Sentiment Analysis

wine_sentiment.R

Dictionary Methods

Count the usage of words from specified lists

Example

LWIC Tausczik and Pennebake (2010),
The Psychological Meaning of Words,
Journal of Language and Social Psychology

Positive and negative emotions

Sources

Essentially make our own later

LIWC developed for various languages

Google for current locations, languages

Software

Methods in other direction: read summary and write article…WSJ

MARKETS

Can You Tell the Difference Between a Robot and a Stock Analyst?

Wall Street tries out research reports written by artificial intelligence

By STEPHANIE YANG
Updated July 9, 2015 2:33 p.m. ET

LIWC Words

Linguistic Inquiry and Word Count (LIWC)

Commercial collection of words

Psychological Processes			
Social processes	social	Mate, talk, they, child	455
Family	family	Daughter, husband, aunt	64
Friends	friend	Buddy, friend, neighbor	37
Humans	human	Adult, baby, boy	61
Affective processes	affect	Happy, cried, abandon	915
Positive emotion	posemo	Love, nice, sweet	406
Negative emotion	negemo	Hurt, ugly, nasty	499
Anxiety	anx	Worried, fearful, nervous	91
Anger	anger	Hate, kill, annoyed	184
Sadness	sad	Crying, grief, sad	101

in category

Sentiment Analysis

Basic version

Identify words that associate with different concepts

- Positive - Negative
- Cruel - Kind
- Red - White wine

Over a corpus of documents, count the prevalence of the different types of words

Use differences in these counts as a measure of the "sentiment" of the document

Application

Words used by judge hearing a case

Word Lists

Established word lists

- Bing Liu's negative/positive words from early paper
- LIWC commercial list (next slide)

Grow your own

- Start with seed words
- Expand using WordNet to find synonyms, antonyms

Issues

- Counting only
- Count "funny" also counts "not funny"
  - Parsing complicates the analysis
- Words that are "negative" may not be negative in every context

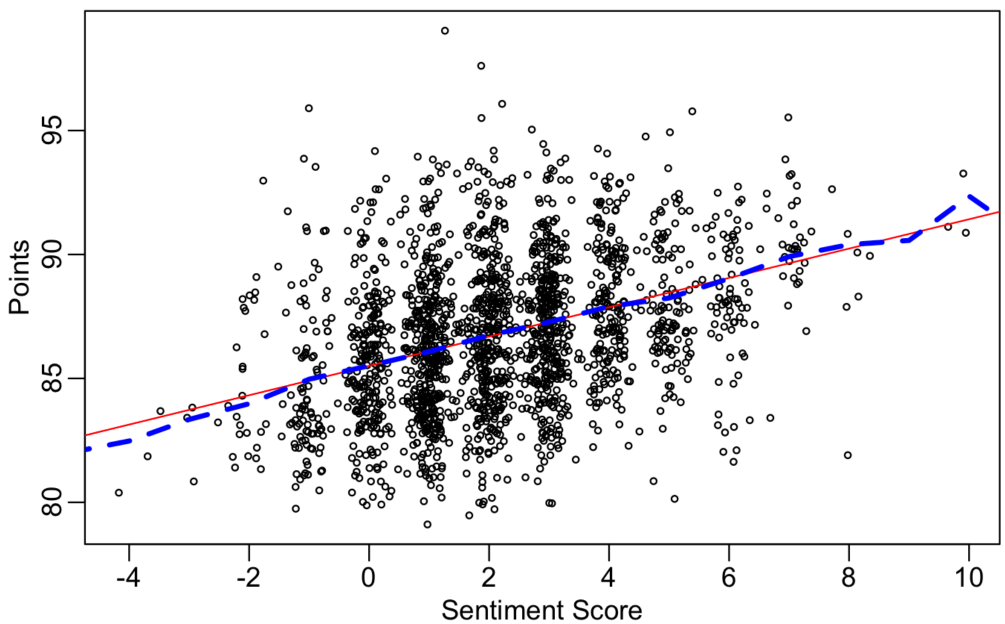
Example with Wines

Relate counts of words to points assigned to wines

Some words clearly not negative are counted as such…
example: lemon

Use counts or proportions

Difference in counts linearly related to points



est points ≈ 85.5 + 0.6 score

RMSE ≈ 3
$R^2 \approx 14\%$

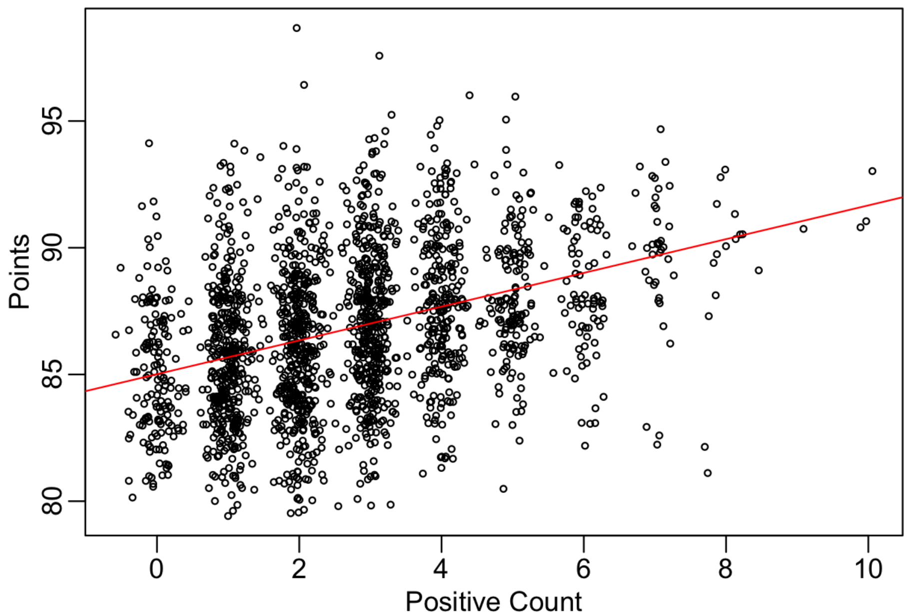
Wharton
Department of Statistics

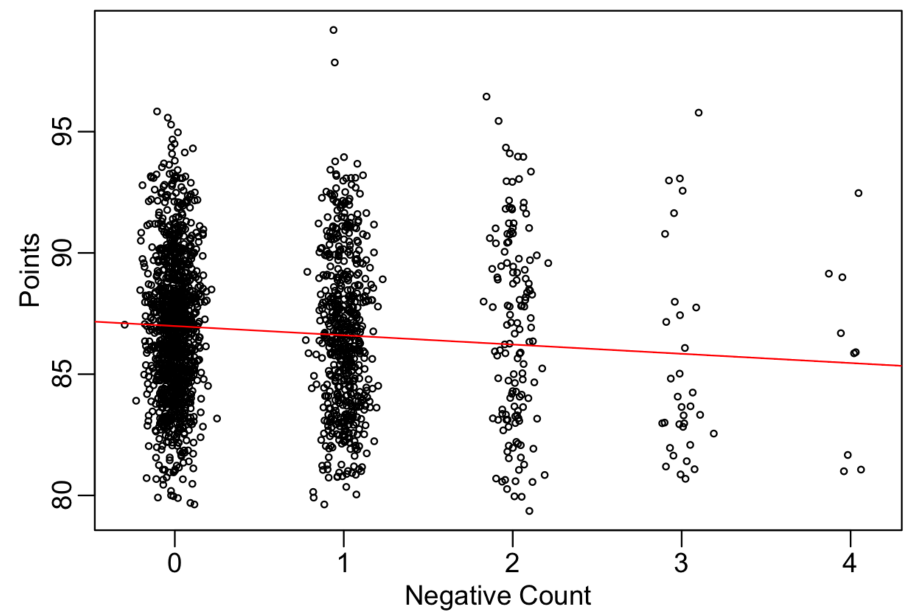
Negative Words less Useful

Role of positive/negative words

Asymmetric association…

Positive words add more than negative words





Multiple regression, however, gives a different impression…

Wharton
Department of Statistics

Combination

Multiple regression with positive and negative

A model with these counts basically repeats the two simple regressions…

These counts are not highly correlated (r ≈ -0.09)

Adding total word count tells a different story

	Estimate	Std. Error	t value
(Intercept)	82.365268	0.065230	1262.70
posCount	0.410919	0.011640	35.30
negCount	-0.577704	0.026823	-21.54
totCount	0.109695	0.002103	52.16

Why so different from prior?

```
Residual standard error: 2.688 on 20325 degrees of freedom
  (179 observations deleted due to missingness)
Multiple R-squared:  0.2497,    Adjusted R-squared:  0.2495
F-statistic:  2254 on 3 and 20325 DF,  p-value: < 2.2e-16
```

Regression
Methods & Examples

wine_regr.R

Regression Analysis

Objective

Find weighted combination of variables that best predicts a response

Application to text

What weighted combination of word counts best predicts the rating point of a wine?

Perspective

Sentiment analysis assigns fixed weight to selected words

Regression assigns weights that are most predictive in the context of the observed corpus

Regression vs Sentiment

Previous sentiment analysis

Common positive weight to "positive" words

Common negative weight to "negative" words

Advantage: no modeling, can do unsupervised

Disadvantage: generic, not adapted to problem

Regression model

Customize the weight for the observed data

Advantage: customized! Better fit, more predictive

Disadvantage: Must be superivised. Which words?

Which words?

How to pick the word features to use?

Variable selection for regression

- Theory
  - Very much like sentiment analysis, but with custom weights
- External sorting
  - Limit the analysis to the most common word types
- Stepwise type selection methods
  - Need criterion like Bonferroni to avoid overfitting
- Lasso type penalized methods
  - Popular, fast alternative to stepwise methods
  - Convex algorithm faster than stepwise search (albeit different search)

Shrinkage Methods

Alternative to subset selection

- Difficult to identify and fit all subsets
  - Consider how many such models are possible...
- Solve a simpler problem that 'shrinks' estimates
  - Careful. Estimates need to be on common scale to combine
- Why shrink? Trade bias to reduce variance
  - Shrinkage allows fitting all the variables even if more variables than cases

Penalized likelihoods

Penalize by a measure of the size of the coefficients.

Fit has to improve by enough (RSS decrease) to compensate for size of coefficients

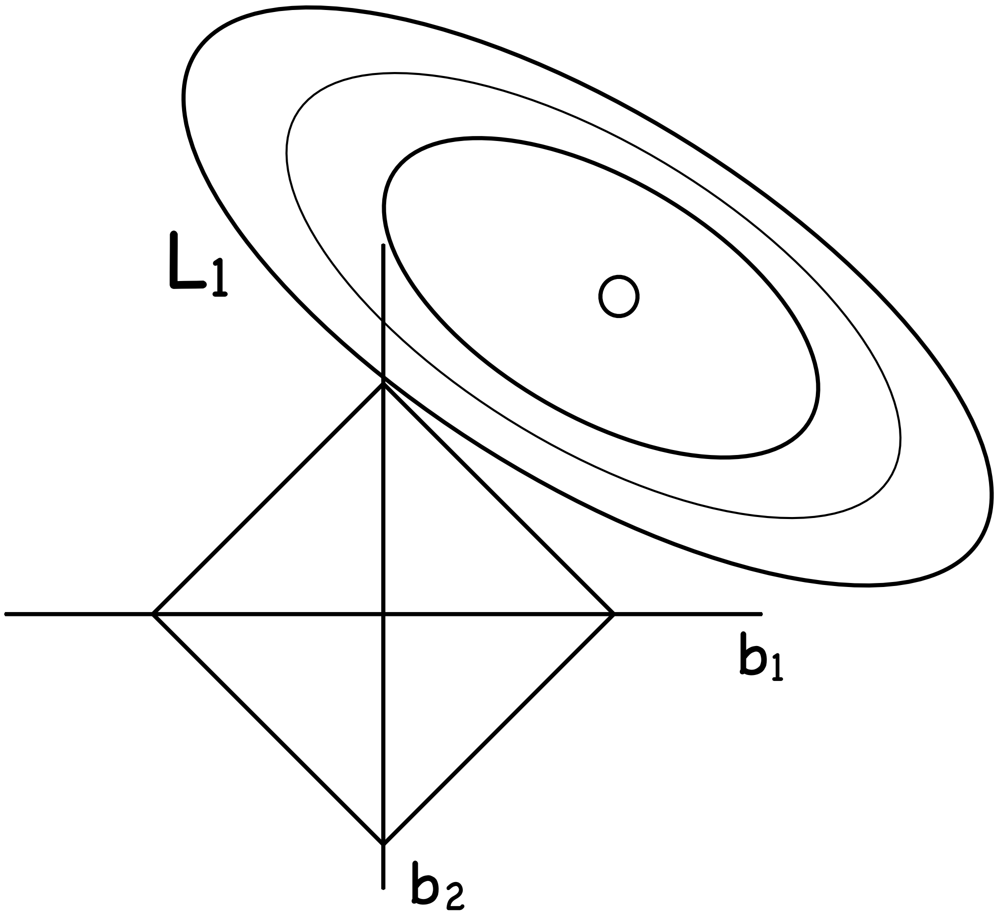
Ridge regression: min RSS + λ_2 b'b

Lasso regression: min RSS + λ_1 Σ|bj|

λ is a tuning parameter that must be chosen by some method usually cross-validation

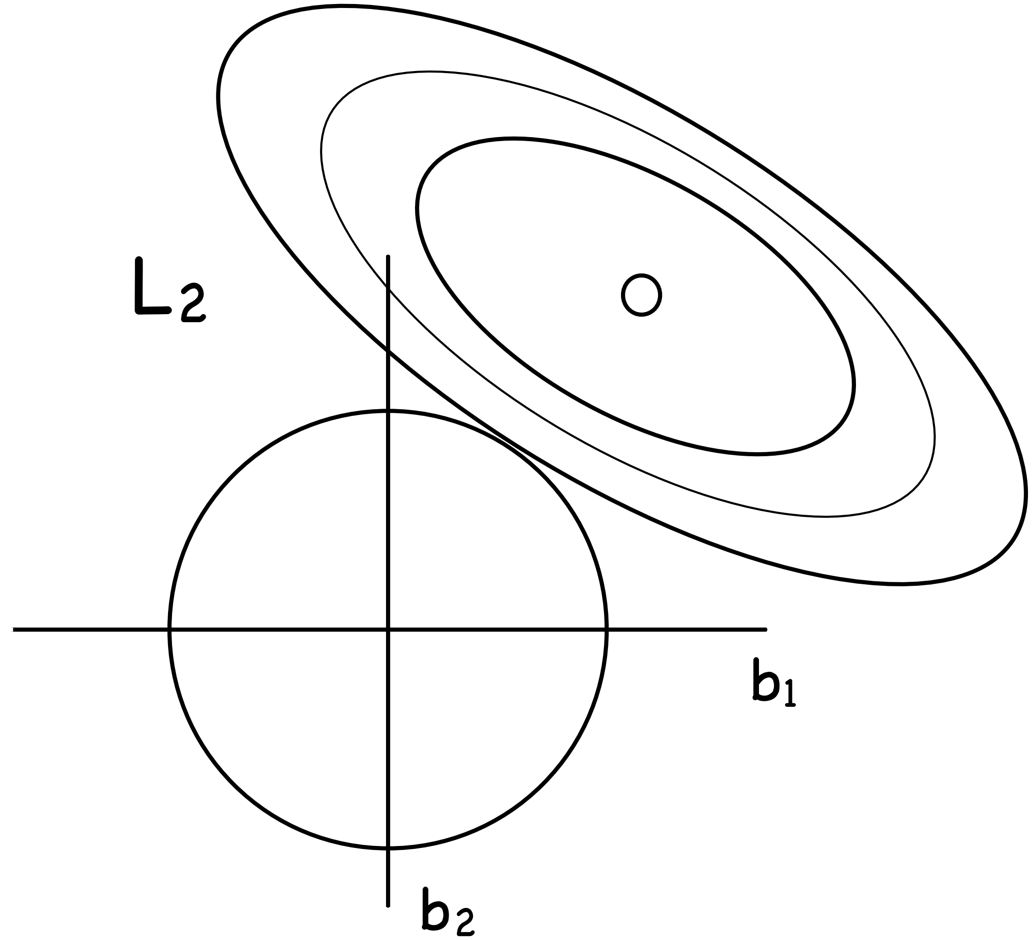
Also have a Bayesian interpretation (see ISL)

L_1 vs L_2 Penalty



min RSS, $\sum|b_j|<c$

Corners produce selection



min RSS, $\sum b_j^2<c$

Figure 6.7, p 222

Interpret λ as Lagrange multiplier.

Cross-Validation

Fundamental, commonly used

- Use part of the data to build a model
- Use a separate, “hidden” part to test the model
- Happens often in practice in consulting

Question: how to partition data?

Remedy

- Repeat the division between the two groups
- K-fold cross-validation partitions data into K parts
- Fit to K-1 folds, validate on 1 fold (K = 5,10)

Missing Data

Always present

- In medical example, only 170 out of 1,200 cases were complete

Often informative

- In bankruptcy model, half of predictors indicate presence of missing data
- Is data ever ‘missing at random’?

Handle as part of the modeling process?

- Offer a simple patch that requires few assumptions

Main idea

- Done as a data preparation step
- Add indicator column for missing values
- Fill the missing value

Handle Missing by Adding Vars

- Add another variable
  - Add indicator column for missing values
  - Fill the missing with average of those seen
- Simple approach, fewer assumptions
  - Expands the domain of the feature search
  - Allows missing cases to behave differently
  - Conservative evaluation of variable
- Part of the modeling process
  - Distinguish missing subsets only if predictive
- Missing in a categorical variable: not a problem
  - Missing define another category

ONLY applies to explanatory variables, never the response

Example

Data frame with missing values

```
>  example.df
  x1 x2         x3 lab  fac
1  1 NA  1.4671553 UVW  ABC
2  1  2 -0.8691613 UVW  ABC
3  1  3  0.1174511 UVW  ABC
4  1 NA -0.3890095 UVW  ABC
5 NA  5  1.2007855 UVW  ABC
6  1 NA  0.3604345 UVW  ABC
7  1  7  0.6692698 UVW  ABC
8  1  8 -1.4056064 UVW  ABC
9  1  9 -1.2858561 UVW <NA>
10 1 10 -0.2103984 UVW <NA>
```

Filled in data with added indicator columns

```
>  fill.missing(example.df)
   x1        x2         x3 lab     fac Missing.x1 Missing.x2
1   1  6.285714  1.4671553 UVW     ABC          0          1
2   1  2.000000 -0.8691613 UVW     ABC          0          0
3   1  3.000000  0.1174511 UVW     ABC          0          0
4   1  6.285714 -0.3890095 UVW     ABC          0          1
5   1  5.000000  1.2007855 UVW     ABC          1          0
6   1  6.285714  0.3604345 UVW     ABC          0          1
7   1  7.000000  0.6692698 UVW     ABC          0          0
8   1  8.000000 -1.4056064 UVW     ABC          0          0
9   1  9.000000 -1.2858561 UVW Missing          0          0
10  1 10.000000 -0.2103984 UVW Missing          0          0
```

Regression for Points

Validation

Set aside 5,000 cases for checking models

Initial model, without words

Note the significant role for the missing indicators

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.181e+02  1.464e+01  -8.069 7.64e-16 ***
alcohol       6.178e-02  1.334e-02   4.631 3.68e-06 ***
vintage       9.924e-02  7.304e-03  13.588  < 2e-16 ***
price         6.157e-02  1.345e-03  45.783  < 2e-16 ***
lengths       1.053e-01  2.086e-03  50.499  < 2e-16 ***
Miss.alcohol -7.718e-01  1.504e-01  -5.133 2.88e-07 ***
Miss.vintage -3.808e-01  6.942e-02  -5.485 4.19e-08 ***
Miss.price    4.866e-01  8.154e-02   5.968 2.46e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.55 on 15321 degrees of freedom
Multiple R-squared:  0.3248, Adjusted R-squared:  0.3245
F-statistic:  1053 on 7 and 15321 DF,  p-value: < 2.2e-16
```

Regression for Points

Initial model, with only words(proportion) and lengths

Just 15 words to get the idea, adding lengths really helps

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  82.363011   0.284349 289.654  < 2e-16 ***
lengths       0.119197   0.003084  38.649  < 2e-16 ***
comma_       -0.574532   0.616589  -0.932 0.351459    
and           8.992627   0.894380  10.055  < 2e-16 ***
period_       1.783183   1.179964   1.511 0.130754    
dash_         1.092387   0.887434   1.231 0.218361    
with          7.568056   1.129717   6.699 2.17e-11 ***
aromas      -17.119562   2.403255  -7.123 1.10e-12 ***
medium       -6.971505   1.967944  -3.543 0.000397 ***
finish        6.354623   1.573823   4.038 5.42e-05 ***
entry       -40.227866   2.332525 -17.246  < 2e-16 ***
fruit        -1.413407   1.414402  -0.999 0.317667    
body         -7.603953   2.869107  -2.650 0.008051 ** 
full         61.308520   1.770266  34.632  < 2e-16 ***
bodied       -8.363368   2.171764  -3.851 0.000118 ***
this        -26.309107   1.958179 -13.435  < 2e-16 ***
leads       -26.526081   2.346595 -11.304  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.495 on 15312 degrees of freedom
Multiple R-squared:  0.3541, Adjusted R-squared:  0.3534 
F-statistic: 524.6 on 16 and 15312 DF,  p-value: < 2.2e-16
```

Regression for Points

Combined…

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.097798  19.796170  -2.127 0.033472 *
alcohol      -0.022784   0.012881  -1.769 0.076934 .
vintage       0.061950   0.009877   6.272 3.66e-10 ***
price         0.046953   0.001309  35.876  < 2e-16 ***
lengths       0.101698   0.003016  33.715  < 2e-16 ***
Miss.alcohol -0.737923   0.149054  -4.951 7.47e-07 ***
Miss.vintage -0.401744   0.067054  -5.991 2.13e-09 ***
Miss.price    0.562938   0.077376   7.275 3.62e-13 ***
comma_       -0.726686   0.594887  -1.222 0.221896
and           8.915556   0.859716  10.370  < 2e-16 ***
period_       3.065042   1.146935   2.672 0.007540 **
dash_         2.841002   0.863970   3.288 0.001010 **
with          7.047177   1.082107   6.512 7.62e-11 ***
aromas      -14.710150   2.353051  -6.252 4.17e-10 ***
medium       -6.627668   1.895310  -3.497 0.000472 ***
finish        4.346500   1.515970   2.867 0.004148 **
```

…more…

```
Residual standard error: 2.386 on 15306 degrees of freedom
Multiple R-squared:  0.4095, Adjusted R-squared:  0.4087
F-statistic: 482.6 on 22 and 15306 DF,  p-value: < 2.2e-16
```

R files build larger models

Dilemma
Get better and better as keep adding more words

Calibration Plot

Check out-of-sample fit is correct on average

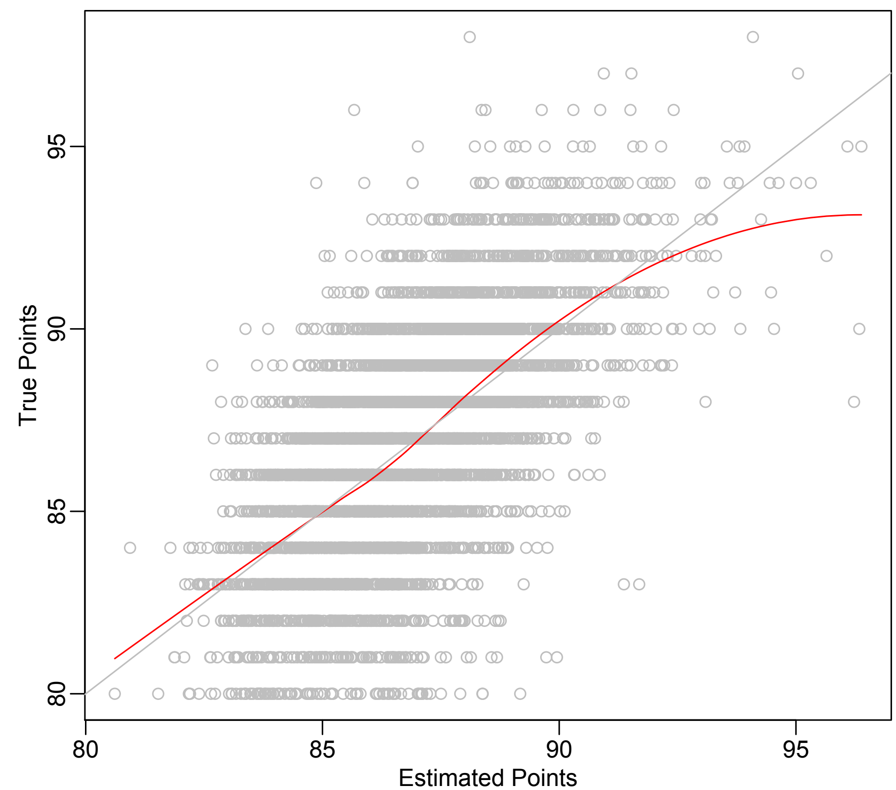
Does out of sample fit match claimed fit of model?

Check that predictions are honest: E(Y|Ŷ) = Ŷ

Common problem

Limited range response

Any wines more than 100 pts?
Less than 80 points?



Checking Claimed Precision

Does model meet claims of precision

Are the predictions of the model for the test data as good as they are when predicting the training data

The training data was used to build the model

Overfitting

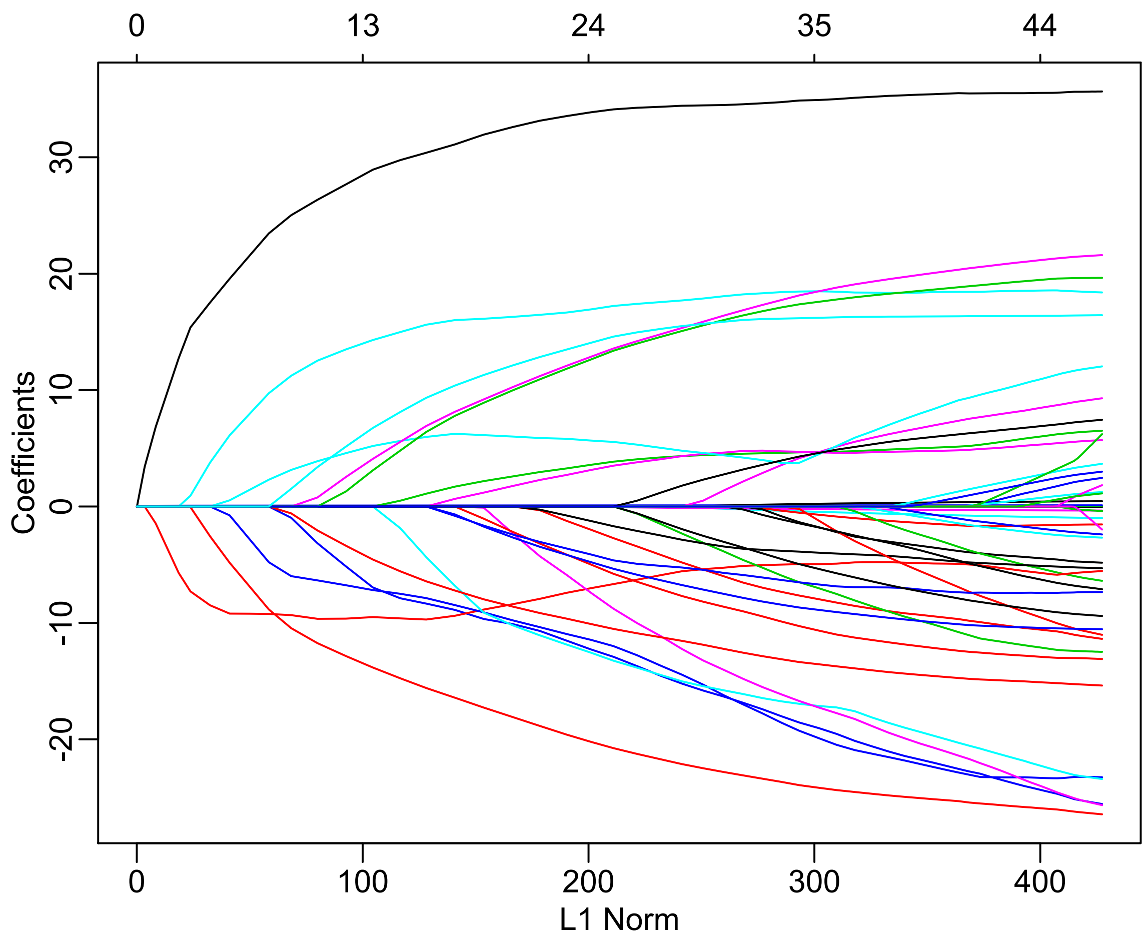
Occurs when model capitalizes on random variation in the training data

Predicts training data better than test data. For example
- Average squared prediction error in test > in training
- Correlation^2(predicted, actual) in test < in training (ie R^2)

Lasso Fit

Which model do you want to keep

Fishbone plot for model with others and words



Wharton
Department of Statistics

Cross-Validation Picks

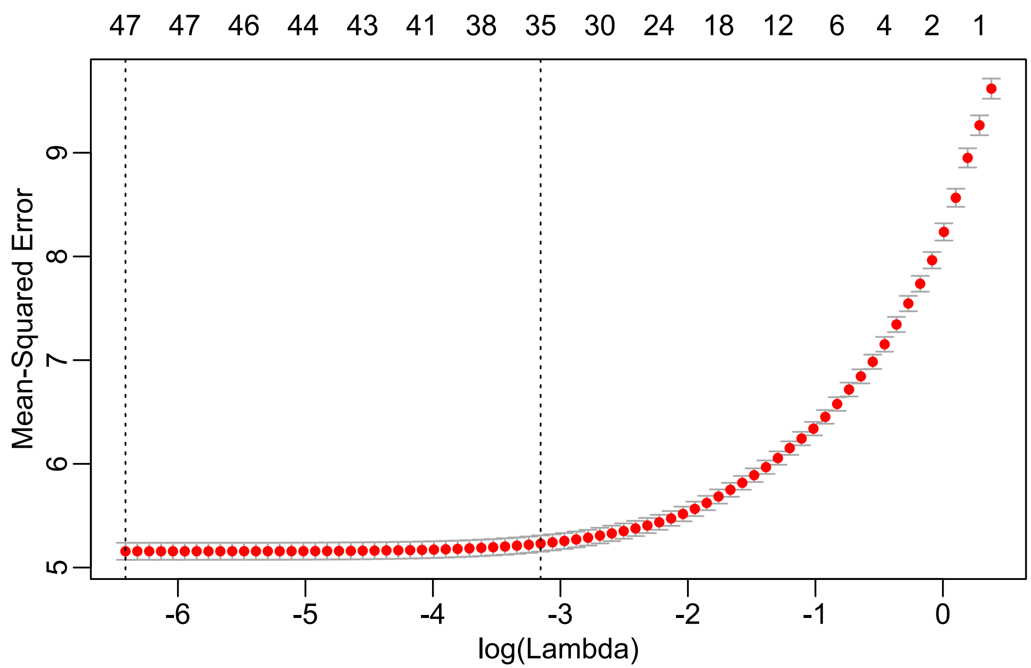
10 fold cross validation

Chooses best value for the tuning parameter

Big model!

Really wants to use them all!

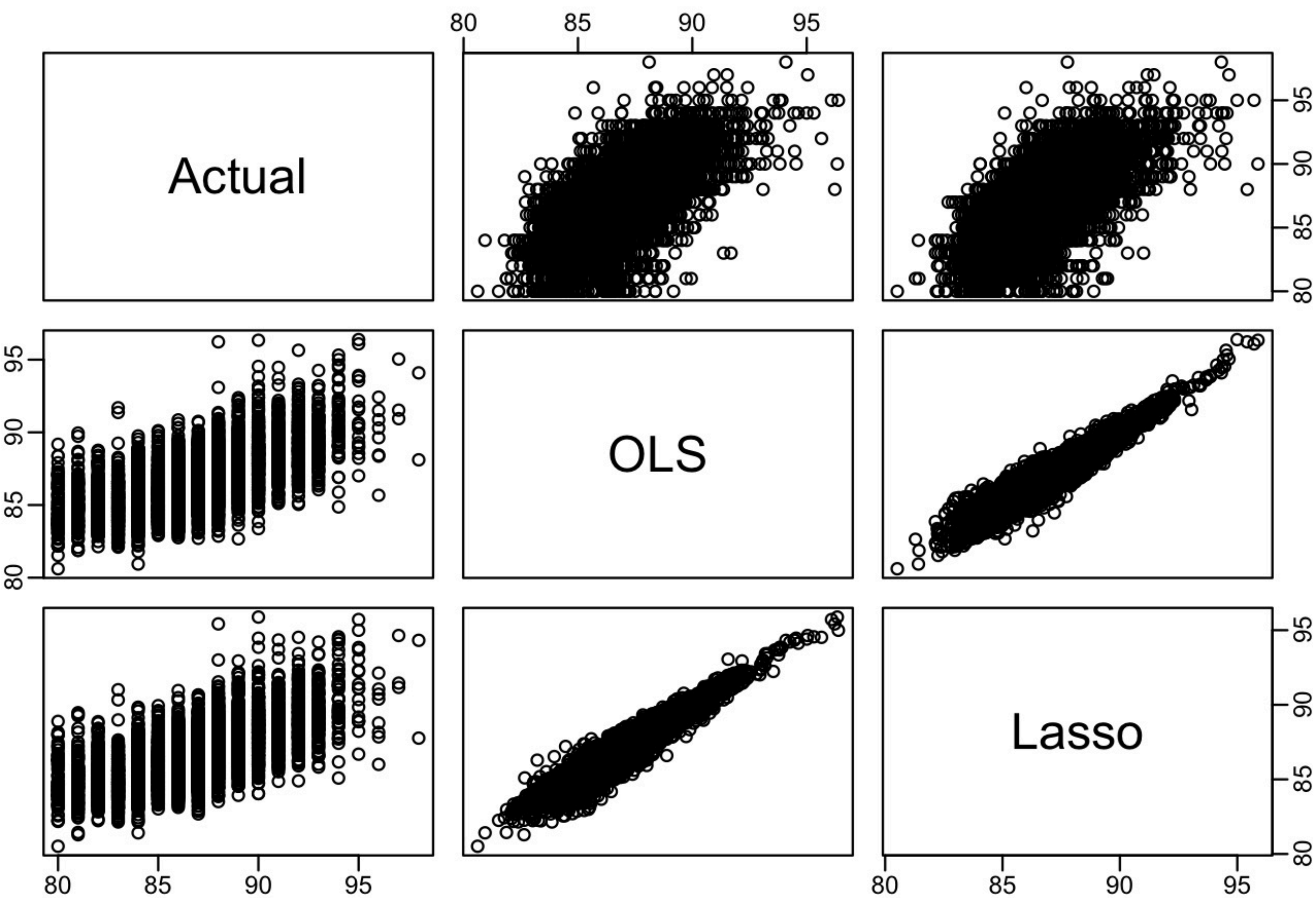
1 SE heuristic picks a simpler model



Comparisons

Scatterplot matrix of the predictions and actual

All in the test sample



Eye Candy

Word cloud

Which words have large coefficients in the lasso model?