Model selection

Strolling through the garden of forking paths

Daniela Palleschi

Humboldt-Universität zu Berlin

2024-02-09

Learning Objectives

Today we will learn about…

- the history of mixed models (again)
- strategies for model selection
- variability in model selection

Resources

- relevant papers for this topic
  - Barr et al. (2013)
  - Bates et al. (2015)
  - Matuschek et al. (2017)
  - Brauer & Curtin (2018)
  - Meteyard & Davies (2020)

Review: random intercepts and slopes

History of LMMs revisited

- recall Clark (1973)'s language-as-fixed-effect fallacy and the issue of generalisability (see also Winter & Grice, 2021; Yarkoni, 2022)
- Baayen et al. (2008) motivated LMM's for linguistic data in the JML special issue
  - effect: everybody adopted random-intercepts only models
- Barr et al. (2013): random-intercepts only models are overconfident, "keep it maximal!"
  - effect: everybody adopted maximal models
- Matuschek et al. (2017) and Bates et al. (2015): maximal models are underconfident and lower statistical power! Use data-driven model selection to find a "parsimonious" model!
  - effect: some people adopt this method, but many psycholinguists just want a "recipe" to follow

2013: Keep it maximal

> A maximal model should optimize generalization of the findings to new subjects and new items.
>
> – Barr et al. (2013), p. 261

- random-intercepts-only models tend to be underpowered
- for this reason, Barr et al. (2013) suggested using a maximal random effects structure justified by the experimental design

2015 & 2017: Parsimonious models

> [W]hile the maximal model indeed performs well as far as Type I error rates were concerned, power decreases substantially with model complexity.
>
> — Matuschek et al. (2017), p. 310-311

- there is a trade-off between Type I (overconfidence) and Type II error (underconfidence)
- i.e., maximal models can lead to over-fitting
  - lowers statistical power, which increases Type II error (false rejection)
- but we should strive for the most *parsimonious* model
  - parsimonous models are a compromise between the maximal model justified by your design and theory, and a given data set
- best way to maintain low Type I and II error: collect lots of data

Model building

> Every statistical model is a description of some real or hypothetical state of affairs in the world.
>
> – Yarkoni (2022), p. 2

- our models reflect not only our hypotheses or effects of interest, but any other plausible or known co-variates
- our predictors should reflect our research questions or theories tested
  - plus any plausible or previously motivated co-variates (e.g., trial order)
  - plus known sources of nonindependence, i.e., our random effects

Choosing predictors

- your model should be defined *a priori*
  - i.e., you should define what predictors you will include and any covariates
  - e.g., if you have a prediction about the effect of phonological neighbours on vowel duration
    - define what phonological characteristics you will include (e.g., place of articulation? manner?)
    - these should be related to specific hypotheses/research questions

Choosing a maximal random effects structure (RES)

- how do we define our maximal model? Some tips from Barr et al. (2013)
  - between-unit factor (e.g., age): include random intercept only
  - within-unit factor with multiple observations per unit-level (e.g., age in longitudinal data): include random slopes
- all factors in an interaction are within-unit: include by-unit random slopes for interaction terms

Example: Biondo et al. (2022)

- Biondo et al. (2022): 2x2 design
  - verb-tense (past, future) and grammaticality (grammatical, ungrammatical)
  - repeated-measures: within-participant and -item design
    - so we should have by-participant and -item random intercepts (multiple observations per unit level)
  - each participant and item contributed multiple data points *per condition* (i.e., tense and grammaticality were manipulated within each unit level)
    - so we should have varying tense and grammaticality slopes by- item and -participant

▶ Code

Observations per cell

- if there is only a single observation per cell, e.g., you collected one observation from each participant per condition, then you can’t fit random intercepts or slopes
- ideally you would have at least 5 observations per cell (per unit level per condition, e.g., each participant has at least 5 observations per condition)
- this is also a question of statistical power

```r
# obvz per sj per condition
df_biondo |>
  filter(roi == 4) |>
  count(sj, verb_t, gramm) |>
  count(n)
```

```
# A tibble: 1 × 2
      n    nn
  <int> <int>
1    16   240
```

```r
# obvz per item per condition
df_biondo |>
  filter(roi == 4) |>
  count(item, verb_t, gramm) |>
  arrange(desc(n)) |>
  count(n)
```

```
# A tibble: 1 × 2
      n    nn
  <int> <int>
1    10   384
```

Data structure

- random effects must be factors/categorical
- single observation per row
  - generally speaking, there should be n(participants) * n(items) rows
  - every fixed or random effect in your model should correspond to a column in your dataset

Variability in methods

- Meteyard & Davies (2020)
  - survey of (psychology) researchers
  - review of papers using LMMs
- insecurity in researchers re: choosing models
- great variation in papers in how models are built and reported

Researcher degrees of freedom

> What we hope to make clear is that there is no single correct way in which LMM analyses should be conducted, and this has important implications for how the reporting of LMMs should be approached.
>
> — Meteyard & Davies (2020), p. 9

- the problem:
  - ‘researcher degrees of freedom’ (Simmons et al., 2011), or ‘the garden of forking paths’ (Gelman & Loken, 2013)
  - the same data can be analysed in a variety of ways
- this leads to insecurity for many researchers

Justify and document

> Replicability and reproducibility are critical for scientific progress, so the way in which researchers have implemented LMM analysis must be entirely transparent. We also hope that the sharing of analysis code and data becomes widespread, enabling the periodic re-analysis of raw data over multiple experiments as studies accumulate over time.
>
> — Meteyard & Davies (2020), p. 9

- the (partial) solution:
  - make model building/selection decisions a priori
  - be transparent
  - share your data and code

Moving forward

- other alternatives that have fewer convergence issues:
  - Julia
    - e.g., in VS Code IDE
  - Bayesian framework (e.g., `brms` R package)
    - also run (G)LMMs, but abandons arbitrary p-values
    - instead quantifies uncertainty
- both are more more computationally powerful
  - the are not (yet) as widely used in the field

Learning objectives 🏁

Today we learned…

- the history of mixed models (again) ✅
- strategies for model selection ✅
- variability in model selection ✅

Important terms

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *arXiv Preprint*, 1–27. https://doi.org/10.48550/arXiv.1506.04967

Biondo, N., Soilemezidi, M., & Mancini, S. (2022). Yesterday is history, tomorrow is a mystery: An eye-tracking investigation of the processing of past and future time reference during sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 1001–1018. https://doi.org/10.1037/xlm0001053

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time*.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092