Lecture 6: Reporting Results

LSA 2013, LI539

Mixed Effect Models

T. Florian Jaeger

Brain and Cognitive Sciences
University of Rochester

July 17, 2013

Data

- Our example data comes from a 2x2 design, where both factors and their interaction are **within-subjects** and **within-items**.

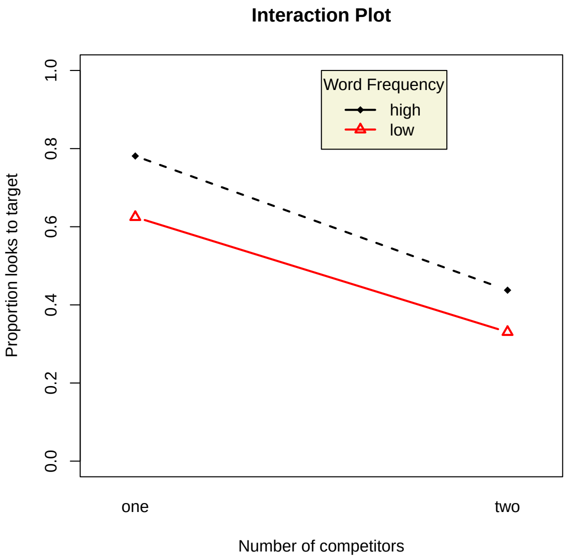
d = read.csv(file = ”eye-tracking-sample.csv”) str(d) ¿ ’data.frame’: 35236 obs. of 11 variables: $Subj : int 1111111111...$ Item : int 1 1 1 1 1 1 1 1 1 1 ... $Sample : int 12345678910...$ Time : num 0.004 0.008 0.012 0.016 0.02 0.024 ... $cTime : num - 0.198 - 0.194 - 0.19 - 0.186 - 0.182...$ Bin : int 1 1 1 1 1 1 1 1 1 1 ...
$CondWordFrequency : Factor w/ 2 levels ”high”, ”low” : 111111...$
cCondWordFrequencyHigh: num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
$CondCompetitors : Factor w/ 2 levels ”one”, ”two” : 22222...$
cCondCompetitionHigh : num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
$LooksToTarget : int 0011001011...$

Repeated measures ANOVA for 2x2 within-factors

- Let's start with the F1 analysis (by-participant). For now we collapse over time.
- For quick help for ANOVA in R, see `http://www.statmethods.net/stats/anova.html`

d.agg = aggregate(d[,c('LooksToTarget')], by= list(Subj = d$Subj, CondWordFrequency = d$CondWordFrequency, CondCompetitors = d$CondCompetitors), FUN = mean) This variables stores average proportions of looks to target (in the original d$
"LooksToTarget" str(d.agg)$ ¿ 'data.frame': 64 obs. of 4 variables: $Subj : int 12345678910...$ CondWordFrequency: Factor w/ 2 levels "high","low": 1 1 1 1 1 ... $CondCompetitors : Factor w/ 2 levels "one","two" : 11111...$ LooksToTarget : num 0.815 0.784 0.772 0.806 0.561 ...

Visualizing the data



F1 and F2 Results

```
m.F1 = aov(LooksToTarget   CondWordFrequency * CondCompetitors +
Error(Subj/(CondWordFrequency * CondCompetitors)), data = d.agg)
summary(m.F1)
```

- ... and similarly for F2 (by aggregating by item) ...

```
[F1:] Df Sum Sq Mean Sq F value Pr(¿F) CondWordFrequency 1 0.08428
0.08428 3.9443 0.0519329 . CondCompetitors 1 0.36136 0.36136 16.9112
0.0001295 *** CondWordFrequency:CondCompetitors 1 0.01749 0.01749
0.8187 0.3694427 Residuals 56 1.19661 0.02137
[F2] Df Sum Sq Mean Sq F value Pr(¿F) CondWordFrequency 1 0.2626
0.26261 10.9929 0.001004 ** CondCompetitors 1 2.1119 2.11193 88.4064 ¡
2.2e-16 *** CondWordFrequency:CondCompetitors 1 0.0415 0.04150
1.7373 0.188284 Residuals 376 8.9822 0.02389
```

Mixed Linear Model

- Here, we are aggregating by subject and item at the same time.

d.agg = aggregate(d[,c('LooksToTarget')], by= list(Subj = d$Subj, Item = d$Item, CondWordFrequency = d$CondWordFrequency, CondCompetitors = d$CondCompetitors), FUN = mean) This variables stores average proportions of looks to target names(d.agg)[length(names(d.agg))] = "LooksToTarget"

Coding the two Factors

- Here, I am first dummy coding and then centering the predictors.
- That's essentially the same as contrast/sum-coding the predictors, which is sometimes also called ANOVA coding.

d.agg$cCondWordFrequencyHigh =$
$myCenter(ifelse(d.agg$CondWordFrequency == ”high”, 1, 0))
d.agg$cCondCompetitorsTwo = myCenter(ifelse(d.agg$CondCompetitors
== ”two”, 1, 0))

Why center?

- Higher-order terms (e.g. interactions or higher order terms of polynomials) are likely to be collinear with the lower order effects.

Using ANOVA-coding

- For a highly balanced data set like the current one (check it for yourself), I could just sum-code the predictors.

contrasts(d.agg$CondWordFrequency) = contr.sum(2) contrasts(d.agg$CondCompetitors) = contr.sum(2)

NB: R will assign the value 1 to the alphabetically first level of the factor and -1 to the second level

NB: Under this coding the distance between the two conditions is *two* units.

contrasts(d.agg$CondWordFrequency)$ [,1] high 1 low -1

- For lots of information on predictor coding, see Maureen Gillespie's tutorial: `http://wiki.bcs.rochester.edu:2525/HlpLab/StatsCourses?action=AttachFile&do=get&target=gillespie-tutorial.pdf`

Mixed Linear Model

- We start with a model with a **full random effect structure** yielding:

Formula: LooksToTarget   cCondWordFrequencyHigh * cCondCompetitorsTwo + (1 + cCondWordFrequencyHigh * cCondCompetitorsTwo — Subj) + (1 + cCondWordFrequencyHigh * cCondCompetitorsTwo — Item) Data: d.agg AIC BIC logLik deviance REMLdev -1016 -917.6 533.2 -1094 -1066 Random effects: Groups Name Variance Std.Dev. Corr Item (Intercept) 0.004607 0.067880 cCondWordFrequencyHigh 0.000439 0.020973 -0.503 cCondCompetitorsTwo 0.000479 0.021897 0.282 -0.392 cCondWFreqHigh:cCondCompTwo 0.000720 0.026849 0.830 0.012 -0.142 Subj (Intercept) 0.017818 0.133484 cCondWordFrequencyHigh 0.000188 0.013739 -1.000 cCondCompetitorsTwo 0.001281 0.035795 0.777 -0.777 cCondWFreqHigh:cCondCompTwo 0.002792 0.052843 0.778 -0.778 0.583 Residual 0.001919 0.043816 Number of obs: 384, groups: Item, 24; Subj, 16

Fixed effects: Estimate Std. Error t value (Intercept) 0.54356 0.03620 15.015 cCondWordFrequencyHigh 0.13165 0.00708 18.596 cCondCompetitorsTwo -0.31924 0.01096 -29.136 cCondWordFrequencyHigh:cCondCompetitorsTwo -0.04928 0.01687 -2.921 [...]

Weighted linear regression over empirical logits

Formula: emplog(LooksToTarget, TotalLooks$x)\ cCondWordFrequencyHigh * cCondCompetitorsTwo + (1 + cCondWordFrequencyHigh * cCondCompetitorsTwo | Subj) + (1 + cCondWordFrequencyHigh * cCondCompetitorsTwo | Item) Data : d.aggAICBIClogLikdevianceREMLdev12561355 - 60311901206 Randomeffects : GroupsNameVarianceStd.Dev.CorrItem(Intercept)5.9687e - 030.0772574cCondWordFrequencyHigh5.4188e - 040.0232783 - 0.077cCondCompetitorsTwo6.7981e - 040.0260731 - 0.172 - 0.413cCondWFreqHigh : cCondCompTwo1.3549e - 040.01164020.1400.968 - 0.331Subj(Intercept)2.3187e - 020.1522720cCondWordFrequencyHigh3.6433e - 060.00190871.000cCondCompetitorsTwo2.8312e - 040.01682621.0001.000cCondWFreqHigh : cCondCompTwo2.5864e - 040.01608220.0870.0870.087Residual4.2018e - 020.2049839[...]EstimateStd.Errortvalue(Intercept)0.2189890.0412845.30cCo$
$1.4831380.008490 - 174.70cCondWordFrequencyHigh : cCondCompetitorsTwo - 0.3772460.011098 - 33.99[...]$

Are all random effects justified?

- Fit reduced model with just random intercepts for subject and item and compare it to full model:

m.full = lmer(emplog(LooksToTarget, TotalLooks$x$) $cCondWordFrequencyHigh * cCondCompetitorsTwo + (1 + cCondWordFrequencyHigh * cCondCompetitorsTwo|Subj) + (1 + cCondWordFrequencyHigh * cCondCompetitorsTwo|Item), d.agg, family = ''gaussian'', weight = emplogweight(LooksToTarget, TotalLooks$x)) m.simple = lmer(emplog(LooksToTarget, TotalLooks$x$) $cCondWordFrequencyHigh * cCondCompetitorsTwo + (1|Subj) + (1|Item), d.agg, family = ''gaussian'', weight = emplogweight(LooksToTarget, TotalLooks$x)) anova(m.full,m.simple) Models: [...] Df AIC BIC logLik Chisq Chi Df Pr(¿Chisq) m.simple 7 1223.3 1250.9 -604.63 m.full 25 1239.8 1338.6 -594.91 19.446 18 0.3648 The full random effect structure does not result in model that fits the data significantly better given the increase in complexity (number of parameter). Can we just stop here?

‘Maximal random effect structure justified by the data’

- Based on model comparison, we find that the following model contains the **maximal random effect structure justified by the data** (see script for details):

emplog(LooksToTarget, TotalLooks$x)\ cCondWordFrequencyHigh * cCondCompetitorsTwo + (1 + cCondCompetitorsTwo|Subj) + (1 + cCondWordFrequencyHigh + cCondCompetitorsTwo|Item)$
AIC BIC logLik deviance REMLdev 1235 1290 -603.5 1191 1207 Random effects: Groups Name Variance Std.Dev. Corr Item (Intercept) 0.00599874 0.077452 cCondWordFrequencyHigh 0.00056475 0.023764 -0.058 cCondCompetitorsTwo 0.00067979 0.026073 -0.181 -0.464 Subj (Intercept) 0.02310328 0.151998 cCondCompetitorsTwo 0.00028816 0.016975 1.000 Residual 0.04252746 0.206222 Number of obs: 384, groups: Item, 24; Subj, 16 [...]

‘Maximal random effect structure justified by the data’

```
[...] Fixed effects: Estimate Std. Error t value (Intercept) 0.219100 0.041236
5.31 cCondWordFrequencyHigh 0.652125 0.007010 93.03
cCondCompetitorsTwo -1.483316 0.008526 -173.98
cCondWordFrequencyHigh:cCondCompetitorsTwo -0.377856 0.010117
-37.35
Correlation of Fixed Effects: (Intr) cCnWFH cCndCT cCndWrdFrqH -0.012
cCndCmpttrT 0.412 -0.251 cCndWFH:CCT -0.006 -0.063 0.051
```

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 ? (?), proportion of fixations to the animal and the object were first empirical logit-transformed [...]*

  [? (?)]

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).*

[? (?)]

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.*

  [? (?)]

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

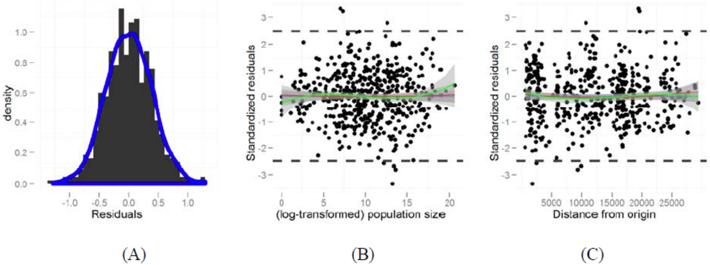
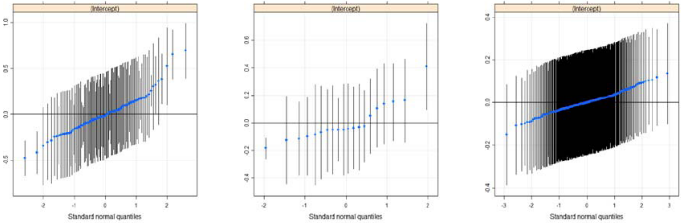


Figure 3 Diagnostic plots for the model described in Equation (E3). The histogram of residuals (the individual-level errors) in (A) suggests normality. Linear (blue), quadratic (red), and cubic fits (green) of log-transformed population size in (B) or distance from the origin in (C) against the standardized 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, Graff, Croft, and Pontillo (in press) – Checking assumptions about the distribution of random effects:

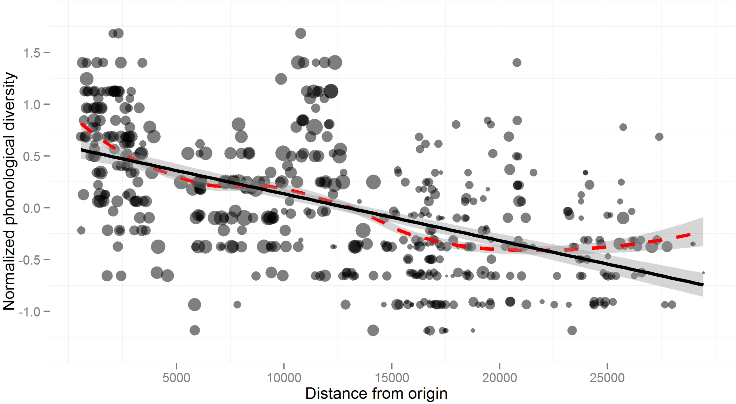


(A) (B) (C)

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

- (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 ($r = -.59$; 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 (↷Lecture 3)

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; ↷Lecture 3).
- 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. ? (?); 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. ? (?)).

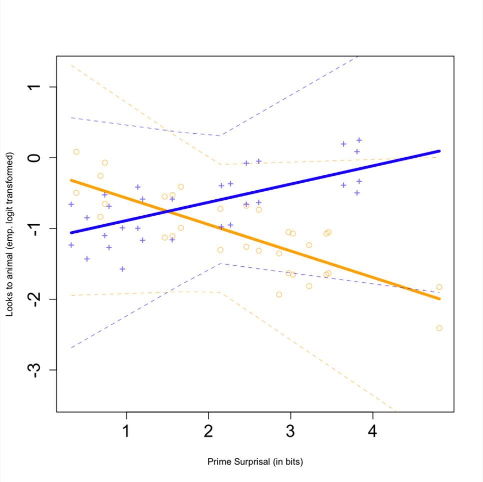
Result Summary: Text Example

> *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.*

[? (?)]

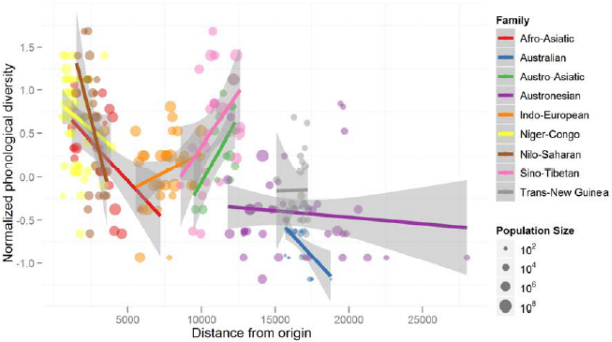
Result Summary: Visualization Example

- From ? (?) – Visualize (preferably on original, interpretable scales):



Result Summary: Visualization Example

- From ? (?) – 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:

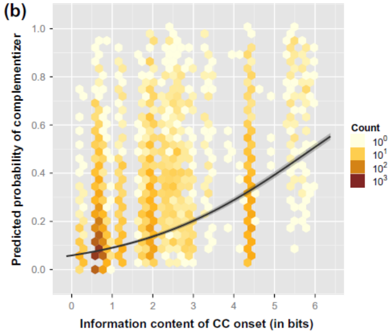
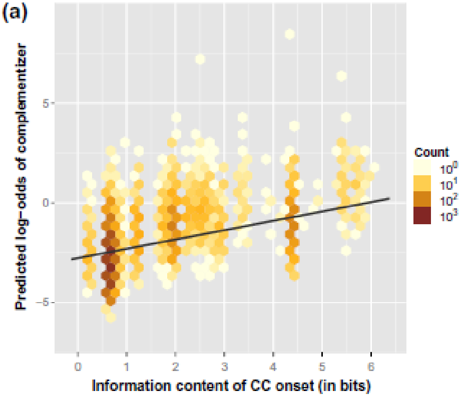


Fig. 4. Effect of information density at the complement clause onset on *that*-mentioning along with 95% CIs (shaded area, which is hard to see because the CIs are very narrow around the predicted mean effect). (a) The effect on the log-odds of complementizer *that* (the space in which the analysis was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



Result Summary: Table Example

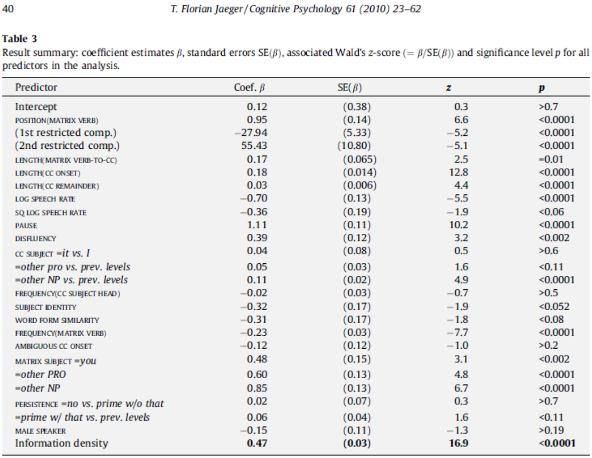


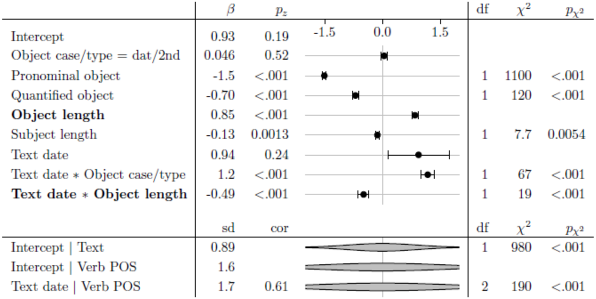
Table 3

Result summary: coefficient estimates β, standard errors SE(β), associated Wald's *z*-score ($= \beta/\mathrm{SE}(\beta)$) 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 Harry Tily's (2011) thesis:



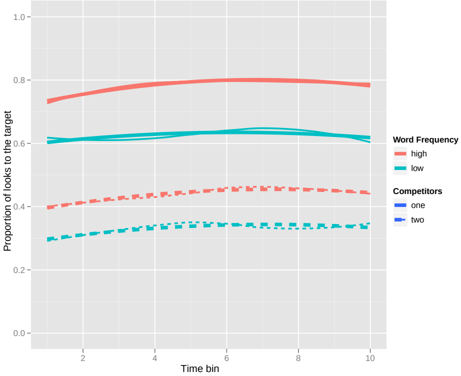
	β	p_z	df	χ^2	p_{χ^2}
Intercept	0.93	0.19			
Object case/type = dat/2nd	0.046	0.52			
Pronominal object	-1.5	<.001	1	1100	<.001
Quantified object	-0.70	<.001	1	120	<.001
Object length	0.85	<.001			
Subject length	-0.13	0.0013	1	7.7	0.0054
Text date	0.94	0.24			
Text date * Object case/type	1.2	<.001	1	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)

Returning to our Example: The Time Course

- Local smoother in general additive model (thin lines) and quadratic fit in binomial GLM (thick lines) for the four conditions over the time bins:



Mixed Logit Model

Formula: LooksToTarget   cCondWordFrequencyHigh * cCondCompetitorsTwo * pol(cTime, 2) + (1 — Subj) + (1 — Item) Data: d AIC BIC logLik deviance 40765 40883 -20368 40737 Random effects: Groups Name Variance Std.Dev. Item (Intercept) 0.11565 0.34007 Subj (Intercept) 0.42845 0.65456 Number of obs: 35236, groups: Item, 24; Subj, 16

Fixed effects: Estimate Std. Error z value Pr(¿—z—) (Intercept) 0.292108 0.178705 1.63 0.1021 cCondWordFrequencyHigh 0.703270 0.036675 19.18 ¡ 2e-16 cCondCompetitorsTwo -1.538732 0.036954 -41.64 ¡ 2e-16 pol(cTime, 2)cTime 0.579599 0.104796 5.53 3.19e-08 pol(cTime, 2)cTime$^2 - 4.5363251.016127 - 4.468.03e - 06cCondWFqHigh : cCondCompetitorsTwo - 0.4764650.073144 - 6.517.31e - 11cCondWFqHigh : cTime0.4184180.2095672.000.0459cCondWFqHigh : cTime^2 - 1.8578472.032027 - 0.910.3606cCondCompTwo : cTime - 0.0048280.209584 - 0.020.9816cCondCompTwo : cTime^2 0.6716212.0322960.330.7410cCondWFqHigh : cCondCompTwo : cTime - 0.5742470.419159 - 1.370.1707cCondWFqHigh : cCondCompTwo : cTime^2 4.9979244.0646801.230.2188$

Disclaimer

- Generated data
  - α = .3
  - $\beta_{WordFrequencyHigh} = .7$
  - $\beta_{CompetitorsTwo} = -1.5$
  - $\beta_{WordFrequencyHigh:CompetitorsTwo} = -.3$
  - $\beta_{time} = .5$ and $\beta_{time^2} = -1.5$
  - $\beta_{WordFrequencHigh:Time} = .4$
  - $\beta_{CompetitorsTwo:Time} = .1$
  - $\beta_{WordFrequencHigh:CompetitorsTwo:Time} = -.8$
  - $\sigma_{\alpha_{Subject}} = 0.5$ and $\sigma_{\alpha_{Item}} = 0.15$
- With data loss rates differing between individual participants ($\mu = 3\%$)

The magic of the ‘original’ scale

- ★ What’s the advantage of having an effect size in familiar units?
  - Comparability across experiments?
  - Intuitive idea of ‘how much’ factor (and mechanisms that predicts it to matter) accounts for?
- ★ But this may be misleadingly intuitive . . .
  - If variables are related in non-linear ways, then *that’s how it is*.
  - If residualization is necessary then it’s applied for a good reason → back-translating will lead to misleading conclusions (there’s only so much we can conclude in the face of collinearity).
  - Most theories don’t make precise predictions about effect sizes on ‘original’ scale anyway.
  - Comparison across experiments/data sets often only legit if similar stimuli (with regard to values of predictors).

Comparing effect sizes

- It ain’t trivial: What is meant by effect size?
  - Change of outcome if ‘feature’ is present? → coefficient
    - per unit?
    - overall range?
  - But that does not capture how much an effect affects language processing:
    - What if the feature is rare in *real language use* (‘availability of feature’)? Could use . . .

    → Variance accounted for (**goodness-of-fit** ↶ improvement associated with factor)

    → **Standardized coefficient** (gives direction of effect)

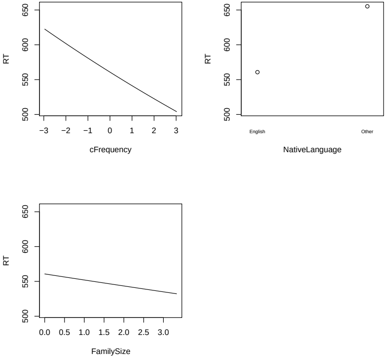
★ **Standardization**: subtract the mean and divide by two standard deviations.

- standardized predictors are on the same scale as binary factors (cf. Gelman & Hill 2006).
- makes all predictors (relatively) comparable.

Plotting coefficients of linear models

Plotting (partial) effects of predictors allows for comparison and reporting of their effect sizes:

- partial fixed effects can be plotted, using `plotLMER.fnc()`. Option `fun` is the back-transformation function for the outcome. Effects are plotted on the same scale, easy to compare their relative weight in the model.



- confidence intervals (obtained by MCMC-sampling of posterior distribution) can be added.

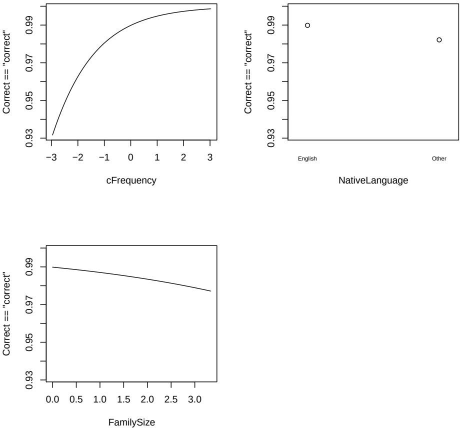
Plotting posterior distributions (for linear mixed models)

- `pvals.fnc()` plots MCMC-sampling posterior distributions, useful for inspection of whether the distributions are well-bounded.

figs/posterior.pdf

Plotting coefficients of mixed logit models

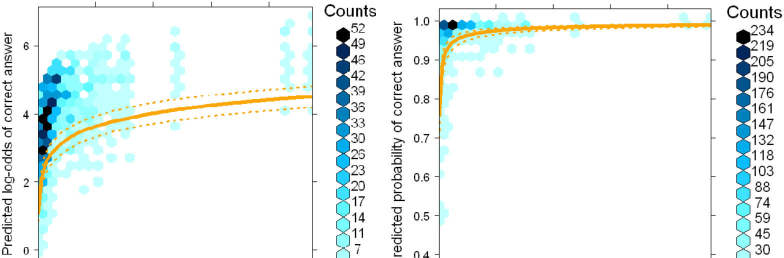
- Log-odd units can be automatically transformed to probabilities.
  - **pros:** more familiar space
  - **cons:** effects are linear in log-odds space, but non-linear in probability space; linear slopes are hard to compare in probability space; non-linearities in log-odd space are hard to interpret



Plotting coefficients of mixed logit models (contd’)

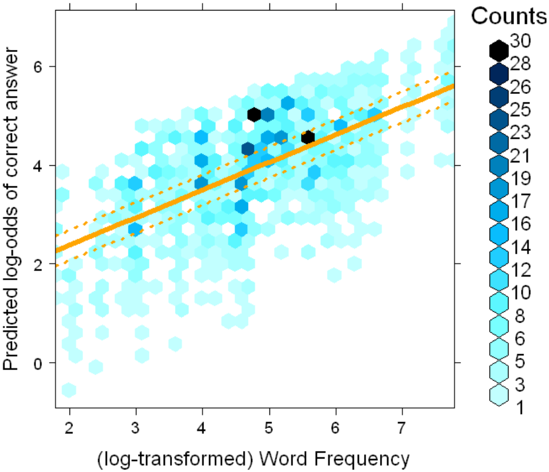
- For an alternative way, see *http://hlplab.wordpress.com/*:

¿ data(lexdec) ¿ lexdec$NativeEnglish = ifelse(lexdec$NativeLanguage == ”English”, 1, 0) ¿ lexdec$rawFrequency = exp(lexdec$Frequency) ¿ lexdec$cFrequency = lexdec$Frequency - mean(lexdec$Frequency) > lexdec$cNativeEnglish = lexdec$NativeEnglish - mean(lexdec$NativeEnglish) ¿ lexdec$Correct = ifelse(lexdec$Correct == ”correct”, T, F) ¿ l¡- lmer(Correct cNativeEnglish + cFrequency + Trial + + (1 — Word) + (1 — Subject), data = lexdec, family=”binomial”) ¿ my.glmerplot(l, ”cFrequency”, predictor= lexdec$rawFrequency, +predictor.centered = T, predictor.transform = log, +name.outcome = ”correctanswer”, xlab = ex, fun = plogis)$



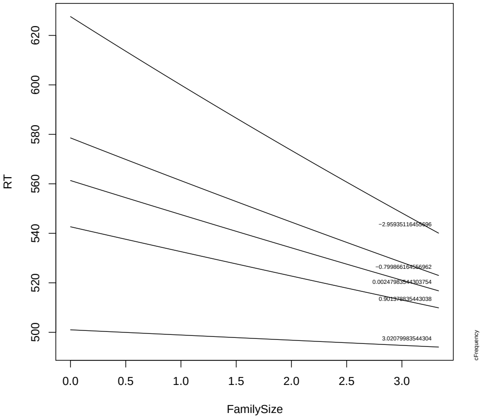
Plotting coefficients of mixed logit models (contd')

- Great for outlier detection. Plot of predictor in log-odds space (actual space in which model is fit):



Plotting interactions

¿ plotLMER.fnc(l, pred = ”FamilySize”, intr = list(”cFrequency”, ¿ quantile(lexdec$cFrequency), ”end”), fun = exp)$



- Can also be plotted as the `FamilySize` effect for levels of `cFrequency`. Plotting and interpretation depends on research hypotheses

Reporting interactions

- Report the p-value for the interaction as a whole, not just p-values for specific contrasts. For linear models, use `aovlmer.fnc()` in `languageR`.

¿ aovlmer.fnc(lmer(RT NativeLanguage + cFrequency * FamilySize + ¿ (1— Subject) + (1—Word), data = lexdec), mcmcm = mcmcSamp) Analysis of Variance Table Df Sum Sq Mean Sq F value F Df2 p NativeLanguage 1 0.20 0.20 6.5830 6.5830 1654.00 0.01 cFrequency 1 1.63 1.63 54.6488 54.6488 1654.00 2.278e-13 FamilySize 1 0.05 0.05 1.6995 1.6995 1654.00 0.19 cFrequency:FamilySize 1 0.03 0.03 1.0353 1.0353 1654.00 0.31

→ `FamilySize` and its interaction with `cFrequency` do not reach significance in the model.

Some thoughts for discussion

- What do we do when *what's familiar* (probability space; original scales such as msecs; linear effects) is not *what's best/better*?
- More flexibility and power to explore and understand complex dependencies in the data do not come for free, they require additional education that is not currently standard in our field.
  - Let's distinguish challenges that relate to complexity of our hypothesis and data vs. issues with method (regression).
  - cf. What's the best measure of effect sizes? What to do when there is collinearity? Unbiased vs. biased variance estimates for ML-fitted models; accuracy of laplace approximation.