On VARBRUL – Or, The Spirit of '74

Kyle Gorman
Department of Linguistics
Institute for Research in Cognitive Science
University of Pennsylvania
kgorman@ling.upenn.edu

October 25, 2009

At NWAV 38's panel on statistical best practices, organized by Sali Tagliamonte, several presenters vehemently argued for the use of VARBRUL, software developed by David Sankoff in the 1970s which performs stepwise logistic regression on multiple categorical predictors using the maximum likelihood method, rather than competing software with more features and a much wider userbase. In this paper, I critique six claims made by those who argue that VARBRUL is superior to generalized hierarchical linear models, a related statistical model available in many software packages. The first three claims deal with empirical issues, and the second three concern the relationship between VARBRUL and the community.

The crucial insight motivating the use of logistic regression for the variable rule research program is the *principle of accountability*. In fact, the earliest paper deploying this principle (Cedergren and Sankoff, 1974, p. 334) explicitly presents these insights as representing a paradigm shift (in the sense of Kuhn, 1962). However, a certain piece of *software* which implements only one of the many analyses available in contemporary statistical software packages is no more of a scientific paradigm than the telescope was for Herschel (or Galileo) a theory of astronomy (cf. Feyerabend, 1975, chap. 9). VARBRUL can do no more than observe association between variable linguistic behaviors and internal and external factors coded by the experimenter.

The inability to handle continuous data is not a virtue

At the time VARBRUL was developed, the mathematics used to fit logistic models including continuous predictors had simply not been invented; to my knowledge, these methods were developed by Krzanowski (1980) and expanded to address missing cells by Little and Schulchter (1985). Roeland van Hout promises (personal communication) to illuminate the

historical context in which VARBRUL is situated in forthcoming work. Crucially, van Hout observes that the fitting method used by VARBRUL was borrowed from biometrics, a field which data gathering generally does not exhibit the hierarchical structure of sociolinguistic sampling discussed in the next two sections.

Proponents of VARBRUL have also argued that linguists' interest in categorical phenomena should lead linguists to only model dichotomized outcomes. In fact, this too is historical residue: in 1974, few means to generate continuous data were available to sociolinguists. But since Matthew Lennig's instrumental measurement of Parisian vowels (Lennig, 1978), sociolinguistics' cup has overflown with continuous data. In fact, in that study Lennig models continuous vowel formants as a factor of age and class. It should be noted, though, that the larger scientific community frowns on dichotomization of continuous scales, because it is known to reduces statistical power (Cohen, 1983). Even in those cases where predictors and outcomes are not linear or monotonic, there are methods for projecting the predictor onto an appropriate plane.

"No word-level effects" is a null hypothesis

Citing the Neogrammarian hypothesis of the regularity of sound change, proponents of VARBRUL have argued that accounting for word-level effects is contrary to the assumptions of the field. This ignores that as a community of practice, sociolinguists frequently *do* look for, and find, effects of lexical identity. And no one can deny the existence of lexically-diffuse changes (e.g., Labov, 1994, Part D). If however, we are to adopt the Neogrammarian program in full, how can the tenet of *gradualness* be resolved with software that deals only with categorical outcomes?

At this point, I must also address the use of the term *null hypothesis*. The choice of a significance point for a statistical test is often opposed with the rate of Type I errors that may result with repeated testing: for instance, it is common to hear that at a significance level of $p = .05$, one out of twenty tests will spuriously reject the null hypothesis. This is incorrect, since there is a strong prior bias on the hypothesis submitted for testing: sociolinguists regularly test for the effect of social class, for example, but not for the effect of it being Monday. By referring to the "no word effects"as a null hypothesis, proponents of VARBRUL imply that they test for it, but since they do not, it has the status rather of an *assumption*.

"No individual-level variation" is a null hypothesis

William Labov's interstitial notes to the 2nd edition of *Social Stratification of English in New York City* suggest that sociolinguists have failed to investigate individual speakers with the appropriate depth. He writes:

> "Many aspects of the NYC study influenced linguists' later work, but one aspect did not. There are no people in most of the sociolinguistic studies that followed—just means, charts, and trends. Although I have campaigned to bring people back into the field of sociolinguistics there has been only a limited response on this front." (Labov, 2006, p. 157)

As I argued in my paper at NWAV 38, interpretation of per-speaker random effects allows researchers to identify exceptional linguistic behavior and compare with the ethnography.

Furthermore, failing to account for speaker-level variation may result in *Simpson's paradox* (Pearson et al., 1899; Yule, 1903; Simpson, 1951; Blyth, 1972), in which variation between different subpopulations in the data overwhelm regularities shared by those subpopulations. Daniel Ezra Johnson made a similar point in his NWAV presentation (Johnson, 2009), showing a study in which samples from as single speaker created the illusion of an apparent-time effect. Thus, the assumption of "no speaker effects" has serious negative empirical consequences. It is `VARBRUL`, much like Galileo's telescope, that is in conflict with naïve realism: in every study in which they are defined, subjects and words do not show complete uniformity. For the empirical consequences of these models, I refer readers to the manuscript version of my NWAV 38 talk, a forthcoming IRCS technical report available in draft form online.[1] Part of the confusion here seems to be that while it is logically possible that speakers in a speech community have different constraints on the use of a variable (Bickerton, 1971), this does not seem to be the case (Guy, 1980). However, a *subject-level intercept* in a hierarchical model allows subjects to have different *input probabilities*, but holds constraint weights constant across all speakers in the sample. It does not, however, force subjects to have drastically different input probabilities.

Other techniques are not difficult to use or interpret

For those who would say that other techniques are difficult, I must ask: "for who(m)?" Surely not for those who champion `VARBRUL`, since they have evaluated the competing methods and software. Is it only these experts' audiences who are unable to make such evaluations themselves? This statement, then, has no other function than to spread fear, uncertainty, and doubt throughout the sociolinguistic community.

`VARBRUL` is harder to teach than alternatives

My experiences, and the experiences of my colleagues, in teaching statistical methods lead me to firmly disagree with the assertion that sociolinguists are incapable of learning new tricks. Students from other subfields are likely to be familiar with statistical packages implementing

[1] `http://ling.upenn.edu/~kgorman/mlm.html`

hierarchical logistic regression, but surely have never used `VARBRUL`. And `VARBRUL` *is* famously difficult to use. To quote Vivian Cook, just "CHECKTOK your tokens, CROSSTAB your cells, TVARB your results, TSORT, MAKECELL and finally MVARB... and the solutions pop out".[2] Cook's satire reminds us that there *is no decision procedure for doing good statistics* and that every model is wrong because every model is necessarily an apotheosis.

`VARBRUL` is part of the larger "community of practice"

My own experience as a researcher speaks to the massive importance of communicating both with members of the immediate community as well as outside that community. A poor understanding of the variationist program in theoretical linguistics and psycholinguistics lies in the failures of sociolinguists to communicate relevant results to those communities. A second critique I would like to level against this assertion is that it conflates descriptive and prescriptive modes. Yes, many sociolinguists use `VARBRUL`, but that doesn't mean that they should, or that they should use nothing else.

Conclusion

I'd like to briefly return to Thomas Kuhn, in particular his discussion of how paradigm shifts are implemented (with the obvious caveat of affording any software or statistical analysis the status of scientific paradigm). It is the young scientists, he writes, with have less invested in the dominant paradigm, who bring about change, rarely attracting adherents to the previous paradigm but rather replacing them as they age. If the papers presented at NWAV 38 are any indication, that the shift has already begun, and furthermore, sociolinguists are already finding "new new" ways to analyze variation beyond methods advocated for here.

Though David Sankoff's observation that sociolinguistics has been able to preserve a remarkable amount of early ideas is a point well-taken, Kuhn's thoughts on the necessity of a bit of forgetfulness in scientific revolution are just as relevant:

> "[T]he sciences, like other professional enterprises, do need their heros and preserve their names. Fortunately, instead of forgetting these heroes, scientists have been able to forget or revise their works." (Kuhn, 1962, p. 139)

Hierarchical linear models are far more like the devil you know (in preserving the fundamental principle of accountability), than the devil you don't (possible averaging effects leading to spurious conclusions). And if not for Sankoff's work, it's hard to imagine the numerical sophistication of the best sociolinguistics today. But it is not the *chevronées* (to borrow the term from Gillian Sankoff's tales of the Montréal corpus), but rather the young and hungry, who will have the final say.

[2] `http://homepage.ntlworld.com/vivian.c/Writings/Reviews/ReviewBayleyPreston.htm`

References

Derek Bickerton. Inherent Variability and Variable Rules. *Foundations of Language*, 7(4): 457–492, 1971.

Colin Blyth. On Simpson's Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

Herietta Cedergren and David Sankoff. Variable Rules: Performance as a Statistical Reflection of Competence. *Language*, 50(2):333–355, 1974.

Jacob Cohen. The cost of dichotomization. *Applied Psychological Measurement*, 7(3):249–253, 1983.

Paul Feyerabend. *Against Method*. Verso, London, 1975.

Gregory R. Guy. Variation in the Group and the Individual: The Case of Final Stop Deletion. In William Labov, editor, *Locating Language in Time and Space*, pages 1–35. Academic Press, New York, 1980.

Daniel Ezra Johnson. Canada's Next Top Model: Mixed models and why sociolinguists should use them. Paper presented at NWAV 38, 2009.

Wojtek Krzanowski. Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 36(3):493–499, 1980.

Thomas Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1962.

William Labov. *Principles of Linguistic Change: Internal Factors*. Blackwell, Oxford, 1994.

William Labov. *The Social Stratification of English in New York City*. Cambridge University Press, Cambridge, 2nd edition, 2006.

Matthew Lennig. *Acoustic Measurements of Linguistic Change: The Modern Paris Vowel System*. Doctoral dissertation, University of Pennsylvania, 1978.

Roderick Little and Mark Schulchter. Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values. *Biometrika*, 72(3):497–512, 1985.

K. Pearson, A. Lee, and L. Bramley-Moore. Genetic (reproductive) selection: Inheritance of fertility in man. *Philosophical Translations of the Royal Statistical Society, Series A*, 173: 534–539, 1899.

Edward Simpson. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, Series B*, 13(2):238–241, 1951.

Udny Yule. Notes on the Theory of Association of Attributes in Statistics. *Biometrika*, 2 (2):121–134, 1903.