Stratified sampling biases models towards nonlinearity

The case of Kallel (2007)

Aaron Ecay
University of Pennsylvania

Motivation Kallel (2005, 2007) has claimed that the time course of the change from a negative concord system to an NPI-based one in the history of English is better modeled by a quadratic function in logistic space than by a linear one. The use of logistic functions to model language change (and specifically syntactic change) was championed by Kroch (1989), and since the publication of that work they are not unfamiliar. The form of models used is generally homogeneous: they are linear in the time variable (reflecting the character of the modeled process of *change*) and include categorial predictors taken from the linguistic and social context of the tokens (reflecting Kroch's thesis that "the rate of use of grammatical options in competition will generally differ across contexts." (1989)) Indeed, if Kallel's proposal is accepted, it constitutes a fundamental challenge for the theory of syntactic change. To wit, it raises the question of the nature of the relation of parabolas (in logistic space) to language change; to have an empirically informed theory of language change such a relation must be understood. However, we will demonstrate that Kallel's result is an artifact of stratified sampling.

Predictive power The models given by Kallel (2007) are replicated in Table 1, using her dataset without binning. It is suspicious that the value of the coefficient of the year^2 term is so small, but the χ^2 test indicates that the reduction in deviance yielded by the quadratic model is significant ($p = 8.119 \times 10^{-6}$). This is Kallel's result. (Kallel 2007, (19)) Having derived these models, it is possible to ask what, if any, additional predictive power the quadratic one affords. To make such a determination, a cross-validation paradigm will be employed. The dataset is split into two pieces, one containing 95% of the tokens and the other containing the other 5%. Linear and quadratic models will be fit to the larger subset of the data. The parameter values thus obtained will then be used to predict the outcomes in the other subset of the data, given the model inputs (year and coordinate context) associated with each token. The results from this trial are given in Table 2. The improvement given by the quadratic model over the linear one is incredibly marginal. Both models perform markedly better than either of two baselines, one of which predicts a random number in the interval (0,1) as the probability of *any* (regardless of year or coordinate status) and the other of which always predicts 0.5.

Simulation In order to further probe the properties of the two models proposed by Kallel and discussed here, we will construct a simulation procedure to generate a dataset similar in structure to Kallel's corpus (consisting of 944 tokens). Bootstrap resampling from the corpus tells us that Kallel's model is not simply overfit to the data, i.e. the observed drop in model deviance is robust *provided that the sample is a good representation of the underlying population.* It is precisely that proviso that is at issue, and the dangers of ignoring the high degree of correlation between linguistic tokens produced by a single speaker has been pointed out by Johnson (2009) and Gorman (2009) in the context of apparent-time sociolinguistic studies (the latter making reference to the negative concord variation in non-Standard Present Day English). It is this variation that we will attempt to capture in our simulation. The base of the simulation will be a generating model – either a linear or a quadratic logistic function – the parameters of which will be taken from the corresponding model fit to Kallel's corpus. We will take the output of this function to be the "average" underlying rate of *any* use. We will then assume that the variation introduced into a corpus by non-random sampling (e.g. sampling more than once from the same author) can be modeled by a noise term added to the output of the generating model; this noise term will be normally distributed in logit space.[1] It is then necessary to determine the standard deviation of the error term. In Table 3 are the average model deviations (a measure of goodness of fit) of a logit-linear model fit to 2000 simulated data sets resultant by varying the standard deviation of the noise in increments of 0.05. The closest match to the linear model fit to Kallel's actual data (deviance = 725.9) is provided by the value 0.55. On the other hand, Kallel's logit-quadratic model (deviance = 699) is most closely approximated by values of the noise standard deviation very close to 0. The variation expected to inhere in the data based on the sampling procedure is not found – it has been absorbed by the quadratic term in the model.

Having determined the value for the noise parameter, it is possible to use the simulation experiment to test our confidence in the χ^2 test. Specifically, we can count the number of simulated logit-linear datasets in which the test gives a significant p-value. If this is greater than the assumed Type I error rate (0.05), it will be evident that the non-independence in the data has invalidated the test procedure. Indeed, with $\sigma_{\text{noise}} = 0.55$, we obtain a p-value less than 0.05 in 10.6% (of 2000 trials). This indicates that the significance of Kallel's result is illusory – in data known to be generated from a logit-linear model with levels of noise comparable to those in the data, the χ^2 test does not provide a reliable indicator of statistical significance.

Future work The question raised by Kallel (2007) about the role of logit-linear models in the understanding of syntactic change remains unanswered. Traditional methods of data collection have been demonstrated to be inadequate for providing a definitive answer. Truly random sampling from the population of speakers should be used to avoid spurious results. The availability of large-scale corpora (Google Books, Project Gutenberg) makes such an undertaking feasible. If the problem of automatically extracting syntactic information from such sources can be addressed, the linearity hypothesis can be put to a fair test.

[1] This is consistent with the Constant Rate Hypothesis; that is it models each "context" for the purposes of sampling as a parallel copy of the generating model.

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
Intercept	-71.09	4.340	-16.38	2.682×10^{-60}
Coordinate	-1.166	0.1922	-6.069	1.289×10^{-09}
Year	0.04687	0.002850	16.44	9.166×10^{-61}

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
Intercept	727.34	191.78	3.79	1.49×10^{-04}
Coordinate	-1.11	0.20	-5.65	1.57×10^{-08}
Year	-1.01	0.25	-3.97	7.30×10^{-05}
Year^2	0.000346	0.0000837	4.14	3.49×10^{-05}

Table 1: Logistic regression models from Kallel (2007): left, linear; right, quadratic

	Linear	Quadratic	Random	Constant
Mean prediction successes	35.174	35.432	23.233	23.496

Table 2: Results of 1,000 cross-validation iterations (max = 49)

	Median Deviance	
σ_{noise}	Logit-linear	Logit-quadratic
0.05	699.7927	698.8479
0.10	701.6719	700.7385
...	...	...
0.40	716.9312	715.3739
0.45	715.7475	714.6969
0.50	721.5591	720.0839
0.55	722.3519	721.1465
0.60	730.7359	728.7819
0.65	737.8180	736.5060
0.70	740.6798	739.4251

Table 3: Determining the standard deviation of the logit-normal noise function, 2,000 simulated datasets

References

Gorman, K. "The consequences of multicollinearity among socioeconomic predictors of negative concord in Philadelphia." *Penn Working Papers in Linguistics*, vol. 16, pp. 66–75 (2009).

Johnson, D. "Getting off the goldvarb standard: Introducing rbrul for mixed-effects variable rule analysis." *Language and linguistics compass*, vol. 3(1), pp. 359–383 (2009).

Kallel, A. *The Lexical Reanalysis of N-words and the Loss of Negative Concord in Standard English*. Ph.D. thesis, Reading University (2005).

Kallel, A. "The loss of negative concord in standard english: Internal factors." *Language Variation and Change*, vol. 19(1), pp. 27–49 (2007).

Kroch, A. "Reflexes of grammar in patterns of language change." *Language variation and change*, vol. 1(3), pp. 199–244 (1989).