Data Mining Summary with Intro to Text Mining

Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania

Wharton
Department of Statistics

Questions

- What's the best way to divide data for CV?
  - Do you want a good model or compare models?
  - For comparison, models tend to be similar, so need lots of test data to distinguish models
- Will I be able to use JMP when I'm not in AA?
  - Get JMP software, iPad graph builder, free trial
  - Mixing these methods into other approaches
  - James text introduces methods in R
- Data mining and theory are compatible...
  - Scaled variables are more predictive in ANES
    - Ideology, political affiliation are important predictors
  - Diagnostic: Is my theory complete?

Comparisons

Regression

- Regression
  - Great tool, particularly when have a theory or strategy to guide choice of explanatory vars
  - P-values protect against over-fitting when used properly to control selection process

    Avoids need for using CV samples to tune model
- Neural networks
  - Best for rich signal, low noise problems
  - Over-fitting likely because so many parameters
  - Need 3-way CV to control modeling

    Train, tune ('validation'), and test to pick form of hidden layers
  - Need 4-way CV to get unbiased estimate of how well chosen network predicts new data

Trees

- Trees
  - Regression with data-driven splitting variables
  - Local averages are grainy, highly variable
  - Model averaging helps by smoothing
    Random forest
    Boosting
- Strategy
  - Theory-based regression gets main structure
  - Tree, NN search for omitted nuances
  - Question of the amount of data available
- Caution about validation
  - Need to repeat the analysis

Don’t Forget About...

- Concepts
  - Bonferroni, multiplicity, and selection bias
  - Missing data, measurement error
  - Causality versus association, interpretation
  - Interaction and linearity in models
  - Model averaging

You can do more of this, such as combining logistic with tree...

- Techniques
  - Plot linking and brushing
  - Spline smoothing
  - Model profiling
  - Calibration
  - ROC curves, confusion matrix, decision threshold

10 Tips for Data Miners

1. Substantive questions propel data analysis
2. Pick the low-hanging fruit
3. Models are not causal without experiments
4. Data is seldom (ever?) missing at random
5. Dependence is prevalent
6. Optimize what you care about
7. Use tools that you understand
8. You cannot validate a model, only a process
9. Cross-validation is optimistic and often unstable
10. Know your audience

Principal Components Analysis

Underlies many ideas in
vector space models for text

Principal Components

- Motivating model
  - Underlying model has few variables (eg, two)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

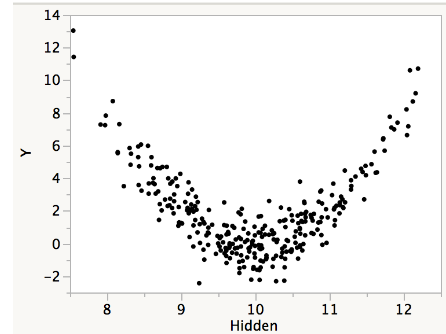
  - But you don't observe X_1 or X_2
- Instead observe noisy versions

$$X_j^* = X_1 + \varepsilon \quad \text{or} \quad X_j^* = X_2 + \varepsilon$$

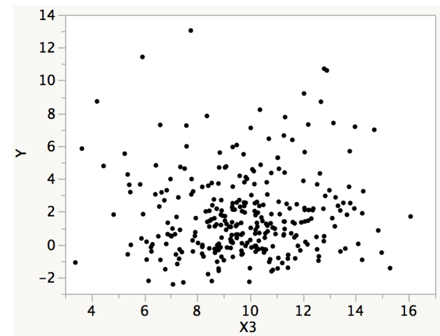
latent variables

- What would you do?
  - Averaging X_j^* works if know which to combine
- Principal component analysis (PCA)
  - Finds uncorrelated clusters of variables
  - Requires variables on common scale (standardized)
  - Derive from eigenvalue/eigenvector calculations

Example of PCA



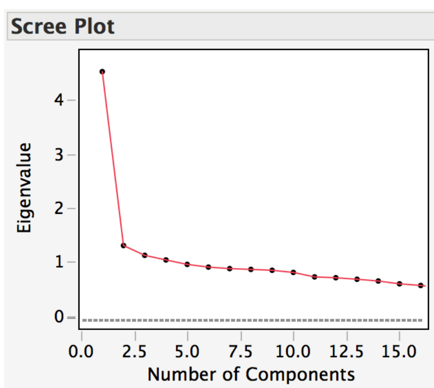
- True structure
  - Y is quadratic in latent variable
- Observed variables
  - None of 20 observed variables is related to the response.



- Finds weighted combination of X's with most variation
  - Only one component suggested

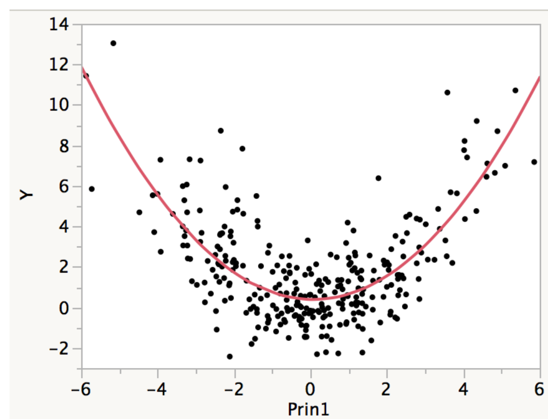
No peeking at Y!

"Unsupervised"



Loading Matri

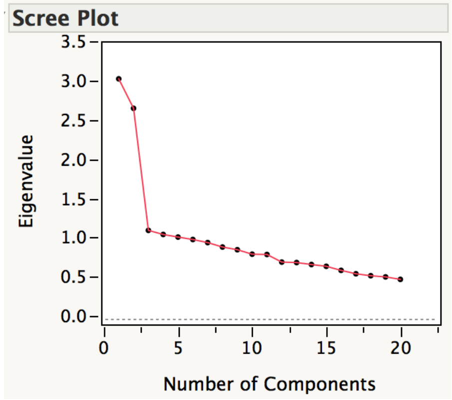
	Prin1
X1	0.50861
X2	0.34985
X3	0.46691
X4	0.45354
X5	0.45324
X6	0.50252
X7	0.41391
X8	0.38666
X9	0.63441
X10	0.46491
X11	0.50693
X12	0.50147
X13	0.55193
X14	0.48366
X15	0.48310
X16	0.45388
X17	0.45365
X18	0.49223
X19	0.50703
X20	0.42771

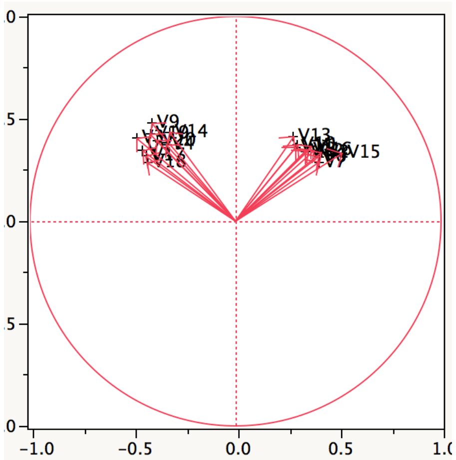


Wharton
Department of Statistics

Example of PCA

- Simulated data
  - n = 500 with 2 underlying variables
- Scree plot shows variance captured by eigenvectors
  - Scree piles at bottom of hillside
- Variables align in two clusters





Loading Matrix

	Prin1	Prin2	Prin3	Prin4
V1	0.36725	0.36446	0.39212	-0.00336
V2	0.41717	0.33465	0.03241	0.08537
V3	-0.47978	0.40362	0.07302	-0.08944
V4	-0.33631	0.37102	-0.06603	0.20634
V5	-0.45054	0.34358	0.25332	-0.30096
V6	0.40554	0.31592	0.15720	0.38004
V7	0.40614	0.28853	0.13687	-0.16689
V8	0.35308	0.34374	-0.26487	0.03047
V9	-0.40655	0.47737	-0.06210	-0.03225
V10	0.30319	0.37220	0.34865	-0.46938
V11	0.29285	0.36064	-0.48268	-0.26186
V12	0.33925	0.34172	0.06733	0.31529
V13	0.28308	0.41295	-0.15789	-0.37687
V14	-0.31960	0.43387	-0.09288	0.33658
V15	0.51770	0.33078	0.27168	0.17809
V16	0.38170	0.34417	-0.46717	0.08215
V17	-0.43661	0.31669	-0.02303	-0.12861
V18	-0.42891	0.28520	0.23079	0.10494
V19	-0.41546	0.42424	0.01612	0.20049
V20	-0.38311	0.39468	-0.20747	-0.05532

Text Mining

Short introduction to
vector space models for text

Text Mining

- What is text mining?
  - Variety of answers for different audiences
  - Focus on using text to predict a response
  - Building explanatory variables
- Applications
  - Interpreting answers to open-ended questions

    Responses to office awareness in ANES 2008

    Illustrate basics in R
  - Prices in real estate listings

    Predict price from the text of a listing, as in the following

    Larger, summarize results

$399000. Stunning skyline views like something from a postcard are yours with this large 2 bedroom, 2 bath loft in Dearborn Tower! Detailed hardwood floors throughout the unit compliment an open kitchen and spacious living-room and dining-room. Huge walk-in closet, steam shower and marble entry. Parking available.

Office Recognition

ANES 2008

- Assess political knowledge
  - Who's Nancy Pelosi?
  - Coded "Yes" or "No" by hand
- Answers are more varied
  - Store text of replies in R vector of strings ('text')

```
[1329] "She's the head of something.  I can't really think right now.   Um, she's head of the House or I don't know."
[1330] "I don't know.   Court system."
[1331] "Nancy Pelosi  Ah, I know this.  I don't even know.  I know it but I don't.   I want to say she's in the Senate."

[1362] "Republican.   The one that was gonna be vice president with McCain.  No.  Who is she, anyway?"
[1363] "I don't know."
[1364] "She's Speaker of the House."
```

- Other ways to capture the patterns and variety of the responses?

Basic Text Processing

- R Packages
  - tm   (short for text mining)
  - text = vector of strings from data file

10_anes_text.R

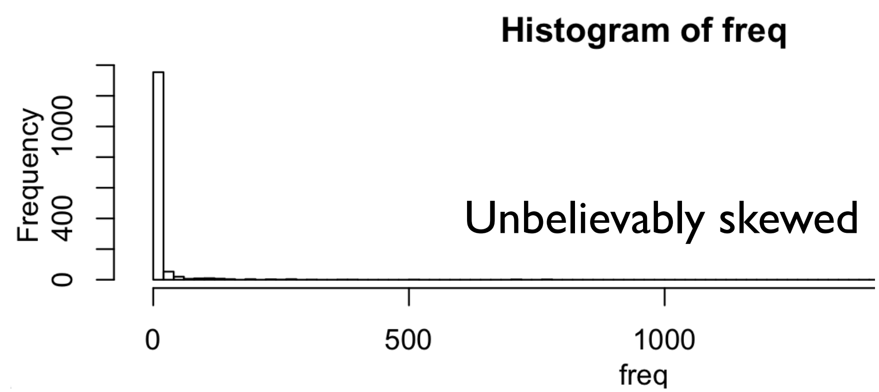
```
library(tm)

(corpus <- VCorpus( VectorSource(text) ))
inspect( corpus[1:2] )

# minimal processing
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, stripWhitespace)
inspect( corpus[1:2] )

# other possible commands ...
# corpus <- tm_map(corpus, removeNumbers)
# corpus <- tm_map(corpus, removeWords, stopwords("english") )
# corpus <- tm_map(corpus, stemDocument)
```

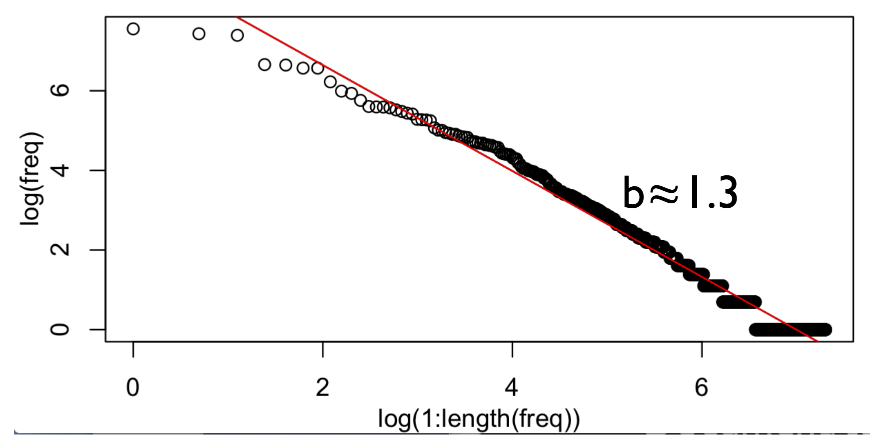
Basic Text Processing

- Document/term matrix
  - Frequencies of word types

"bag of words"

```
# --- construct document/type matrix
dtm <- DocumentTermMatrix( corpus )
dim(dtm); inspect(dtm[1:3,1:10])

# --- Zipf distribution
freq <- colSums(as.matrix(dtm))
hist(freq, breaks=100)


# log/log scale with fitted line
freq <- sort(freq, decreasing=TRUE)
plot(log(1:length(freq)),log(freq))

lf <- log(freq); lx <- log(1:length(freq))
abline(regr <- lm(lf ~ lx), col="red")
summary(regr)
```





Singular Value Decomposition

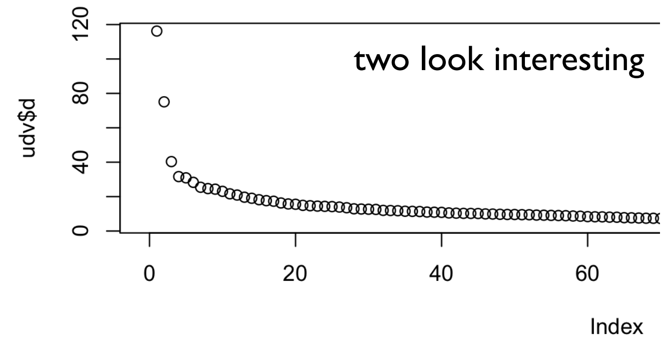
- Modern way to do PCA
  - Factors data matrix directly, X = UDV'
    Avoids direct computation of covariance matrix
  - D is diagonal, U and V are orthogonal
    U holds principal components, V holds the weights

```
X <- as.matrix(dtm.sparse)

# --- divide each row by square root of sum (variance stability)
X <- (1/sqrt(rowSums(X))) * X
# --- divide each column by square root of sum (variance stability)
X <- t ( t(X) * 1/sqrt(colSums(X)) )

udv <- svd(X); names(udv)

# --- plot diagonal elements, first few principal components
plot(udv$d)
pairs(udv$u[,1:4]); cor(udv$u[,1:4])
```

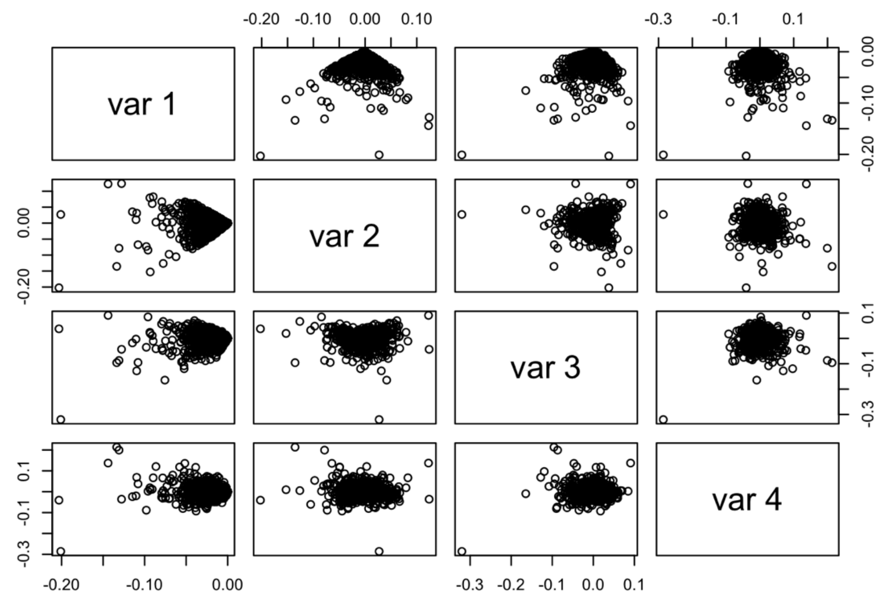
Optional

Plots of Components

- Singular values
  - Square root of usual eigenvalue of covariance



- Coordinates
  - Extreme points, skewness
  - Produce leverage points in subsequent regression



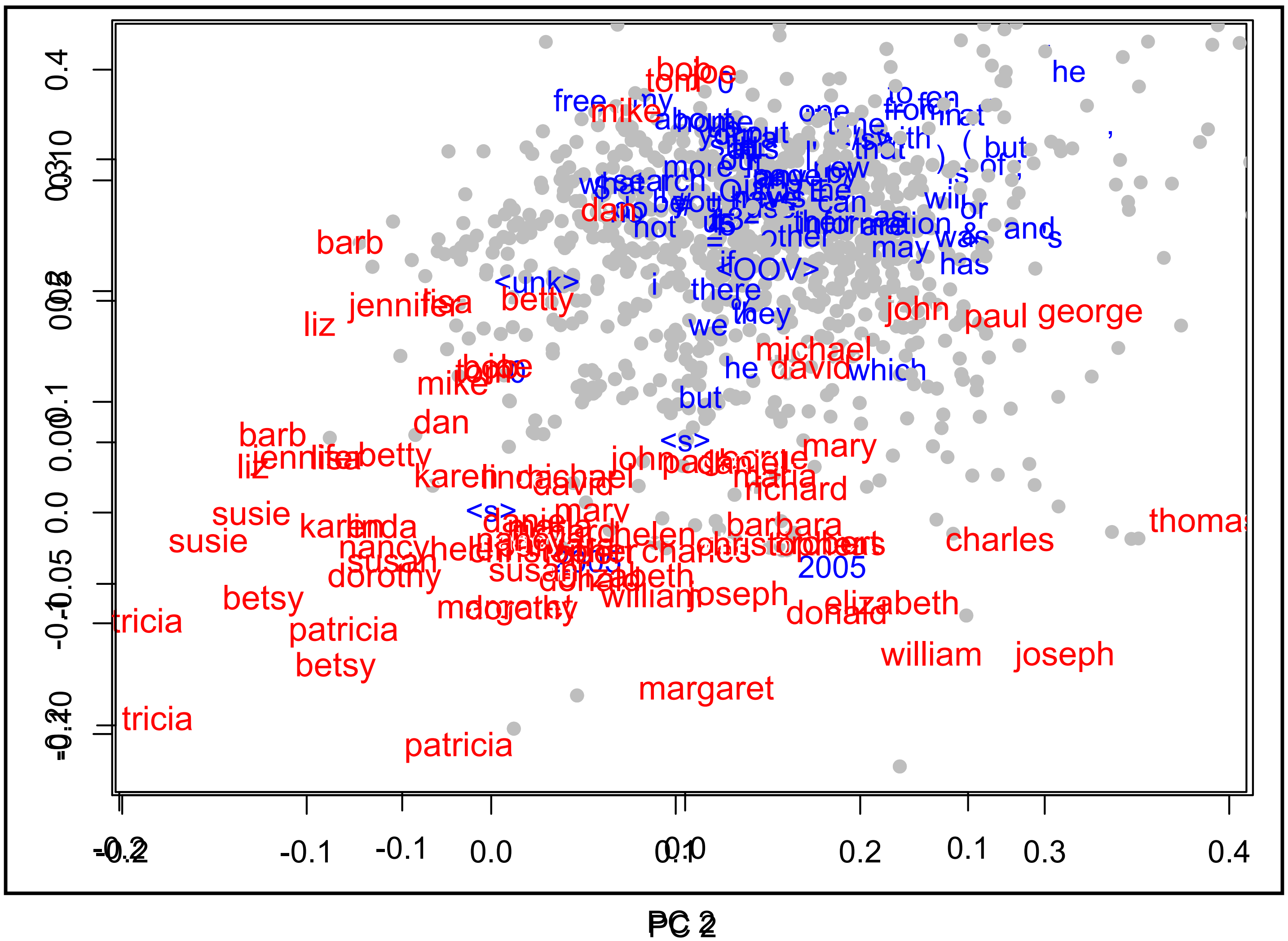
Eigenwords

- What are those principal components...
  - Inspect word types with large coefficients
  - Relatively small corpus for interpretation

	word type	coefficient/weight
1	the	-0.63465131
2	know	0.45578481
3	dont	0.45190270
4	house	-0.26423410
5	speaker	-0.24268062
6	shes	-0.16950095
7	what	0.06851086

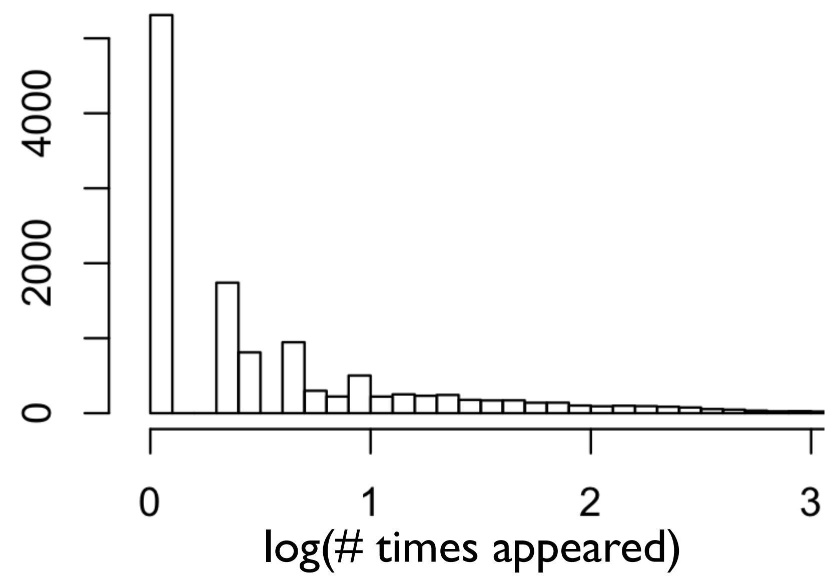
Example of Eigenwords

Larger corpus (Google)



Predictive Modeling

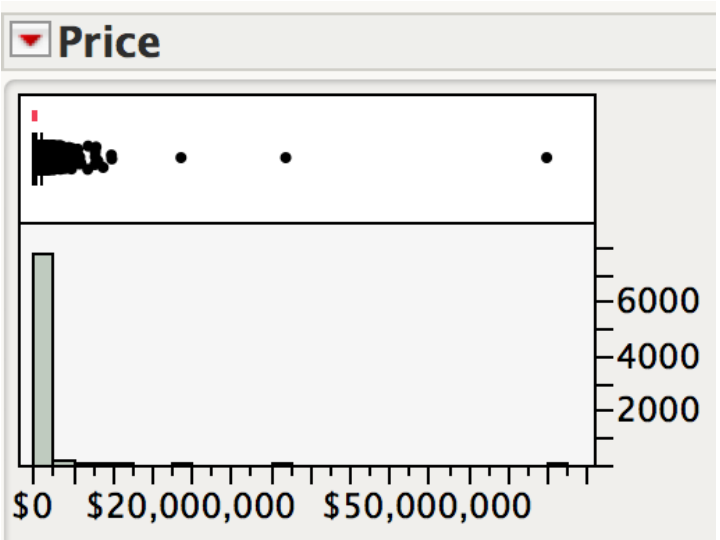
- Idea
  - Use unsupervised PCA to build Xs, predictive features
- Data
  - 7,773 listings
  - 571,000 tokens, 12,400 types
  - Avg listing has 75 words

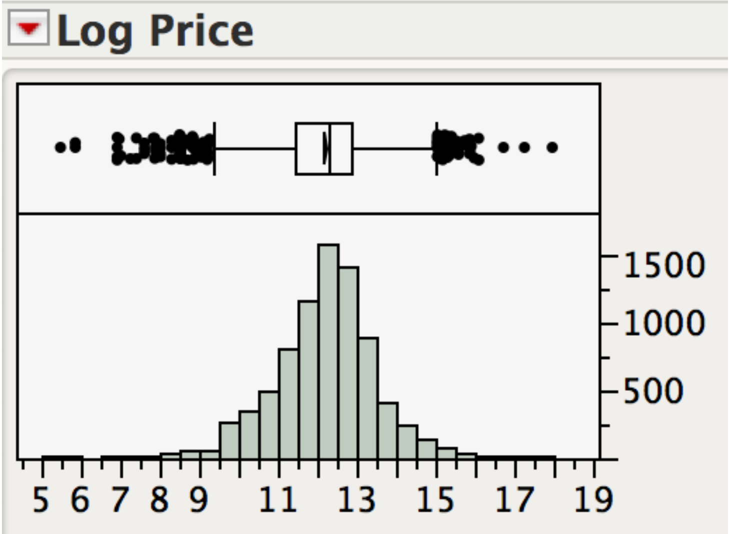


- Rare types
  - 5310 types used once
  - 1740 types used twice
  - 811 three times...
  - Use those appearing at least 3 times, keeping 5,400

Response: Home Prices

- Highly skewed distribution
  - mean price is \$390,000, max is \$65,000,000
  - most variation obtained by 'explaining' outliers
- Log scale normally distributed (lognormal)
  - regression response





Regression Model

- Stepwise picks eigenword associations over the extracted, constructed prior variables
  - Improves prior fit by significant margin
  - Length of description, bathrooms
  - Square footage? Probably too few.

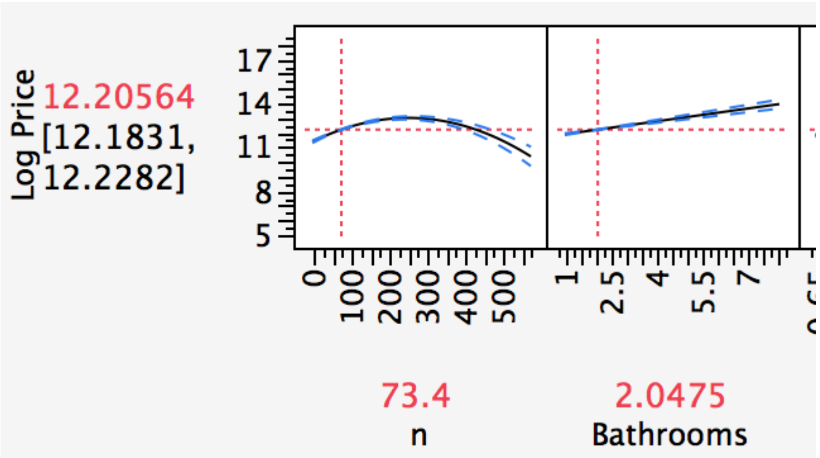
Final model explains about 2/3 of variation

SSE	DFE	RMSE	RSquare	RSquare Adj
7569.4865	7762	0.9875211	0.3274	0.3265

Current Estimates

Step History

Step	Parameter	Action	"Sig Prob"
1	X51	Entered	0.0000
2	X42	Entered	0.0000
3	X6	Entered	0.0000
4	n	Entered	0.0000
5	X35	Entered	0.0000
6	X88	Entered	0.0000
7	(n-73.4002)*(n-73.4002)	Entered	0.0000
8	X82	Entered	0.0000
9	Bathrooms	Entered	0.0000
10	X1	Entered	0.0000



Text Mining Suggestions

- Find a good question
  - New techniques are enough to learn without having to learn a new substantive area too
- Learn linguistics
  - Or work with a linguist
- Lots of data
  - Text has lower predictive power per 'item' than labeled numerical or categorical data
- Automated methods
  - Getting better, often a useful supplement
  - Combine with DM tools like trees and NN

Literature

- HUGE
  - Just google ‘text mining’
  - Google scholar
- Two illustrative examples
  - Monroe,Colaresi, Quinn (2008) Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict.
  - Netzer, Feldman, Goldenberg, Fresko, (2012). Mine your own business: Market-structure surveillance through text mining.

That's all for now...
Thanks for coming!