Language and Linguistics Compass 3/1 (2009): 359–383, 10.1111/j.1749-818x.2008.00108.x

Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis

Daniel Ezra Johnson*
University of York

Abstract

The variable rule program is one of the predominant data analysis tools used in sociolinguistics, employed successfully for over three decades to quantitatively assess the influence of multiple factors on linguistic variables. However, its most popular current version, GoldVarb, lacks flexibility and also isolates its users from the wider community of quantitative linguists. A new version of the variable rule program, Rbrul, attempts to resolve these concerns, and with mixed-effects modelling also addresses a more serious problem whereby GoldVarb overestimates the significance of effects. Rbrul's superior performance is demonstrated on both simulated and real data sets.

Introduction

The variable rule was introduced in Labov's (1969) discussion of the regularly conditioned patterns of contraction and deletion observed for the African-American Vernacular English copula.[1] The next decade saw the development of the variable rule program for estimating the parameters of such rules (Cedergren and Sankoff 1974; Rousseau and Sankoff 1978a).

The variable rule, as originally conceived, is no longer a preferred theoretical concept for accounting for linguistic variation (Fasold 1991); indeed, much of current phonological theory has moved away from rules in general. But the name has persisted, often abbreviated as VARBRUL, to refer to a type of quantitative variationist analysis, as well as the computer programs that make it possible.

A variable rule program evaluates the effects of multiple factors on a binary linguistic 'choice' – the presence or absence of an element, or any phenomenon treated as an alternation between two variants. The factors can be internal (linguistic), such as phonological or syntactic environment, or external (social), for example, speaker gender or social class. The program identifies which factors significantly affect the response variable of interest, in what direction, and to what degree.

The mathematical underpinnings of the variable rule method were refined during the 1970s, but in the three subsequent decades it has remained

fairly constant. The method has proven extremely popular: it is one of the tools of choice for those who study linguistic variation quantitatively (Tagliamonte 2006). By way of illustration, over the period 2005–2008, some 40% of the articles published in the journal *Language Variation and Change* employed variable rule analysis.

The version of the program used most often today, GoldVarb X (Sankoff et al. 2005), is essentially an attractive, user-friendly implementation of VARBRUL 2 (Sankoff 1975). Thus, it retains some of the idiosyncrasies of its predecessors, although this helps make its results comparable with earlier work. Several desirable features were added to VARBRUL 3, but this version was never implemented 'for personal computers' (Sankoff 2004: 1157).

Today's younger sociolinguists may have never even seen the type of hardware VARBRUL 3 could run on, but they do have access to powerful software packages for statistical analysis. These include commercial platforms such as SPSS and SAS, as well as the free, open-source, user-extendable statistical software environment R, which is being used more and more by linguists (Baayen 2008).

Notwithstanding these other platforms, for some sociolinguists, performing quantitative analysis has remained equivalent to using GoldVarb, with its limited range of functions. At the same time, some other linguists – not to mention our potential collaborators in other fields – may have broad statistical backgrounds without being familiar with GoldVarb's output, and may not understand why we still need such a venerable piece of single-purpose software, no matter how cutting-edge it was in 1975.

From GoldVarb to Rbrul

The procedure at the heart of GoldVarb – multiple logistic regression[2] – is available in any statistical package. However, GoldVarb presents the results of the regression in a format that is rarely seen elsewhere, and using a slightly different terminology.

Imagine that we were looking at the effect of speech style on the variable (ing) in English – the use of [n] instead of [ŋ] at the end of words like *working*.[3] In the variable rule tradition, style would be called a factor group and the individual styles being studied – spontaneous speech, reading passage, wordlist – would be called factors. Given a set of observations of (ing) across the styles, GoldVarb would return an input probability representing the overall likelihood of [n] in the data,[4] and another probability, called a factor weight, for each style factor.

Suppose that the input probability came out as 0.4, and within the style factor group, reading passage had a weight of 0.5, spontaneous speech 0.6, and wordlist 0.3. We would conclude that [n] for (ing) is somewhat disfavored in the data overall, and that a token occurring in a reading passage is no more or less likely to be realized with [n]; a factor weight

factor weight (probability)	log-odds
.000	–∞
.100	–2.197
.200	–1.386
.300	–0.847
.400	–0.405
.500	0
.600	+0.405
.700	+0.847
.800	+1.386
.900	+2.197
1.000	+∞

Fig. 1. Some factor weights (probabilities) and the corresponding log-odds.

of 0.5 is equivalent to no effect. Spontaneous speech tokens are somewhat more likely to occur with [n], while wordlist tokens are considerably less likely.

Most other statistical software reports logistic regression results differently. First of all, what GoldVarb calls factor groups are usually called factors, and they are divided into levels. One method of reporting factor effects is very similar to GoldVarb; this is called sum contrasts, where each coefficient represents a deviation from the mean. Another method is treatment contrasts, where one level of each factor is chosen as the baseline, and is given a coefficient of 0. Each of the other levels is then assigned a coefficient representing the effect on the response of switching from the baseline level to the 'treatment' level in question (the terminology clearly derives from an experimental paradigm).

Another difference is the units in which the coefficients are expressed. Rather than being probabilities ranging from 0 to 1, they are in units called log-odds which can be any positive or negative number. We obtain log-odds from probabilities by taking the natural (base *e*) logarithm of the odds, where the odds are the probability of an event occurring, divided by the probability of it not occurring. The formula is $\ln[p/(1-p)]$; a positive value is a favoring effect, a negative value disfavoring, and a value of 0 is neutral. Figure 1 gives a comparison between factor weights (probabilities) and log-odds. We see that if there were a binary factor group with weights of 0.400 and 0.600, this would correspond to log-odds of −0.405 and +0.405 (as in sum contrasts), or a difference of 0.810 between the two levels (as in treatment contrasts).

The differences noted above are fairly superficial, and there are advantages to both forms of presentation. Individual probabilities are perhaps easier to interpret, but when they combine, log-odds are preferable because they can simply be added together. If we were to include age, social class, dialect region, and grammatical category as well as speech style in our

model for (ing), the prediction of the model for, say, the spontaneous speech of a 65-year-old, working-class, Southern US speaker in progressive verbal forms would simply be the sum of the log-odds coefficients for those particular levels, plus the value for the intercept. If we had GoldVarb's factor weights and input probability instead, the only way to form a joint probability would be to convert the values into log-odds, combine them, and convert them back into probabilities.[5]

If quantitative sociolinguistics were starting from scratch, reporting regression coefficients only in log-odds might be preferable. But since so much previous research has been conducted with GoldVarb, the field could perhaps benefit best from software that can display results in both formats. We may continue to think of effects in terms of factor weights, but with a more mainstream presentation alongside them, our work will be much more comprehensible to psycholinguists, psychologists, statisticians, and so forth.

The new program Rbrul, written by the author and available for download at http://www.danielezrajohnson.com, has been designed, among other things, to replicate the functionalities and factor-weight-based output of GoldVarb, while also presenting results in log-odds with the option of sum or treatment contrasts.

Rbrul is a text-based interface to existing functions in the R environment, particularly the model-fitting functions *glm* and *glmer* (Bates and Sarkar 2008).[6] It is designed for current or potential users of GoldVarb who want the benefit of powerful modern statistical techniques, without having to learn to use an entirely unfamiliar platform. In this, Rbrul shares the goals of R-Varb (Paolillo 2002b), but it offers a number of specific advantages.[7]

Rbrul over GoldVarb: Other Advantages

GoldVarb requires its input to be in a dedicated token file, with each factor level represented by a single-character code. Rbrul can read comma- or tab-delimited spreadsheets, with no need to abbreviate the content of the fields. Users can thus interpret results with less head-scratching, and switch back and forth more easily between Rbrul and program like Excel.

Like the never-implemented VARBRUL 3, Rbrul can handle continuous numeric predictors (for which it is at best dubious statistical practice to 'bin', or convert into factors). For example, if we included speaker age in a model, the program would report that for each year older a speaker is, the likelihood of the response increases by a particular amount.[8]

As noted, variable rule analysis carries out logistic regression, dealing with binary response variables representing discrete linguistic alternatives. However, there is no reason why the same software should not also be able to perform linear regression, with continuous responses: vowel formant measurements, for example. Rbrul makes it possible to estimate the effects of multiple predictors on data of this type, too.[9]

While it is possible in GoldVarb to detect and model interactions between factors (Paolillo 2002a), Rbrul makes the process easier, and optionally a part of the same automatic procedure that identifies significant main effects.[10]

GoldVarb uses a fixed 0.05 threshold for determining factor group significance; in Rbrul, this value can be adjusted, as may be called for if many predictors are under consideration. For example, if there are five potential predictors, testing each with a threshold of 0.01 keeps the overall error level at 0.05 (the Bonferroni correction).[11]

Rbrul is also more forgiving with regard to 'knockouts', situations where the response is invariant – either 0% or 100% – in a subset of the data. To avoid knockouts, the Rbrul user can group factors together or exclude them as in GoldVarb, but doing so is rarely obligatory (although good practice may still require their exclusion; see Guy 1988).

Grouping Structure, Significance, and Mixed-Effects Modelling

The improvements discussed in the previous section might not be sufficient to lead the average GoldVarb user to abandon the program in favor of Rbrul. But GoldVarb also suffers from a more serious problem, related to the way it evaluates the significance of factor groups.

One of the assumptions underlying regression analysis is that the observations making up the data are independent of each other.[12] But in most linguistic data sets, the tokens are not independent. In particular, they are naturally grouped according to the individual speakers who produced them.

As it is usually run, without a factor group for speaker, GoldVarb necessarily ignores the grouping and treats each token as if it were an independent observation. This leads the program to overestimate – potentially drastically – the significance of external effects, those of social factors like gender and age. Indeed, GoldVarb will often include one or more external effects in its best stepwise regression run even if the differences involved are really quite likely to be due to individual variation combining with chance.[13]

On the other hand, if we do include an individual-speaker factor group, GoldVarb (like any regression software) will effectively underestimate the significance of speaker-external effects, so that they are always eliminated from the best run, even when they are truly significant over and above individual variation.

These complementary shortcomings have never been fully recognized in the variable rule literature (Young and Bayley 1996; Paolillo 2002a; Sankoff 2004; Tagliamonte 2006; but see Sigley 1998), although in psycholinguistics an analogous statistical issue has been extensively discussed since Clark (1973).

External factors such as age, gender and social class are properties of speakers, and so the true significance of such effects depends on the

patterning of speakers, not linguistic tokens.[14] As an extreme example, if a preliminary study of only two men and two women suggested that a certain linguistic variable was correlated with gender, collecting more and more tokens from the same speakers would not help settle the question; we would have to collect data from other men and women.

Or imagine that we have transcribed 1,000 tokens of words with a historical post-vocalic /r/; half the tokens come from men, and half from women. Suppose that 60% of the men's tokens are /r/-less, compared with 40% of those from women. Given such a distribution, GoldVarb would identify gender as a highly significant factor group.[15]

And that conclusion would be perfectly justified if the data came from 40 speakers, 20 men ranging between 45% and 75% /r/-lessness, and 20 women ranging between 25% and 55%. Here, while men and women both show considerable diversity, and some women are even more /r/-less than some men, the men are more /r/-less overall. And with so many speakers in each group, the difference is very unlikely to be due to chance.[16]

But if the same 1,000 tokens had come from only eight speakers – four men with 45%, 55%, 65% and 75% /r/-lessness, and four women with 25%, 35%, 45% and 55% – we would not have sufficient evidence for a gender effect. Here, too, the average man is more /r/-less than the average woman, but the number of speakers is small enough that the difference could have arisen by chance.[17]

By leaving out the speaker variable entirely, GoldVarb is not equipped to distinguish between these possibilities. Therefore, it cannot accurately assess external effects' significance. In another context, Sankoff (2004: 1159) suggests first running GoldVarb with a speaker factor group, and then evaluating the effects of external factors using other statistical techniques. This is sensible, but Rbrul uses the R mixed-effects modelling function *glmer* to obtain similar results in a single step.

Mixed-effects modeling is 'a flexible and powerful tool for the analysis of grouped data . . . includ[ing] longitudinal data, repeated measures, blocked designs and multilevel data. The increasing popularity of mixed-effect models is explained by the flexibility they offer in modeling the within-group correlation often present in grouped data, by the handling of balanced and unbalanced data in a unified framework, and by the availability of reliable and efficient software for fitting them' (Pinheiro and Bates 2000: vii).

Only in recent years has software been developed to fit generalized linear mixed models, including, for example, for data with binary responses (Breslaw and Clayton 1993; Bates and Sarkar 2008). Such models are now being used in psycholinguistics (see Jaeger 2008, its references, and the entire special issue of the *Journal of Memory and Language* on Emerging Data Analysis and Inferential Techniques), and to a lesser extent, in sociolinguistics (Jaeger and Staum 2005).

Mixed models make a distinction between two types of factor that can affect a response. Fixed effects are factors with a fairly small number of

possible levels, for example, male/female, stressed/unstressed, or following vowel/consonant/pause. These factors are usually the direct object of interest, and their levels would be replicable in a further study.

Factors drawn from a larger population, such as the speakers in a study, are called random effects. These are usually not replicable – two studies of the same linguistic phenomenon might both involve men and women, but probably not the same individuals. For random effects, accounting for the variation in the population is more important than knowing the exact values of individual effects (although these are also estimated).

For a fixed effect like gender, the mixed model gives the familiar set of coefficients associated with the differences between factor levels. For a random effect like speaker, it estimates a single parameter representing the amount of inter-speaker variation.[18]

Including a speaker random effect takes into account that some individuals might favor a linguistic outcome while others might disfavor it, over and above (or 'under and below') what their gender, age, social class, etc. would predict.

Unlike an ordinary regression model with speaker included, a mixed model does not directly fit a parameter to each speaker's data. Because of this, it can still capture external effects, but only when they are strong enough to rise above the inter-speaker variation. If there is a lot of individual variation, chance can create the appearance of external effects, and Rbrul raises its standards accordingly.

Misidentifying a chance effect as a real one is called a Type I error, and Rbrul's Type I error rate stays close to the theoretical value of 0.05 in many situations where GoldVarb's greatly exceeds it. However, this more conservative behavior has a trade-off: in some situations, Rbrul is more likely than GoldVarb to make a Type II error by failing to identify an effect that really does exist. The following section uses simulated data sets to compare the performance of GoldVarb and Rbrul, focusing on their Type I and Type II error rates.

Tests with Simulated Data

Simulated data files were created in R, each consisting of a number of tokens of a hypothetical binary variable, which we will call (ing). The number of 'speakers' was either 10, 20, 40 or 80; half were 'male', half 'female'. Every speaker was represented by 20, 40, 80 or 160 tokens, so the number of tokens per file was between 200 and 12,800; most real variable rule analyses would fall within this range.

The effect of an external 'gender' factor was set to either 0, 0.2, 0.4, 0.8 or 1.6, these numbers being the difference in log-odds between men, who favored [n] for (ing), and women, who disfavored [n]. The amount of individual speaker variation on top of any gender effect was also manipulated; its standard deviation was set to 0, 0.2, 0.4, 0.8 or 1.6.[19]

Figure 2 illustrates what these gender and speaker parameters mean. Each dot stands for a speaker, with the x-coordinate being that speaker's proportion of [n] for (ing). The y-axis indicates how many speakers there are expected to be with that proportion, given the mean for the speaker's gender and the amount of individual variation.

In the leftmost column of the figure, where the gender effect is zero, the male (blue) and female (red) distributions are identical. As the gender effect increases from left to right, there is more separation between the distributions of male and female speakers.

In the bottom row, individual variation is zero, so all men and women follow their gender means exactly. But as the speaker standard deviation increases from bottom to top, the two distributions become wider, and there is more overlap between them.

The crossing of the four parameters led to 800 different data sets. For each, the simulation was run 500 times. In each run, the tokens of (ing) were generated randomly, with [n] or [ŋ] chosen based on the combined probabilities for gender and speaker.

Two logistic regressions were performed on each file: an Rbrul mixed-effect regression with a fixed effect for gender and a random effect for speaker, and a fixed-effect regression with gender only, as in GoldVarb.[20] In both cases, the gender effect was considered significant whenever the software returned a p-value less than 0.05.

The Type I error rate was defined as the proportion of times that gender was identified as significant, when the gender parameter was in fact zero. When the gender parameter was greater than zero, the proportion of runs returning no significant effect (or a spurious one in the wrong direction) constituted the Type II error rate.

The leftmost column of Figure 3 compares the Type I error rates for the two programs. As speaker standard deviation increases from bottom to top, GoldVarb's Type I error rate increases substantially. The more speakers vary within the gender groups being compared, the more often GoldVarb confuses mere sampling error with an external factor effect. Meanwhile, regardless of speaker variation, Rbrul's Type I error rate remains close to the theoretical value of 0.05.

GoldVarb's Type I error rate also increases from left to right within each panel, as the number of tokens per speaker goes up. This is because GoldVarb treats the observations from the same speaker as independent, and here they are not. When a small amount of data departs from expectations, things will normally even out in time (the law of large numbers). But if a sample of men and women do happen to deviate from identical population means, collecting more data from the same individuals will only seem to confirm a gender difference, if we ignore the grouping. As shown by its nearly flat Type I error rate within each panel, Rbrul takes note of the dependency between tokens and avoids this trap.

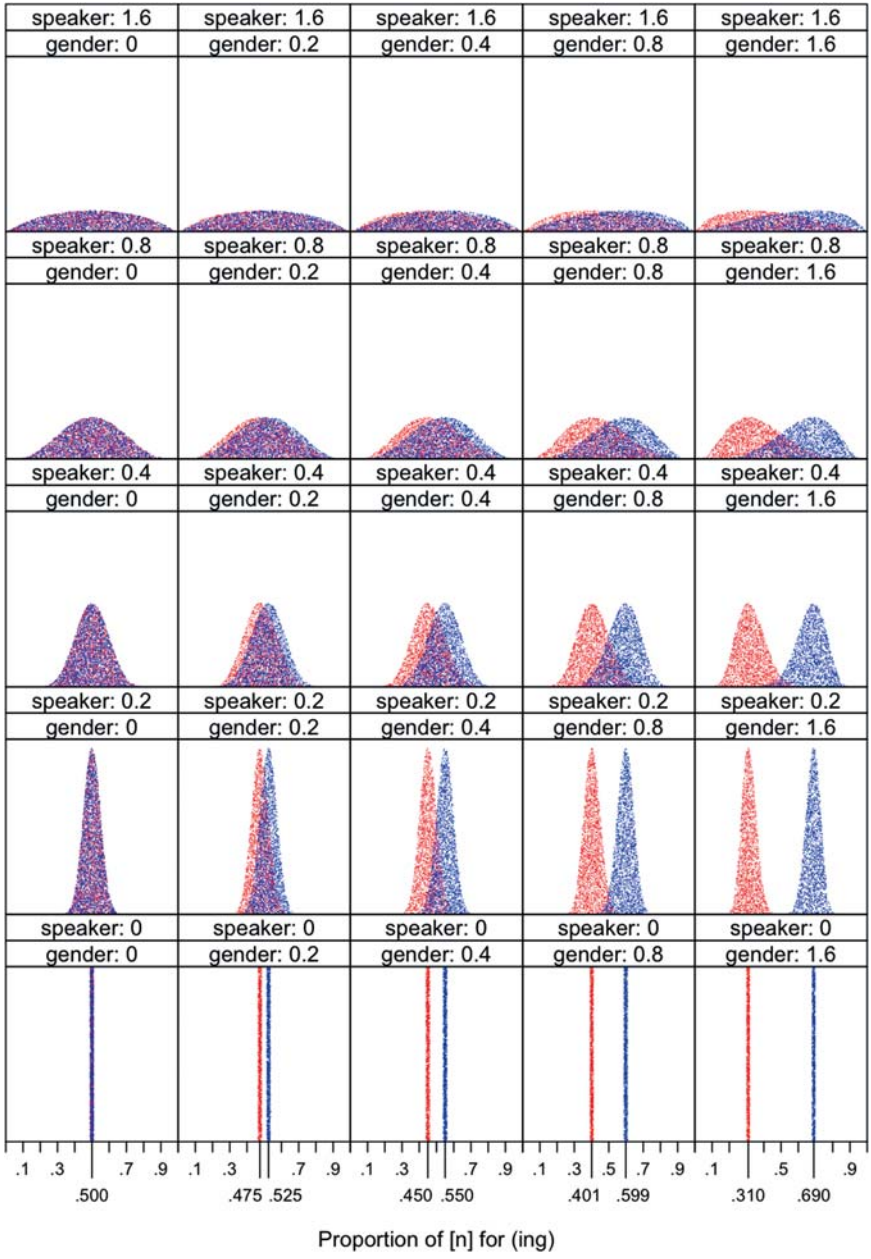

Fig. 2. Probability densities for the proportion of [n] for (ing) for males (blue) and females (red), plotted by gender effect size (columns) and speaker standard deviation (rows).

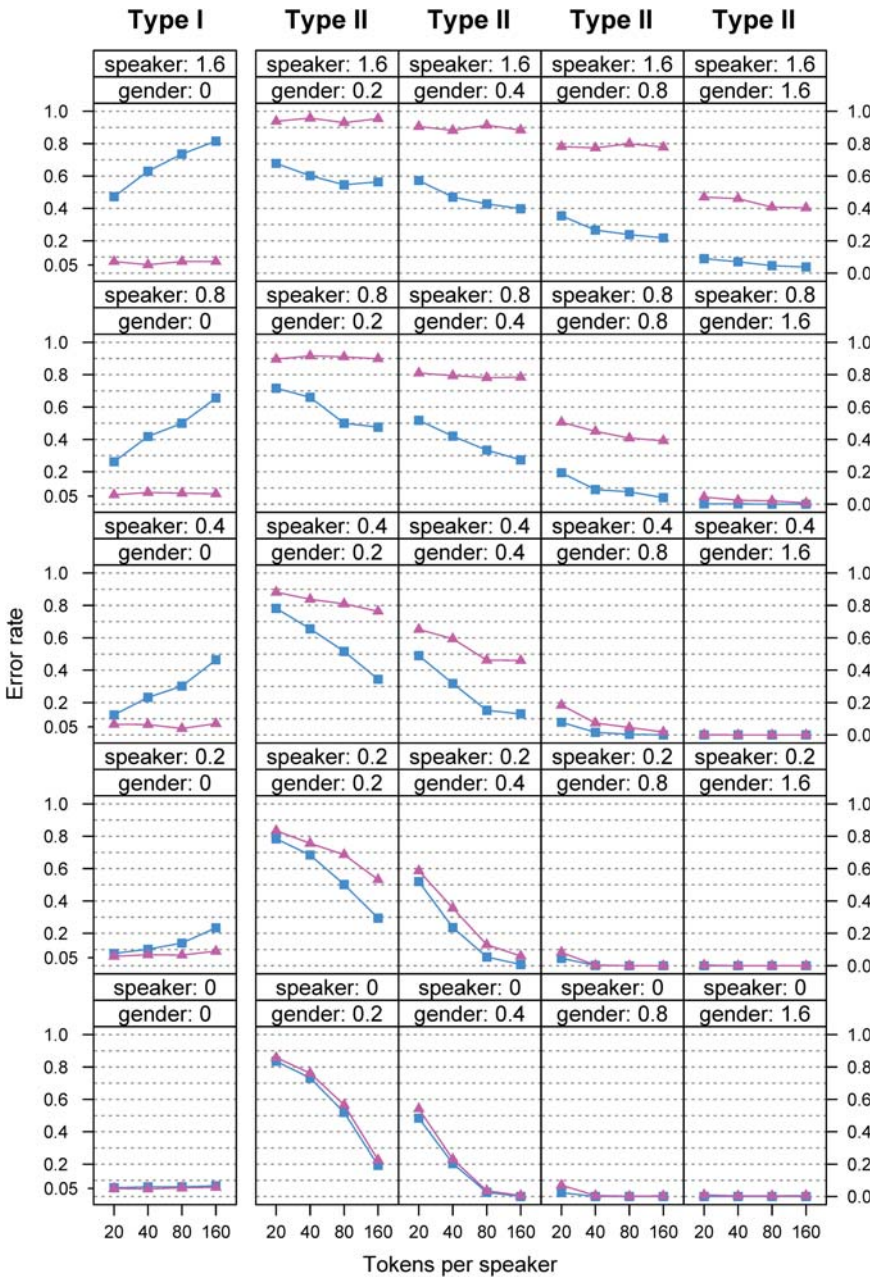

Fig. 3. Type I and Type II error rates for Rbrul (pink triangles) versus GoldVarb (blue squares), plotted by gender effect size (columns), speaker standard deviation (rows), and number of tokens per speaker (within panels). Number of speakers: 20 (ten males, ten females). Runs per condition: 500.

The remaining four columns of Figure 3 show Type II error. There is little or no difference between the two programs' output in the bottom rows of the figure, where between-speaker variation is low, but as individual variation increases, Rbrul's error rate grows faster than GoldVarb's. Rbrul is less likely to identify the real underlying gender effect in these situations.[21–23]

The tradeoff between Type I and Type II error is an ever-present issue in statistical analysis, and has no simple solution. Most researchers would probably endorse a conservative approach, arguing that it is better to overlook something that does exist than to report something that does not. This attitude would lead us to prefer Rbrul.

However, if we had prior reason to suspect that a gender effect existed in our population – say, if one had been reported in previous studies – then we might proceed more confidently, using GoldVarb or raising the significance threshold in Rbrul. After all, a high, non-significant p-value is not evidence against an effect. It only means that the difference in the data could easily have arisen by chance, not that it actually did.

Rbrul's p-values are actually more accurate than GoldVarb's, even when they lead to fairly high rates of Type II error. We can see why if we return to Figure 2, and imagine repeatedly picking ten blue dots and ten red dots at random from the top right panel, where the gender effect is large. Rbrul's Type II error rate says that 40–50% of the time, the same pattern of dots could easily have come from the top left panel, where the population-level gender effect is zero.

Figure 4 illustrates this point. On the left side, each row is a sample of ten 'men' and ten 'women' taken from the top left panel of Figure 2. They have been put in order vertically according to the male–female difference they exhibit, purely by chance. Of the 100 samples, five are Type I errors (three pairs of filled circles at the top and two at the bottom of the figure); they were found significant at $p < 0.05$ by a two-sample t-test. With similar data, Rbrul's Type I error rate stayed below 10%, while GoldVarb's went as high as 80%.

On the right side of Figure 4, the samples have been drawn from the top right panel of Figure 2, where the underlying gender difference is 1.6 in log-odds, or 0.310–0.690 in terms of probability. When the observed gender difference is at least that large, the t-test finds it significant. But many of the samples show a smaller effect, and for them, significance is usually not reached. Because these samples overlap with many of the ones on the left side of the figure, we are right to call them chance-level effects.

In fact, only 66 of the 100 samples on the right side of the figure show significance: the high level of individual speaker variation has quite often 'canceled out' the gender effect, leading to a 34% rate of Type II error. With similar data, Rbrul's Type II error rate also exceeded 30%, while GoldVarb's remained less than 10%. Figure 4 shows why this latter figure is unreasonable; lowering the bar in order to reduce Type II error would inevitably produce a very high level of Type I error.

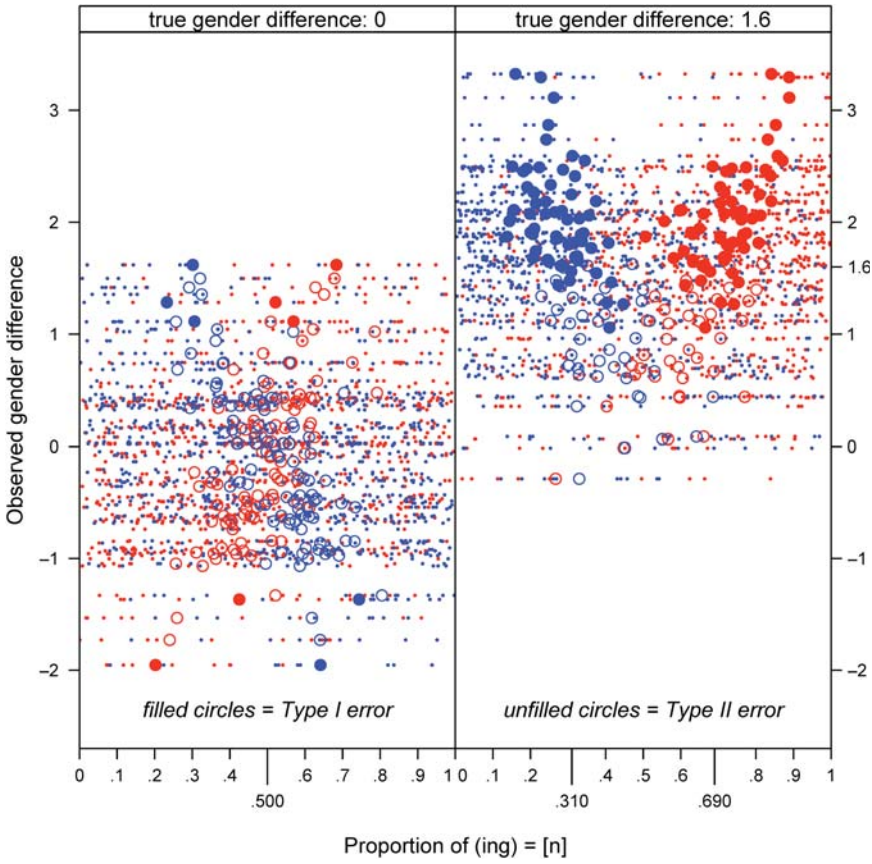

Fig. 4. The relationship between Type I and Type II errors. One hundred random samples of ten males (red) and ten females (blue). Large circles are the means for each sample. Filled circles are significant at $p < 0.05$ by t-test.

In summary, GoldVarb's p-values are too low, resulting in a Type I error rate that is too high; its behavior is anti-conservative. Rbrul's Type I rate is close to optimal, and while its Type II rate exceeds GoldVarb's when speaker variability is high, this directly follows from its more realistic Type I rate.[24]

If users might in some cases still prefer the results generated by GoldVarb, Rbrul can emulate the behavior of that program exactly, simply by omitting the speaker random effect. On the other hand, GoldVarb users who would like more accurate significance estimates have little recourse within that software.[25]

Tests with Real Data

The simulated data sets were perfectly balanced, meaning that the same amount of data was generated for each speaker. We saw that Rbrul handled

```
Input 0.517
Group # 2 -- =: 0.687, *: 0.266
Group # 3 -- +: 0.305, $: 0.730, !: 0.634
Log likelihood = -298.915
```

Fig. 5. GoldVarb's best run for loanword stress shift in Hønefoss Norwegian. Group 2 is age: = signifies older adults, * young adults. Group 3 is education: + means university graduates, $ vocational training, ! high school only.

this balanced data better than GoldVarb. When data are unbalanced, such as when different speakers are represented by different numbers of tokens (and this is the rule, not the exception, in sociolinguistic studies), mixed-model analysis using Rbrul gives better estimates not only for the significance (p-values) of external effects, but for their sizes (coefficients), as well.[26]

This section takes two real data sets and analyzes them in GoldVarb and Rbrul. The output of the two programs can thus be compared in both format and content. The first example concerns the placement of stress on loanwords in Norwegian (Hilton 2007). The second example concerns /r/-vocalization in New York City (Becker 2007).

Loanword Stress Shift in Hønefoss Norwegian

In the Norwegian dialect spoken in the city of Hønefoss, some foreign borrowings can be stressed in two different ways. Speakers can stress the same syllable that is stressed in the lending language, or they can shift stress to the initial syllable, which is where stress always occurs in the native vocabulary.

The data, from Hilton (2007), consists of 565 tokens of this variable, collected in interviews with 20 different speakers. The number of tokens per speaker ranges from 8 to 72, making the data quite unbalanced from this point of view.[27,28]

The only factor groups considered here are external ones: gender, age and education. A GoldVarb analysis concludes that age and education are significant, each at the '0.000' level, with age being added first. Adding gender is associated with a p-value of 0.079, so it is not retained. Figure 5 reproduces GoldVarb's output for the best run (which is identical between stepping-up and stepping-down).

Once we decipher the single-character factor codes, we see from Figure 5 that loanword stress shift is favored by older adults, the vocationally trained, and those with only a high school education. Younger adults and university graduates disfavor the shift.

As shown in Figure 6, running Rbrul without a speaker random effect gives exactly the same output, give or takes 0.001 in some cases. Rbrul reports the factor effects in log-odds (here using sum contrasts), the number of tokens and proportion of stress shift for each factor level, and the overall


```
best step-up model is with: age (6.39e-25) + education (7.96e-18)

                factor logodds tokens proportion  uncentered weight
age
                   old   0.903    318       0.701              0.688
                 young  -0.903    247       0.271              0.266

education
           vocational   0.756    205       0.751              0.731
          high school   0.310     67       0.642              0.635
           university  -1.066    293       0.317              0.305

deviance     df     intercept  total        mean          input prob
  597.83      4         0.196    565       0.513               0.517
```

Fig. 6. Rbrul's best fixed-effect model for loanword stress shift in Hønefoss Norwegian.

data mean, as well as the factor weights and input probability. With proper labels instead of symbols for factors – for example, 'university' instead of '+' – the results are easier to parse.[29]

Rbrul tells us that it first added the age factor group, with $p = 6.4 \times 10^{-25}$, then education, with $p = 4.8 \times 10^{-9}$. These are infinitesimal p-values; recall that, during its stepwise runs, GoldVarb reported 'significance 0.000' for both. Rbrul's deviance of 597.83 is equal to −2 times GoldVarb's log-likelihood of −298.915, demonstrating that these models are exactly equivalent. However, when we add a subject random effect, Rbrul's output changes considerably, as seen in Figure 7.

In the mixed-model case, Rbrul begins with the speaker random effect. It then adds education, with a p-value close to 0.01. This p-value is more than a million times higher than previously, but it still falls under the 0.05 threshold. But adding age to the model gives $p = 0.28$. This is well above the threshold, so age is not added.

Education is the only fixed factor in the best model; we can say that taking speaker into account has stopped age from appearing as a significant effect. A second difference between the fixed-effect and mixed-effect models can be seen in the factor weights for education. Compared to Figure 6, the education weights of Figure 7 are more extreme. The

```
best step-up model is with: speaker (random) + education (0.00996)

                factor logodds tokens proportion uncentered weight
education
           vocational   2.281    205       0.751              0.939
          high school   0.237     67       0.642              0.665
           university  -2.518    293       0.317              0.112

speaker std dev
         2.736

deviance     df      intercept  total        mean         input prob
 396.553      4          0.489    565       0.513              0.510
```

Fig. 7. Rbrul's best mixed-effect model for loanword stress shift in Hønefoss Norwegian.

weight for 'high school' is similar, but 'university' is much lower and 'vocational' much higher.[30]

By inspecting the data (shown in part in Figure 10) with an eye on speaker patterns, we can understand these two differences, and see why the mixed-model estimates are more sensible for this data set.

First, take the possible effect of age. Averaging over tokens, the young adult group shifts stress 27% of the time and the older adult group shifts it 70% of the time: a very large difference. But if we average over speakers instead, the difference is smaller: 45% (young adults) compared to 69% (older adults).[31] This is because those younger speakers who stress-shift less have many more tokens in the data, dragging down their group's raw mean.[32] The mixed model incorporates individual differences like these, and estimates an age effect on top of them. The adjusted effect size for age is less than half as large (in log-odds), and as a result, it does not reach significance.

A different explanation is needed for the several-fold increase in effect size observed for education. The education effect does appear larger if we average over speakers rather than tokens, but only slightly. Unbalanced data is not the primary issue here; rather, it is the way that speakers pattern within the education groups.

The average stress shift rate for the eight university-educated speakers is 31%, and for the eight vocationally trained speakers it is 80%. This difference, equivalent to 2.19 in log-odds, is not much greater than the difference estimated between the two groups by the fixed-effect model, which is 1.82 (from Figure 6: 0.756 minus −1.066).

But a closer look at each group reveals that their score distributions are severely skewed; half of the university-educated speakers stress-shifted less than 7%, while half the vocationally trained speakers stress-shifted more than 95%.

These details of speaker distribution mean nothing to the fixed-effect model, which does not include speaker and works only with grouped token averages. But a mixed model strikes a balance between group (fixed) and individual (random) effects; in the fitting process, there is a penalty on the size of the random effects (Bates, personal communication). Here, estimating a larger fixed effect for education – 4.80 log-odds, corresponding to 12% shift for university versus 94% for vocational – allows most of the individual deviations to be quite small. Thus, the model fits worse for a few speakers, but better for most.[33] If we were to observe new Hønefoss speakers from the different education levels, the mixed model is likely to make better predictions for them.

/r/–Vocalization in New York City English

Our second example focuses on a familiar sociolinguistic variable: the vocalization of post-vocalic /r/ in New York City English. This data set,

from Becker (2007), consists of 3,000 tokens of (r) from seven natives of the Lower East Side of Manhattan. With fewer speakers than the Norwegian example, the data are also better balanced, having between 248 and 591 tokens per speaker.

A GoldVarb analysis identifies many significant factors, of which only a few will be discussed here.[34] Speaker age and social class are classic *external* factors, which we know can behave quite differently in a mixed model. The phonological environment following (r) – consonant, vowel or pause – is a classic internal factor. Speech style – casual speech, reading passage or wordlist – is also a speaker-internal factor, because each speaker provides data in more than one style. The same applies to the topic factor, which refers to whether or not speakers were talking about the Lower East Side.

Taking speaker variation into account in a mixed model can change the estimates of *internal* factors in at least two ways. The first is when a factor level is produced disproportionately often by certain speakers, for whom the response variable is particularly favored or disfavored. A fixed-effect model will always attribute the effect to the factor, while the mixed model will also consider holding the speaker responsible.

A second possibility is that different speakers have different internal constraints. This relates to a controversy dating from the 1970s, when variable rule analysts' alleged assumption of a single 'community grammar' caused considerable debate (Kay and McDaniel 1979; Sankoff and Labov 1979).

The Rbrul program will eventually allow for more complex mixed models that allow individuals' constraints to vary around a community norm (random slopes). In the mixed models being fit here, individuals differ only by their input probabilities (random intercepts).

For the (r) data, the best fixed-effect model (Figure 8) is again equivalent to GoldVarb's output. We see the expected large difference between a following consonant and vowel; /r/ is rarely dropped before a vowel. We also see the expected ordering of styles, with casual speech showing the greatest tendency to vocalize /r/ and wordlist the least. Topic, too, is significant, with topics related to the Lower East Side having a small disfavoring effect on /r/ articulation; that is to say, the Lower East Side topics favor /r/ vocalization.

One external effect is as expected: younger speakers favor (r), in accordance with the known, slow reintroduction of the consonant into New York City speech over the past half-century or so. The small but significant social class effect, however, is not in the expected direction. Lower social class is apparently associated with a higher use of (r). This is not in accordance with the literature on this topic, which considers the rise of (r) to be a change from above (Labov 1966).

Figure 9 shows the best model using a speaker random factor. The same factor groups have been retained with little change, except for social class,


```
best step-up model with: following (1.02e-55) + age (5.68e-54) +
     preceding (8.38e-28) + style (8.12e-14) + topic (0.00637) +
     class (0.0123)

                factor logodds tokens proportion   uncentered weight
following
                 vowel   1.708    326       0.782               0.905
             consonant  -0.803   2269       0.332               0.436
                 pause  -0.905    405       0.301               0.411

age
               younger   0.698    946       0.571               0.722
                 older  -0.698   2054       0.287               0.392
...

style
              wordlist   0.756    124       0.669               0.764
               reading  -0.231    572       0.409               0.546
                casual  -0.525   2304       0.353               0.473

topic
                 other   0.136   1615       0.415               0.531
          neighborhood  -0.136   1385       0.331               0.463

class
                 lower   0.108   1538       0.416               0.526
                higher  -0.108   1462       0.335               0.472

deviance     df      intercept  total        mean          input prob
3260.984     16          1.059   3000       0.377               0.360
```

Fig. 8. Rbrul's best fixed-effect model for the articulation of post-vocalic /r/ in New York City English.

which is no longer significant. We may be content to conclude that the earlier 'backwards' social class effect was actually a Type I error. There are only seven speakers, and they differ only marginally along social class lines: 34% versus 42% (r). Because of the variation *within* each social class group (see Figure 10), the mixed-model p-value of 0.347 for class is more plausible than GoldVarb's 0.012.

There are other questions we could address with mixed-model analysis, such as: while taking into account whatever imbalances exist among the other significant factors, which speakers depart most from the norm? Inspecting the random effect estimates (not shown) tells us that of the New York City speakers, six are close to their predicted levels; the seventh uses (r) quite a bit more often than would be expected given her age. In the Hønefoss data, there were more speakers who departed substantially from their education group's prediction (in general, there was more individual variation for Hønefoss stress shift, as seen by the higher speaker standard deviation estimate: 2.736 versus 0.488 for New York /r/). Figure 10 helps illustrate the two cases.

In both data sets, Rbrul's mixed model found one fewer significant factor than GoldVarb's fixed-effect model. Neither age in Hønefoss nor social class in New York City had a significant effect on the variable of

```
best step-up model with: speaker (random) + following (4.46e-65) +
     preceding (1.31e-28) + style (1.86e-10) + topic (0.000563) +
     age (0.0102)

               factor logodds tokens proportion   uncentered weight
following
                vowel   1.792    326       0.782               0.914
            consonant  -0.840   2269       0.332               0.433
                pause  -0.952    405       0.301               0.405
...

style
             wordlist   0.717    124       0.669               0.745
              reading  -0.287    572       0.409               0.517
               casual  -0.430   2304       0.353               0.481

topic
                other   0.177   1615       0.415               0.541
         neighborhood  -0.177   1385       0.331               0.452

age
              younger   0.699    946       0.571               0.722
                older  -0.699   2054       0.287               0.392

speaker std dev
          0.488

deviance      df   intercept  total        mean           input prob
 3193.78      16       1.096   3000       0.377                0.372
```

Fig. 9. Rbrul's best mixed-effect model for the articulation of post-vocalic /r/ in New York City English.

interest, once speaker variation was taken into account. In either case, only data from more speakers could help us decide conclusively whether we are dealing with Type I error by GoldVarb or Type II error by Rbrul.

Conclusions

Variable rule analysis is an essential tool for sociolinguists, whose data, unlike experimentalists', are usually unbalanced across the factors of interest. For binary response variables, VARBRUL allowed sociolinguists to carry out multiple logistic regression when this statistical procedure was fairly new. But if VARBRUL and its successor GoldVarb were cutting-edge when introduced, this is no longer the case. Researchers outside a fairly narrow tradition are hard-pressed to understand GoldVarb's output, and it remains fairly idiosyncratic (with its factor weights and input probabilities) and inflexible (handling interactions and continuous variables with difficulty).

This article has pointed out some advantages of a new variable rule program, Rbrul, and some disadvantages of GoldVarb. Of the latter, the most serious involves the common situation where linguistic tokens are not independent, but grouped. Both simulations and real data demonstrated that GoldVarb overestimates the statistical significance of external factors such as age and gender, whenever individual speakers vary in their behavior

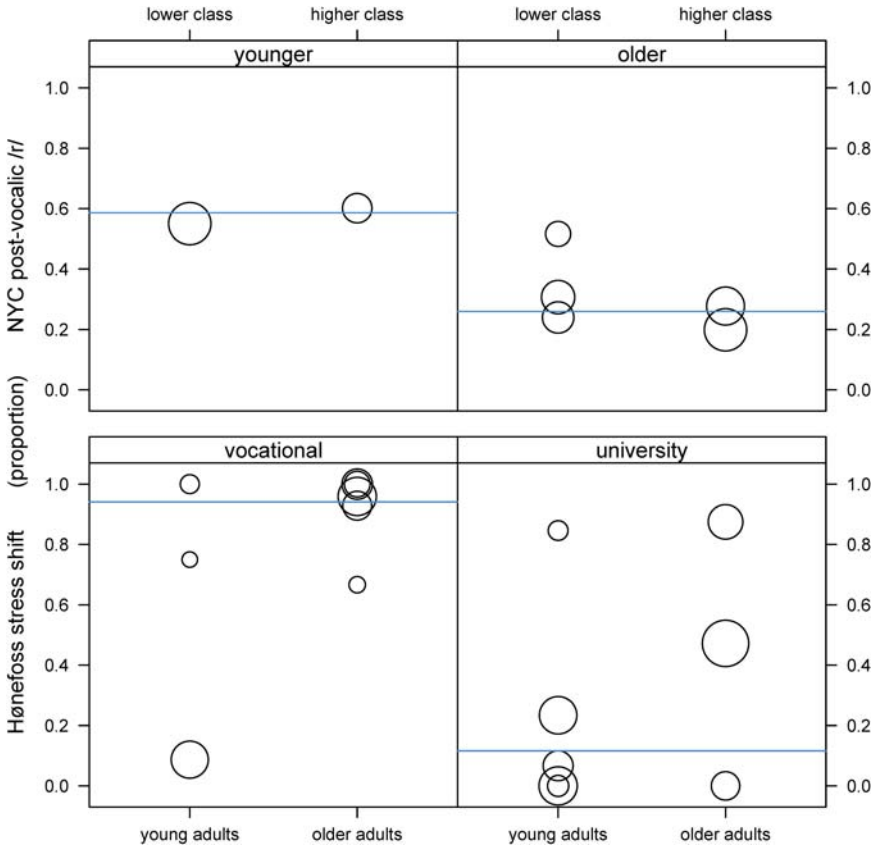

Fig. 10. Speaker means compared with Rbrul group predictions. Bottom: stress shift in Hønefoss Norwegian; top: post-vocalic /r/ in New York City (NYC) English. Across panels, definitely significant factors: education (Hønefoss), age (NYC). Within panels, possibly significant factors: age (Hønefoss), social class (NYC) (within Hønefoss or NYC, the size of each point corresponds to the number of tokens observed for that speaker).

over and above the factors considered in an analysis. Rbrul, by taking speaker grouping into account, provides more accurate results.

It might seem uncontroversial that individual speakers can favor or disfavor a particular linguistic outcome, but in fact there has been much debate on the issue; for an overview, see Wolfram and Thomas (2002: 161–5). Relaxing what they call the 'homogeneity assumption', we can incorporate the possibility of individual variation into our analyses directly, using mixed models. This helps us avoid the criticism that individual speaker agency is lost in quantitative analyses using social categories. With Rbrul, the use of mixed models can proceed in tandem with other approaches that divide the data by speaker or speaker group, fitting and comparing separate fixed-effect models (D'Arcy 2005).

Whether individual lexical items can favor or disfavor a linguistic outcome or process is a matter of at least as much theoretical disagreement. Certainly, exemplar theory (Pierrehumbert 2006) with its 'word-specific phonetics' would strongly predict that they can. If so, then using mixed models would also lead to more accurate conclusions for internal effects like phonological context or lexical frequency.

The Rbrul program is open-source and written in R, which should allow interested parties to improve it and add capabilities that are not implemented in the initial version. For example, there is a middle ground between ignoring the individual completely and assuming that every speaker is different; it might therefore be worth implementing. Rousseau and Sankoff's (1978b) automated speaker-grouping algorithm, which builds speaker groups from the bottom up, rather than assessing the differences between pre-existing social categories. Another useful challenge (Van Herk, personal communication) would be modeling autocorrelation in linguistic data – the so-called parallel processing effect whereby tokens tend to resemble closely preceding tokens of the same variable.

Preparing a GoldVarb token file can only lead to one type of analysis, and one which usually ends up displayed as a table of numbers. Once operating in the R environment, Rbrul users will be able to explore hundreds of other functions for statistical analysis, and in particular for graphical display. R makes graphics like this article's trellis plots relatively easy to create. A number of basic plotting options are also built into Rbrul.

GoldVarb has been hailed as a program 'specifically set up to receive the type of data generated in studies of language variation, and which calculate[s] the results in a form most useful in these studies' (Sankoff 2004: 1157) simply by retaining GoldVarb's capability for logistic regression with categorical predictors, and extending it with support for interactions and continuous independent and/or dependent variables, a program like Rbrul might find some use (although these advantages were not focused on here).

GoldVarb cannot fit mixed models, and the ability to do so fairly easily is the most significant (no pun intended) advantage of Rbrul. While most commercial statistical packages also support mixed-model analysis, the free software R offers the newest and most powerful tools, and Rbrul provides an interface to them.

Whatever statistical software sociolinguists go on to use, much of our data will continue to consist of unbalanced, repeated measurements from different speakers – and of different lexical items. If we proceed without mixed models, grouping heterogeneous data together in fixed-efect regressions, we are endorsing the homogeneity assumption in despite of the facts, and our results will be less accurate, in several respects, than they could be. If individual people (and/or words) do vary, our analyses will profit by embracing (and modeling) this variation, not ignoring it.

Acknowledgements

Thanks to Kara Becker and Nanna Haug Hilton for generously sharing their data, and to Alex D'Arcy, Gregory Guy, Bill Haddican, Florian Jaeger, Nicholas Johnson, Tyler Kendall, Geoff Morrison, John Paolillo, Peter Patrick, Fabian Scheipl, Gerard Van Herk, and two anonymous reviewers for their helpful comments.

Short Biography

Daniel Ezra Johnson received his Ph.D. from the University of Pennsylvania in 2007. His dissertation was about geographic and temporal stability and change in the vowel systems along a dialect boundary in Massachusetts and Rhode Island. As part of that research he analyzed survey responses with mixed-effect modeling, the same technique advocated here for more conventional sociolinguistic data. He currently works on the Accent and Identity on the Scottish-English Border (AISEB) project at the University of York.

Notes

* Correspondence address: Daniel Ezra Johnson, Department of Language and Linguistic Science, University of York, Heslington, York, YO10 5DD, UK. E-mail: dej500@york.ac.uk.

[1] The earliest variable rule formalism is in Weinreich et al. (1968), but the term 'variable rule' itself is first found in Labov's (1969) paper in *Language*.

[2] For an informative, entertaining historical 'tour' of categorical data analysis, including logistic regression, see Agresti (2007: 325–31). The same text offers extensive coverage of ordinary logistic regression, and a shorter chapter on mixed-model logistic regression.

[3] In practice, regression analysis is unnecessary if there is only one contextual factor whose influence is to be assessed. This example is for ease of presentation.

[4] The input probability can be thought of as the predicted probability of the response, averaged over all factor combinations (or cells). If each factor combination is represented in the data by an equal number of tokens, the input probability will be equal to the overall proportion of the response.

[5] Converting from a probability p into log-odds x is called the logistic transformation. The first step, $p/(1-p)$, turns a probability into an odds, a non-negative value centered on 1. The second step, $x = \ln[p/(1-p)]$, takes the natural logarithm of the odds, hence the name log-odds. To convert from log-odds back into probability involves the inverse logistic transformation, $p = e^x/(1+e^x)$.

[6] The examples and simulations were run with R 2.7.1 for Mac OS, using version 0.999375 of the mixed-effect modelling package *lme4*. This package includes functions for fitting linear mixed models and generalized linear (logistic) mixed models, the latter using the Laplace approximation.

[7] Morrison (2005) shows how GoldVarb-like output can be obtained from SPSS, with the advantage of significance testing for interactions and individual factor coefficients. R-Varb (Paolillo 2002b) emulates GoldVarb much more closely than Rbrul does, and has a command-line interface rather than interacting with the user through menus and questions. These points may have contributed to sociolinguists' reluctance to adopt R-Varb. Rbrul's intention is to improve on the functionality of GoldVarb, hopefully enough to make up for any unfamiliarity.

[8] For binary responses, Rbrul reports continuous predictors' effects in log-odds units only. There is no logical way to report the effect of a continuous predictor in terms of factor weights, because a fixed log-odds increase does not always correspond to the same increase in probability.

[9] In the case of continuous responses, factor effects are expressed in the units of the response variable, rather than log-odds or factor weights. Another situation common in sociophonetics, where the response variable is discrete but has more than two important variants and so cannot be collapsed to a dichotomy, calls for a more advanced analysis of a type not yet implemented in Rbrul.

[10] Interactions between effects should not be confused with multicollinearity, which is when substantial correlations exist between two or more of the independent variables in a regression. In such situations, estimates of both significance and effect size can be highly unreliable and unstable (a small change in the data could produce a large change in the coefficients). Neither GoldVarb nor Rbrul tests for multicollinearity, so care must be taken when predictors are correlated. More about multicollinearity, and suggestions for dealing with it, can be found in Baayen (2008).

[11] Like GoldVarb, Rbrul's default is to use a likelihood-ratio chi-squared test to assess the significance of effects, and this procedure was followed for the simulations in this article. A simulation-based significance test is also available, as for mixed models the chi-squared test gives anti-conservative results (i.e. p-values that are too low, but nowhere near as much so as those from fixed-effect models applied to grouped data). See Pinheiro and Bates (2000) for details.

[12] More precisely, it is the errors which are assumed to be independent, meaning that the observations are independent within each cell, or combination of the independent variables. If a model predicts (ing) from speaker gender (among other things), the observations in the female gender category will probably not be independent if they derive from, say, five different women.

[13] GoldVarb uses a 'step-up, step-down' algorithm to decide on the best logistic regression model. Stepping up, it starts with no predictors and adds the most significant factor group, if there is one, before repeating the procedure. Stepping down, it starts with all possible predictors and removes the one that contributes least to the model, and then repeats this until all remaining predictors are significant. Building regression models through automated stepwise procedures is generally frowned upon in today's statistical community, but this is more or less a separate issue from those which will be explored here.

If speaker is included as a possible predictor, it will usually be added before any of the external factor groups, stepping up. Once speaker is in the model, none of the external factor groups will provide additional information, so they will not be selected as significant. For the same reason, all external factor groups will be eliminated when stepping down. The speaker group makes them redundant.

In Tagliamonte's (2006) words, 'any combination of the factor group encoding individual speaker will be non-orthogonal with any social factor. This dictates removing one of the factor groups from the analysis – typically, individual speaker' (p. 182). But as we shall see, this solution brings further problems.

[14] In experimental design terms, speakers are 'nested' within external factors such as gender. This means that all the observations for a given speaker have the same value for gender. The same goes for any other between-speaker factor.

In many sociolinguistic data sets, there is also another type of grouping structure, whereby the word (or lexeme) is nested within some of the internal factors. For example, the individual lexical item might well be nested within a factor representing grammatical category, and also nested within some factors representing phonological context. Just as ignoring grouping by speaker can lead to spuriously significant external effects, ignoring grouping by lexical item can lead to spuriously significant internal effects.

[15] For this example, GoldVarb states the significance as '0.000'. A standard test of equal proportions in R gives $P < 2.2 \times 10^{-16}$, or tantamount to zero.

[16] A t-test for equality of means was repeatedly performed on two sets of 20 randomly generated speaker scores, one centered on 0.60 (and normally distributed so that 95% of the scores fell between 0.45 and 0.75), the other centered on 0.40 (with 95% of scores between 0.25 and 0.55). The p-value was rarely greater than 10^{-8}, and often several orders of magnitude lower.

[17] Given the sets (0.25, 0.35, 0.45, 0.55) and (0.45, 0.55, 0.65, 0.75), a two-sample t-test for equality of means returns a p-value of 0.07 (a result of 0.07 also obtains if the scores are first transformed to log-odds).

[18] This is equivalent to saying that the input probability varies from speaker to speaker, and estimating the magnitude of that variation. The individual speaker estimates (or other random effects) are not formally parameters of the model, but they behave similarly, and can be inspected in a mixed model's output. In this article, we focus more on the fixed-effect side of the mixed models.

[19] The assumption of the simulation is that when individual speakers favor or disfavor a linguistic variable, their deviations in log-odds from the group mean are normally distributed. Normality of random effects is also an assumption of mixed-model analysis. In practice, the mixed model does not require its random effects to be normally distributed. If they are not, however, the quality of inference that can be made from the model suffers.

[20] Actually, both the mixed-model analyses and the GoldVarb-style analyses were conducted with Rbrul, operating with different settings. Before beginning the analyses, it was verified that in its fixed-effect mode, Rbrul provided nearly identical output to the actual GoldVarb program.

[21] The Type II columns from left to right represent an increase in the size of the underlying gender effect that the two programs are attempting to detect. Perhaps unsurprisingly, this variable has a large effect. When the underlying gender effect is a small 0.2, corresponding to factor weights of 0.475–0.525, the Type II error rate for both programs is relatively high, approaching 100% in some conditions. But when the gender effect is a robust 1.6, corresponding to 0.310–0.690 in factor weights, we observe practically no Type II error, unless speaker standard deviation is also at its highest value. However, increasing the gender parameter does not have a consistent effect on the *difference* between GoldVarb's performance and Rbrul's.

[22] The number of tokens per speaker mainly affects Type II error when the gender effect is small. It makes sense that if men average 52.5% and women 47.5% on a variable, 20 observations per person is not enough to discriminate between them, but 160 probably is. If individual variation is high, the problem is different. Then, even if speaker probabilities are estimated precisely with many tokens, their distribution may not support a gender difference above chance level.

[23] The data in Figure 3 come from runs with 20 simulated speakers: ten males and ten females. Other simulations showed that with more speakers, Type II error rates declined, but the difference between GoldVarb and Rbrul was not greatly affected.

[24] If making a Type I error is not a concern for a particular effect – if we know the effect exists and our goal is only to measure it – then significance testing (known in statistics as hypothesis testing) is not necessary.

[25] Paolillo (personal communication) has recently shown that inter-speaker variation can be modeled within GoldVarb by using a number of interaction factor groups. Nevertheless, the current author believes that dedicated mixed-effect modelling software, such as the *glmer* function used by Rbrul, offers a better means to the same end.

[26] With the balanced simulation data, the average effect size sometimes differed between GoldVarb and Rbrul, but only because the calculation only took into account significant effects, and some runs, with smaller effect sizes, were significant in GoldVarb only. For a given data set, the effect size estimate was always nearly the same.

[27] The interviews were of comparable length, but not all Norwegian speakers use loanwords to the same extent, so there is no obvious natural and efficient way to obtain balanced data for this variable.

[28] In addition to being grouped by speaker, these tokens are grouped by word. The most frequent loanword in the corpus is *dialekt*, which occurs 35 times. On the other hand, some 200 words only occur once each. If *dialekt* behaved more or less idiosyncratically, it would not make sense to weight it 35 times more heavily than any of the 200 or so words that only occur once. Therefore, a thorough mixed-model analysis of this data would include a random effect for word. This would lead to different estimates of speaker-internal factors such as orthography and word frequency. The presentation here sets these issues aside.

[29] The deviance, reported by Rbrul, is defined as −2 times the log-likelihood, reported by GoldVarb. In both cases, values closer to zero represent better-fitting models. When two models are compared, the more complicated model almost always fits better, but the improvement is not always worth the greater complication. To obtain the p-value, the change in deviance is tested against a chi-squared distribution, with degrees of freedom equal to the difference in the number of parameters between the two models.

[30] This is true even if age were to be included in the model.

[31] Strictly speaking, it may not be correct to take arithmetical averages of speaker scores, or of proportions between zero and one more generally. Instead, we might average the log-odds values, so that if one speaker scored 0.1 and another scored 0.2 on a variable, the average would come out 0.143, not 0.15. Convenience and tradition has trumped strict accuracy on this point, which becomes more important later. See note 33.

[32] As Hilton (2007) points out, it is probably no coincidence that Hønefoss speakers who use foreign borrowings more often are also more likely to pronounce them as they are pronounced in the original languages. However, the proper statistical analysis of this type of correlation – between token frequency and realization – is likely to be quite complex.

[33] In essence, the mixed model estimates an average speaker score for each educational group, but does so on the log-odds scale. In the example, there are two university-educated speakers with 0% stress shift and three vocationally trained speakers with 100% stress shift. The corresponding log-odds values would be infinite, but the mixed-model software adjusts for this and arrives at an average that – on the probability scale – appears weighted towards more extreme speakers (imagine two speakers, one shifting 5% and one shifting 65%; their combined *tokens*, if balanced, would average 35%; the mean of the two *speaker scores*, averaged via log-odds, would only be 24%). The point is that the speakers in the different education groups really differ more than is recognized by GoldVarb's method of taking group averages over all tokens.

[34] For the New York City data, the factor group for the phonological environment preceding (r) was found to be highly significant. It was included in the models, but in the interest of a more economical presentation it was omitted from Figures 8 and 9 and from discussion.

Works Cited

Agresti, Alan. 2007. An introduction to categorical data analysis, 2nd edition. Hoboken, NJ: Wiley.

Baayen, Harald R. 2008. Analyzing linguistic data: a practical introduction to statistics using R. Cambridge, UK: Cambridge University Press.

Bates, D., and D. Sarkar. 2008. lme4: linear mixed-effects models using S4 classes. http://cran.r-project.org. R package, version 0.999375-8.

Becker, Kara. 2007. /r/, place and identity on the Lower East Side of New York City. University of Colorado at Boulder, paper presented at CLASP.

Breslow, N. E., and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88(421).9–25.

Cedergren, Henrietta J., and David Sankoff. 1974. Variable rules: performance as a statistical reflection of competence. Language 50(2).333–55.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12.335–59.

D'Arcy, Alex. 2005. The development of linguistic constraints: phonological innovations in St. John's English. Language Variation and Change 17(3).327–55.

Fasold, Ralph W. 1991. The quiet demise of variable rules. American Speech 66(1).3–21.

Guy, Gregory. 1988. Advanced VARBRUL analysis. Linguistic Change and Contact, ed. by Kathleen Ferrara et al., pp. 124–36. Austin, TX: University of Texas, Department of Linguistics.

Hilton, Nanna Haug. 2007. The variation of stress assignment in Hønefoss Norwegian. Philadelphia, PA: Poster presented at NWAV 36.

Jaeger, T. Florian. 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language. doi: 10.1016/j.jml.2007.11.007.

Jaeger, T. Florian, and Laura Staum. 2005. That-omission beyond processing: stylistic and social effects. New York, NY: Paper presented at NWAV 34.

Kay, Paul, and Chad K. McDaniel. 1979. On the logic of variable rules. Language in Society 8.151–87.

Labov, William. 1966. The social stratification of English in New York City. Washington, DC: Center for Applied Linguistics.

——. 1969. Contraction, deletion and inherent variability of the English copula. Language 45(4).715–62.

Morrison, Geoffrey Stewart. 2005. Dat is what the PM said: a quantitative analysis of Prime Minister Jean Chrétien's pronunciation of English voiced dental fricatives. Cahiers linguistiques d'Ottawa 33.1–21.

Paolillo, John C. 2002a. Analyzing linguistic variation: statistical models and methods. Stanford, CA: CSLI Publications.

——. 2002b. R-Varb. <http://info.slis.indiana.edu/~paolillo/projects/varbrul/rvarb/>

Pierrehumbert, Janet B. 2006. The next toolkit. Journal of Phonetics 34.516–30.

Pinheiro, José C., and Douglas M. Bates. 2000. Mixed-effect models in S and S-PLUS. New York, NY: Springer.

Rousseau, Pascale, and David Sankoff. 1978a. Advances in variable rule methodology. Linguistic variation: models and methods, ed. by David Sankoff, 57–69. New York, NY: Academic Press.

——. 1978b. A solution to the problem of grouping speakers. Linguistic variation: models and methods, ed. by David Sankoff, 97–117. New York, NY: Academic Press.

Sankoff, David. 1975. VARBRUL version 2. Unpublished program and documentation.

——. 2004. Variable rules. Sociolinguistics: an international handbook of the science of language and society, 2nd edition, ed. by Ulrich Ammon et al., 1150–63. Berlin, Germany: Walter de Gruyter.

Sankoff, David, and William Labov. 1979. On the uses of variable rules. Language in Society 8.189–222.

Sankoff, David, Sali Tagliamonte, and Eric Smith. 2005. GoldVarb X: a variable rule application for Macintosh and Windows. <http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm>

Sigley, Robert. 1998. Quoted by Mario Cal Varela in Sum: GoldVarb (addendum). LINGUIST List 9.1461.

Tagliamonte, Sali A. 2006. Analysing sociolinguistic variation. Cambridge, UK: Cambridge University Press.

Weinreich, Uriel, William Labov, and Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. Directions for historical linguistics, ed. by W. P. Lehmann and Yakov Malkiel, 95–195. Austin, TX: University of Texas Press.

Wolfram, Walt, and Erik R. Thomas. 2002. The development of African American English. Oxford, UK: Blackwell.

Young, Richard, and Robert Bayley. 1996. VARBRUL analysis for second language acquisition research. Second language acquisition and linguistic variation, ed. by Robert Bayley and Dennis R. Preston, 253–306. Amsterdam, The Netherlands: John Benjamins.

