Regression Modeling Strategy

Basic Strategy:

- Look at the data.
- Look at the data.
- Look at the data.
- Look at the data. Select preliminary variables of interest.
- *Then* try a model.

- Check the model for accuracy, overfitting, . . .
- Check the model assumptions (nonlinearities? collinearity?).
- Eliminate unnecessary variables; transform nonlinear variables.
- Refit the model. Iterate above as necessary.
- *Then* interpret the model.

Look at the data –

- use Design library summary plots; use trellis graphics plots ('lattice' package in R)
- explore the data graphically, inspecting predictors and their relation to the response and each other

Choose type of model…

- logistic regression: for dichotomous responses ('NP NP' vs. 'NP PP'; presence vs. absence of *that*) use lrm() in Design library or glm()
- ordinal logistic regression: for discrete ordered responses ('perfect', 'marginal', 'ungrammatical') use lrm()

- ordinal logistic regression: for discrete ordered responses ("etymological age": Dutch < West Germanic < Germanic < Indo-European)
  use lrm()
- multinomial regression: for nominal responses (4 alternative possessive constructions in modern Low Saxon)
  use multinom() in nnet library or poisson regression with glm/lrm; see Venables and Ripley's MASS for examples

In textbook:

R. Harald Baayen. 2006. *Practical Data Analysis for the Language Sciences with R* (forthcoming)

- Multiple linear regression:
  - *lm*, model specification, interaction terms, sequential anova
  - *ols* (ordinary least squares in Design package), R-squared, residuals, plotting partial effects, nonlinearities (pol, rcs), collinearity, simultaneous anova and fastbw, which.influence, bootstrap validation with validate()

- Multiple logistic regression (Generalized linear models):
  - *glm* for logistic regression on tabular data (proportions), anova(..., test = “Chisq”) for binomial link function
  - *lrm* (Design package) for logistic regression on individual observations (single outcomes), simultaneous anova on partial effects, penalized maximum likelihood, AIC

Mixed models

- advanced topic
- fewer user-friendly tools
- you can roll your own

How to evaluate generalized linear mixed models

- quality of model: Baayen's concordance.fnc()
- how to validate model assumptions (to be continued)