Data Mining with Regression

Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania

Some Details

- Office hours
  - Let me know and we can meet at Newberry
  - stine@wharton.upenn.edu
- Class notes
  - http://www-stat.wharton.upenn.edu/~stine/mich/
- Data
  - Will post ANES and others on Z drive
- JMP software
  - Depends on your school

Topics for Today

- Review from last time
  - Any questions, comments?
- Growing regression models
  - Deciding which variables improve a model
  - Standard errors and significance
- Missing data
- Stepwise regression

Why use regression?

- Claim
  - Regression is capable of matching the predictive performance of black-box models
  - Just a question of having the right X's
- Regression is familiar
  - Recognize then fix problems
  - Shares problems with black-boxes

    Opportunity to appreciate what happens in less familiar, more complex models with more flexible structure.
- Familiarity allows improvements
  - Patches in Foster and Stine 2004

Review ANES Example

- Start with simple regr, expand to multiple
  - Post FT Obama on Pre FT Obama
  - Add ‘Happy/Sad’ and ‘Care Who Wins’
  - Include interaction effect
- Visual exploration of model form
  - Show the effects of an interaction
  - What’s the interaction mean
- Calibration
  - Being right on average
- Tests and inference
  - Which terms are significant? What’s that mean?

profiling

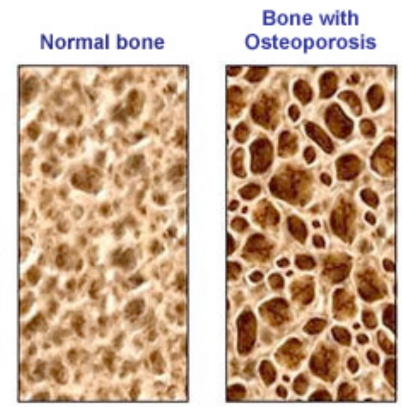
$\text{avg}(Y|\hat{Y})=\hat{Y}$

Modeling Question

- How do we expand a regression model
  - Reach beyond obvious variables
  - Find subtle but important features
- Automate typical manual procedure
  - Iterative improvement
  - Try variable, diagnose, try another, diagnose…
- Computing allows more expansive search
  - Open modeling process to allow a surprise
  - Example: Include interactions

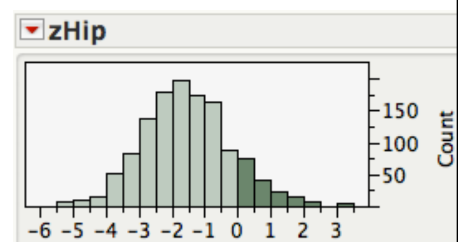
    transformations, combinations (e.g. ratios), bundles (e.g. prin comp)
  - Magnified scope also magnifies problems

Medical Example



- Numerical response
- Diagnosing severity of osteoporosis
  - Brittle bones due to loss of calcium
  - Leads to fractures and subsequent complications
  - Personal interest

- Response
  - X-ray measurement of bone density
  - Standardized to N(0,1) for normal
  - Possible to avoid expense of x-ray, triage?



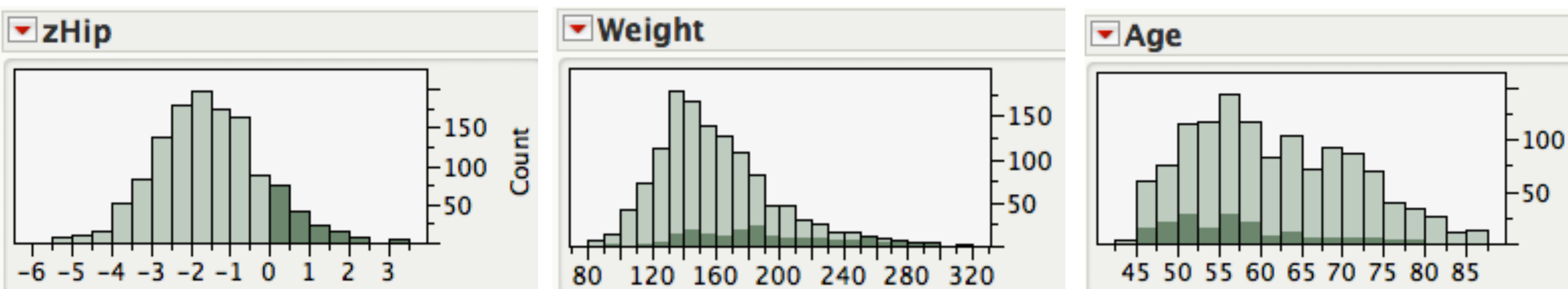
- Explanatory variables
  - Data set designed by committee
    doctors, biochemists, epidemiologists

Osteoporosis Data

- Sample of postmenopausal women
  - 1,232 women with 127 columns
  - Nursing homes in NE… Dependence? Bias?
  - Presence of missing data
  - Measurement error

ideal data?

- Marginal distributions
  - X-ray scores (zHip), weight, age...



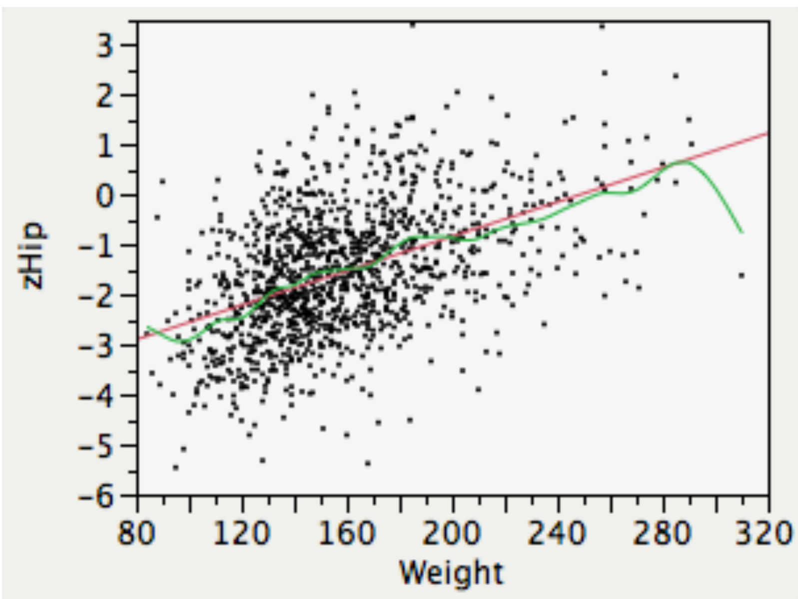
Initial Osteo Model

- Simple regression
  - zHip on which variable?
  - How would you decide…

pick largest correlation

consult science

- Impact of weight



RSquare	0.221923
RSquare Adj	0.22129
Root Mean Square Error	1.140076
Mean of Response	-1.55801
Observations (or Sum Wgts)	1230

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-4.27558	0.14880	-28.73	<.0001*
Weight	0.01722	0.00092	18.71	<.0001*

Interpretation?

Wharton
Department of Statistics

Expanding Model

- What to add next?
  - Residual analysis
  - Add others and see what sticks
- Add them all?
  - Singularities imply redundant combinations
  - Summary of fit
    Impressive R^2 until you look at the sample size.

RSquare	0.9882
RSquare Adj	0.9620
Root Mean Square Error	0.2280
Mean of Response	-1.5767
Observations (or Sum Wgts)	171.0000

Missing Data

- Fit changes when add variables
  - Collinearity among explanatory variables
  - Different subsets of cases
- What to do about the missing cases
  - Exclude

    "Listwise deletion"

    "Pairwise deletion"
  - Impute. Fill them in, perhaps several times
- Imputation relies on big assumption

  Missing cases resemble those included.

Real data is seldom (if ever)
missing at random

Handle Missing Data

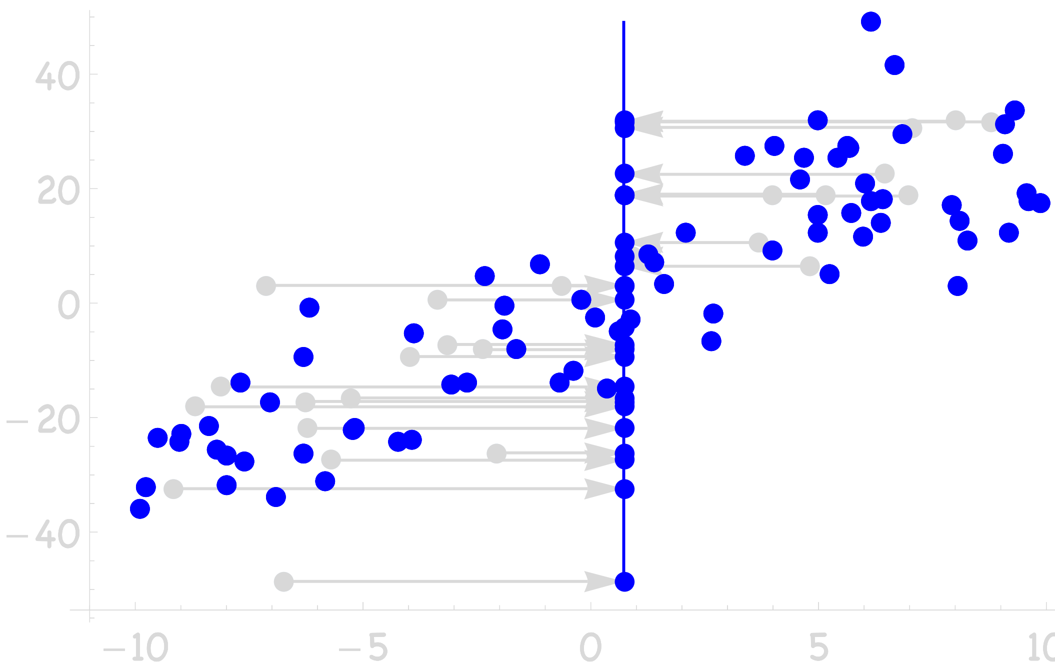
- Add another variable
  - Add indicator column for missing values
  - Fill the missing value with average of those seen
- Simple, reduced assumption approach
  - Expands the domain of the feature search
  - Allows missing cases to behave differently
  - Conservative evaluation of variable

Leads to complaints about lack of power

- Part of the modeling process
  - Distinguish missing subsets only if predictive
- Categorical: not a problem
  - Missing form another category

Example of Procedure

- Simple regression, missing at random
  - Conservative: unbiased estimate, inflated SE
  - $n=100$, $\beta_0=0$, $\beta_1=3$
  - 30% missing at random, $\beta_1=3$



Complete

	Est	SE
b_0	-0.25	1.0
b_1	3.05	0.17

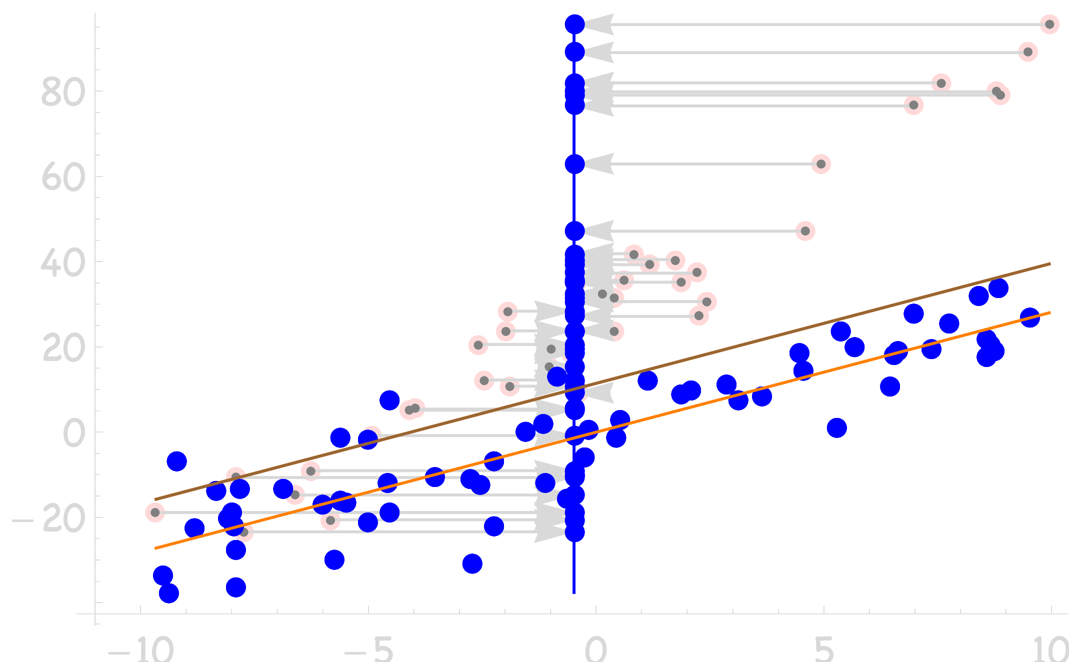
Filled In

	Est	SE
b_0	-1.5	1.4
b_1	3.01	0.27

Example of Procedure

- Simple regression, not missing at random
  - Conservative: unbiased estimate, inflated SE
  - n=100, β_0=0, β_1=3
  - 30% missing follow steeper line

Requires robust variance estimate



Filled In

	Est	SE
b_0	-0.02	2.6
b_1	2.82	0.44

Wharton
Department of Statistics

Example from R

Data frame with missing values

```
>   example.df
   x1 x2          x3 lab  fac
1   1 NA -0.9532650 UVW  ABC
2   1  2 -2.8903951 UVW  ABC
3   1  3 -0.1693143 UVW  ABC
4   1 NA -0.8343432 UVW  ABC
5  NA  5  1.0919509 UVW  ABC
6   1 NA  1.3706193 UVW  ABC
7   1  7 -1.7155066 UVW  ABC
8   1  8  0.6355785 UVW  ABC
9   1  9  0.7014913 UVW <NA>
10  1 10  0.4994391 UVW <NA>
```

Filled in data with added indicator columns

```
>   fill.missing(example.df)
   x1        x2         x3 lab     fac Miss.x1 Miss.x2
1   1  6.285714 -0.9532650 UVW     ABC       0       1
2   1  2.000000 -2.8903951 UVW     ABC       0       0
3   1  3.000000 -0.1693143 UVW     ABC       0       0
4   1  6.285714 -0.8343432 UVW     ABC       0       1
5   1  5.000000  1.0919509 UVW     ABC       1       0
6   1  6.285714  1.3706193 UVW     ABC       0       1
7   1  7.000000 -1.7155066 UVW     ABC       0       0
8   1  8.000000  0.6355785 UVW     ABC       0       0
9   1  9.000000  0.7014913 UVW Missing       0       0
10  1 10.000000  0.4994391 UVW Missing       0       0
```

No cheating: You don't get to fill in the y's!

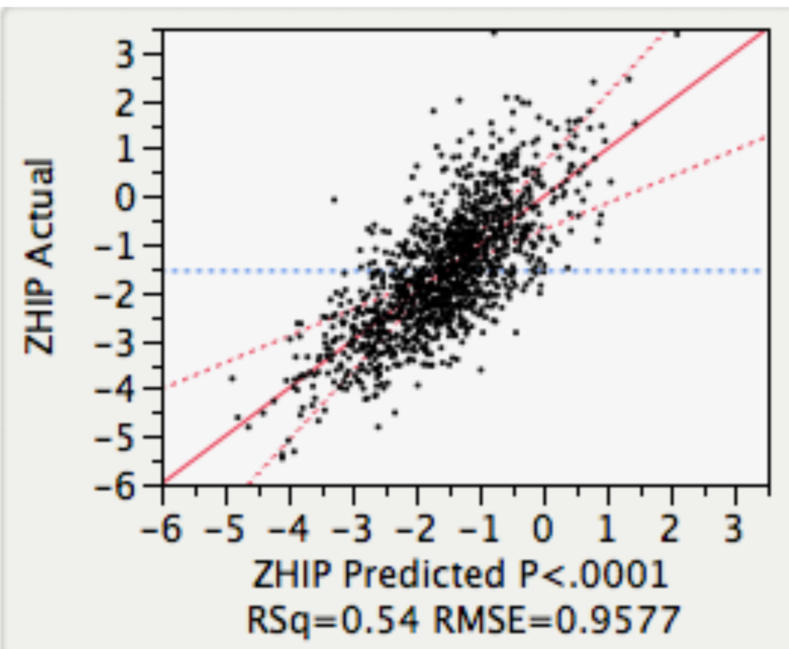
Background of Procedure

- Been around for a long time
  - Well suited to data mining when need to search for predictive features
- Reference
  - Paul Allison's Sage monograph on Missing Data (Sage # 136, 2002).
- For a critical view, see Jones, M. P. (1996)
  - J Amer. Statist. Assoc., 91, 222–230
  - He's not too fond of this method, but he models missing data as missing at random.

Expanded Osteo Data

- Fill in missing data
  - Grows from 126 to 208 possible Xs
- Saturated model results
  - Full sample but so few significant effects

Do in R



Still missing interactions

RSquare	0.541046
RSquare Adj	0.45095
Root Mean Square Error	0.957692
Mean of Response	-1.55821
Observations (or Sum Wgts)	1232

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	202	1112.5810	5.50783	6.0052
Error	1029	943.7711	0.91717	**Prob > F**
C. Total	1231	2056.3521		<.0001*

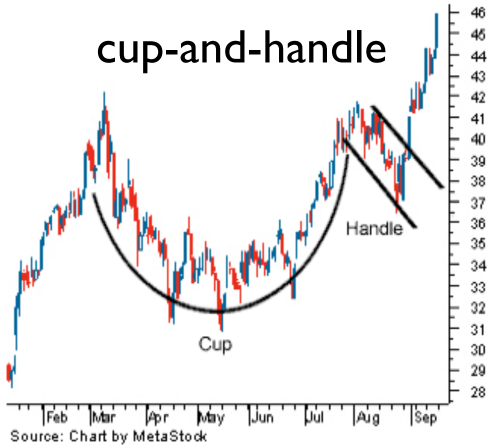
Stepwise Regression

- Need a better approach
  - Cannot always fit the saturated model
  - Saturated model excludes transformations such as interactions that might be useful
- Mimic manual procedure
  - Find variable that improves the current model the most
  - Add it if the improvement is significant.
- Greedy search
  - Common in data mining with many possible X's
  - One step ahead, not all possible models
  - Requires caution to use effectively

Stepwise Example

- Predict the stock market
- Response
  - Daily returns (essentially % change) in the S&P 500 stock market index through April 2014
- Goal
  - Predict returns in May and June using data from January through April
- Explanatory variables
  - 15 technical trading rules based on observed properties of the market
  - Designed to be easy to extrapolate

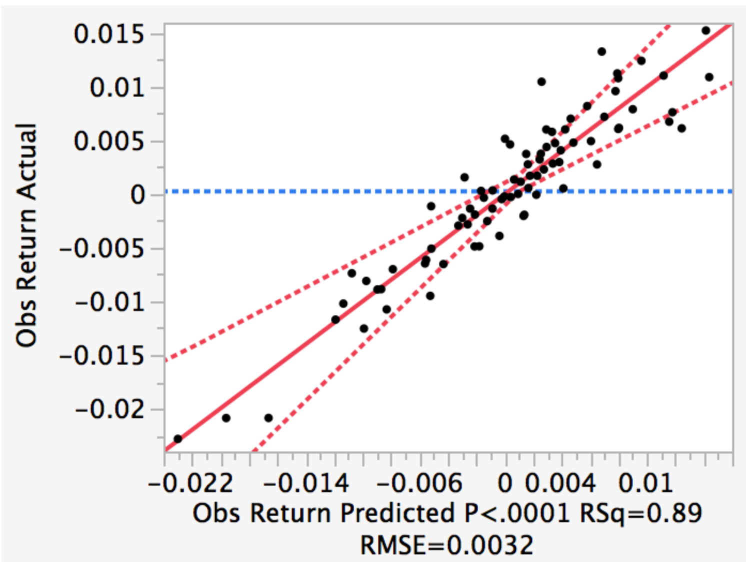
cup-and-handle



Wharton
Department of Statistics

Results

- Model has quite a few X's but is very predictive and highly stat significant.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	29	0.00424379	0.000146	14.1056
Error	52	0.00053947	0.000010	**Prob > F**
C. Total	81	0.00478325		<.0001*

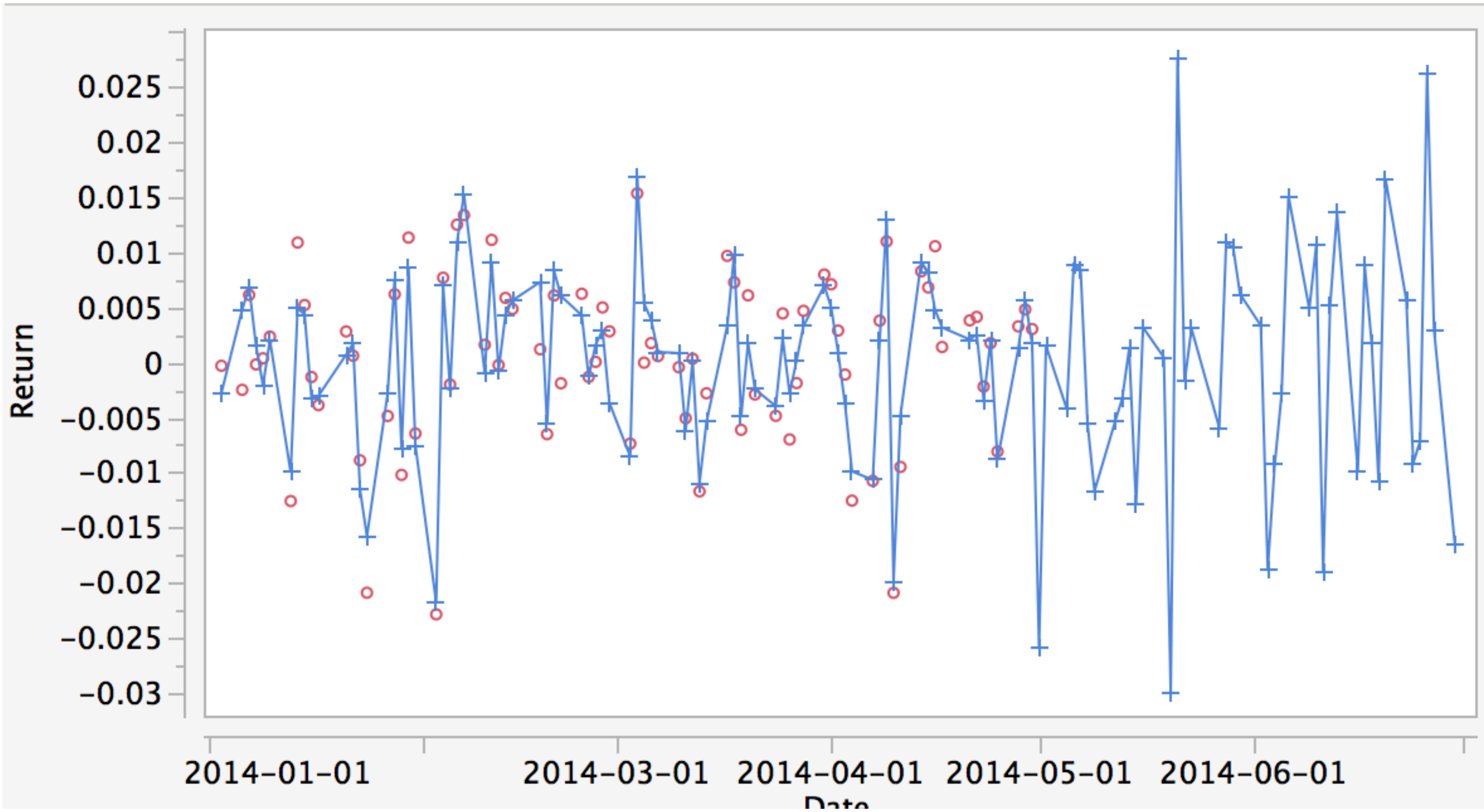
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0047436	0.000834	5.69	<.0001*
Trading Rule 02	-0.002382	0.000526	-4.53	<.0001*
Trading Rule 06	-0.001643	0.000473	-3.47	0.0010*
Trading Rule 07	-0.002415	0.000501	-4.82	<.0001*
Trading Rule 10	0.0014874	0.000401	3.71	0.0005*
Trading Rule 11	0.0020475	0.000434	4.72	<.0001*
(Trading Rule 01+0.16029)*(Trading Rule 02-0.03684)	0.0024829	0.000449	5.53	<.0001*
(Trading Rule 03+0.10456)*(Trading Rule 03+0.10456)	-0.001174	0.000349	-3.37	0.0014*
(Trading Rule 01+0.16029)*(Trading Rule 04-0.05089)	0.0023611	0.000424	5.56	<.0001*
(Trading Rule 01+0.16029)*(Trading Rule 05+0.10883)	-0.00283	0.000488	-5.80	<.0001*
(Trading Rule 02-0.03684)*(Trading Rule 05+0.10883)	-0.002749	0.000533	-5.15	<.0001*
(Trading Rule 04-0.05089)*(Trading Rule 06-0.13398)	-0.00102	0.000367	-2.78	0.0076*
(Trading Rule 07-0.08816)*(Trading Rule 07-0.08816)	-0.001282	0.000333	-3.85	0.0003*
(Trading Rule 06-0.13398)*(Trading Rule 08-0.06525)	-0.002597	0.000468	-5.55	<.0001*
(Trading Rule 05+0.10883)*(Trading Rule 09-0.00019)	0.0013912	0.000419	3.32	0.0017*
(Trading Rule 06-0.13398)*(Trading Rule 09-0.00019)	-0.002956	0.000431	-6.87	<.0001*
(Trading Rule 08-0.06525)*(Trading Rule 09-0.00019)	-0.002402	0.000563	-4.27	<.0001*
(Trading Rule 09-0.00019)*(Trading Rule 09-0.00019)	0.0021271	0.000338	6.30	<.0001*
(Trading Rule 08-0.06525)*(Trading Rule 10-0.17487)	-0.001669	0.00066	-2.53	0.0145*
(Trading Rule 09-0.00019)*(Trading Rule 10-0.17487)	-0.003865	0.000433	-8.93	<.0001*
(Trading Rule 08-0.06525)*(Trading Rule 11+0.00907)	0.0011033	0.000471	2.34	0.0231*
(Trading Rule 11+0.00907)*(Trading Rule 11+0.00907)	0.0014265	0.000298	4.79	<.0001*
(Trading Rule 02-0.03684)*(Trading Rule 12+0.11888)	-0.002147	0.000634	-3.39	0.0014*
(Trading Rule 01+0.16029)*(Trading Rule 13-0.12776)	-0.003254	0.000506	-6.43	<.0001*
(Trading Rule 07-0.08816)*(Trading Rule 13-0.12776)	0.0024976	0.00036	6.94	<.0001*
(Trading Rule 01+0.16029)*(Trading Rule 14+0.0272)	-0.004153	0.000476	-8.73	<.0001*
(Trading Rule 08-0.06525)*(Trading Rule 14+0.0272)	0.0022315	0.000745	2.99	0.0042*
(Trading Rule 14+0.0272)*(Trading Rule 14+0.0272)	-0.003191	0.000381	-8.38	<.0001*
(Trading Rule 08-0.06525)*(Trading Rule 15-0.12571)	-0.005382	0.000672	-8.01	<.0001*
(Trading Rule 09-0.00019)*(Trading Rule 15-0.12571)	-0.003577	0.000528	-6.78	<.0001*

Wharton
Department of Statistics

Residuals diagnostics check out fine...

Predictions

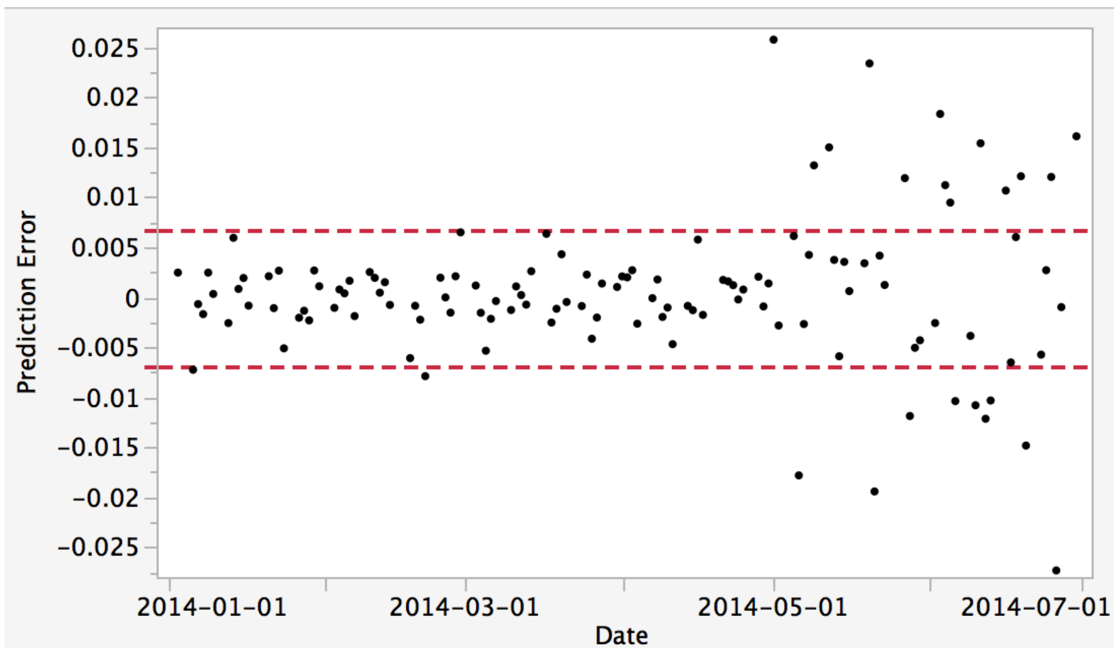
- Plot of predictions with actual
- Fit anticipates turning points.



Evaluating the Model

- Compare claimed to actual performance
  - R^2 = 89% with RMSE = 0.0032
  - How well does it predict May and June?
- SD of prediction errors much larger than model claimed

±2 RMSE



What went wrong?

Wharton
Department of Statistics

Forward Stepwise

- Allow all possible interactions, 135 possible
  - Start with 15 X's
  - Add 15 squares of X's
  - Add $^{15*14}/_{2}$ = 105 interactions
  - Principle of marginality?
- Forward search
  - Greedy search says to add most predictive
  - Problem is when to stop?
- Use statistical significance?
  - What threshold for the p-value?
  - Follow convention and set α=0.05 or larger?

Response surface
in JMP

Explanation of Problem

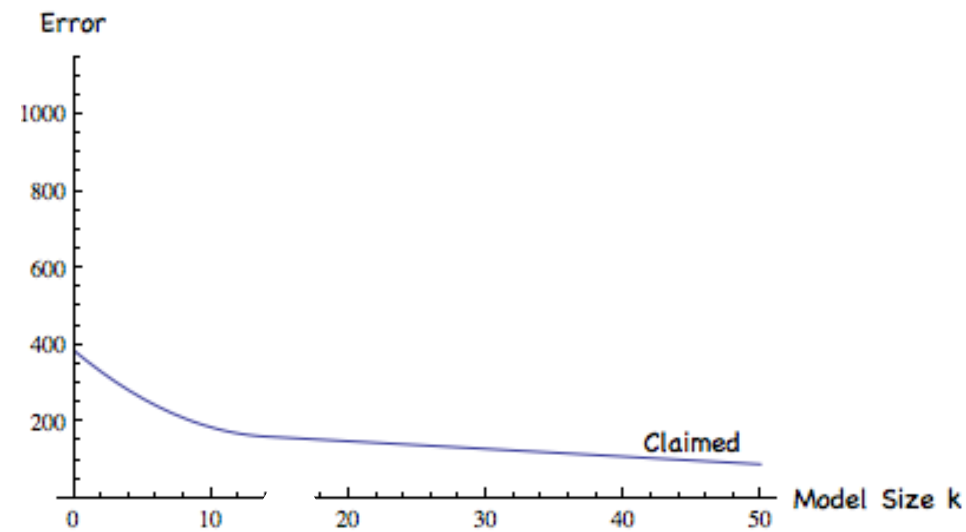
- Examine the definition of the technical trading rules used in the model

Random Normal[ ]

- Why did the stepwise get this so wrong?
  - Problem is classic example of over-fitting
  - Tukey "Optimization capitalizes on chance"
- Problem is not with stepwise
  - Rather it lies with our use of classical statistics
  - α=0.05 intended for one test, not 135

Over-Fitting

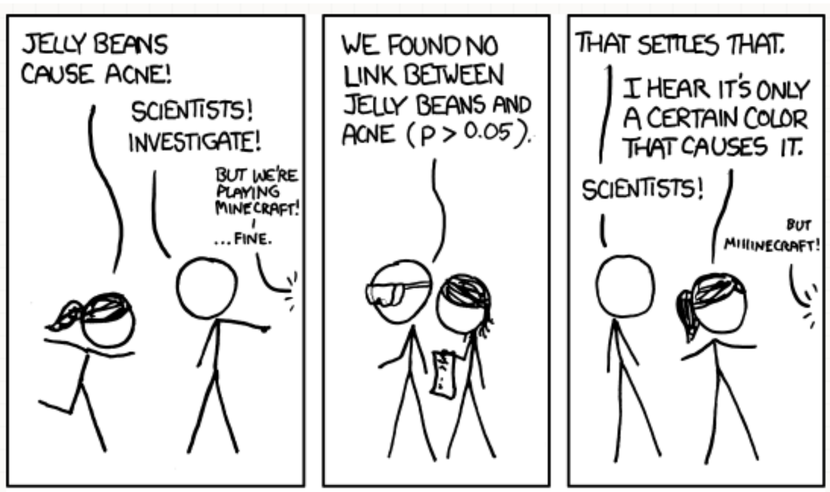
- Critical problem in data mining
  - Caused by an excess of potential explanatory variables (predictors)
- Claimed error steadily shrinks with size of the model
- "Over-confident"
  - Model claims to predict new cases better than it will.

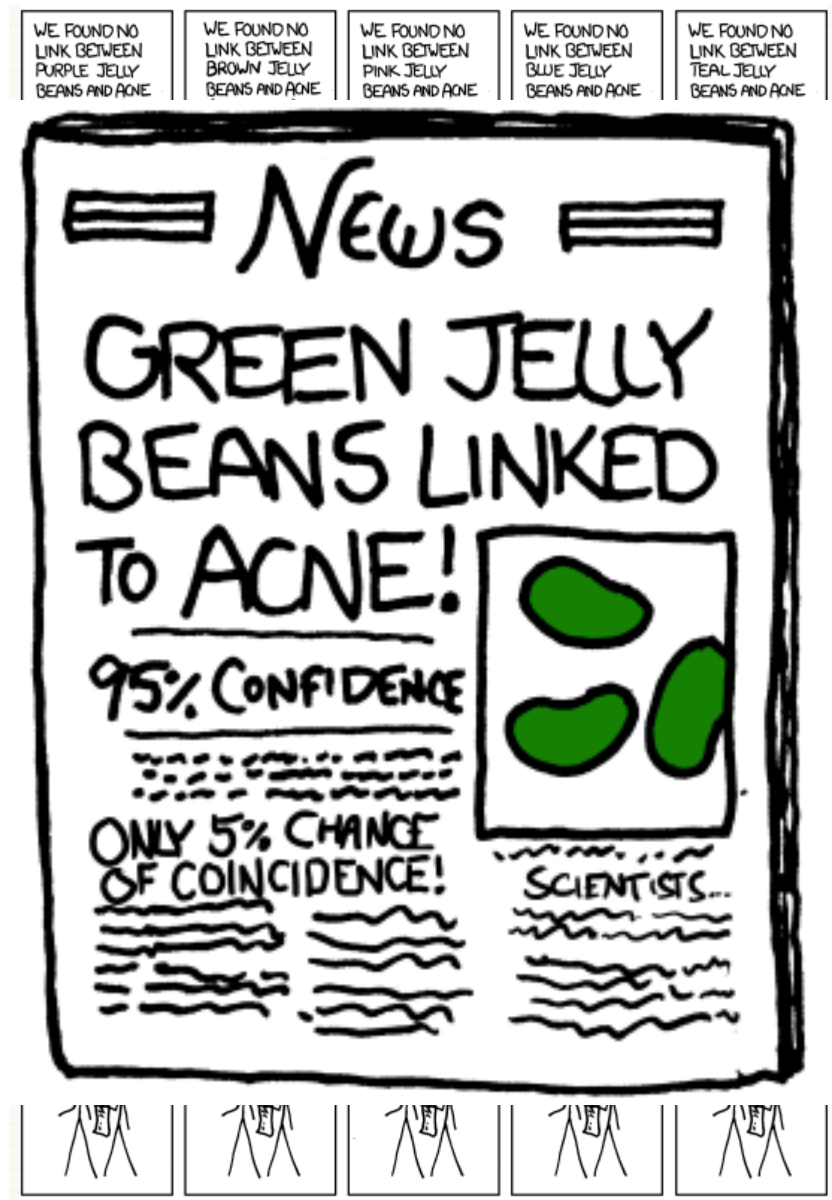


- Challenge
  - Select predictors that produce a model that minimizes the prediction error without over-fitting.

Problem in Science

xkcd





- Source of publication bias in journals
- Statistics rewards persistence

How to get it right?

- Three approaches
  - Avoid stepwise (and similar methods) altogether
  - Reserve a validation sample (cross-validation)
  - Be more choosy about what to add to model
- Bonferroni rule
  - Set the p-value based on the scope of the search
  - Searching 135 variables, so set the threshold to $0.05/135 \approx 0.00037$
  - Result of stepwise search?

Bonferroni gets it right…
Nothing is added to the model!

Take-Aways

- Missing data
  - Fill in with an added indicator for missingness
- Over-fitting
  - Model includes things that appear to predict the response but in fact do not
- Stepwise regression
  - Illustrative greedy search for features that mimics what we do manually when modeling
  - Expansive scope that includes interactions
  - Bonferroni: Set p-to-enter = 0.05/(# possible)

Assignment

- Missing data
  - What do you do with them now?
- Try doing stepwise regression with your own software.
  - Does your software offer robust variance estimates (aka White or Sandwich estimates)
- Take a look at the ANES data

Next Time

- Review of over-fitting
  - What it is and why it matters
  - Role of Bonferroni
- Other approaches to avoiding over-fitting
  - Model selection criteria: AIC, BIC, …
  - Cross-validation
  - Shrinkage and the lasso