Rutgers Linguistics Workshop on Mixed Effects Models — Common issues / solutions —

09/17/2016

Judith Degen
Stanford University

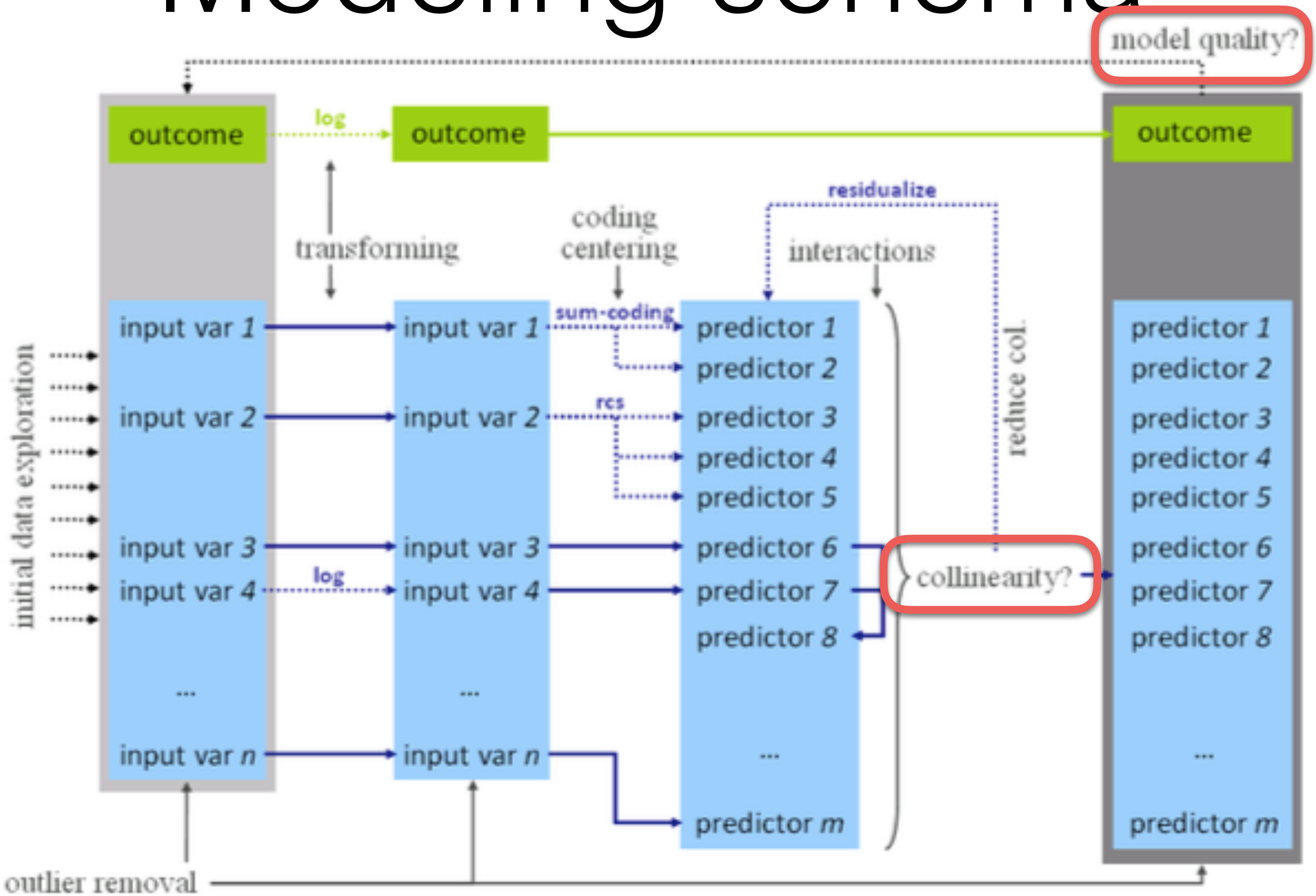
Hypothesis testing in psycholinguistic research

- often, we make predictions not just about the **existence**, but also about the **direction** of the effect

- sometimes, we're also interested in effect **shapes** (e.g., non-linearities)

- unlike ANOVA, regression analyses test hypotheses about effect **direction**, **shape**, and **size** without requiring post-hoc analyses
  - **if predictors are coded appropriately**
  - **if the model can be trusted**

Modeling schema



from Jaeger 2011

For data exploration, variable selection, transformation, coding, centering: other tutorials (e.g. on Florian Jaeger's HLPlab website)

Today

- towards a model with interpretable coefficients: dealing with *collinearity*
- model evaluation
- model comparison

Collinearity

Collinearity: predictors are collinear with each other if there are high (partial) correlations between them

Even if a predictor is not highly correlated with any single other predictor, it can be highly collinear with a combination of predictors —> collinearity will affect the predictor

This is common

- in models with many predictors
- when several somewhat related predictors are included in the model (e.g., word length & word frequency or subjecthood and information status)

Consequences of collinearity

- standard errors (SE) of collinear predictors are biased (inflated), leading to underestimation of significance (increased risk of Type II error) but sometimes to overestimation as well (Type I error)

- coefficients of collinear predictors hard to interpret
  - 'bouncing betas': minor changes in data may have major impact on βs
  - coefficients may flip sign, double, half

- model R^2 may be inflated or deflated

No conclusions about coefficients to be drawn!

Extreme collinearity: example

`meanWeight` (rating of the weight of an object denoted by the word, averaged across subjects) and **`meanSize`** (average rating of object size) in **`lexdec`**

Look at it in R….

Collinearity example

- unusually heave objects for their size tend to also be more frequent
- both effects disappear when frequency is included (though you could residualize…)
- What is the effect of collinearity?
  - Type II error increase (power loss)
  - There can be mild Type I error increases (but small differences between highly correlated predictors can be highly correlated with another predictor and create “apparent effects”, see example)

When coefficients are unstable, check for mediated effects!

Detecting collinearity

- inspect correlation matrix (partial correlations of fixed effects in the model)
- use pairscor.fnc() for visualization
- formal tests of collinearity: variance inflation factor (VIF)
    - VIF > 4 start being problematic, VIF > 10: collinearity highly likely

Dealing with collinearity

- Good news: estimates are only problematic for the collinear predictors
  - If collinearity is in the control/nuisance predictors, nothing needs to be done
- Somewhat good news: if collinear predictors are of interest but we're not interested in effect direction, we can use **model comparison** to decide which predictor, if any, to include
- If collinear predictors *are* of interest and we are interested in direction of effect, we need to reduce collinearity

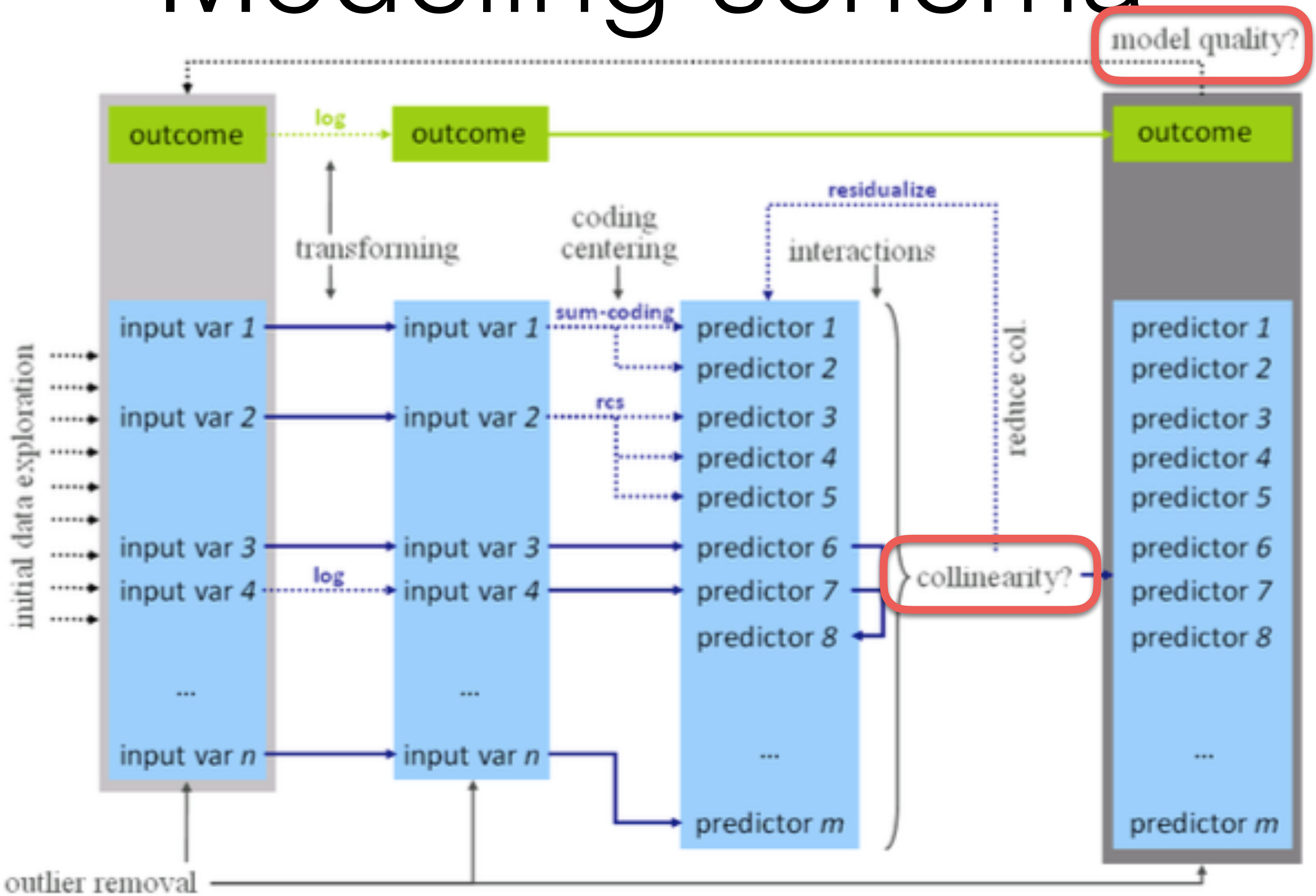
Reducing collinearity

- **center** predictors: reduces collinearity of predictor with intercept and higher level terms involving the predictor —> highly recommended (easy to do and interpret, often improves interpretability of effects)

- **re-express variable** based on conceptual considerations (not always applicable)

- **residualize:** regress collinear predictor against combination of correlated predictors (using `lm()`)
  - pro: systematic way of dealing with collinearity, directionality of effect interpretable
  - cons: effect sizes hard to interpret; judgment calls (what to residualize against what?)

Modeling schema



from Jaeger 2011

Overfitting

- **overfitting**: fit might be too tight due to excessive number of parameters (coefficients). The maximal number of predictors that a model allows depends on their distribution and distribution of outcome

- **rules of thumb:**
  - **linear models**: > 20 observations per predictors
  - **logit models**: the less frequent outcome should be observed > 10 times more often than there are predictors in the model
  - how to count predictors: one per random effect, one per fixed effect predictor, one per interaction

Validation & goodness of fit measures

- **goodness-of-fit** measures assess the relation between fitted (predicted) values and observed outcomes
  - linear models: numerical outcomes
  - logit models: predicted log-odds (and probabilities) of outcomes

Goodness-of-fit measures for 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

Data likelihood measures

- data likelihood: probability of the data given the model (ie, given the predictors and the best parameter values)

- standard model output often includes such measures, e.g.:

AIC	BIC	logLik	deviance	df.resid
498.6	536.5	-242.3	484.6	1652

- **log-likelihood**: simply the maximized model's log data likelihood. Problem: no correction for number of parameters. **Larger (closer to zero) is better.** Log-likelihoods should always be negative, the others positive.

Data likelihood measures

- measures that trade of goodness-of-fit (data likelihood) and model complexity (number of parameters)
  - **deviance** = -2 times log-likelihood ratio
  - **Akaike Information Criterion (AIC)** = k -2ln(L), where k is number of parameters
  - **Bayesian Information Criterion (BIC)** = k*ln(n) - 2ln(L), where k n is number of observations
  - **For all: smaller is better!**

Model comparison

- models can be compared for performance using any goodness-of-fit measures

- to test whether one model is significantly better than another one: **likelihood ratio tests (for nested models only!)**

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.

What to report

- goodness-of-fit measures for **linear models**: R^2; possibly additionally amount of variance explained by fixed effects over and above random effects)

- goodness-of-fit measures for **logit models**: increase in classification accuracy over and above baseline model

- for **model comparison**: p-value of log-likelihood test