Closely aligning our quantitative methods with our sociolinguistic theories

Josef Fruehwald
4/19/2017

Intro

Outline

- A brief review of how statistical models and sociolinguistic theory were related in the good (& bad) old days.
- In intro to Stan, a system for writing and estimating Bayesian models
- A few examples of bespoke Stan models
- Pros, Cons & Resoures

The good (& bad) old days

Quantitative Methods & Theory

Early variationist work had tight relationship between our hypothesized linguistic theories & our statistical models:

$$(91) \qquad ə \longrightarrow (\emptyset)\ /[\text{*pro}]\ \#\#\begin{bmatrix} \text{—} \\ +\text{T} \end{bmatrix} \overset{\text{C}_0^1}{\text{[*nas]}}\ \#\#\begin{bmatrix} \alpha\text{Vb} \\ \beta\text{gn} \end{bmatrix}$$

which is automatically read as:

$$(92) \qquad \varphi = 1 - \left(\frac{-1 * 1}{-2}\right)(k_0 - \alpha k_1 - \beta k_2 \cdots \nu k_n).$$

Labov(1969)

Quantitative Methods & Theory

To some extent this is still true for, e.g. Stochastic OT & Maximum Entropy Grammars

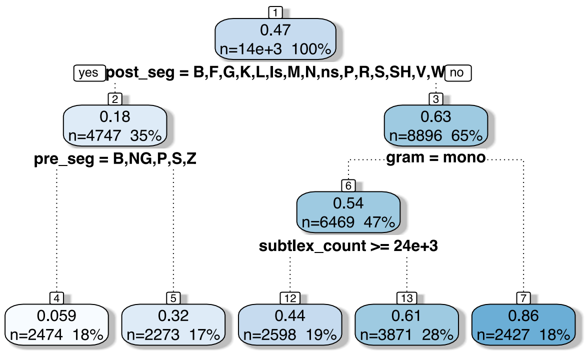
(22)

	2	1	1	H	e^H	p
/guddo/	Id-Vce	OCP-Vce	*Vce-Gem			
guddo		–1	–1	–2	0.14	.50
gutto	–1			–2	0.14	.50

Coetzee & Pater (2011)

Today - Better and fancier statistics

… but less tightly connected to theory



TD Retention

Bayesian Models & MCMC

What is Bayesian Modelling & MCMC?

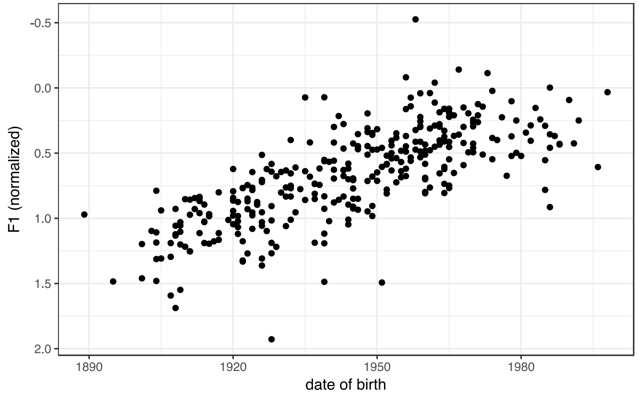
Bayesian statistics is a paradigm of statistical modelling and inference that takes into account prior beliefs about the model parameters being estimated.

- a coefficient of ±5 is about as big as it gets in logistic regression (Gelman et al, 2007)
- variance estimates will skew leftwards, but have a fat right tail (Gelman, 2006)

MCMC, and related methods, are ways of estimating the parameters of a Bayesian model. In this talk, I'll be using Stan (Stan Development Team, 2016).

A Basic Linear Model

Pre-voiceless /ay/ raising in Philadelphia (Fruehwald, 2017).



Writing a Stan model

Composing a model in Stan consists of writing a program in which you:

- Declare the data to be modeled
- You define the parameters of the model,
- You define the statistical constraints (a.k.a. priors) on the parameters,
    - "The intercept is drawn from a normal distribution with mean 0 and sd 100"
- You define the relationship between the parameters and the data.

Stan Model - Data and Parameters

```
data{
  int <lower=0> N;
  real y[N];
  real x[N];
}

parameters{
  real intercept;
  real slope;
  real<lower=0> sigma;
}
```

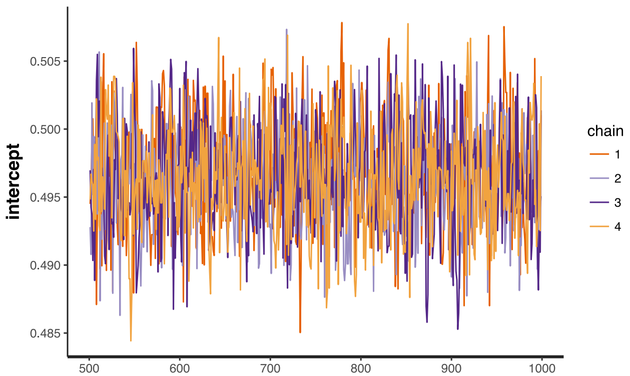
Stan Model - The Model

```
model{
  real mus[N];

  for(i in 1:N){
    mus[i] = intercept + (slope * x[i]);
  }

  intercept ~ normal(0, 100);
  slope ~ normal(0, 100);
  sigma ~ cauchy(0, 100);

  y ~ normal(mus, sigma);
}
```

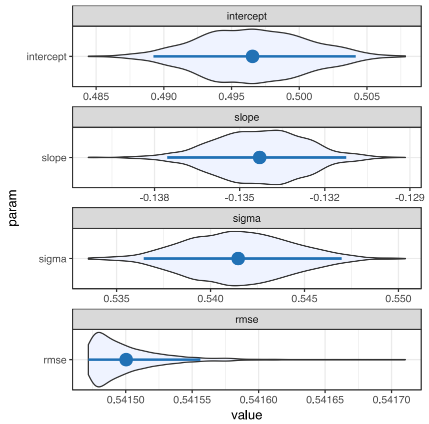

Fitting a the Stan Model

```
ay0 %>%
  mutate(dob0 = (dob-1950)/10)->ay0_to_use

ay0_dat <- list(N = length(ay0_to_use$F1_n),
                y = ay0_to_use$F1_n,
                x = ay0_to_use$dob0)

linear_model_fit <- stan(file = "../stan/linear_model.stan",
                         data = ay0_dat,
                         chains = 4, iter = 1000)
```

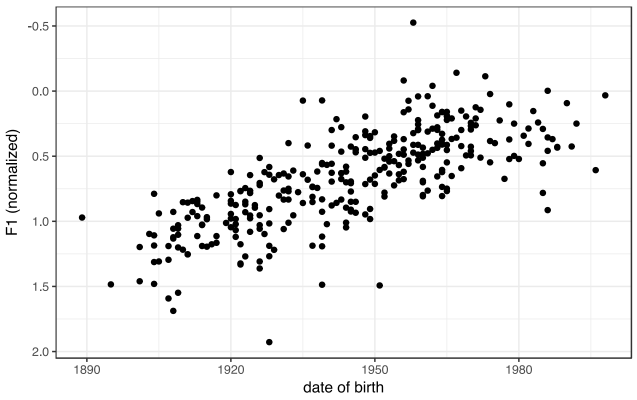

Results



Coefficients



Fits



A mixed effects model

The simple linear model was a "flat" model, but as we all know, we should be including random effects, at least of speaker and word.

Stan - Mixed Effects Data

```
data{
  int <lower=0> N;
  real y[N];
  real x[N];
  int speaker[N];
  int max_speaker;
  int word[N]
  int max_word;
}
```

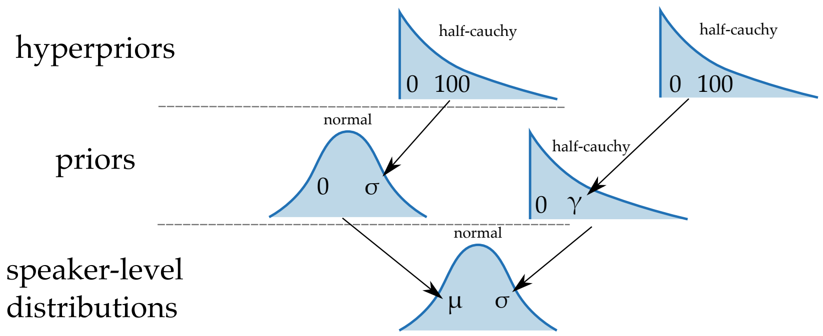

Stan - Mixed Effects Parameters

```
parameters{
  real intercept;
  real slope;
  real<lower=0> sigma;

  real speaker_effect[max_speaker];
  real<lower=0> speaker_sigma;
  real<lower=0> sigma_per_speaker[max_speaker];

  real word_effect[max_word];
  real<lower=0> word_sigma;
}
```

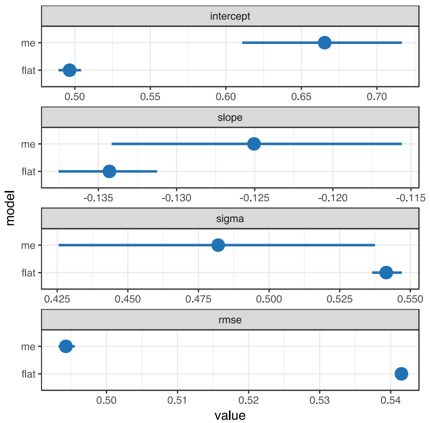
Stan - Mixed Effects Model

```
model{
  real mus[N];
  sigmas mus[N];
  for(i in 1:N){
    mus[i] = intercept + (slope * x[i]) +
              speaker_effect[speaker[i]] + word_effect[word[i]];
    sigmas[i] = sigma_per_speaker[speaker[i]];
  }
  intercept ~ normal(0, 100);
  slope ~ normal(0, 100);
  sigma ~ cauchy(0, 100);

  sigma_per_speaker ~ cauchy(0, sigma);
  speaker_effect ~ normal(0, speaker_sigma);
  speaker_sigma ~ cauchy(0, 100);
  word_effect ~ normal(0, word_sigma);
  word_sigma ~ cauchy(0, 100);
  y ~ normal(mus, sigmas);
}
```

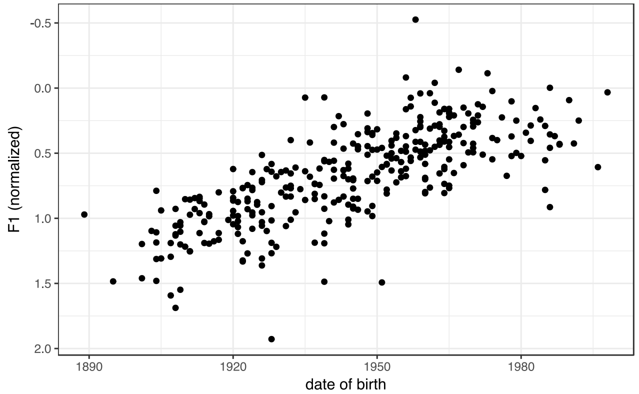

Stan - Model



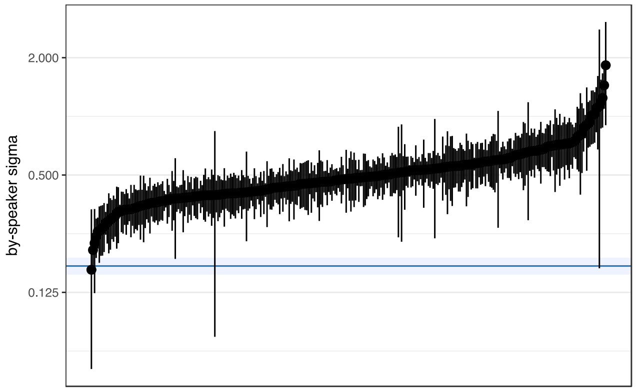
Mixed Effects Comparison



Fits



Speaker sigmas



Matching Models to Theories

Exponential Model

Guy (1991) proposed the following model of TD Retention

level	monomorpheme	semiweak	past
stem	*mist*	*kep[t]*	*miss*
word	*mist*	*kept*	*miss[ed]*
phrase	*mist*	*kept*	*missed*

Exponential Model

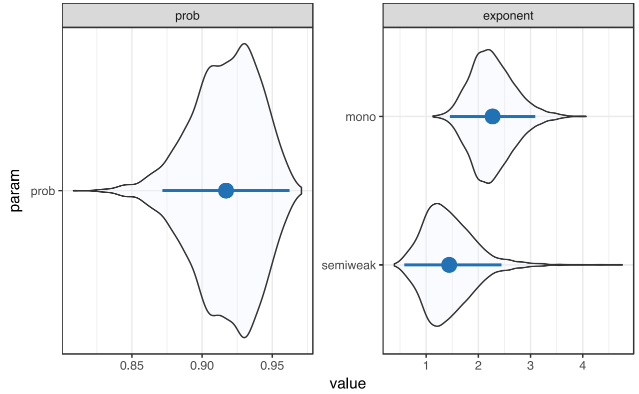
level	monomorpheme	semiweak	past
stem	p_{ret}		
word	p_{ret}	p_{ret}	
phrase	p_{ret}	p_{ret}	p_{ret}
total retention	p_{ret}^3	p_{ret}^2	p_{ret}

Exponential Stan Model

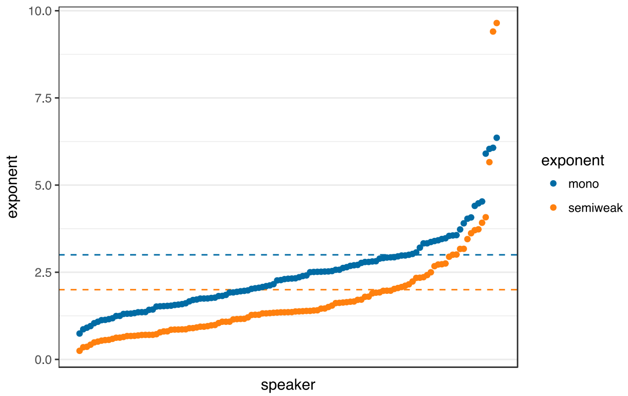
- Estimate a community level and speaker level p_{ret}
- Estimate a community & speaker level exponent j for semiweak verbs
- Estimate a community & speaker level exponent k for monomorphemes
- Estimate random word effects
- Estimate preceding segment effects
- Estimate following segment effects

Data from Tamminga (2014)

Results



Results



Lifecycle & Two Step

Bermúdez-Otero (2010) proposed a slightly different model of TD Retention

level	monomorpheme & semiweak	past
stem	p_{stem}	
word	p_{word}	p_{word}
total	$p_{stem} \times p_{word}$	p_{word}

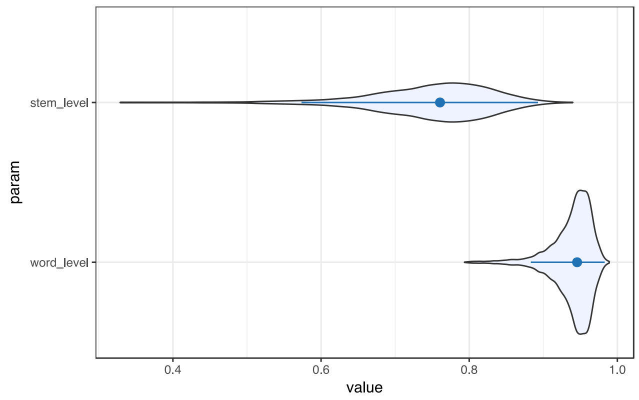
$p_{word} < p_{stem}$

The Model

Using just pre-vocalic /t d/ tokens:

- Estimate community and speaker level word level retention rates
- Estimate community and speaker level stem level retention rates
- Random effects of word
- Preceding segment effects

Results



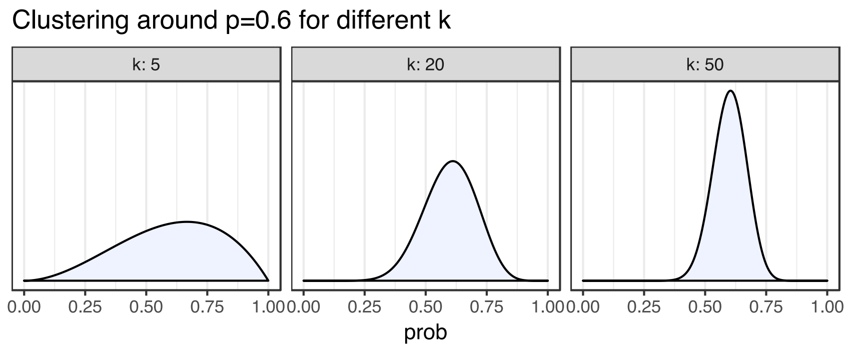
Kohesion

It's also possible to model external factors. For example, if we were interested in how tightly clustered each speaker's $p_{speaker}$ was around some communuty norm p, we could model it like so:

$$p_{speaker} \sim beta(p \times \kappa, (1 - p) \times \kappa)$$

As κ increases, the more tightly clustered speakers' probabilities will be around the community norm.

Illustation



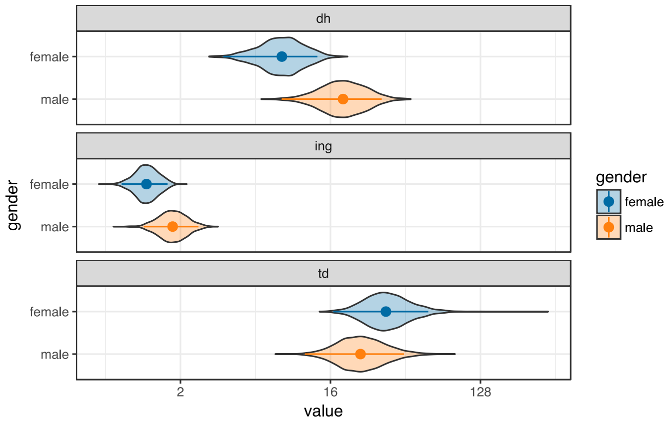
Kohesion Comparison

- TD Deletion
  - monomophemes only
- ING
  - progressive only
- DH

Random effects of word for all variables. Separate kohesion estimates for male and female speakers.

Data from Tamminga (2014)

Results



The Pros & Cons

Pros

We can more directly evaluate our theories if our statistical models closely match them.

With bespoke statistical models, the capabilities of of-the-shelf models need not be the horizion of our analyses.

Cons

Writing fully fledged models can get complex.

- Requires learning more about statistical distributions.
- If you think your R code needs debugging…

Fitting the models can be cumbersome.

- "You set it to run, and go get a hambuger for lunch" - Labov (p.c.) on GoldVarb
- If you thought convergence was an issue in `glmer()`…

It requires explaining the full model, not just "I fit a mixed effects logistic regression".

- But, it has been done (Fruehwald, 2016).

Resources

- Kruschke (2014), *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*
  - a.k.a. "The Dog Book"
  - It's very good
- The Stan manual (http://mc-stan.org/documentation/)
- `rstanarm`
  - An R package that creates Stan models using the familiar R formula interfaces.

The End

References

- Bermúdez-Otero, R. (2010). Currently available data on English t/d-deletion fail to refute the classical modular feedforward architecture of phonology. The 18th Manchester Phonology Meeting. Retrieved from www.bermudez-otero.com/18mfm.pdf
- Coetzee, A. W., & Pater, J. (2011). The Place of Variation in Phonological Theory. In J. Goldsmith, J. Riggle, & A. C. L. Yu (Eds.), The Handbook of Phonological Theory (2nd ed., pp. 401–434). Blackwell.
- Fruehwald, J. (2016). The early influence of phonology on a phonetic change. Language, 92(2), 376–410. http://doi.org/10.1353/lan.2016.0041
- Fruehwald, J. (2017). Generations, lifespans, and the zeitgeist. Language Variation and Change, 29(1), 1–27. http://doi.org/10.1017/S0954394517000060
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, (3), 515–533.
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics, 2(4), 1360–1383. http://doi.org/10.1214/08-AOAS191
- Guy, G. (1991). Explanation in variable phonology: An exponential model of morphological constraints. Language Variation and Change, 3(1), 1–22. Retrieved from http://journals.cambridge.org/abstract_S0954394500000429
- Labov, W. (1969). Contraction, Deletion, and Inherent Variability of the English Copula. Language, 45(4), 715–762.
- Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.14.1. http://mc-stan.org/.
- Tamminga, M. (2014). Persistence in the Production of Linguistic Variation. University of Pennsylvania.