Stratified sampling biases models towards nonlinearity
The case of Kallel (2007)

Aaron Ecay

The University of Pennsylvania

June 3, 2011
Introduction

- This talk’s theme: non-linearities in corpora
- Plan:
  - Background on modeling language change
  - A concrete example of a non-linearity discovered in a corpus
  - Ways of understanding competing models of this non-linearity
  - What does this tell us about diachronic syntax and how to study it?
Table of Contents

Introduction

Background
   The logistic curve
   The challenge

Model testing
   Higher order models
   Simulation
   Crossvalidation

Conclusion
   Nonlinearities in corpora
Introduction

Background

Model testing

Conclusion

The logistic curve

History

- The observation that linguistic changes have a characteristically S-shaped pattern is an old one, dating from at least the mid-20th century

- The identification of the S-shaped curve with the logistic function dates from the early 1980s (Bailey 1973; Kroch 1989; Weinreich, Labov, and Herzog 1968)
History

- The logistic belongs to the family of Lotka-Volterra models, named after their discoverers: Lotka (1925) and Volterra (1926).
- The logistic equation in particular describes the dynamic of two species, one of greater fitness than the other, competing for finite ecological resources.
The use of a logistic function as the model of the time course of syntactic change answers a fundamental question about what language change is:

- The **species** are grammars: abstract mental objects that structure linguistic competence
- The **resource** for which they compete is use in linguistic expression
- Language change is **competition among grammars for use**
The logistic regression procedure fits a logistic function of various predictors to a body of binary data.

It does this via a mathematical transformation (“logit”) that maps logistic-shaped curves onto straight lines and a linear model (slope(s) and intercept(s)).

It has become a standard tool in the analysis of syntactic change: see Santorini (1993), Frisch (1997), and many others.
Misuse

- The use of the logistic regression for the study of language is not statistically “pure”
- The logistic regression model assumes that every data point is absolutely independent of all the others
- Linguists sample from the same speaker more than once
  - Two tokens from the same speaker are not independent of each other!
Other views on language change

Kallel (2005, 2007) explores the change in Middle English from a grammar with negative concord (NC) to one which lacks NC:

1. I would *not* for *no* good (The Lisle Letters, Vol. V: 305)
2. we cannot certaynelye larne by any meanes (The Letters of Sir Francis Hastings L.3)
Other views on language change

- In the cited publications, Kallel explores the hypothesis that the logit-transformed shape of NC loss is not a linear function of time, but rather a quadratic.
- Kallel accepts the hypothesis.

<table>
<thead>
<tr>
<th>Source</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage</td>
<td>1</td>
<td>19</td>
<td>3.86</td>
<td>0.0644</td>
<td>3.86</td>
<td>0.0496</td>
</tr>
<tr>
<td>Type</td>
<td>1</td>
<td>19</td>
<td>14.43</td>
<td>0.0012</td>
<td>14.43</td>
<td>0.0001</td>
</tr>
<tr>
<td>Stage * Stage</td>
<td>1</td>
<td>19</td>
<td>22.93</td>
<td>0.0001</td>
<td>22.93</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Function</td>
<td>1</td>
<td>19</td>
<td>0.57</td>
<td>0.4592</td>
<td>0.57</td>
<td>0.4500</td>
</tr>
</tbody>
</table>
The data

Kallel’s NC corpus, with regression fits

Fit
- Quad.
- Lin.

n
- 10
- 20
- 30
- 40

Context
- Coord
- Noncoord
Other views on language change

- If Kallel’s conclusion is to be believed, it poses a serious question to historical syntax.
- “Constancy across time [= linearity of regression fit – AE], this study shows, is not a requirement; it may or it may not be obtained” (Kallel 2007, p. 47)
  - Linear logistic regression came to be used because language change was thought to be like ecological competition (in the relevant sense). If there is no linear logistic in general, then what is language change?
  - Put another way, years are a plausible regression predictor of language usage because usage patterns change over time. If we accept years-squared as a predictor also, we owe ourselves an explanation of what makes them meaningful in the context of language.
Back from the brink

- I will argue that Kallel’s conclusion is not warranted
- Further statistical analysis of the models used by Kallel fails to show a meaningful difference between the empirical power of the linear and quadratic models
  - By Ockham’s Razor, the traditional linear model is preferred
- The statistical patterns in the data noticed by Kallel may instead be attributable to familiar sources of non-homogeneity in language change: inter-speaker and inter-dialect variation
  - The effects of this variation on the model could come about in several ways
Why not higher-order models?

- Accepting *arguendo* Kallel’s hypothesis that the quadratic logistic is a better fit for NC-loss, it is possible to ask whether a higher-order function might yield a still-better fit.

- An ANOVA test with significance levels derived from the $\chi^2$ test (analogous to Kallel’s procedure) reveals the following results:

| Model    | Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|----------|-----------|------------|----|----------|---------|
| Linear   | 928       | 724.56     | —  | —        | —       |
| Quadratic| 927       | 701.85     | 1  | 22.70    | 0.0000  |
| Cubic    | 926       | 697.45     | 1  | 4.40     | 0.0359  |
| Quartic  | 925       | 694.64     | 1  | 2.81     | 0.0938  |
| Quintic  | 924       | 693.07     | 1  | 1.57     | 0.2098  |
Why not higher-order models?

- A different model-comparison technique, the AIC, yields similar results:

<table>
<thead>
<tr>
<th>Linear</th>
<th>Quadratic</th>
<th>Cubic</th>
<th>Quartic</th>
<th>Quintic</th>
</tr>
</thead>
<tbody>
<tr>
<td>730.56</td>
<td>709.85</td>
<td>707.45</td>
<td>706.64</td>
<td>707.07</td>
</tr>
</tbody>
</table>

- The AIC is a measure of the *likelihood* of the model, penalized by the *number of parameters*. Of a pair of models, the one with a lower AIC is taken to be a better fit.

- This is an even more worrying result than Kallel’s original: we have further increased the explanatory tension between our statistical procedure and our scientific hypothesis about language change.
Structured datasets

- In the area of sociolinguistics, where rich data about speaker background is available, attempts have been made to determine the effect of corpus structure on regression results (Gorman 2009; Johnson 2009)
  - Structure could mean sampling over-extensively from one social class, gender, ethnicity, etc.
  - But it also means sampling more than once from the same speaker’s idiolect, *simpliciter*
- These authors conclude that structured datasets can wreck statistical inference, leading to spurious results
- In the present context, we can understand the variation in the data as composed of two parts:
  - that which is generated by the model
  - that which is generated by unmodeled structure
Simulating structure

- The “deviance” of a model measures how much sample variance a model fails to capture.
- We can quantify the structure in Kallel’s data set with a simulation experiment:
  1. Start by generating a dataset from a model (either linear or quadratic).
  2. Add some noise to this data.
  3. Fit the model to the noisy data, measure the deviance.
  4. The amount of noise that most closely approximates the deviance of the model fit to real data will be a quantitative measure of the amount of structure in said real data.
Simulating structure

Here are the results of the simulation experiment, with 1000 simulated datasets per $\sigma$:

The deviances of the data-fitted models are 724 for the linear model, and 698 for the quadratic.

These are most closely approximated by 0.55 and 0.1 respectively.
Simulating structure

- What these results suggest it that the quadratic model has almost completely soaked up the structural variance in the model
- We know from the sociolinguistic studies that structured variation exists in similar corpora
- So, Kallel’s quadratic term is modeling real variation, but in the wrong way
  - It is not that years-squared are important in themselves
  - Rather, it is necessary to look for the true causes of structured variation in the data
Crossvalidation

- Another paradigm for comparing the linear and quadratic models is to ask the question “How well do the models predict usage?”
- One way to answer this question: collect new data, compare to model predictions
  - ...maybe there isn’t any more data
  - ...maybe data collection methodologies differ
- Easier answer: split the data set in two
  - **Training**: fit the model to one subset of the data
  - **Test**: evaluate the model’s predictions on the other set
Crossvalidation

- Over 100 iterations, with a test set of size 10% of the data, the linear model makes an average of 7.1 errors, whereas the quadratic model makes on average 6.5.
- By comparison, a straight line (not S-curve) fit to the data makes 11 errors; guessing a random probability in (0, 1) results in a mean of 23 errors. The total number of tokens in the training corpus was 93.
- Given the other problems with the quadratic model, this is not impressive enough a result to justify the model’s adoption.
Nonlinearities in corpora

- The regression procedure used by Kallel (2007) was successful in finding a non-linearity in the corpus.

- The analysis of this fact must take into account the understanding of competition in a population that underlies the statistical technique.

- The non-linearity might be due to:
  - Dialect variation (Ingham 2006)
  - Properties of the corpus used? ...
Using Ellegård’s *do*-support corpus and the parsed corpora of Early Modern English (PPCEME, PCEEC), it is possible to calculate a quadratic regression. Those results are shown here:

<table>
<thead>
<tr>
<th>Ellegård (sz. corr):</th>
</tr>
</thead>
</table>
| Intercept                   
| \(-2.114\)                  |
| \(p : 0.000\)               |
| Sent. type: Aff. Q/Neg. Decl |
| \(0.683\)                   |
| \(p : 0.000\)               |
| Sent. type: Neg. Q/Neg. Decl |
| \(1.900\)                   |
| \(p : 0.000\)               |
| Time                        |
| \(56.404\)                  |
| \(p : 0.000\)               |
| Time\(^2\)                  |
| \(-8.520\)                  |
| \(p : 0.029\)               |
The quadratic in other data sets

Using Ellegård’s *do*-support corpus and the parsed corpora of Early Modern English (PPCEME, PCEEC), it is possible to calculate a quadratic regression. Those results are shown here:

<table>
<thead>
<tr>
<th>Parsed corpora:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−2.013</td>
</tr>
<tr>
<td></td>
<td>( p : 0.000 )</td>
</tr>
<tr>
<td>Sent. type: Aff. Q/Neg. Decl.</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>( p : 0.000 )</td>
</tr>
<tr>
<td>Sent. type: Neg. Q/Neg. Decl.</td>
<td>2.104</td>
</tr>
<tr>
<td></td>
<td>( p : 0.000 )</td>
</tr>
<tr>
<td>Time</td>
<td>60.437</td>
</tr>
<tr>
<td></td>
<td>( p : 0.000 )</td>
</tr>
<tr>
<td>Time²</td>
<td>−7.842</td>
</tr>
<tr>
<td></td>
<td>( p : 0.117 )</td>
</tr>
</tbody>
</table>
There are several messages to be distilled from this exploration:

- An understanding of the logistic function as a model of the behavior of a population is essential to historical syntax.
- Using general-purpose corpora might help minimize the impact of non-independence on statistical procedures.
- It is possible to compare models based on their predictive power, not (only) their abstract statistical properties.
- The more data we get, the more likely we are to find non-linearities.
  - These are best understood not as a failure of the competition model of change, but rather an invitation to study further the ways in which linguistic and biological change differ.
Thanks are due to the following individuals and organizations:

- Tony Kroch
- The authors of the parsed corpora used in this study
- Kyle Gorman, for digitizing Kallel’s corpus
- Anthony Warner, for digitizing Ellegård’s corpus
- Mark Liberman, for helpful discussion

Remaining lacunae are, naturally, attributable to the author alone
Questions?
Bibliography I


Bibliography II


