Measuring Rhetoric

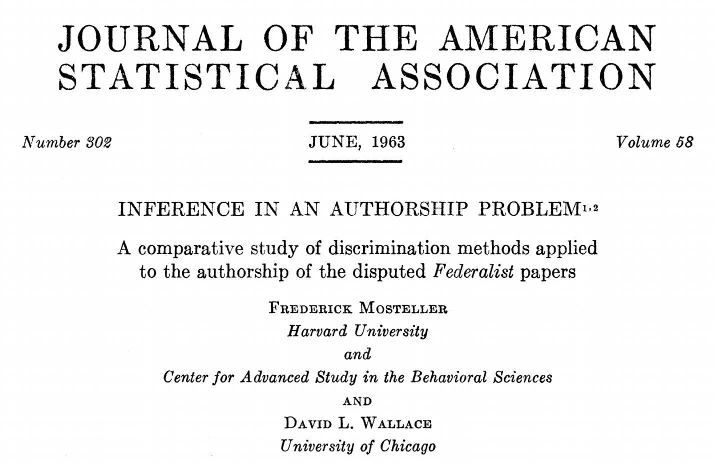
Statistical Language Models in Social Science

Matt Taddy, Chicago Booth

`faculty.chicagobooth.edu/matt.taddy/research`

History: Text as data

Social science text-as-data from the 1960s:
author identification in the Federalist papers.



JOURNAL OF THE AMERICAN
STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied
to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

M+W count words in papers by Hamilton and Madison,

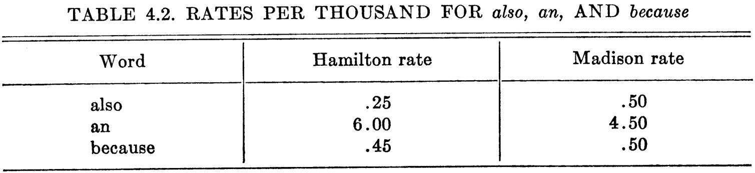


TABLE 4.2. RATES PER THOUSAND FOR *also*, *an*, AND *because*

Word	Hamilton rate	Madison rate
also	.25	.50
an	6.00	4.50
because	.45	.50

then fit models for counts|author (essentially what I use today!), and use Bayes rule to predict authors|counts on disputed work.

$$\mathrm{p}(\text{Hamilton} \mid \text{text}) \approx \frac{\mathrm{p}(\text{text} \mid \text{Hamilton})}{\mathrm{p}(\text{text} \mid \text{Madison}) + \mathrm{p}(\text{text} \mid \text{Hamilton})}$$

The ‘bag of words’

A ‘word’ is a self-contained meaningful token...

- Actual words: ‘all’, ‘world’, ‘stage’, ‘:-)’, ‘`#textdata`’.
- n-grams: ‘merely players’ (bi), ‘men and women’ (tri)
- complicated clauses: parts of speech, act-of-god.
- user selections on a website, domain ids in browser history

All we do is count them.

The remains state of the art!

Treat tokens for each doc as an i.i.d. sample.

Document i is summarized by counts c_{ij} for tokens $j = 1...d$.

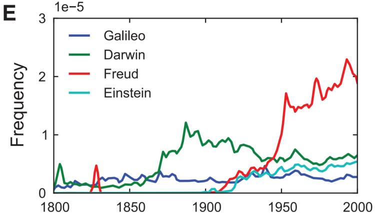
Dumb but works: extra rules aren’t worth their complexity.

Text as data in Social Science

There's been an explosion of interest from social scientists.

Until very recently, one used pre-defined dictionaries.

Picking words: culturomics, Michel et al, Science 2011.



Psychosocial dictionaries, such as Harvard GI in *Tetlock 2007, Giving Content to Investor Sentiment* and others:

able, abundant, accept vs abandon, abrupt, absurd.

Topic Models

Techniques from stats and ML are beginning to filter through and researchers are estimating relationships *from the data*.

A large area of research has developed around *topic models*

$$\mathbf{c}_i \sim \mathrm{MN}(\omega_{i1}\boldsymbol{\theta}_1 + \ldots + \omega_{iK}\boldsymbol{\theta}_K, m_i)$$

a multinomial with probabilities $\sum_k \omega_{ik}\boldsymbol{\theta}_k$ and size $m_i = \sum_j c_{ij}$.

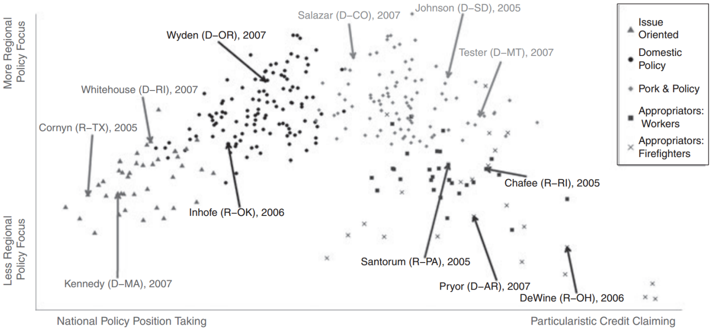
(Latent Dirichlet Allocation; Blei, Ng, Jordan 2003)

This is a factor model for count data.

Topics provide low-D structure, which the SS interprets.

Especially common in PoliSci; King, Grimmer, Quinn, ...

FIGURE 1 A Typology of Home Styles in the U.S. Senate



Grimmer 2013: fit latent topics in press releases
(e.g., ‘Iraq’, ‘Infastructure’) then investigate who uses what topic.

Structured topic models

The basic topic model finds *dominant sources of variation* in **C**.

In SocSci, this is often not what we're seeking (needle + haystack).

There is a huge industry on extensions to topic models that push the topics to be relevant or interpretable for specific questions.

supervised TM, dynamic TM, structural TM, IR TM ...

These model weights (ω) and topics (θ) as functions of covariates.

Lots of good and interesting work.

Multinomial Regression

Instead of jumping straight to latent structure, perhaps we can answer our questions by regressing the text onto observables '**v**'.

Massive response logistic regressions:

$\mathbf{c}_i \sim \mathrm{MN}(\mathbf{q}_i, m_i)$ with $q_{ij} = e^{\eta_{ij}} / \sum_l e^{\eta_{il}}$

$\eta_{ij} = \alpha_j + \mathbf{v}_i' \boldsymbol{\varphi}_j$ is a 'log intensity' $\approx \mathrm{E} \log(c_{ij}/m_i)$

This is a regression like any other.

We will be estimating *partial correlations*, can build *random effects* and *interactions* into **v**, ... all our familiar regression ideas apply.

MN Inverse Regression + rejoinder, 2013. Political Sentiment on Twitter, 2013. Distributed MN Regression, 2015.

Distributed Multinomial Regression

A regression like any other, except the response is super HD.

We approximate the MN likelihood with *independent* Poissons:

$$c_{ij} \sim \mathrm{Po}(\ m_i e^{\eta_{ij}}\ )$$

$\Rightarrow$ you can estimate each regression fully independently!

This works because MN dependence is *only induced by totals*.

DMR is equivalent to MN logit in a variety of simple examples, and is shown empirically to perform well in more complex settings.

Everything in distribution: estimation, penalization, selection ...

More precisely, start from the Poisson:

$$c_{ij} \overset{ind}{\sim} \mathrm{Pois}\left(\exp\left[\mu_i + \eta_{ij}\right]\right)$$

where μ_i is a ‘verbosity’ nuisance parameter.

This model leads to

$$\Pr\left(\mathbf{c}_i \mid m_i\right) = \frac{\prod_j \mathrm{Po}\left(c_{ij}; \exp\left[\mu_i + \eta_{ij}\right]\right)}{\mathrm{Po}\left(m_i; \sum_l \exp\left[\mu_i + \eta_{il}\right]\right)} = \mathrm{MN}(\mathbf{c}_i;\ \mathbf{q}_i, m_i)$$

Thus, given m_i, Poisson and MN imply the same model.

DMR fixes $\hat{\mu}_i = \log m_i$, so LHD factorizes to independent Poissons.

More generally: for Big Data, consider using plug-in [marginal] estimates of parameters about which you have little uncertainty. Focus computation on the bits that are hard to measure.

Yelp Reviews

We'll illustrate using publicly available review data from Yelp.

- $n = 215{,}879$ reviews on 11,535 businesses by 43,873 users.
- taken around Phoenix AZ on January 19, 2013.
- $d = 13{,}938$ words in more than 20 reviews.

The reviews are marked with review, business, and user attributes: number of stars, user and business star averages, business type (333 overlapping), and the number of funny/useful/cool votes.

Each *word-j* intensity regression equation has

$$\eta_{ij} = \alpha_j + \mathbf{a}_i'\varphi_j^a + \mathbf{b}_i'\varphi_j^b$$

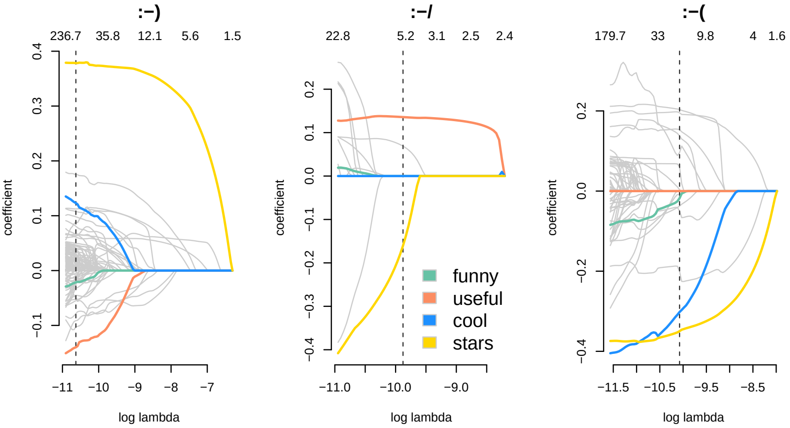
where we've resolved the meta-data attributes $\mathbf{V} = [\ \mathbf{A}\ \mathbf{B}\ ]$ into variables of primary interest and those viewed as controls

- **a:** star and vote counts; 11 dimensional.
- **b:** 400 categories and 11,500 business random effects.

Estimate parameters to minimize the penalized Poisson deviance

$$\hat{\alpha}_j, \hat{\boldsymbol{\varphi}}_j = \mathrm{argmin}\left\{ l(\alpha_j, \boldsymbol{\varphi}_j) + n\lambda \left[\sum_k \omega_{jk}^a |\varphi_{jk}^a| + \frac{1}{\tau} \sum_k \omega_{jk}^b |\varphi_{jk}^b| \right] \right\}$$

one-step estimator paths for concave regularization, 2015.



Poisson regression regularization paths under relative weight $\tau = 2$. AICc selection is marked.

Resolving correlated effects

Bigger τ gives fewer but *cleaner* nonzero terms.

	τ	$\lvert\hat{\varphi}\rvert_0$	**top ten words by loading**
+stars	marg		*great love amaz favorite deliciou best awesome alway perfect excellent*
	2	8440	*unmatch salute :-)) prik laurie pheonix trove banoffee exquisite sublime*
	20	3077	*heaven perfection gem divine amaz die superb phenomenal fantastic deliciousnes*
	200	508	*gem heaven awesome wonderful amaz fantastic favorite love notch fabulou*
-stars	marg		*not worst ask horrib minut rude said told would didn*
	2	8440	*rude livid disrespect disgrace inexcusab grosse incompet audacity unmelt acknowl*
	20	3077	*rude incompet unaccept unprofession inedib worst apolog disrespect insult acknowl*
	200	508	*worst horrib awful rude inedib terrib worse tasteles disgust waste*
funny	marg		*you that know like your yelp ... what don who*
	2	6508	*dimsum rue reggae acne meathead roid bong crotch peni fart*
	20	1785	*bitch shit god dude boob idiot fuck hell drunk laugh*
	200	120	*bitch dear god hell face shit hipst dude man kidd*
useful	marginal		*that yelp you thi know biz-photo like all http ://*
	2	5230	*fiancee rife dimsum maitre jpg poultry harissa bureau redirect breakdown*
	20	884	*biz-photo meow harissa www bookmark :-/ http :// (?), tip*
	200	33	*www http :// com factor already final immediate ask hope*

I think $\tau = 20$ strikes a good balance here.

Sufficient Reduction

What is the funny/useful/cool content of a review?

Coefficients $\boldsymbol{\Phi}$ are a linear map from text to attribute space.
They provide a *sufficient reduction*. For example,

$$\mathbf{a}_i \perp\!\!\!\perp \mathbf{c}_i \mid \boldsymbol{\Phi}^a \mathbf{c}_i, \mathbf{b}_i, m_i$$

where $\boldsymbol{\Phi}^a$ are loadings relevant to our 'primary interest' covariates.

In words: the 11 dimensional $\mathbf{z}_i = \boldsymbol{\Phi}^a \mathbf{c}_i$ contains all the information in the text that is *directly* relevant to $\mathbf{a}_i$, controlling for $\mathbf{b}_i$ and m_i.

Funniest and most useful 50-100 word review, as voted by Yelp users (votes normalized by square root of review age).

I use to come down to Coolidge quite a bit and one of the cool things I use to do was come over here and visit the ruins. A great piece of Arizona history! Do you remember the Five C's? Well, this is cotton country. The Park Rangers will tell you they don't really know how old the ruins are, but most guess at around 600 years plus. But thanks to a forward thinking US Government, the ruins are now protected by a 70 foot high shelter. Trust me, it comes in handy in July and August, the two months I seem to visit here most. LOL. I would also recommend a visit to the bookstore. It stocks a variety of First Nation history, as well as info on the area. http://www.nps.gov/cagr/index.htm. While you are in Coolidge, I would recommend the Gallopin' Goose for drinks or bar food, and Tag's for dinner. Both are great!

50-100 word review with the most funny content,
as measured by SR projection $z_{\texttt{funny}} = \hat{\phi}'_{\texttt{funny}}\mathbf{c}$.

Dear La Piazza al Forno: We need to talk. I don't quite know how to say this so I'm just going to come out with it. I've been seeing someone else. How long? About a year now. Am I in love? Yes. Was it you? It was. The day you decided to remove hoagies from your lunch menu, about a year ago, I'm sorry, but it really was you...and not me. Hey... wait... put down that pizza peel... try to stay calm... please? [Olive oil container whizzing past head] Please! Stop throwing shit at me... everyone breaks up on social media these days... or haven't you heard? Wow, what a Bitch!

most funny by $z_{\texttt{funny}}/m$: *Holy Mother of God*

50-100 word review with the most useful content,
as measured by SR projection $z_{\texttt{useful}} = \hat{\phi}'_{\texttt{useful}}\mathbf{c}$.

We found Sprouts shortly after moving to town. There's a nice selection of Groceries & Vitamins. It's like a cheaper, smaller version of Whole Foods. [biz-photo] [biz-photo] We shop here at least once a week. I like their selection of Peppers....I like my spicy food! [biz-photo][biz-photo][biz-photo] Their freshly made Pizza isn't too bad either. [biz-photo] Overall, it's a nice shopping experience for all of us. Return Factor - 100%

most useful by $z_{\texttt{useful}}/m$: *Ask for Nick!*

The SR projections are based on *partial correlations*.

E.g., compare the correlation matrices

*attributes (**v**)*

	f	u	c	⋆
funny	1	0.7	0.8	0
useful	0.7	1	0.9	0
cool	0.8	0.9	1	0
stars	0	0	0	1

*text projections (**z**)*

	f	u	c	⋆
funny	1	-0.1	-0.7	-0.4
useful	-0.1	1	0.1	-0.2
cool	-0.7	0.1	1	0.5
stars	-0.4	-0.2	0.5	1

SR projections make great inputs to prediction algorithms: MNIR.

Confounder Adjustment

Text is also a useful, but high-dimensional, control.

For example: does a user's experience effect their review ranking?
Do they get more positive, say because of a yelp community effect?

Given the same review, do experienced users give more/less stars then a newbie? We can answer by *controlling* for review content.

Treatment effect of experience on rating

- response attribute, v_{iy}, is *star rating*.
- treatment, v_{id}, is the log *number of reviews* by the author.
- controls are $\mathbf{v}_{i,-yd}$ (everything else) and $\mathbf{c}_i$.

Given our MN language model,

$$v_{iy}, v_{id} \perp\!\!\!\perp \mathbf{c}_i \mid z_{iy}, z_{id}, m_i, \mathbf{v}_{i,-yd}$$

So the *joint distribution* of treatment and response is independent of review content ($\mathbf{c}$) given the SR projections on each (z_{iy}, z_{id}).

We can control for this content, *and its interaction with business classification*, and estimate treatment effect γ in

$$\mathrm{E}[v_{iy}] = \gamma v_{id} + f(z_{iy}, z_{id}, m_i, \mathbf{v}_{i,-yd})$$

This gives $\hat{\gamma} = 0.02$, vs a marginal effect near zero and an effect of 0.015 conditioning on $\mathbf{v}_{i,-yd}$ alone. Interacting $\mathbf{c}_i$ with biz class gives 5 million coefficients ⇒ not enough degrees of freedom.

Measuring segregation in high dimensions

With Gentzkow+Shapiro, the congressional record since 1873 and a question: "How has partisanship of rhetoric changed over time?"

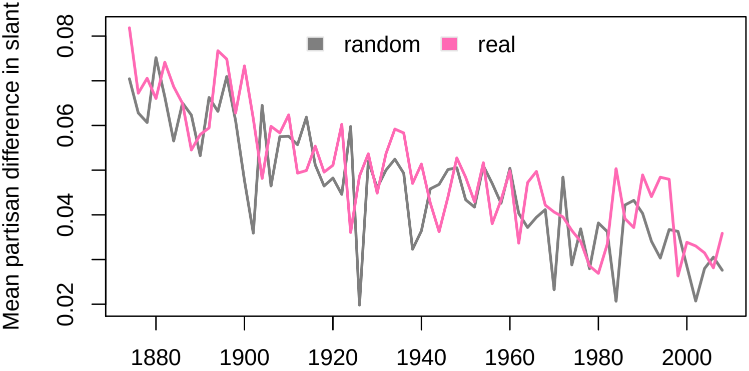
The concept here is of speakers across parties:

- using different words to describe the same thing
  `tax.cut`/`tax.break`, `war.on.terror`/`war.in.iraq`
- choosing to focus on different substantive topics
  `stem.cell`, `african.american`, `soldier.sailor`.

And doing this because of party membership or ideology.

In HD, indices of segregation can have strange properties: variation in the index is dominated by your ability to measure it.

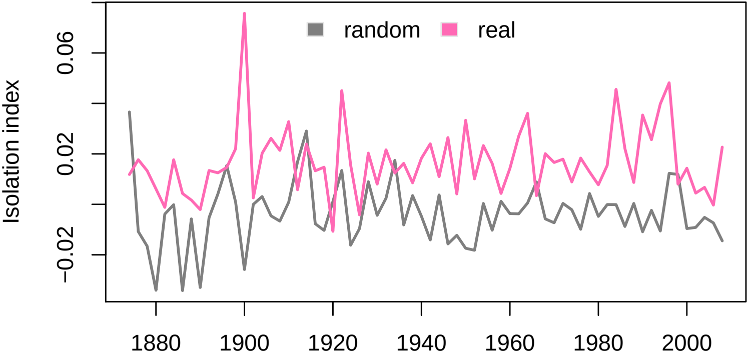
G+S 2013, *Slant index* on 10k most common phrases



Real is Dem v. GOP. Random is a random permutation.

In HD, indices of segregation can have strange properties: variation in the index is dominated by your ability to measure it.

G+S 2011, *Isolation index* on 10k most common phrases



Real is Dem v. GOP. Random is a random permutation.

Instead, we write down a multinomial choice model with intensities

$$\eta_{ijt} = \alpha_{jt} + \mathbf{u}'_{it}\gamma_j + \varphi'_{jt} r_{it}$$

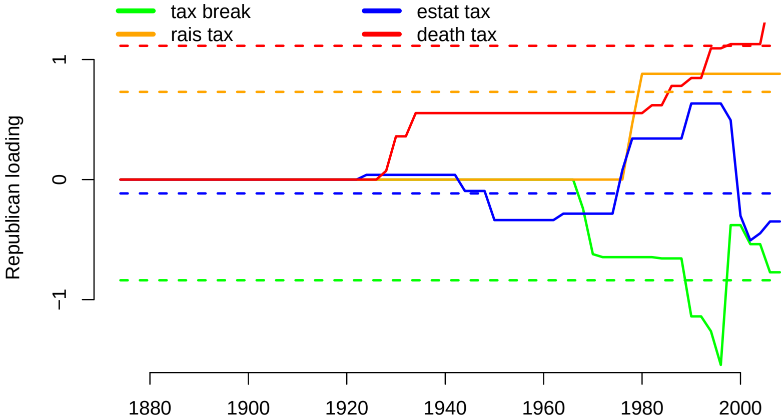
So that η_{ijt} is the mean utility of phrase j for speaker i in session t.

$\mathbf{u}_{it}$ are measured attributes of speaker i in session t (e.g., state, chamber), excluding the Republican party membership indicator r_{it}.

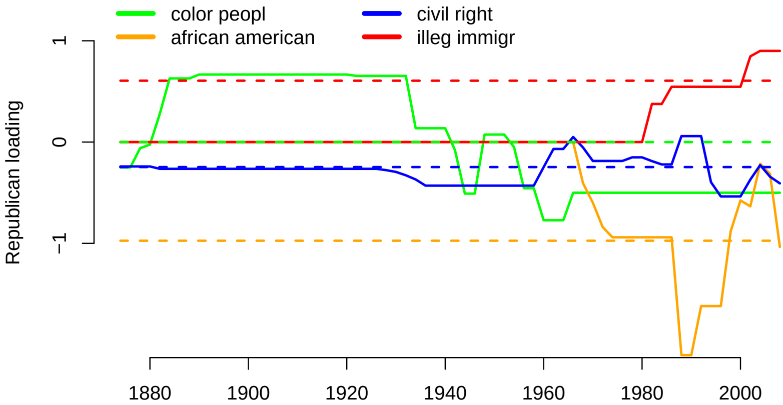
The SR projection for (Republican) partisanship is $z_{it} = \boldsymbol{\varphi}'_t \mathbf{c}_{it}$.

Here, this is the expected utility gain to a Republican relative to a Democrat from speaking exactly like speaker i in session t.

We use a spline model to allow φ_t to change slowly in time.

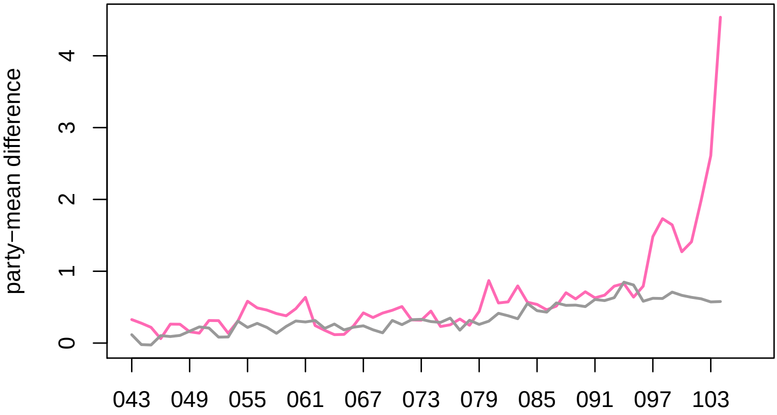


We use a spline model to allow φ_t to change slowly in time.



SR measures *preferences* in a structural model.

Segregation of preferences is $\bar{z}_{\mathtt{gop},t} - \bar{z}_{\mathtt{dem},t}$.



We see a clear pattern: partisanship explodes after Carter.

Messaging and Negotiation on eBay.com

eBay has a ‘best offer’ button; a buyer can use this to circumvent the auction or fixed price sale, and make a direct offer to the seller.

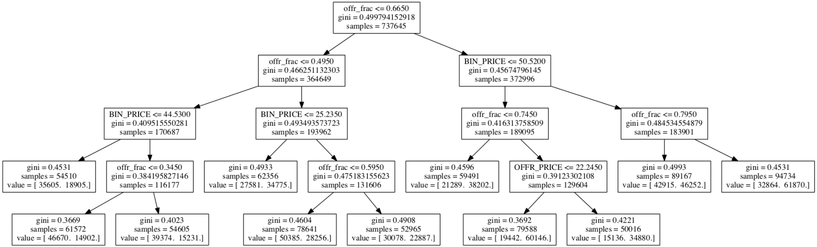
We track the communication.

After controlling for parameters of the transaction (item, buyer offer, seller/buyer info, ...) what language leads to a higher probability of a seller responding with a deal or counteroffer?

We can use the results to educate buyers, give templates/examples, or generate hypotheses on bargaining behavior.

To isolate the targeted effect, we first remove what was predictable from the 0/1 response, y : ‘did seller respond to buyer?’

Fit $p_i = \mathrm{p}(y_i = 1|\mathbf{x}_i)$ as an *Empirical Bayesian Forest* ($\approx$ an RF).
$\mathbf{x}$: buyer, seller, item attributes (even includes sale title topics).



Then we look at the SR projection from text onto $r_i = y_i - \hat{p}_i$.

First, without the synthetic treatment residual:

Log intensity $\eta_{ij} = \alpha_j + \varphi_j^y y_i + \mathbf{x}_i'\boldsymbol{\beta}_{\boldsymbol{j}}$, SR $z_i^y = \mathbf{c}_i'\boldsymbol{\varphi}^y$.

The highest z_i all showed evidence of previous deals and bundling. The lowest z_i are agressive offers, driving a hard bargain:

know this low but its based on what individual [product] selling

give you two cash you ship it free its very common low end [product] has broken pin good pin youd be lucky get this real its worth melt price dont believe me its really not worth listing fees that you paid

hi my offer basically what price guide lists individual [product] it may seem too low you but it never hurts place offer regards [name]

do you know that clubhouse sigs fakes worthless ridiculous fake sigs thats why other ball sold they were real sigs real sigs lot better its like gehrigs wife signing his name not worth anything

Now, with residual r_i as treatment:

Log intensity $\eta_{ij} = \alpha_j + \varphi_j r_i + \mathbf{x}_i' \boldsymbol{\beta}_{\mathbf{j}}$, SR $z_i = \mathbf{c}_i' \boldsymbol{\varphi}$.

The highest z_i are still previous deals and bundling.
But the lowest z_i are now pleading:

kcollect [product] know that my offer not close what you asking cant afford pay much more than my offer but really want this pin please let me know you can make counteroffer this offer not acceptabl

you dont like offer feel free tell me what lowest youll go on these [product] like you ive been collect since was im now hope we can reach deal wich fair both us by way they some awsome [product] !!

hello my friend hope my offer good enough really want [product] ... cgc going be there they grading comics also do you know how much they charge grade comic thkas greg

last one sold on dec 26th ,, can match that price im also interested [product] was wondering how much pair ?? can have money ur account tonight we can reach deal thanks your time ,,

wrap up

Big picture: Give regression a chance!

Everything here – random effects, synthetic controls, lasso variable selection, utility interpretations – is common in regression.
We can apply the same ideas to text via DMR.

Future pitch: I've been learning about the recently popular 'deep' distributed language models (e.g., word2vec) that have a word's probability dependent upon its neighbors.

It is pretty straightforward to add covariates into these 'models', and it should lead to even cleaner SR projections.

Thanks!