matrix of the fixed-effects covariance matrix contains a lot of information about confidence level in the parameters.
- It can be used to determine a t-statistic for each parameter.
- As with GLMs (but not as with LMs), the properties of this statistic are asymptotic—it is asymptotically normal.
- Likewise, the likelihood-ratio test can be used to compare models differing in fixed effects structure alone.
- It’s slightly anticonservative, but not too bad in practice.
- Finally, models differing in random effects structure alone can in principle be compared with likelihood-ratio tests.
  - However, these results can be either conservative or anti-conservative, so take them with a grain of salt.
Results for the nonword-recognition experiment

```r
> dat$X <- dat$neighbors
> m2 <- lmer(time ~ X + (1 + X | participant) + (1|target), dat,REML=F)
> print(m2,corr=F)
```

Linear mixed model fit by maximum likelihood
Formula: time ~ X + (1 + X | participant) + (1 | target)
Data: dat

- AIC: 22452
- BIC: 22490
- logLik: -11219
- deviance: 22438
- REMLdev: 22428

Random effects:
- Groups: participant
  - (Intercept) Variance: 5795.392, Std.Dev.: 76.1275
  - X Variance: 23.062, Std.Dev.: 4.8023
- Groups: target
  - (Intercept) Variance: 649.806, Std.Dev.: 25.4913
- Residual Variance: 18431.635, Std.Dev.: 135.7632

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

Fixed effects:
- X: Estimate: 8.986, Std. Error: 2.124, t value: 4.23
Principles of random-effects specification I

There has been disagreement/unclarity regarding how to specify random-effects structure for one’s model
Principles of random-effects specification I

- There has been disagreement/unclarity regarding how to specify random-effects structure for one’s model
  - Random intercepts are enough?
There has been disagreement/unclarity regarding how to specify random-effects structure for one’s model

- Random intercepts are enough?
- Start with random intercepts and then use model selection?
There has been disagreement/unclarity regarding how to specify random-effects structure for one’s model

- Random intercepts are enough?
- Start with random intercepts and then use model selection?
- Maximal random effect structure, backing off to random intercepts if there are convergence problems?
There has been disagreement/unclarity regarding how to specify random-effects structure for one’s model:

- Random intercepts are enough?
- Start with random intercepts and then use model selection?
- Maximal random effect structure, backing off to random intercepts if there are convergence problems?

In Barr et al. (2013) we have taken a strong but, we believe, traditional stand (really following Clark, 1973):

*Random-effect structure should be maximal with respect to the theoretically critical questions you are posing of your data.*
Principles of random-effects specification II

- For traditional, balanced designs with a small number of theoretically critical predictor, this means:
Principles of random-effects specification II

- For traditional, balanced designs with a small number of theoretically critical predictor, this means:

  - For every theoretically critical fixed-effect term in your model that varies *between* clusters (e.g., subjects or items), include a random intercept for that clustering
Principles of random-effects specification II

- For traditional, balanced designs with a small number of theoretically critical predictor, this means:
  - For every theoretically critical fixed-effect term in your model that varies *between* clusters (e.g., subjects or items), include a random intercept for that clustering
  - For every theoretically critical fixed-effect term in your model that varies *within* clusters, include a random slope for that clustering
A controlled experiment I

- Rohde et al. (2011) used self-paced reading to assess the real-time deployment of discourse knowledge in syntactic ambiguity resolution
Rohde et al. (2011) used self-paced reading to assess the real-time deployment of discourse knowledge in syntactic ambiguity resolution.

Sample item:
John babysat the children of the musician who...
A controlled experiment I

- Rohde et al. (2011) used self-paced reading to assess the real-time deployment of discourse knowledge in syntactic ambiguity resolution.

- Sample item:
  John babysat the children of the musician who...

  - ...was generally arrogant and rude.
A controlled experiment I

- Rohde et al. (2011) used self-paced reading to assess the real-time deployment of discourse knowledge in syntactic ambiguity resolution

- Sample item:
  John babysat the children of the musician who...

  - ... *was* generally arrogant and rude.
  - ... *were* generally arrogant and rude.
A controlled experiment I

- Rohde et al. (2011) used self-paced reading to assess the real-time deployment of discourse knowledge in syntactic ambiguity resolution

- Sample item:
  John babysat the children of the musician who...
  - ...was generally arrogant and rude.
  - ...were generally arrogant and rude.

- Sample item in implicit causality condition:
  John detested the children of the musician who...
Rohde et al. (2011) used self-paced reading to assess the real-time deployment of discourse knowledge in syntactic ambiguity resolution.

Sample item:
John babysat the children of the musician who...
  
  ...*was* generally arrogant and rude.
  
  ...*were* generally arrogant and rude.

Sample item in *implicit causality* condition:
John *detested* the children of the musician who...
  
  ...*was* generally arrogant and rude.
A controlled experiment I

- Rohde et al. (2011) used self-paced reading to assess the real-time deployment of discourse knowledge in syntactic ambiguity resolution

- Sample item:
  John babysat the children of the musician who...
  - ... *was* generally arrogant and rude.
  - ... *were* generally arrogant and rude.

- Sample item in *implicit causality* condition:
  John *detested* the children of the musician who...
  - ... *was* generally arrogant and rude.
  - ... *were* generally arrogant and rude.
A controlled experiment I

Rohde et al. (2011) used self-paced reading to assess the real-time deployment of discourse knowledge in syntactic ambiguity resolution.

Sample item:
John babysat the children of the musician who...

- ... *was* generally arrogant and rude.
- ... *were* generally arrogant and rude.

Sample item in **implicit causality** condition:
John *detested* the children of the musician who...

- ... *was* generally arrogant and rude.
- ... *were* generally arrogant and rude.

The question of theoretical interest for our data is whether the processing penalty induced by disambiguation of the RC attachment would show up immediately (before potentially biasing semantic content of the RC shows up).
In self-paced reading, many kinds of word properties show up primarily in reading times *one or more words downstream* ("spillover" effects)
A controlled experiment II

In self-paced reading, many kinds of word properties show up primarily in reading times one or more words downstream ("spillover" effects)

Thus we focus on statistical analysis of the word immediately after disambiguation:

\[
\text{John babysat/detested the children of the musician who was/were generally arrogant and rude}
\]
In self-paced reading, many kinds of word properties show up primarily in reading times *one or more words downstream* ("spillover" effects)

Thus we focus on statistical analysis of the word immediately after disambiguation:

\[
\text{John babysat/detested the children of the musician who was/were generally arrogant and rude}
\]

We’ll abbreviate the type of verb (implicit causality or not) the \( V \) factor and the RC’s attachment level (high or low) the \( A \) factor
A controlled experiment II

- In self-paced reading, many kinds of word properties show up primarily in reading times *one or more words downstream* (“spillover” effects)
- Thus we focus on statistical analysis of the word immediately after disambiguation:

  \[
  \text{V} \quad \text{John babysat/detested the children of the musician}
  \]

  \[
  \text{B} \quad \text{who was/were generally arrogant and rude}
  \]

- We’ll abbreviate the type of verb (implicit causality or not) the \textbf{V} factor and the RC’s attachment level (high or low) the \textbf{A} factor
- These factors are crossed in the experiment, and both within-subject
A controlled experiment III

Results of a maximal LME fit:

Linear mixed model fit by maximum likelihood
Formula: rt ~ V * A + (V * A | subj) + (V * A | item)
Data: d

AIC      BIC  logLik deviance REMLdev
12528    12648 -6239   12478    12447

Random effects:
Groups   Name     Variance  Std.Dev.  Corr
subj     (Intercept) 16769.275  129.4962
         V          315.422   17.7601 -1.000
         A          20.165    4.4906 -1.000  1.000
         V:A        11372.255 106.6408 -0.511  0.511  0.511
item     (Intercept) 1510.106   38.8601
         V          2068.089  45.4762 -0.803
         A         1674.812  40.9245  0.054 -0.638
         V:A        5546.927  74.4777  0.038  0.565 -0.996
Residual                              38616.850 196.5117

Number of obs: 919, groups: subj, 55; item, 20

Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>470.4038</td>
<td>20.5847</td>
<td>22.852</td>
</tr>
<tr>
<td>V</td>
<td>-33.6708</td>
<td>16.7822</td>
<td>-2.006</td>
</tr>
<tr>
<td>A</td>
<td>-0.2227</td>
<td>16.0099</td>
<td>-0.014</td>
</tr>
<tr>
<td>V:A</td>
<td>-84.7617</td>
<td>34.2489</td>
<td>-2.475</td>
</tr>
</tbody>
</table>
A controlled experiment III

- Likelihood-ratio-based hypothesis testing for a fixed effect:

```r
> rt.lmer.null <- lmer(rt ~ V + A + (V*A | subj) + (V*A | item),
+   data=d,REML=F)

> print(anova(rt.lmer.full,rt.lmer.null))

Data: d
Models:
rt.lmer.null: rt ~ V + A + (V * A | subj) + (V * A | item)
rt.lmer.full: rt ~ V * A + (V * A | subj) + (V * A | item)

             Df AIC    BIC logLik Chisq Chi Df Pr(>Chisq)
rt.lmer.null 24 12532 12647 -6241.7
rt.lmer.full 25 12528 12648 -6238.9  5.5719 1  0.01825 *
```

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Mixed logit models

Recall the inverse logit function that we used for logistic regression:

\[ \mu = \frac{e^{\eta}}{1 + e^{\eta}} \]
Mixed logit models

- A generalized linear mixed model (GLMM) works exactly the same as an LME model; the cluster-level variables contribute to the linear predictor
Mixed logit models

- A generalized linear mixed model (GLMM) works exactly the same as an LME model; the cluster-level variables contribute to the linear predictor.
- A mixed logit model thus has the logit link function:

\[
\eta = \log \left( \frac{\mu}{1 - \mu} \right)
\]
Mixed logit models

- A generalized linear mixed model (GLMM) works exactly the same as an LME model; the cluster-level variables contribute to the linear predictor.
- A mixed logit model thus has the logit link function:

\[ \eta = \log \frac{\mu}{1 - \mu} \]

- Bernoulli noise distribution around predicted mean \( \mu \):

\[ P(Y = y|\mu) = \begin{cases} 
\mu & y = 1 \\
1 - \mu & y = 0 \\
0 & \text{otherwise}
\end{cases} \]
Mixed logit models

- A generalized linear mixed model (GLMM) works exactly the same as an LME model; the cluster-level variables contribute to the linear predictor.
- A mixed logit model thus has the logit link function:
  \[ \eta = \log \frac{\mu}{1 - \mu} \]
- Bernoulli noise distribution around predicted mean \( \mu \):
  \[ P(Y = y|\mu) = \begin{cases} 
  \mu & y = 1 \\
  1 - \mu & y = 0 \\
  0 & \text{otherwise}
\end{cases} \]
- And linear predictor
  \[ \eta = X\beta + Zb \]
  where \( b \) is multivariate-normal distributed:
  \[ b \sim N(0, \Sigma_b) \]


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 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 $H_0$. 
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!
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).
Alternative: Bayesian hypothesis testing, which is symmetric:

\[
\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)P(H_0)}{P(D|H_1)P(H_1)}
\]
A note on $p$-values and philosophy of science

- Alternative: Bayesian hypothesis testing, which is symmetric:

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

- 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:

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

- 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:

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

- 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:

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

- 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).