Fitting, Evaluating, and Reporting Mixed Models for Groningen

T. Florian Jaeger

August 23, 2011
Getting Help

- Subscribe to ling-R-lang: https://mailman.ucsd.edu/mailman/listinfo/ling-r-lang-l
- Great list of points to various FAQs: http://glmm.wikidot.com/faq
- In R: try ?foo or help(foo) first
- For more HLP Lab materials, check out:
  - http://www.hlp.rochester.edu/
  - http://wiki.bcs.rochester.edu:2525/HlpLab/StatsCourses
  - http://hlplab.wordpress.com/ (e.g. multinomial mixed models code)
- Subscribe to our paper feed: http://rochester.academia.edu/tiflo/Papers
Hypothesis testing in psycholinguistic research

- Typically, we make predictions not just about the existence, but also the *direction* of effects.
- Sometimes, we’re also interested in effect *shapes* (non-linearities, etc.)
- Unlike in ANOVA, regression analyses reliably test hypotheses about effect direction and shape without requiring post-hoc analyses provided (a) *the predictors in the model are coded appropriately* and (b) *the model can be trusted*.
- **Today:** Provide an overview of (a) and (b).
Building an interpretable model
Collinearity
What is collinearity?
Detecting collinearity
Dealing with collinearity
Model Evaluation
Beware overfitting
Detecting overfitting: Validation
Goodness-of-fit
Aside: Model Comparison
Random effect structure
A note on p-value estimation
What to report?
Model Description
Model Assumptions
Model Fit and Evaluation
Reporting Results
References

Modeling schema
[from Jaeger (2011)]
For data exploration, variable selection, transformation, coding, and centering, please see earlier tutorials (e.g. Jaeger and Kuperman (2009))
Overview

- Towards a model with interpretable coefficients:
  - collinearity

- Model evaluation:
  - fitted vs. observed values
  - model validation
  - investigation of residuals
  - case influence, outliers

- Model comparison

- Reporting the model:
  - comparing effect sizes
  - back-transformation of predictors
  - visualization
Data 1: Lexical decision RTs

- **Outcome:** log lexical decision latency RT
- **Inputs:**
  - factors Subject (21 levels) and Word (79 levels),
  - factor NativeLanguage (English and Other)
  - continuous predictors Frequency (log word frequency),
    and Trial (rank in the experimental list).

<table>
<thead>
<tr>
<th>Subject</th>
<th>RT</th>
<th>Trial</th>
<th>NativeLanguage</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1</td>
<td>6.340359</td>
<td>23</td>
<td>English</td>
<td>owl</td>
</tr>
<tr>
<td>2</td>
<td>A1</td>
<td>6.308098</td>
<td>27</td>
<td>English</td>
<td>mole</td>
</tr>
<tr>
<td>3</td>
<td>A1</td>
<td>6.349139</td>
<td>29</td>
<td>English</td>
<td>cherry</td>
</tr>
<tr>
<td>4</td>
<td>A1</td>
<td>6.186209</td>
<td>30</td>
<td>English</td>
<td>pear</td>
</tr>
<tr>
<td>5</td>
<td>A1</td>
<td>6.025866</td>
<td>32</td>
<td>English</td>
<td>dog</td>
</tr>
<tr>
<td>6</td>
<td>A1</td>
<td>6.180017</td>
<td>33</td>
<td>English</td>
<td>blackberry</td>
</tr>
</tbody>
</table>
Data 2: Lexical decision *response*

- **Outcome:** Correct or incorrect response (*Correct*)
- **Inputs:** same as in linear model

```r
> lmer(Correct == "correct" ~ NativeLanguage +
  + Frequency + Trial +
  + (1 | Subject) + (1 | Word),
  + data = lexdec, family = "binomial")
```

**Random effects:**

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>(Intercept)</td>
<td>1.01820</td>
<td>1.00906</td>
</tr>
<tr>
<td>Subject</td>
<td>(Intercept)</td>
<td>0.63976</td>
<td>0.79985</td>
</tr>
</tbody>
</table>

Number of obs: 1659, groups: Word, 79; Subject, 21

**Fixed effects:**

| (Intercept)     | Estimate | Std. Error | z value | Pr(>|z|) |
|-----------------|----------|------------|---------|---------|
| NativeLanguageOther | -5.726e-01 | 4.639e-01 | 1.234   | 0.217104 |
| Frequency       | 5.600e-01  | 1.570e-01  | -3.567  | 0.000361 *** |
| Trial           | 4.443e-06  | 2.965e-03  | 0.001   | 0.998804 |

---

A note on p-value estimation
**Definition of collinearity**

- **Collinearity**: A predictor is collinear with other predictors in the model if there are high (partial) correlations between them.
- Even if a predictor is not highly correlated with any single other predictor in the model, it can be highly collinear with the combination of predictors → collinearity will affect the predictor.
- This is not uncommon!
  - in models with many predictors
  - when several somewhat related predictors are included in the model (e.g. word length, frequency, age of acquisition)
Consequences of collinearity

→ standard errors $SE(\beta)$s of collinear predictors are biased (inflated).
  → \textit{tends} to underestimate significance (but see below)

→ coefficients $\beta$ of collinear predictors become hard to interpret (though not biased)
  • ‘bouncing betas’: minor changes in data might have a major impact on $\beta$s
  • coefficients will flip sign, double, half

→ coefficient-based tests don’t tell us anything reliable about collinear predictors!
Extreme collinearity: An example

**Drastic example of collinearity:** `meanWeight` (rating of the weight of the object denoted by the word, averaged across subjects) and `meanSize` (average rating of the object size) in `lexdec`.

```
lmer(RT ~ meanSize + (1 | Word) + (1 | Subject), data = lexdec)
Fixed effects:
  Estimate Std. Error  t value
(Intercept)  6.3891053  0.0427533  149.44
meanSize     -0.0004282  0.0094371    -0.05
```

- n.s. correlation of `meanSize` with RTs.
- similar n.s. weak negative effect of `meanWeight`.
- The two predictors are highly correlated (`r > 0.999`).
If the two correlated predictors are included in the model …

\[
\text{> lmer(RT} \sim \text{meanSize} + \text{meanWeight} + \\
\quad + \quad (1 \mid \text{Word}) + (1 \mid \text{Subject}), \text{ data = lexdec)}
\]

<table>
<thead>
<tr>
<th>Fixed effects:</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5.7379</td>
<td>0.1187</td>
<td>48.32</td>
</tr>
<tr>
<td>meanSize</td>
<td>1.2435</td>
<td>0.2138</td>
<td>5.81</td>
</tr>
<tr>
<td>meanWeight</td>
<td>-1.1541</td>
<td>0.1983</td>
<td>-5.82</td>
</tr>
</tbody>
</table>

Correlation of Fixed Effects:

| (Intr) meanSz           | -0.949 |
| meanSize                | 0.942  |
| meanWeight              | -0.999 |

- SE(\beta)\text{s} are hugely inflated (more than by a factor of 20)
- large and highly significant significant counter-directed effects (\beta\text{s}) of the two predictors
  \rightarrow collinearity needs to be investigated!
Extreme collinearity: An example (cnt’d)

- Objects that are perceived to be unusually heavy for their size tend to be more frequent (accounts for 72% of variance in frequency).
- Both effects apparently disappear though when frequency is included in the model (but cf. residualization → meanSize or meanWeight still has small expected effect beyond Frequency).

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>6.64846</td>
<td>0.06247</td>
<td>106.43</td>
</tr>
<tr>
<td>cmeanSize</td>
<td>-0.11873</td>
<td>0.35196</td>
<td>-0.34</td>
</tr>
<tr>
<td>cmeanWeight</td>
<td>0.13788</td>
<td>0.33114</td>
<td>0.42</td>
</tr>
<tr>
<td>Frequency</td>
<td>-0.05543</td>
<td>0.01098</td>
<td>-5.05</td>
</tr>
</tbody>
</table>
Generalized Linear Mixed Models

T. Florian Jaeger

Building an interpretable model

What is collinearity?

Detecting collinearity

Dealing with collinearity

Collinearity

Model Evaluation

Beware overfitting

Detect overfitting: Validation

Goodness-of-fit

Aside: Model Comparison

Random effect structure

A note on p-value estimation

What to report?

Model Description

Model Assumptions

Model Fit and Evaluation

Reporting Results

References

So what does collinearity do?

- Type II error increases $\rightarrow$ power loss

```r
h <- function(n) {
  x <- runif(n)
  y <- x + rnorm(n, 0, 0.01)
  z <- ((x + y) / 2) + rnorm(n, 0, 0.2)

  m <- lm(z ~ x + y)
  signif.m.x <- ifelse(summary(m)$coef[2,4] < 0.05, 1, 0)
  signif.m.y <- ifelse(summary(m)$coef[3,4] < 0.05, 1, 0)

  mx <- lm(z ~ x)
  my <- lm(z ~ y)
  signif.mx.x <- ifelse(summary(mx)$coef[2,4] < 0.05, 1, 0)
  signif.my.y <- ifelse(summary(my)$coef[2,4] < 0.05, 1, 0)
  return(c(cor(x,y),signif.m.x,signif.m.y,signif.mx.x, signif.my.y))
}
result <- sapply(rep(M,n), h)
print(paste("x in combined model:", sum(result[2,])))
print(paste("y in combined model:", sum(result[3,])))
print(paste("x in x-only model:", sum(result[4,])))
print(paste("y in y-only model:", sum(result[5,])))
print(paste("Avg. correlation:", mean(result[1,])))
```
So what does collinearity do?

- Type II error increases $\rightarrow$ power loss
- Type I error does not increase much (5.165% Type I error for two predictors with $r > 0.9989$ in joined model vs. 5.25% in separate models; 20,000 simulation runs with 100 data points each)

```r
set.seed(1)
n <- 100
M <- 20000
f <- function(n) {
x <- runif(n)
y <- x + rnorm(n,0,0.01)
z <- rnorm(n,0,5)
m <- lm(z ~ x + y)
mx <- lm(z ~ x)
my <- lm(z ~ y)
signifmin <- ifelse(min(summary(m)$coef[2:3,4]) < 0.05, 1, 0)
signifx <- ifelse(min(summary(mx)$coef[2,4]) < 0.05, 1, 0)
signify <- ifelse(min(summary(my)$coef[2,4]) < 0.05, 1, 0)
signifxory <- ifelse(signifx == 1 | signify == 1, 1, 0)
return(c(cor(x,y),signifmin,signifx,signify,signifxory))
}
result <- sapply(rep(n,M), f)
sum(result[2,])/M # joined model returns >=1 spurious effect
sum(result[3,])/M
sum(result[4,])/M
sum(result[5,])/M # two individual models return >=1 spurious effect
min(result[1,])
```
So what does collinearity do?

- Type II error increases → power loss
- Type I error does not increase (much)
- But small differences between highly correlated predictors can be highly correlated with another predictors and create ‘apparent effects’ (like in the case discussed).
  → Can lead to misleading effects (not technically spurious, but if they we interpret the coefficients causally we will have a misleading result!).
- This problem is not particular to collinearity, but it frequently occurs in the case of collinearity.
- When coefficients are unstable (as in the above case of collinearity) treat this as a warning sign - check for mediated effects.
Detecting collinearity

• Mixed model output in R comes with correlation matrix (cf. previous slide).
  • Partial correlations of fixed effects in the model.
• Also useful: correlation matrix (e.g. cor(); use Spearman option for categorical predictors) or pairs.cor.fnc() in languageR for visualization.
  • apply to predictors (not to untransformed input variables)!

```
> cor(lexdec[,c(2,3,10, 13)])

     RT    Trial Frequency     Length
RT 1.000000 -0.0524113 -0.2132495  0.1467381 0.14673811
Trial -0.0524113 1.0000000 -0.0068491  0.0098658  0.00986581
Frequency -0.2132495 -0.0068491 1.0000000 -0.4273381 -0.42733813
Length  0.1467381  0.0098658 -0.4273381  1.0000000  1.00000000
```
Formal tests of collinearity

- Variance inflation factor (VIF, \( \text{vif}() \)).
  - generally, \( VIF > 10 \) \( \rightarrow \) absence of absolute collinearity in the model cannot be claimed.
  - \( VIF > 4 \) are usually already problematic.
  - but, for large data sets, even VIFs > 2 can lead inflated standard errors.

- Kappa (e.g. \( \text{collin.fnc}() \) in \text{languageR})
  - generally, c-number (\( \kappa \)) over 10 \( \rightarrow \) mild collinearity in the model.

- Applied to current data set, \( \ldots \)

\[
> \text{collin.fnc(lexdec[,c(2,3,10,13)]})$cnumber
\]
  - \( \ldots \) gives us a kappa > 90 \( \rightarrow \) Houston, we have a problem.
Dealing with collinearity

What is collinearity?
Detecting collinearity
Dealing with collinearity

Beware overfitting
Detect overfitting:
Validation
Goodness-of-fit
Aside: Model Comparison
Random effect structure
A note on p-value estimation

What to report?
Model Description
Model Assumptions
Model Fit and Evaluation
Reporting Results
Dealing with collinearity

- **Good news:** Estimates are only problematic for those predictors that are collinear.
  
  → If collinearity is in the nuisance predictors (e.g. certain controls), nothing needs to be done.

- **Somewhat good news:** If collinear predictors are of interest but we are *not* interested in the direction of the effect, we can use \( \Leftrightarrow \) model comparison (rather than tests based on the standard error estimates of coefficients).

- If collinear predictors are of interest and we *are* interested in the direction of the effect, we need to reduce collinearity of those predictors.
Reducing collinearity

- **Centering**: reduces collinearity of predictor with intercept and higher level terms involving the predictor.
  - **pros**: easy to do and interpret; often improves interpretability of effects.
  - **cons**: none?

- **Re-express the variable** based on conceptual considerations (e.g. ratio of spoken vs. written frequency in `lexdec`; rate of disfluencies per words when constituent length and fluency should be controlled).
  - **pros**: easy to do and relatively easy to interpret.
  - **cons**: only applicable in some cases.
Reducing collinearity (cnt’d)

- **Stratification**: Fit separate models on subsets of data holding correlated predictor A constant.
  - **pros**: Still relatively easy to do and easy to interpret.
  - **cons**: harder to do for continuous collinear predictors; reduces power, → extra caution with null effects; doesn’t work for multicollinearity of several predictors.

- **Principal Component Analysis (PCA)**: for \( n \) collinear predictors, extract \( k < n \) most important orthogonal components that capture \( > p\% \) of the variance of these predictors.
  - **pros**: Powerful way to deal with *multicollinearity*.
  - **cons**: Hard to interpret (→ better suited for control predictors that are not of primary interest); technically complicated; some decisions involved that affect outcome.
Reduce collinearity (cnt’d)

- **Residualization**: Regress collinear predictor against combination of (partially) correlated predictors
  - usually using ordinary regression (e.g. `lm()`, `ols()`).
  - **pros**: systematic way of dealing with multicollinearity; directionality of (conditional) effect interpretable
  - **cons**: effect sizes hard to interpret; judgment calls: what should be residualized against what?
An example of moderate collinearity (cnt’d)

- Consider two moderately correlated variables \( r = -0.49 \), (centered) word length and (centered log) frequency:

```r
> lmer(RT ~ cLength + cFrequency +
    + (1 | Word) + (1 | Subject), data = lexdec)
<...>
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>6.385090</td>
<td>0.034415</td>
<td>185.53</td>
</tr>
<tr>
<td>cLength</td>
<td>0.009348</td>
<td>0.004327</td>
<td>2.16</td>
</tr>
<tr>
<td>cFrequency</td>
<td>-0.037028</td>
<td>0.006303</td>
<td>-5.87</td>
</tr>
</tbody>
</table>

Correlation of Fixed Effects:

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>cLength</th>
<th>cFrequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1</td>
<td>0.000</td>
<td>0.429</td>
</tr>
<tr>
<td>cLength</td>
<td>0.000</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>cFrequency</td>
<td>0.000</td>
<td>0.429</td>
<td></td>
</tr>
</tbody>
</table>
<...>
```

- Is this problematic? Let’s remove collinearity via residualization
Residualization: An example

- Let’s regress word length vs. word frequency.

```r
> lexdec$rLength = residuals(lm(Length ~ Frequency, data = lexdec))
```

- `rLength`: difference between actual length and length as predicted by frequency. Related to actual length ($r > 0.9$), but crucially not to frequency ($r \ll 0.01$).

- Indeed, collinearity is removed from the model:

```r
<...>
Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.385090  0.034415  185.53
rLength       0.009348  0.004327   2.16
cFrequency   -0.042872  0.005693  -7.53

Correlation of Fixed Effects:
            (Intr)  rLength
cFrequency  0.000  0.000
cFrequency  0.000  0.000  
<...>
```

→ $\text{SE}(\beta)$ estimate for frequency predictor decreased
→ larger $t$-value
Residualization: An example (cnt’d)

- **Q:** What precisely is \( r_{\text{Length}} \)?

- **A:** Portion of word length that is not explained by (a linear relation to log) word frequency.

→ Coefficient of \( r_{\text{Length}} \) needs to be interpreted as such

- No trivial way of back-transforming to Length.

- **NB:** We have granted frequency the entire portion of the variance that cannot unambiguously attributed to either frequency or length!

→ If we choose to residualize frequency on length (rather than the inverse), we may see a different result.
Understanding residualization

- So, let’s regress frequency against length.
- Here: no qualitative change, but word length is now *highly* significant (random effect estimates unchanged)

```r
> lmer(RT ~ cLength + rFrequency +
+       (1 | Word) + (1 | Subject), data = lexdec)
<...>
Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>6.385090</td>
<td>0.034415</td>
<td>185.53</td>
</tr>
<tr>
<td>cLength</td>
<td>0.020255</td>
<td>0.003908</td>
<td>5.18</td>
</tr>
<tr>
<td>rFrequency</td>
<td>-0.037028</td>
<td>0.006303</td>
<td>-5.87</td>
</tr>
</tbody>
</table>

Correlation of Fixed Effects:

<table>
<thead>
<tr>
<th>(Intr)</th>
<th>cLength</th>
<th>rFrequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>cLength</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>rFrequency</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
<...>
```

→ Choosing what to residualize, changes interpretation of $\beta$s and hence the hypothesis we’re testing.
Extreme collinearity: ctn’d

- we can now residualize meanWeight against meanSize and Frequency, and
- and residualize meanSize against Frequency.
- include the transformed predictors in the model.

```r
> lexdec$rmeanSize <- residuals(lm(cmeanSize ~ Frequency + cmeanWeight, 
+ data=lexdec))
> lexdec$rmeanWeight <- residuals(lm(cmeanWeight ~ Frequency, 
+ data=lexdec))
> lmer(RT ~ rmeanSize + rmeanWeight + Frequency + (1|Subject) + (1|Word), 
+ data=lexdec)

(Intercept)  6.588778  0.043077  152.95
rmeanSize   -0.118731  0.351957  -0.34
rmeanWeight  0.026198  0.007477   3.50
Frequency   -0.042872  0.005470  -7.84
```

- NB: The frequency effect is stable, but the meanSize vs. meanWeight effect depends on what is residualized against what.
Residualization: Which predictor to residualize?

- What to residualize should be based on conceptual considerations (e.g. rate of disfluencies $=\text{number of disfluencies} \sim \text{number of words}$).

- **Be conservative** with regard to your hypothesis:
  - If the effect only holds under some choices about residualization, *the result is inconclusive*.
  - We usually want to show that a hypothesized effect holds *beyond what is already known* or that it *subsumes other effects*.

  \[ \rightarrow \text{Residualize effect of interest}. \]

  - E.g. if we hypothesize that a word’s predictability affects its duration beyond its frequency $\rightarrow$ residuals(lm(Predictability $\sim$ Frequency, data)).

  - (if effect *direction* is not important, see also $\rightarrow$ model comparison)
Modeling schema

- Outcome
- Input var 1
- Input var 2
- Input var 3
- Input var 4
- Input var n
- Log
- Transforming
- Centering
- Coding
- Interactions
- Residualize
- Reduce collinearity
- Predictor 1
- Predictor 2
- Predictor 3
- Predictor 4
- Predictor 5
- Predictor 6
- Predictor 7
- Predictor 8
- Predictor m
- Outlier removal
- Model quality?
Overfitting: Fit might be too tight due to the exceeding number of parameters (coefficients). The maximal number of predictors that a model allows depends on their distribution and the distribution of the outcome.

- **Rules of thumb:**
  - linear models: $> 20$ observations per predictor.
  - logit models: the less frequent outcome should be observed $> 10$ times more often than there predictors in the model.
  - Predictors count: one per each random effect $+$ residual, one per each fixed effect predictor $+$ intercept, one per each interaction.
Validation allows us to detect overfitting:

- How much does our model depend on the exact data we have observed?
- Would we arrive at the same conclusion (model) if we had only slightly different data, e.g. a subset of our data?
- **Bootstrap-validate** your model by repeatedly sampling from the population of speakers/items with replacement. Get estimates and confidence intervals for fixed effect coefficients to see how well they generalize (Baayen, 2008:283; cf. `bootcov()` for ordinary regression models).
Visuallyize validation

- Plot predicted vs. observed (averaged) outcome.
- E.g. for logit models, plot.logistic.fit.fnc in languageR or similar function (cf. http://hlplab.wordpress.com)

- The following shows a badly fitted model:

```r
> lexdec$NativeEnglish = ifelse(lexdec$NativeLanguage == "English", 1, 0)
> lexdec$cFrequency = lexdec$Frequency - mean(lexdec$Frequency)
> lexdec$cNativeEnglish = lexdec$NativeEnglish - mean(lexdec$NativeEnglish)
> lexdec$Correct = ifelse(lexdec$Correct == "correct", T, F)
> l <- glmer(Correct ~ cNativeEnglish * cFrequency + Trial +
+ (1 | Word) + (1 | Subject),
+ data = lexdec, family="binomial")
```

![Graph showing a badly fitted model]
Fitted values

So far, we’ve been worrying about coefficients, but the real model output are the fitted values.

Goodness-of-fit measures assess the relation between fitted (a.k.a. predicted) values and actually observed outcomes.

- **linear models**: Fitted values are predicted numerical outcomes.

<table>
<thead>
<tr>
<th>RT</th>
<th>fitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.340359 6.277565</td>
</tr>
<tr>
<td>2</td>
<td>6.308098 6.319641</td>
</tr>
<tr>
<td>3</td>
<td>6.349139 6.265861</td>
</tr>
<tr>
<td>4</td>
<td>6.186209 6.264447</td>
</tr>
</tbody>
</table>

- **logit models**: Fitted values are predicted log-odds (and hence predicted probabilities) of outcome.

<table>
<thead>
<tr>
<th>Correct</th>
<th>fitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 correct</td>
<td>0.9933675</td>
</tr>
<tr>
<td>2 correct</td>
<td>0.9926289</td>
</tr>
<tr>
<td>3 correct</td>
<td>0.9937420</td>
</tr>
<tr>
<td>4 correct</td>
<td>0.9929909</td>
</tr>
</tbody>
</table>
Goodness-of-fit measures: Linear Mixed Models

- $R^2 = \text{correlation(observed, fitted)}^2$.
- Random effects usually account for much of the variance → obtain separate measures for partial contribution of fixed and random effects Gelman and Hill (2006, 474).
- E.g. for

```r
> cor(l$RT, fitted(1mer(RT ~ cNativeEnglish * cFrequency + Trial +
+ (1 | Word) + (1 | Subject), data = l)))^2
```

- ... yields $R^2 = 0.52$ for model, but only 0.004 are due to fixed effects!
Measures built on data likelihood

- **Data likelihood**: What is the probability that we would observe the data we have given the model (i.e. given the predictors we chose and given the ‘best’ parameter estimates for those predictors).

- **Standard model output** usually includes such measures, e.g. in R:

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>REMLdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>-96.48</td>
<td>-63.41</td>
<td>55.24</td>
<td>-123.5</td>
<td>-110.5</td>
</tr>
</tbody>
</table>

- **log-likelihood**, $\logLik = \log(L)$. This is the maximized model’s log data likelihood, no correction for the number of parameters. **Larger (i.e. closer to zero) is better.** The value for log-likelihood should always be **negative**, and AIC, BIC etc. are positive.
Measures built on data likelihood (cont’d’)

- Other measures trade off goodness-of-fit (data likelihood) and model complexity (number of parameters; cf. Occam’s razor; see also model comparison).
  - Deviance: -2 times log-likelihood ratio. **Smaller is better**.
  - Aikaike Information Criterion, $AIC = k - 2ln(L)$, where $k$ is the number of parameters in the model. **Smaller is better**.
  - Bayesian Information Criterion, $BIC = k \times ln(n) - 2ln(L)$, where $k$ is the number of parameters in the model, and $n$ is the number of observations. **Smaller is better**.
  - also Deviance Information Criterion
Likelihood functions used for the fitting of linear mixed models

- Linear models:
  - Maximum Likelihood function, ML: Find $\theta$-vector for your model parameters that maximizes the probability of your data given the model’s parameters and inputs. Great for point-wise estimates, but provides biased (anti-conservative) estimates for variances.
  - Restricted or residual maximum likelihood, REML: default in \texttt{lmer} package. Produces unbiased estimates for variance.
  - In practice, the estimates produced by ML and REML are nearly identical Pinheiro and Bates (2000, 11).

$\rightarrow$ hence the two deviance terms given in the standard model output in \texttt{R}. 

\[ \hat{\text{Likelihood functions used for the fitting of linear mixed models}} \]
Goodness-of-fit: Mixed Logit Models

- Best available right now:
  - some of the same measures based on data likelihood as for mixed models

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>499.1</td>
<td>537</td>
<td>-242.6</td>
<td>485.1</td>
</tr>
</tbody>
</table>

- but no known closed form solution to likelihood function of mixed logit models → current implementations use Penalized Quasi-Likelihoods or better Laplace Approximation of the likelihood (default in \texttt{R}; cf. Harding & Hausman, 2007)

- Discouraged:
  - 
    - pseudo-$R^2$ a la Nagelkerke (cf. along the lines of [http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm](http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm))
  - classification accuracy: If the predicted probability is $< 0.5$ → predicted outcome $= 0$; otherwise 1. Needs to be compared against baseline. (cf. Somer’s $D_{xy}$ and $C$ index of concordance).
• Models can be compared for performance using any goodness-of-fit measures. Generally, an advantage in one measure comes with advantages in others, as well.

• To test whether one model is significantly better than another model:
  • likelihood ratio test (for nested models only)
  • (DIC-based tests for non-nested models have also been proposed).
Likelihood ratio test for nested models

-2 times ratio of likelihoods (or difference of log likelihoods) of nested model and super model.

Distribution of likelihood ratio statistic follows asymptotically the $\chi^2$-square distribution with $DF(model_{super}) - DF(model_{nested})$ degrees of freedom.

$\chi^2$-square test indicates whether sparing extra df’s is justified by the change in the log-likelihood.

- in R: `anova(model1, model2)`
- NB: use restricted maximum likelihood-fitted models to compare models that differ in random effects.
Example of model comparison

> super.lmer = lmer(RT ~ rawFrequency + (1 | Subject) + (1 | Word), data = lexdec)
> nested.lmer = lmer(RT ~ rawFrequency + (1 + Trial | Subject) + (1 | Word), data = lexdec)
> anova(super.lmer, nested.lmer)

<table>
<thead>
<tr>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>super.lmer</td>
<td>5</td>
<td>-910.41</td>
<td>-883.34</td>
<td>460.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nested.lmer</td>
<td>7</td>
<td>-940.71</td>
<td>-902.81</td>
<td>477.35</td>
<td>34.302</td>
<td>2 3.56e-08 ***</td>
</tr>
</tbody>
</table>

→ change in log-likelihood justifies inclusion Subject-specific slopes for Trial, and the correlation parameter between trial intercept and slope.
Determining the random effect structure

- It is *crucial* to evaluate hypotheses under an adequate random effect structure.
- For example, often it is *not* enough to simply include random intercepts in the model. Random slopes might also be required.
- For suggestions as to how to determine the maximum random effect structure justified by the data, see http://hlplab.wordpress.com/2011/06/25/more-on-random-slopes/
• From Jaeger, Graff, Croft, and Pontillo (in press):
Evaluating p-values

- **MCMC-sampling** (e.g. `mcmcsamp()`, `pvals.fnc()`, etc.)
- **Parametric bootstrap:**
  - Fit model without fixed effect predictor(s) of interest
  - Repeatedly simulate data from this reduced (‘null’) model
  - For each sample compare fit of null model against fit of model with predictors (on simulated data). E.g. calculate the difference in deviance.
  - Compare the difference in deviance for the null model and model with predictor on the actual data against the distribution of deviance differences from the repeated simulations based on the null model.
What to report?

- Describe your model
- State enough for readers and reviewers to assess whether they can trust the model
- Summarize your results
• State the outcome variable (e.g. for a binomial model, what is the value of the outcome you are predicting)
• Describe the predictors (incl. random effects)
• State what you did you about outliers
• State the outcome variable (e.g. for a binomial model, what is the value of the outcome you are predicting:

[...] our dependent variable is the proportion of fixations, during the ambiguous region, to the animal (the potential recipient, e.g., the horse). This captures the degree to which participants expect the recipient rather than the theme. [...] Following Barr (2008), proportion of fixations to the animal and the object were first empirical logit-transformed [...]

[Fine and Jaeger (submitted)]
Model Description

- State the predictors (incl. random effects)
- Transformations, centering, (potentially standardizing), coding, residualization should be described as part of the predictor summary.
  - Where what you did isn't already standard (e.g. unlike a log-transform for frequency), give theoretical, and/or empirical arguments for any decision made.
  - Consider reporting scales for outputs, inputs and predictors (e.g., range, mean, sd, median).
Main effects of prime structure, the surprisal of the first and second primes, target structure, and the bias of the target verb (probability that the target verb occurs in the DO version of the dative alternation) were included in the analysis. Additionally, the interaction between the surprisal of the first prime and prime structure, as well as the interaction between the surprisal of the second prime and prime structure were included. The model included the maximal random effect structure justified by the data (cf. Jaeger, 2011).

[Fine and Jaeger (submitted)]
State what you did you about outliers and whether this affected your results:

*Two trials containing primes with very large surprisal values (values that exceeded 6 bits; mean surprisal value=2.25, SD=1.4) were removed. The results below do not depend on this removal.*

[Fine and Jaeger (submitted)]
Sometimes it can be crucial to be clear about what assumptions the analysis you conducted makes. (also, remind yourself of those assumptions – your conclusions about theories only hold under those assumptions, cf. linearity!).

At least for yourself, you should also check model assumptions (residuals, etc.), but those are not usually reported. Sometimes, it is worth reporting these tests, though usually this would go into an appendix (it can easily get rather expansive).
Diagnostic plots - residuals

- From Jaeger, Graff, Croft, and Pontillo (in press) – Checking assumptions about the distribution of residuals in a linear mixed model:

![Diagnostic plots](image.png)

- Figure 3: Diagnostic plots for the model described in Equation (E3). The histogram of residuals (the individual-level errors) in (A) suggests normality. Linear (blue), quadratic (red), and cubic fits (green) of log-transformed population size in (B) or distance from the origin in (C) against the standardized residuals reveal no correlations (the shaded 95% confidence intervals include the zero line at all times). Only six data points fall outside the interval of -2.5 to 2.5 standardized residuals (indicated by the dashed lines). Excluding these languages (Austronesian: Iaai, Po-Ai; Niger-Congo: Bisa; Nilo-Saharan: Koyra Chiimi; Sino-Tibetan: Garo, Naxi, and Newari) strengthens both the population and the distance effect.
Diagnostic plots - random effects

- From Jaeger, Graff, Croft, and Pontillo (in press) – Checking assumptions about the distribution of random effects:

Figure 4: Quantile to quantile plot of random intercepts by language family (A), subfamily (B), and genus (C) in a linear mixed model with the main effects and interaction of (log-transformed) language population and distance from best fit single-origin. Theoretical quantiles – what would be expected under a normal distribution - are shown on the x-axis. The y-axis shows the best linear unbiased predictors (BLUPs) for each level of the random effect. Intervals around dots represent the 95% highest posterior density intervals.
(Jaeger, Graff, Croft, and Pontillo, in press) – check linearity assumption, e.g. by means of local smoothers:
Model Evaluation

- State to what extent you tested whether collinearity was an issue and what you did about it. Did this in any way affect your results? E.g.

  Collinearity was observed between prime structure and the surprisal of the second prime ($r = -0.59$; all other fixed effect correlations $r < .2$). Leave-one-out model comparison confirmed that collinearity did not affect any of the significant effects reported below. An ANCOVA over the difference scores yields the same results as those reported below.

  [Fine and Jaeger, submitted to Cognitive Science]
Model Evaluation: Quality of Fit

- Often it can be informative to say something about the model quality
  - For **linear models**: report $R^2$. Possibly, also the amount of variance explained by fixed effects over and beyond random effects, or predictors of interest over and beyond the rest of predictors.
  - For **logistic models**: report $D_{xy}$ or concordance C-number. Report the increase in classification accuracy over and beyond the baseline model.

**NB:** Be cautious, classification accuracy and its derivatives can be very misleading!

- Plots illustrating classification accuracy based on values of predictors (see above)
Some considerations for good science

• If at all possible, know and state whether whatever you did in terms of coding, transformation, and data exclusions affected the results.

• **Do not** report effects that heavily depend on the choices you have made;

• **Do not** fish for effects. There should be a strong theoretical motivation for what variables to include and in what way.

• To the extent that different ways of entering a predictor are investigated (without a theoretical reason), **do** make sure your conclusions hold for *all* ways of entering the predictor or that the model you choose to report is superior (\(\sim\) model comparison).
Result Summary

- **Standard textual summary**
  - Describe effects in your own words and provide coefficient, either SE or $t/z$-statistics, and p-value. Some things you might want to mention:
    - **Effect size** (What is that actually?)
    - Effect direction
    - Effect shape (tested by significance of non-linear components & superiority of transformed over un-transformed variants of the same input variable); plus visualization
  - Illustrate effect size, especially for continuous variables (e.g. predicted difference in outcome for 5th and 95th quantile of continuous predictor, perhaps on its original scale; see above).
- **Visualize**, especially for interactions.
- If you have many predictors in the model, you might want to provide a table of results.
Result Summary: Terminological Suggestions

- In regression studies, it is common to talk about predictors (independent variables) and outcomes (dependent variables).
- ‘the maximal random effect structure justified by the data’ (e.g. Jaeger, Graff, Croft, and Pontillo (in press); also http://hlplab.wordpress.com/2009/05/14/random-effect-structure/ and http://hlplab.wordpress.com/2011/06/25/more-on-random-slopes/).
- “random by-subject intercepts and slopes for frequency as well as neighborhood density” (cf. Jaeger et al. (in press)).
The main effect of prime structure remained only marginally significant when prime surprisal and the prime structure-prime surprisal interactions were included in the model ($\beta = .34$, $SE=.34$, $p= .1$), but was statistically significant when these terms were left out ($\beta = .43$, $SE=.21$, $p< .05$), replicating Thothathiri and Snedeker (2008). The reason for the reduced significance of the main effect of priming is that the effect of prime structure is carried by the high-surprisal primes, discussed below.

As expected, no main effect of the surprisal of either the first or the second prime was observed ($ps> .5$). Crucially, we found the predicted two-way interaction between the surprisal of the first prime and prime Structure ($\beta = .53$, $SE=.24$, $p< .05$)-for DO primes, as prime surprisal increased, fixations to the animal relative to the object increased; for PO primes, as prime surprisal increased, fixations to the animal relative to the object decreased. The interaction between the surprisal of the second prime and prime structure was not significant ($p=.9$). The significant interaction of prime structure and prime surprisal for prime 1 is shown in Figure 2.

[Fine and Jaeger (submitted)]
Result Summary: Visualization

Example

- From Fine and Jaeger (submitted) – Visualize (preferably on original, interpretable scales):

![Graph](attachment:example.png)
Result Summary: Visualization

Example

- From Jaeger et al. (in press) – Consider using smoothers to explore and visualize local fits:

Figure 5  (A) Distribution of the nine largest language families in the sample (at least 16 languages each). Circles represent languages. The size of the circle reflects the number of speakers of that language as reported in WALS. The color of the circle reflects the language family. (B) Normalized phonological complexity plotted against distance from the origin for the same subset of languages. Solid colored lines show the best fit linear trend with 95% confidence intervals (shaded area) by language family.
Result Summary: Continuous Predictors

- estimate the effect in ms across the frequency range and then the effect for a unit of frequency.

```r
> intercept = as.vector(fixef(lexdec.lmer4)[1])
> betafreq = as.vector(fixef(lexdec.lmer4)[3])
> eff = exp(intercept + betafreq * max(lexdec$Frequency)) - 
> exp(intercept + betafreq * min(lexdec$Frequency))
[1] -109.0357 #RT decrease across the entire range of Frequency
> range = exp(max(lexdec$Frequency)) - 
> exp(min(lexdec$Frequency))
[1] 2366.999
```

- Report that the full effect of Frequency on RT is a 109 ms decrease.

⭐ But in this model there is no simple relation between RTs and frequency, so resist to report that “the difference in 100 occurrences comes with a 4 ms decrease of RT”.

```r
> eff/range * 100
[1] -4.606494
```
The increase in 1 log unit of cFrequency comes with a -0.039 log units decrease of RT.

Utterly uninterpretable!

To get estimates in sensible units we need to back-transform both our predictors and our outcomes.

- decentralize cFrequency, and
- exponentially-transform logged Frequency and RT.
- if necessary, we de-residualize and de-standardize predictors and outcomes.
Often there is a trade-off between visualizing fit and using an intuitive scale:

Fig. 4. Effect of information density at the complement clause onset on that-mentioning along with 95% CIs (shaded area, which is hard to see because the CIs are very narrow around the predicted mean effect). (a) The effect on the log-odds of complementizer that (the space in which the analysis was conducted). (b) The effect transformed back into probability space. Hexagons indicate the distribution of information density against predicted log-odds (a) and probabilities (b) of that, considering all predictors in the model. Fill color indicates the number of cases in the database that fall within the hexagon.
Result Summary: Visualization Example

![Graph showing predicted log-odds of complementizer vs. information content of CC onset (in bits)]
Table 3
Result summary: coefficient estimates $\beta$, standard errors SE$(\beta)$, associated Wald's z-score ($= \beta/SE(\beta)$) and significance level $p$ for all predictors in the analysis.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef. $\beta$</th>
<th>SE$(\beta)$</th>
<th>$z$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.12</td>
<td>0.38</td>
<td>0.3</td>
<td>&gt;0.7</td>
</tr>
<tr>
<td>POSITION/MATRIX VERB</td>
<td>0.35</td>
<td>0.51</td>
<td>1.9</td>
<td>0.051</td>
</tr>
<tr>
<td>(1st restricted comp.)</td>
<td>-27.94</td>
<td>5.33</td>
<td>6.6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>(2nd restricted comp.)</td>
<td>55.43</td>
<td>10.80</td>
<td>5.1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LENGTH/MATRIX VERB-TO-CC</td>
<td>0.17</td>
<td>0.065</td>
<td>2.6</td>
<td>0.01</td>
</tr>
<tr>
<td>LENGTH(CC ONSET)</td>
<td>0.18</td>
<td>0.014</td>
<td>12.8</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LENGTH(CC REMAINDER)</td>
<td>0.03</td>
<td>0.006</td>
<td>4.4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LOG SPEECH RATE</td>
<td>-0.70</td>
<td>0.13</td>
<td>-5.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SQ LOG SPEECH RATE</td>
<td>-0.36</td>
<td>0.19</td>
<td>-1.9</td>
<td>&lt;0.06</td>
</tr>
<tr>
<td>FAUSE</td>
<td>1.11</td>
<td>0.11</td>
<td>10.2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>DISPLACEMENT</td>
<td>0.39</td>
<td>0.12</td>
<td>3.2</td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>CC SUBJECT = it vs. I</td>
<td>0.04</td>
<td>0.08</td>
<td>0.5</td>
<td>&gt;0.6</td>
</tr>
<tr>
<td>=other pro vs. prev. levels</td>
<td>0.05</td>
<td>0.03</td>
<td>1.6</td>
<td>0.11</td>
</tr>
<tr>
<td>=other NP vs. prev. levels</td>
<td>0.11</td>
<td>0.02</td>
<td>4.9</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>FREQUENCY/CC SUBJECT HEAD</td>
<td>-0.02</td>
<td>0.03</td>
<td>-0.7</td>
<td>&gt;0.5</td>
</tr>
<tr>
<td>SUBJECT IDENTITY</td>
<td>-0.32</td>
<td>0.17</td>
<td>0.1</td>
<td>&lt;0.052</td>
</tr>
<tr>
<td>WORD FORM SIMILARITY</td>
<td>-0.31</td>
<td>0.17</td>
<td>-1.9</td>
<td>&lt;0.08</td>
</tr>
<tr>
<td>FREQUENCY/MATRIX VERB</td>
<td>-0.23</td>
<td>0.03</td>
<td>-7.7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>AMBIGUOUS CC ONSET</td>
<td>-0.12</td>
<td>0.12</td>
<td>-1.0</td>
<td>&gt;0.2</td>
</tr>
<tr>
<td>MATRIX SUBJECT = you</td>
<td>0.48</td>
<td>0.15</td>
<td>3.1</td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>=other PRO</td>
<td>0.60</td>
<td>0.13</td>
<td>4.8</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>=other NP</td>
<td>0.85</td>
<td>0.13</td>
<td>6.7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>PERSISTENCE = to vs. prime w/o that</td>
<td>0.02</td>
<td>0.07</td>
<td>0.3</td>
<td>&gt;0.7</td>
</tr>
<tr>
<td>=prime w/ that prev. levels</td>
<td>0.06</td>
<td>0.04</td>
<td>1.6</td>
<td>0.11</td>
</tr>
<tr>
<td>MALE SPEAKER</td>
<td>-0.15</td>
<td>0.11</td>
<td>-1.3</td>
<td>&gt;0.19</td>
</tr>
<tr>
<td>Information density</td>
<td>0.47</td>
<td>0.03</td>
<td>16.9</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Result Summary: Table Example

- From a draft of Tily (2010):

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$p_z$</th>
<th>df</th>
<th>$\chi^2$</th>
<th>$p_{\chi^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.93</td>
<td>0.19</td>
<td>-1.5</td>
<td>0.0</td>
<td>1.5</td>
</tr>
<tr>
<td>Object case/type = dat/2nd</td>
<td>0.046</td>
<td>0.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronominal object</td>
<td>-1.5</td>
<td>&lt;.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantified object</td>
<td>-0.70</td>
<td>&lt;.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object length</td>
<td>0.85</td>
<td>&lt;.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject length</td>
<td>-0.13</td>
<td>0.0013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text date</td>
<td>0.94</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text date * Object case/type</td>
<td>1.2</td>
<td>&lt;.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text date * Object length</td>
<td>-0.49</td>
<td>&lt;.001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$p_z$</th>
<th>df</th>
<th>$\chi^2$</th>
<th>$p_{\chi^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Text</td>
<td>0.89</td>
<td></td>
<td>1</td>
<td>980 &lt;.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>Verb POS</td>
<td>1.6</td>
<td>0.61</td>
<td>2</td>
<td>190 &lt;.001</td>
</tr>
</tbody>
</table>

Table 3: Final model for VO/OV order (positive outcome is VO)


References


References
