Mixed-effects models and unbalanced sociolinguistic data: The need for caution

Joseph Roy (*University of Illinois at Urbana Champaign*)
Stephen Levey (*University of Ottawa*)

Recent discussion (e.g. Saito, 1999; Johnson, 2009; Gorman & Johnson, 2013; cf. Paolillo, 2002; Baayen, 2008) of the optimal techniques for analyzing quantitative data in sociolinguistics have resulted in a call for an expansion in the use of existing statistical models. A prominent component of this debate is the premium attached by certain researchers to the use of speaker-as-random effect models in the place of more traditional models based on Goldvarb analyses (Johnson, 2009; Draegar and Hay, 2012) in order to control for speaker variation claimed to be present in Goldvarb's estimates of significance and effect size for social effects.

For this study, we compare the differences in results produced by Goldvarb (Sankoff, Tagliamonate, and Smith, 2012) and those based on random effects model with intercepts for each speaker. Results from three widely studied linguistic variables are presented drawing on data from a mainstream urban variety of contemporary Canadian English: -ing variation (16 speakers, 800 tokens), quotative use (19 speakers, 857 tokens) and relative variation (37 speakers, 1077 tokens). Each dataset differs in terms of the relative balance of the data in each, constituting an ideal test case for the efficacy of mixed-effects models. The –ing data set is balanced (50 tokens per speaker). The quotative data set is moderately unbalanced (with 6 speakers having less than 20 tokens). The relative clause data set is the most unbalanced (18 speakers having less than 20 tokens). Varying degrees of data imbalance explain the difference between the estimates produced by the two statistical models across the data sets. For the most balanced data (ing), the statistical models produce identical results for the social and linguistic effects in the selection of factor groups as statistically significant, the size of the effects and even the estimates of the factor weights themselves. For the moderately unbalanced data (quotatives), both models select as statistically significant the same set of independent variables. In these data, however, there are differences in the magnitude of effect for some factor groups. Finally, for the highly unbalanced data set (relatives), the two statistical models not only select a different set of significant factor groups for both the linguistic and social factor groups, but also reverse the direction of effect within one shared linguistic factor group. We conclude that the reversal for this linguistic effect, based on the linguistic literature and marginal distribution of the data itself, is in fact an error in the direction of effect for the mixed effects model. Moreover, it is known in other fields that use subjects as random effects that analyses with less than 30-50 tokens per speaker with at least 30-50 speakers vastly overestimate variance (Moineddin, Matheson and Glazier, 2007).

The results of our comparative analysis drawing on different statistical models converge in demonstrating that sociolinguists should be cautious in applying mixed-effects models to highly unbalanced data.

References

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge University Press.

Drager, Katie, and Jennifer Hay. 2012. Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change* 24(01): 59-78.

Johnson, Daniel Ezra. 2009. Getting off the Goldvarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass* 3(1): 359–383.

Gorman, Kyle, and Daniel Ezra Johnson. 2012. Quantitative analysis. In Robert Bayley, Richard Cameron and Ceil Lucas (eds.), *The Oxford Handbook of Sociolinguistics*, pp. 214-240.

Moineddin, R., F.I. Matheson & R.H. Glazier. 2007. A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology* 7(1). 34.

Paolillo, John. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford: Center for the Study of Language and Information Publications.

Saito, Hidetoshi. 1999. Dependence and interaction in frequency data analysis in SLA research. *Studies in Second Language Acquisition*, 21:453-75.

David Sankoff, Sali A. Tagliamonte, and Eric Smith (2012) *Goldvarb Lion: A multivariate analysis application for Macintosh*.