Lecture 6: Common Issues and Solutions in GLM/GLMM modeling

LSA 2013, LI539

Mixed Effect Models

T. Florian Jaeger

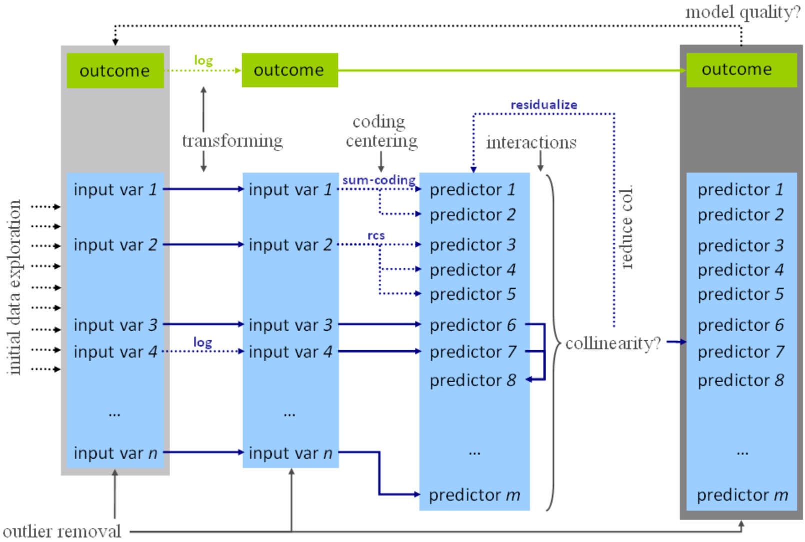
Brain and Cognitive Sciences
University of Rochester

July 20, 2013

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).

Modeling schema



model quality?
outcome
log
outcome
outcome
transforming
coding
centering
residualize
interactions
initial data exploration
input var *1*
input var *2*
input var *3*
input var *4*
...
input var *n*
input var *1*
input var *2*
input var *3*
input var *4*
...
input var *n*
sum-coding
rcs
log
predictor *1*
predictor *2*
predictor *3*
predictor *4*
predictor *5*
predictor *6*
predictor *7*
predictor *8*
...
predictor *m*
reduce col.
collinearity?
predictor *1*
predictor *2*
predictor *3*
predictor *4*
predictor *5*
predictor *6*
predictor *7*
predictor *8*
...
predictor *m*
outlier removal

[from Jaeger (2011)]

Data exploration

- 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

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 $\rightarrow$ 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 $\mathrm{SE}(\beta)$s of collinear predictors are biased (*in*flated).

  → *tends* to underestimate significance, although overestimation can also occur, depending on the specifics of a) the multicollinearity and b) the correlations of the fixed effects with the outcome. For an excellent treatment of these questions, see Friedman and Wall (2005)

→ coefficients β of collinear predictors become hard to interpret. Coefficient estimates might over- or underestimate the true fixed effect correlations with the outcome (Friedman & Wall, 2005).

  - 'bouncing betas': minor changes in data might have a major impact on βs
  - coefficients will flip sign, double, half

- The model's R^2 might be inflated or deflated (Friedman & Wall, 2005)

→ 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).

Extreme collinearity: An example (cont'd)

- If the two correlated predictors are included in the model . . .

```
> lmer(RT ~ meanSize + meanWeight +
+          (1 | Word) + (1 | Subject), data = lexdec)
Fixed effects:
            Estimate Std. Error t value
(Intercept)   5.7379     0.1187   48.32
meanSize      1.2435     0.2138    5.81
meanWeight   -1.1541     0.1983   -5.82

Correlation of Fixed Effects:
           (Intr) meanSz
meanSize   -0.949
meanWeight  0.942 -0.999
```

- $\mathrm{SE}(\beta)$s are hugely inflated (more than by a factor of 20)
- large and highly significant significant counter-directed effects (βs) of the two predictors

→ collinearity needs to be investigated!

Extreme collinearity: An example (cont'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.

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  6.64846    0.06247  106.43
cmeanSize   -0.11873    0.35196   -0.34
cmeanWeight  0.13788    0.33114    0.42
Frequency   -0.05543    0.01098   -5.05
```

So what does collinearity do?

- Type II error increases → power loss

```
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 → 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)

```
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**.

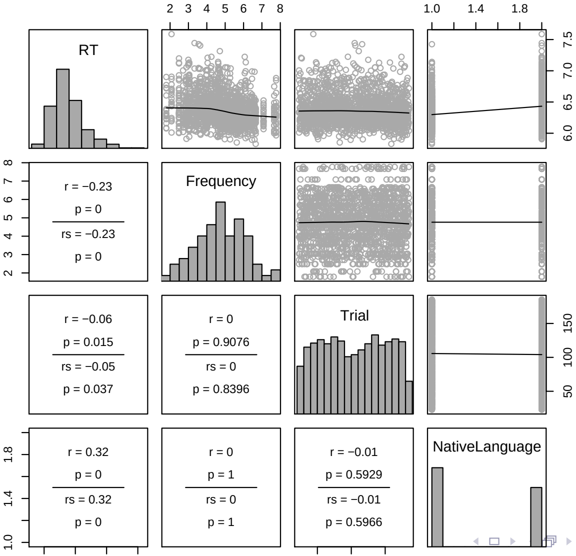
Detecting collinearity

- 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 `pairscor.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         1.0000000 -0.052411295 -0.213249525  0.146738111
Trial     -0.0524113  1.000000000 -0.006849117  0.009865814
Frequency -0.2132495 -0.006849117  1.000000000 -0.427338136
Length     0.1467381  0.009865814 -0.427338136  1.000000000
```

Visualizing collinearity

- `pairscor.fnc()` in `languageR`



Formal tests of collinearity

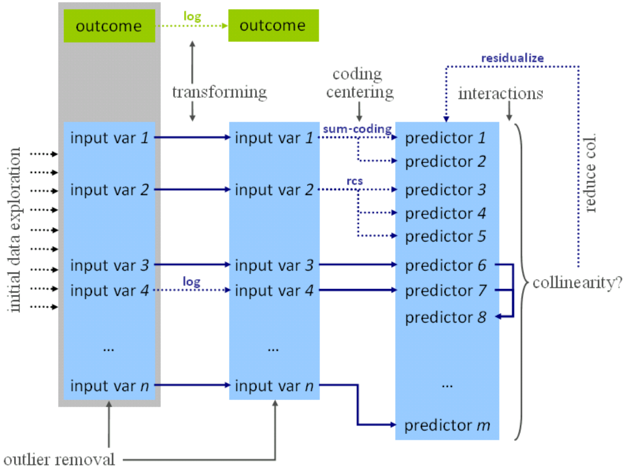
- Variance inflation factor (VIF, `vif()`).
  - generally, VIF > 10 → absence of absolute collinearity in the model cannot be claimed.
  - ★ VIF > 4 are usually already problematic.
  - ★ but, for large data sets, even VIFs > 2 can lead inflated standard errors.
- Kappa (e.g. `collin.fnc()` in `languageR`)
  - generally, c-number (κ) over 10 → mild collinearity in the model.

- Applied to current data set, …

```
> collin.fnc(lexdec[,c(2,3,10,13)])$cnumber
```

  - …gives us a kappa > 90 → Houston, we have a problem.

Dealing with collinearity



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 ↷**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 (cont'd)

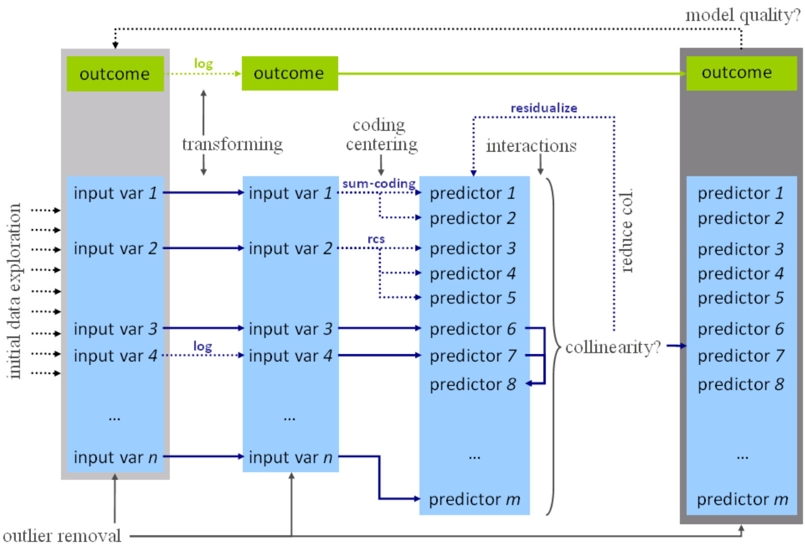
- **Stratification**: Fit separate models on **subsets** of data holding correlated predictor A constant.
- If effect of predictor B persists → effect is probably real.
  - **pros:** Still relatively easy to do and easy to interpret.
  - **cons:** harder to do for continuous collinear predictors; reduces 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 *multi*collinearity.
  - **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 (cont'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
  - **cons:** effect sizes hard to interpret; judgment calls: what should be residualized against what? directionality of (conditional) effect not always easily interpretable;

NB: the residualized predictor is *not* the same as the original predictors "after controlling for the collinear effects"

Modeling schema



model quality?

outcome

log

outcome

outcome

transforming

coding
centering

residualize

interactions

initial data exploration

input var *1*
input var *2*
input var *3*
input var *4*
...
input var *n*

log

input var *1*
input var *2*
input var *3*
input var *4*
...
input var *n*

sum-coding

rcs

predictor *1*
predictor *2*
predictor *3*
predictor *4*
predictor *5*
predictor *6*
predictor *7*
predictor *8*
...
predictor *m*

reduce col.

collinearity?

predictor *1*
predictor *2*
predictor *3*
predictor *4*
predictor *5*
predictor *6*
predictor *7*
predictor *8*
...
predictor *m*

outlier removal

Overfitting

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

Validation allows us to detect **overfitting**:

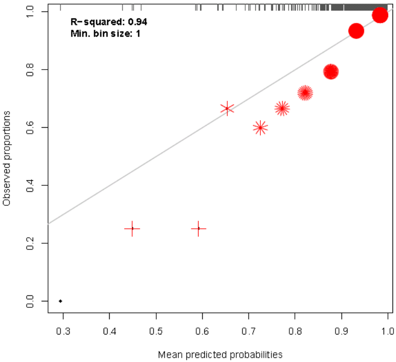
- How much does our model depend on the exact data we have observed?
- Would we arrive at the same conclusion (model) if we had only slightly different data, e.g. a subset of our data?

- **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).

Visualize 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:

```
> 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")
```



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.

```
         RT   fitted
1 6.340359 6.277565
2 6.308098 6.319641
3 6.349139 6.265861
4 6.186209 6.264447
```

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

```
  Correct    fitted
1 correct 0.9933675
2 correct 0.9926289
3 correct 0.9937420
4 correct 0.9929909
```

Goodness-of-fit measures: Linear Mixed Models

- R^2 = 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

```
> cor(l$RT, fitted(lmer(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`:

```
   AIC     BIC logLik deviance REMLdev
-96.48 -63.41  55.24   -123.5  -110.5
```

- **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 is usually *negative*, and AIC, BIC etc. are positive. It can, however, happen that we observe positive likelihoods

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 $= k * ln(n) - 2ln(L)$, 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 θ-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 `lmer` package. Produces unbiased estimates for variance.
  - In practice, the estimates produced by ML and REML are nearly identical Pinheiro and Bates (2000, 11).

→ hence the two deviance terms given in the standard model output in `R`.

Goodness-of-fit: Mixed Logit Models

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

```
   AIC BIC logLik deviance
499.1 537 -242.6    485.1
```

  - 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 `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*)
  - **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).

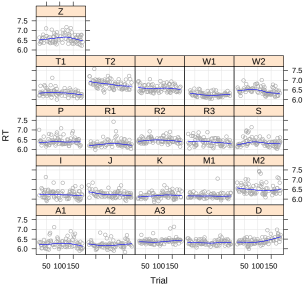
Model comparison

- 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 χ-square distribution with $DF(model_{super}) - DF(model_{nested})$ degrees of freedom.
- χ-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)
            Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
super.lmer   5 -910.41 -883.34  460.20
nested.lmer  7 -940.71 -902.81  477.35 34.302      2   3.56e-08 ***
```

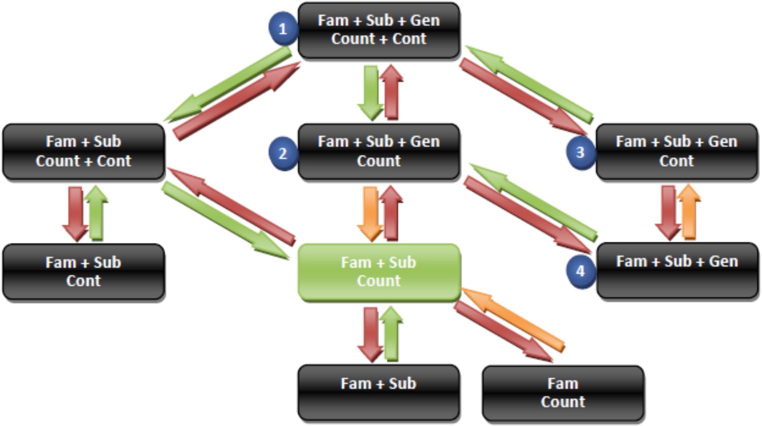
→ change in log-likelihood justifies inclusion Subject-specific **slopes** for Trial, and the **correlation parameter** between trial intercept and slope.

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.
- See lecture on **Random effects** (introduction to GLMMs).

An example

- From Jaeger, Graff, Croft, and Pontillo (2011):



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.
  - `http://www.agrocampus-ouest.fr/math/useR-2009/slides/SanchezEspigares+Ocana.pdf`

What to report?

- Describe your model
- State enough for readers and reviewers to assess whether they can trust the model
- Summarize your results

Model Description

- 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

Model Description

- 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 Florian Jaeger (2013)]

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).

Model Description - Example

> *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 Florian Jaeger (2013)]

Outlier Exclusion

- 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.*

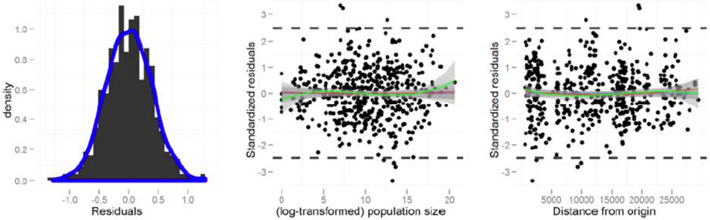
  [Fine and Florian Jaeger (2013)]

Model assumptions

- *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 et al. (2011) – Checking assumptions about the distribution of residuals in a linear mixed model:

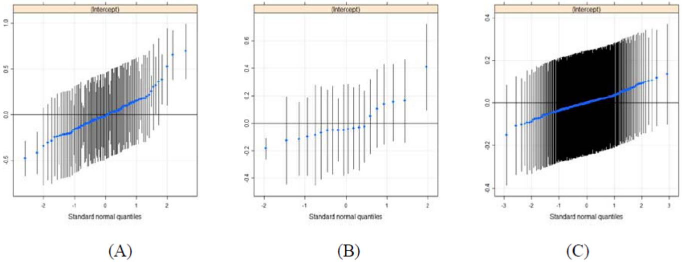


(A) (B) (C)

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 Chiini; Sino-Tibetan: Garo, Naxi, and Newari) strengthens both the population and the distance effect.

Diagnostic plots - random effects

- From Jaeger et al. (2011) – Checking assumptions about the distribution of random effects:

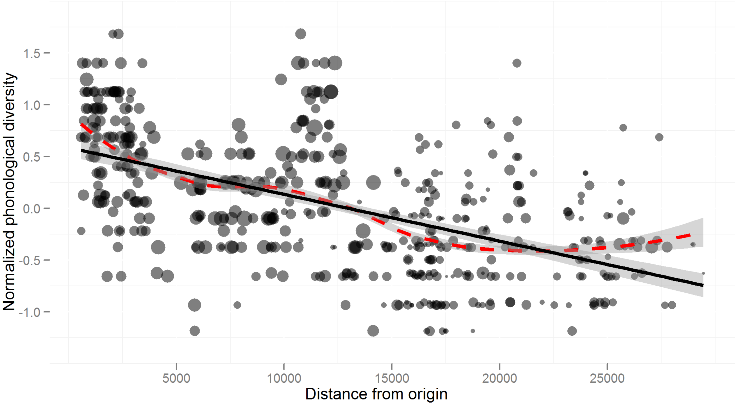


(A) (B) (C)

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.

Diagnostic plots - linearity

- From Jaeger et al. (2011) – 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 ($r = -.59$; 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 Florian Jaeger (2013)]

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 (↷**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)).

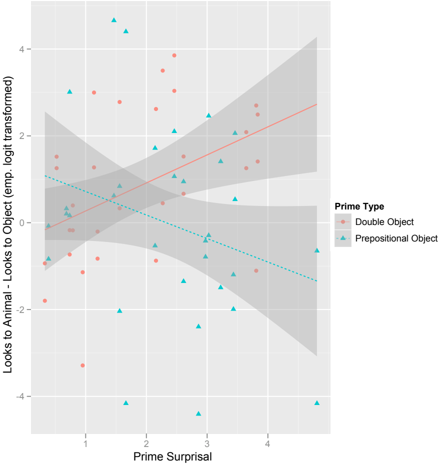
Result Summary: Text Example

> *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 Florian Jaeger (2013)]

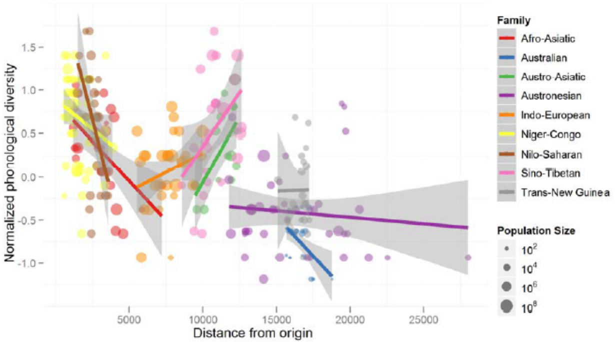
Result Summary: Visualization Example

- From Fine and Florian Jaeger (2013) – Visualize (preferably on original, interpretable scales):



Result Summary: Visualization Example

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



(B)

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.

```
> 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`”.

```
> eff/range * 100
[1] -4.606494
```

‘Back-transforming coefficients’

```
 Fixed effects:
                     Estimate Std. Error t value
(Intercept)           6.323783   0.037419  169.00
NativeLanguageOther   0.150114   0.056471    2.66
cFrequency           -0.039377   0.005552   -7.09
```

- 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.

Result Summary: Visualization Example

- Often there is a trade-off between visualizing fit and using an intuitive scale:

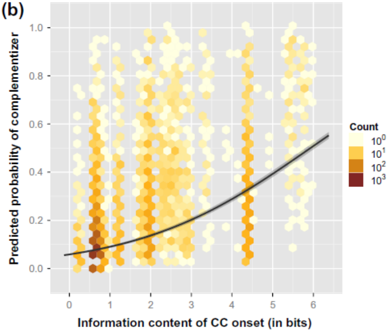
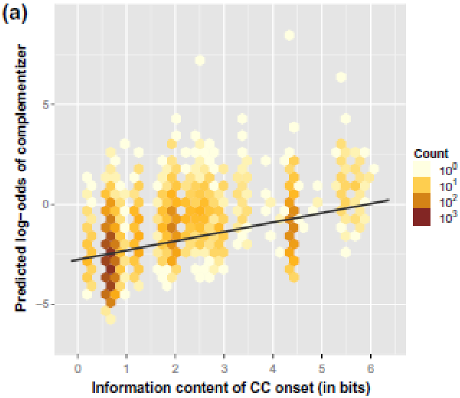


Fig. 4. Effect of information density at the complement clause onset on *that*-mentioning along with 95% CIs (shaded area, which is hard to see because the CIs are very narrow around the predicted mean effect). (a) The effect on the log-odds of complementizer *that* (the space in which the analysis was 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

38 *T. Florian Jaeger / Cognitive Psychology 61 (2010) 23–62*



Result Summary: Table Example

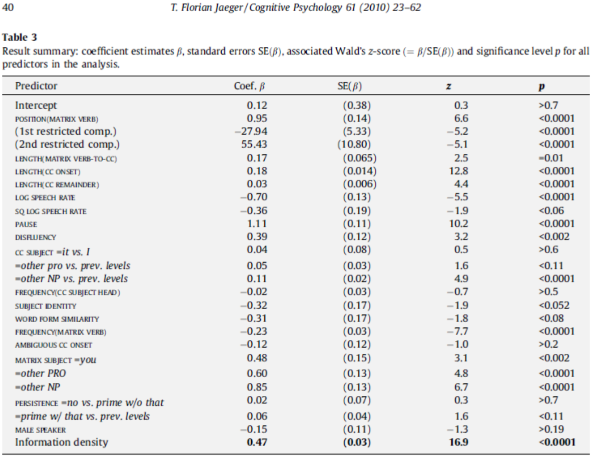
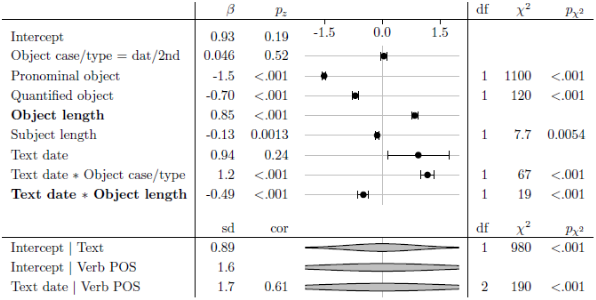


Table 3
Result summary: coefficient estimates β, standard errors SE(β), associated Wald's *z*-score (= β/SE(β)) and significance level *p* for all predictors in the analysis.

Predictor	Coef. β	SE(β)	*z*	*p*
Intercept	0.12	(0.38)	0.3	>0.7
POSITION(MATRIX VERB)	0.95	(0.14)	6.6	<0.0001
(1st restricted comp.)	−27.94	(5.33)	−5.2	<0.0001
(2nd restricted comp.)	55.43	(10.80)	−5.1	<0.0001
LENGTH(MATRIX VERB-TO-CC)	0.17	(0.065)	2.5	=0.01
LENGTH(CC ONSET)	0.18	(0.014)	12.8	<0.0001
LENGTH(CC REMAINDER)	0.03	(0.006)	4.4	<0.0001
LOG SPEECH RATE	−0.70	(0.13)	−5.5	<0.0001
SQ LOG SPEECH RATE	−0.36	(0.19)	−1.9	<0.06
PAUSE	1.11	(0.11)	10.2	<0.0001
DISFLUENCY	0.39	(0.12)	3.2	<0.002
CC SUBJECT =*it vs. I*	0.04	(0.08)	0.5	>0.6
=*other pro vs. prev. levels*	0.05	(0.03)	1.6	<0.11
=*other NP vs. prev. levels*	0.11	(0.02)	4.9	<0.0001
FREQUENCY(CC SUBJECT HEAD)	−0.02	(0.03)	−0.7	>0.5
SUBJECT IDENTITY	−0.32	(0.17)	−1.9	<0.052
WORD FORM SIMILARITY	−0.31	(0.17)	−1.8	<0.08
FREQUENCY(MATRIX VERB)	−0.23	(0.03)	−7.7	<0.0001
AMBIGUOUS CC ONSET	−0.12	(0.12)	−1.0	>0.2
MATRIX SUBJECT =*you*	0.48	(0.15)	3.1	<0.002
=*other PRO*	0.60	(0.13)	4.8	<0.0001
=*other NP*	0.85	(0.13)	6.7	<0.0001
PERSISTENCE =*no vs. prime w/o that*	0.02	(0.07)	0.3	>0.7
=*prime w/ that vs. prev. levels*	0.06	(0.04)	1.6	<0.11
MALE SPEAKER	−0.15	(0.11)	−1.3	>0.19
Information density	**0.47**	**(0.03)**	**16.9**	**<0.0001**

Result Summary: Table Example

- From a draft of Tily (2010):



	β	p_z	df	χ^2	p_{χ^2}
Intercept	0.93	0.19			
Object case/type = dat/2nd	0.046	0.52			
Pronominal object	-1.5	<.001	1	1100	<.001
Quantified object	-0.70	<.001	1	120	<.001
Object length	0.85	<.001			
Subject length	-0.13	0.0013	1	7.7	0.0054
Text date	0.94	0.24			
Text date * Object case/type	1.2	<.001	1	67	<.001
Text date * Object length	-0.49	<.001	1	19	<.001
	sd	cor	df	χ^2	p_{χ^2}
Intercept | Text	0.89		1	980	<.001
Intercept | Verb POS	1.6				
Text date | Verb POS	1.7	0.61	2	190	<.001

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

Barr, D. (2008). Analyzing ‘visual world’eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474.

Fine, A. B., & Florian Jaeger, T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive science*.

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, *59*(2), 127–136.

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

Jaeger, T. F. (2011). Corpus-based research on language production: Information density and reducible subject relatives. In E. M. Benders & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage, and processing. studies in honor of tom wasow* (p. 161-197). Stanford: CSLI Publications.

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

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (in press). Mixed effect models for genetic and areal dependencies in linguistic typology:

Commentary on atkinson. *Linguistic Typology*.

Jaeger, T. F., & Kuperman, V. (2009). *Standards in fitting, evaluating, and interpreting regression models.* UC Davis. Available from `http://hlplab.wordpress.com` (Presentation give at the Workshop on Ordinary and Multilevel Modeling)

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in s and s-plus*. Springer Verlag.

Tily, H. (2010). *The role of processing complexity in word order variation and change*. Unpublished doctoral dissertation.