Data Mining Introduction

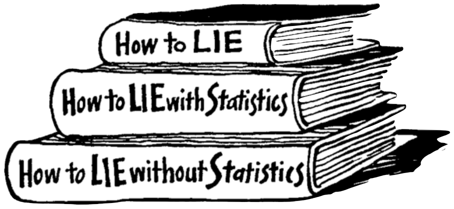
Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania

www-stat.wharton.upenn.edu/~stine

Wharton
Department of Statistics

What is data mining?

- An insult?



- Predictive modeling
  - Large, wide data sets, often unstructured
  - Automatic, complex models

    Networks, trees, ensembles… “black boxes”

    Exploit results from theory...

    universal models, random projections, multiview learning
  - Prediction rather than explanation

    Association and prediction rather than cause and effect

What is magic?

- Testable claims
  - Science requires making claims that are testable
  - Claimed predictive accuracy provides such a test

Data Mining in Social Sci

- Poor match to social science?
  - Empiricism run wild, lack of theory or hypotheses
  - Post hoc inference
- Response
  - Need to leverage technology

    Tukey comments on cost of theory vis-a-vis cost of computing
  - Honest

    A better match to what most of us do in practice
  - Diagnostic

    Have I missing something?
  - Deep connections

    Multidimensional scaling, likelihood, modern regression

Plan

- Week 1
  - Data mining with regression, logistic regression
  - Illustrate key ideas in familiar context
- Week 2
  - Alternative methods
  - Trees, networks, ensemble methods

    Boosting and bagging
- Syllabus
  - Hands-on: Lab sessions each week
  - Annotated bibliography
  - July 4

Software

- Must do statistics to learn statistics
- Modern computing provides
  - New ways to look at old things, like regression
  - New approaches to data analysis
- Packages
  - JMP from SAS

    Front-end to SAS Enterprise Miner

    Available on Newberry systems
  - R
  - Others: Stata, SPSS, Weka,…

My Background

- Time series analysis
  - Effects of modeling on forecast accuracy
  - Bootstrap resampling
- Model selection in general
  - Predictive models in credit, health
- Recent
  - Alternative methods for building regression
  - Combining traditional data and text
- Long time ‘friend’ of Summer Program
  - Political science and voting behavior

t-shirts

Research Questions

- What question do you want to answer?
  - Can your data provide an answer?

Question to guide analysis

Ideal data?

- Questions from science, business
  - Who's most at risk of a disease?
  - What's going to happen in financial markets?
  - Are any of these people dishonest?
- Social science questions: voting behavior
  - Will this person vote if I get them to register?
  - Whom will this registered voter choose?
  - Whom would those who didn't vote choose?

2008 ANES Survey

- Background of survey ICPSR #25383
  - Two waves, every two years
- Questions
  - Categorical responses
    Did you vote? For whom?
  - Numerical responses
    How much do you like this candidate
- Why are these interesting?
  - Get out the vote, phone banks
  - Role of participation in election…
    Would those who didn't vote change things?
- Is the ANES ideal data?

90/10 rule

  - Missing data, self-reported, interviewer effects...

2012 ANES Survey

R data file

- Background ICPSR #34808
  - Mix of in-person interview, internet panels
  - Fewer variables, less detail than in 2008
  - More cases than in 2008
- Questions
  - Key responses: Did you vote? For whom?
  - No numerical responses
    Recoded into bins (e.g., age ranges)
  - Want numerical variables?
    Role for theory (example follows)
- Issues remain
  - Prevalent missing data, manipulating labels
  - Not a simple random sample (50.6% Obama vs 58% in anes)

Data Browsing

- Spirit of EDA, exploratory data analysis
  - Know your data
  - Know your tools
- ANES 2008 data table in JMP
  - Load directly from SPSS sav file 25383-0001-Data.sav
  - Almost square: 2,323 cases x 1956 variables
  - Sampling weights
  - Virtually all categorical, with many missing
  - Feeling thermometers (B1), 'moderators' (N5)
- Variable creation
  - No algorithm is as good (yet) as the modeler who knows how to build predictive features

Browsing ANES

- Marginal distributions in 2008 data
  - Interactive graphics: Plot linking and brushing
- Interesting variables
  - Participation, political interest (A1-A10)

    prevalence of missing data. Problem for categorical?
  - Feeling thermometer (FT, B1 group)

    numbers or categories? Missing a problem?
- Other interesting relationships to explore
  - Spending bundle and scaling (P1 group)

    Likert scales, ordinal-interval-ratio measurement
  - Intention to vote (A6, Q1 in first wave)

    Repeats prior question, reliability of data
  - Choice in election (C6 in second wave)

    Importance of sampling weights (65.5% in sample, 53% in election)

JMP treatment
of numerical/
categorical

Browsing ANES

- Bivariate relationships

  Special scatterplot
  if mix types

  - Contingency tables, scatterplots
  - Asymmetry of roles: explanatory vs response
- Consistency of responses: scatterplot
  - FT rating of Dem candidate pre/post election B1/D1
- War and voter choice: table, mosaic plot

  A14f/C6

  - Choice and opinion of war in Iraq
- Feelings and voter choice: logistic regression

  D2/C6

  - Choice and rating of candidate

Models

- What is a statistical model?
- Model
  - Simplification of reality
  - Facilitate answering specific types of questions
  - Example: Maps

    Map for driving directions versus subway map

"All models are wrong,
but some are useful"
Box

- What is a statistical model?
  - Data generating process
  - Probability model describing a random mechanism
- Link to theory
  - Test theory's claims for features of model

Assumptions

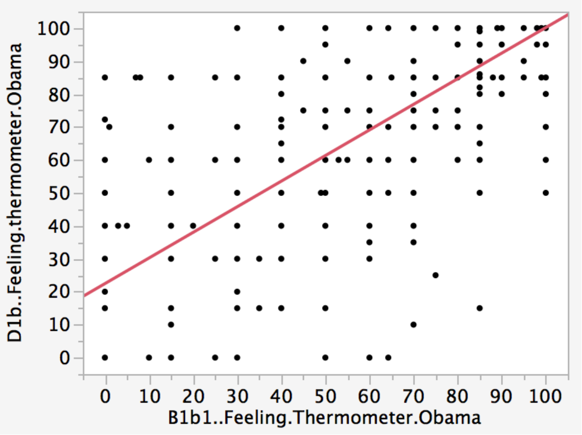
- Models make two types of assumptions
  - systematic structure

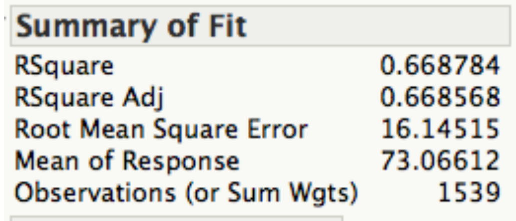
    linear equation in regression
  - “unexplained” variation

    [✓ (a) Independent]
    (b) Equal uncertainty
    (c) Bell shaped
- Which make sense within the context of this model using the ANES data?
  - Does it matter if the assumption is not met?
- Why do we make such assumptions?

Simple Model

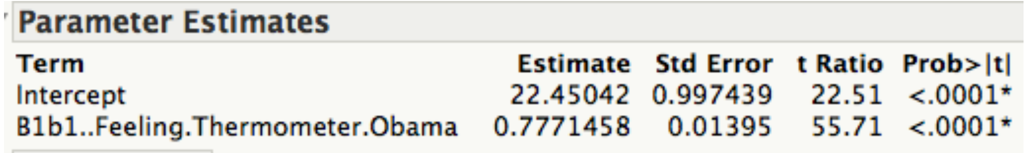
- Bandwagon model
  - Affiliation with winner
- Relate to SRM
  - $Y = \beta_0 + \beta_1 X + \varepsilon$
  - H0: $\beta_0=0$, $\beta_1=1$
- Tests, inference
  - Confidence interval
  - Hypothesis test
  - Standard error
  - t-statistic
  - p-value





Summary of Fit

RSquare	0.668784
RSquare Adj	0.668568
Root Mean Square Error	16.14515
Mean of Response	73.06612
Observations (or Sum Wgts)	1539



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	22.45042	0.997439	22.51	<.0001*
B1b1..Feeling.Thermometer.Obama	0.7771458	0.01395	55.71	<.0001*

Conclude?

Analysis using just voters, anes_2008_voters.jmp

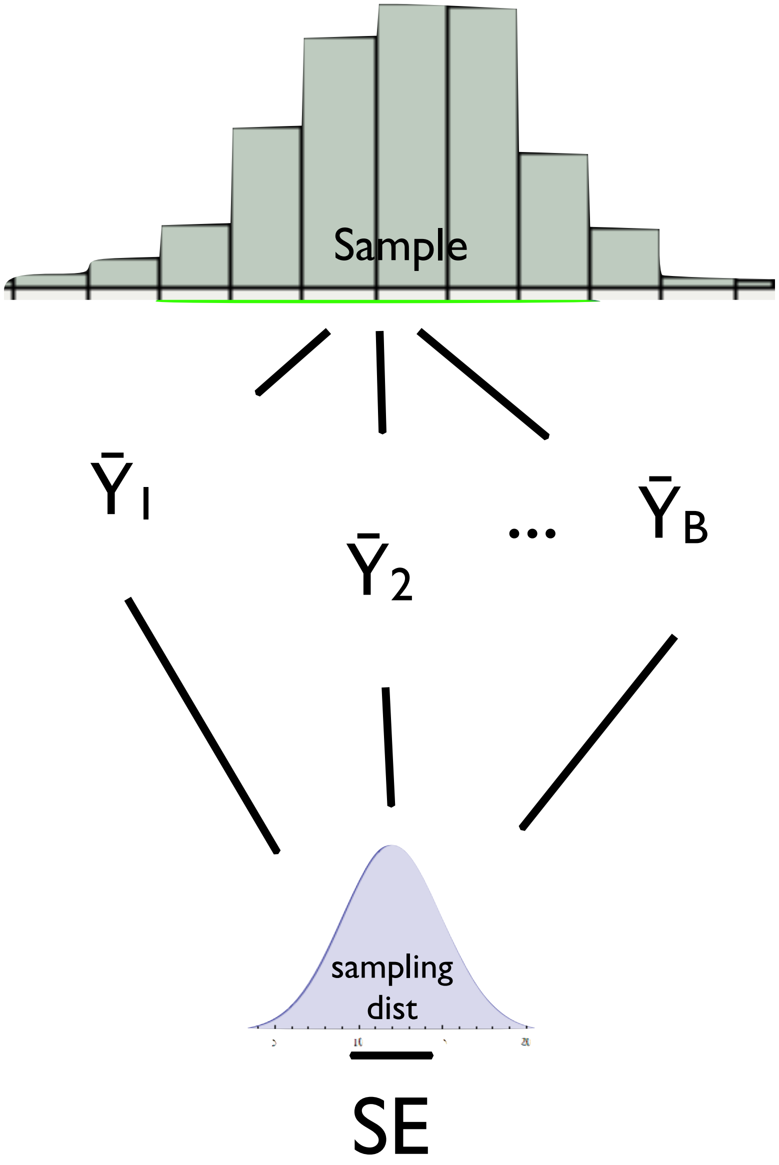
Bootstrapping

- Standard error is key to inference
  - What are standard errors?
- BS is alternative method for obtaining standard errors and confidence intervals
  - Estimates standard error by simulation
  - Sampling with replacement from observed distribution of data
- Implementation
  - R ‘bootstrap’ package - also easy to do yourself
  - Throughout JMP

    Control click.

Bootstrap Sampling

- Standard error
  - Standard deviation of statistic
  - Repeated samples from the population
- Bootstrap standard errors
  - Simulate standard error
  - Draw B samples from the observed sample itself.
  - Sampling is done with replacement times
  - Collection of stats estimates sampling distribution



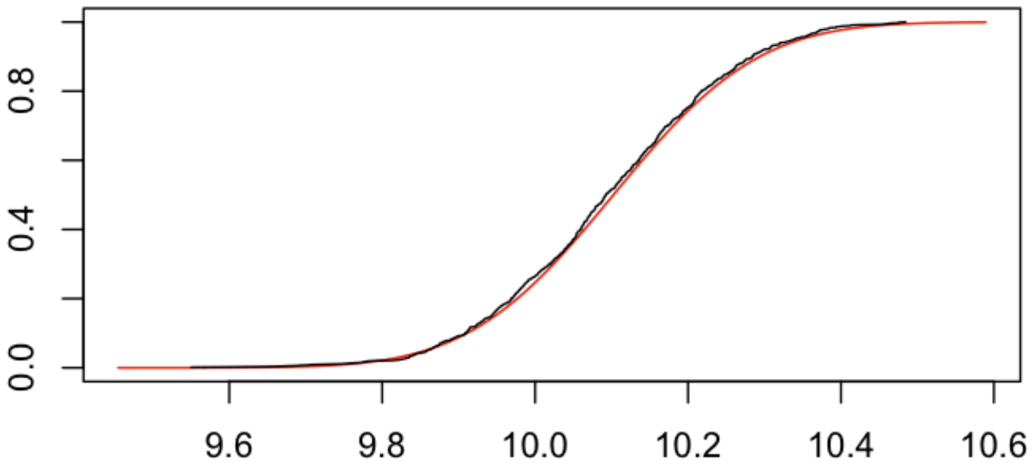
Wharton
Department of Statistics

Bootstrap Example

- Bootstrap problem with known answer
  - Normal population with mean μ and var σ^2.
  - Sampling distribution of the mean is $N(\mu,\sigma^2/n)$
  - Simple to do in R since easy to script

    Several R packages implement extensive bootstrap methods

- Bootstrap sampling distribution
  - Matches theory without the math

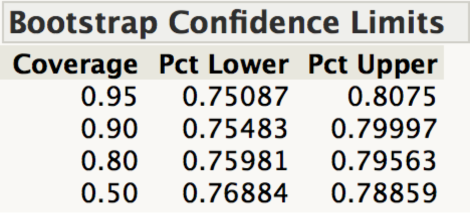


bootstrap.R

Bootstrap in Regression

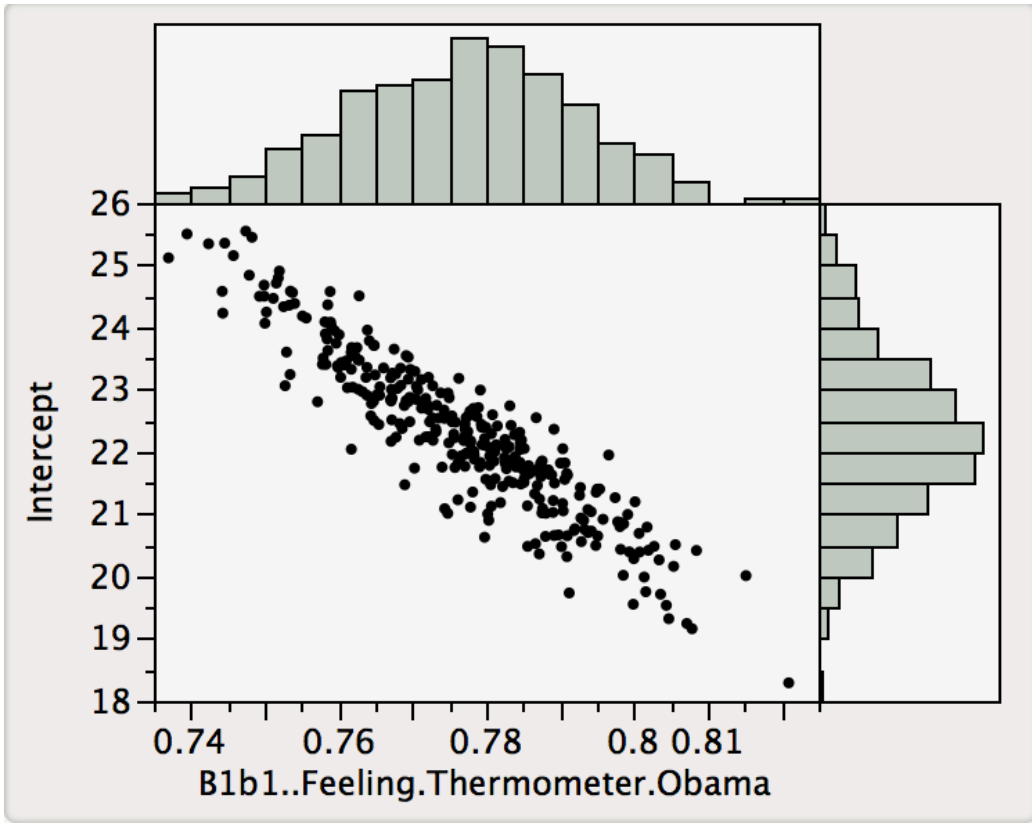
- Distributions of slope and intercept in BS samples estimate sampling distributions

Can get confidence intervals too



Bootstrap Confidence Limits

Coverage	Pct Lower	Pct Upper
0.95	0.75087	0.8075
0.90	0.75483	0.79997
0.80	0.75981	0.79563
0.50	0.76884	0.78859

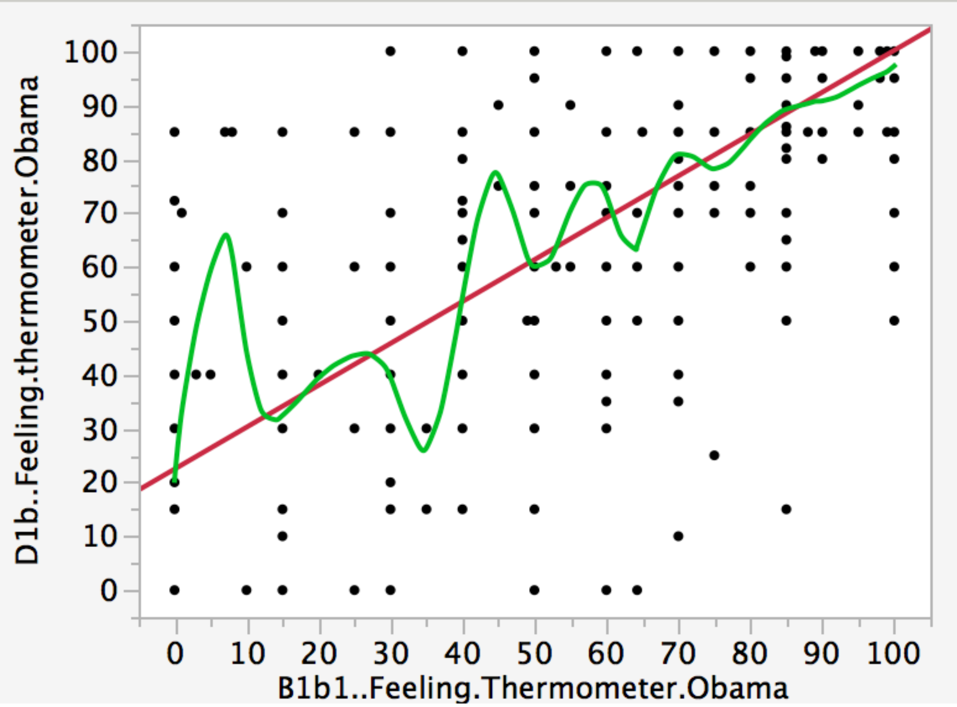


Notice the correlation between the estimated slope and intercept.

Cntl-Click on estimates table

Model Diagnostics

- Residual diagnostics
- Calibration
  - Is the model correct on average: $E(Y|\hat{Y}) = \hat{Y}$
  - Check by smoothing Y on X or Y on $\hat{Y}$



Interactive tool
for spline in JMP

Wharton
Department of Statistics

Multiple Regression

- Does one explanatory variable provide a complete description of the response?
  - What other factors affect association between pre-election rating and post rating?

    Media

    Emotional interest in outcome

    Attitude to Iraq war, economy,…
  - How do these factors contribute to model

    Additive as another explanatory variable

    Affecting other factors (interaction)
- How should we decide which?
  - Trial and error by adding to multiple regression?
  - Use of t-statistics and p-values to decide

Multiple Regression Model

- Grow to a multiple regression model
  - Underlying model has assumptions
  - Key assumption is the larger equation

$$E(Y|X) = \beta_0 + \beta_1\ X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

  - Same assumptions for the unexplained variation

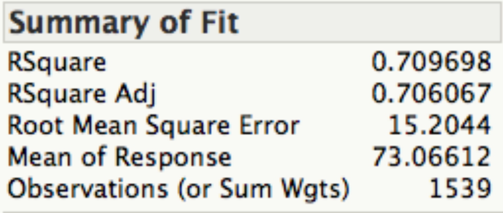
X's are known
and additive

- Evaluating explanatory variables
  - Which do we keep, which do we exclude?
- Use of t-statistics, F-statistics in this setting
  - How many variables did you try?
  - What made you try those?
  - What about other correlated variables?

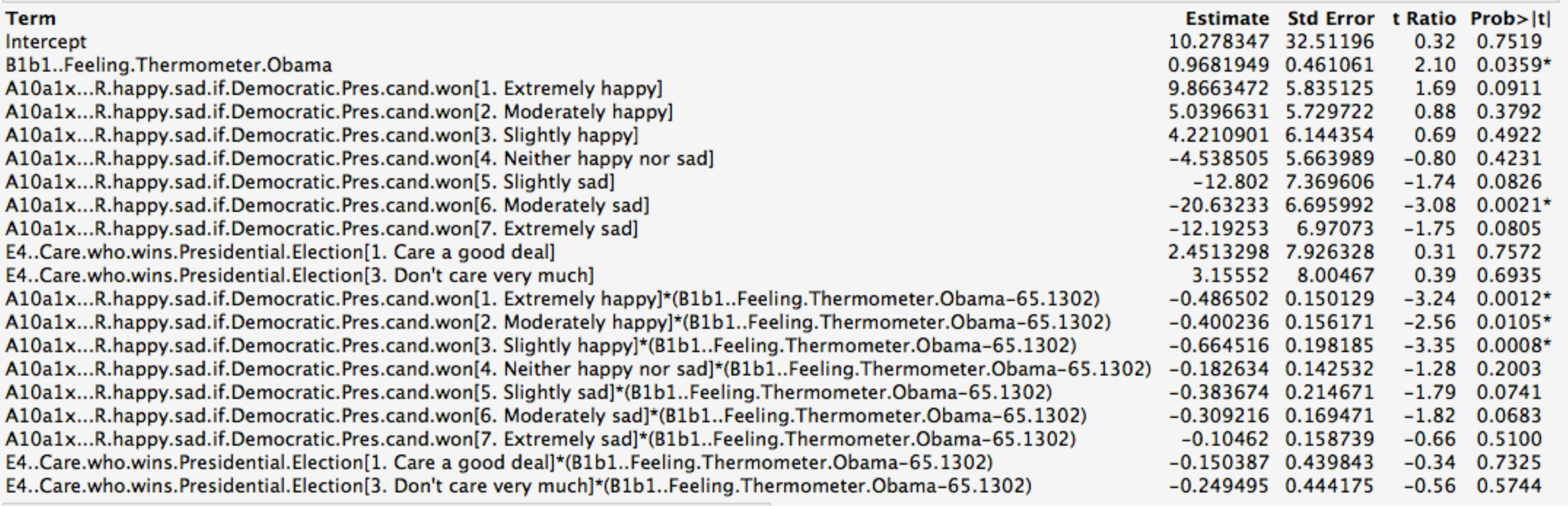
Statistics
rewards
persistence!

Possible Model

- Grow simple regression into a multiple regression model that includes interactions
  - Add Happy/Care, 'care who wins'
  - Interaction: flexibility vs complexity
  - What does all of this tell you?



Summary of Fit	
RSquare	0.709698
RSquare Adj	0.706067
Root Mean Square Error	15.2044
Mean of Response	73.06612
Observations (or Sum Wgts)	1539

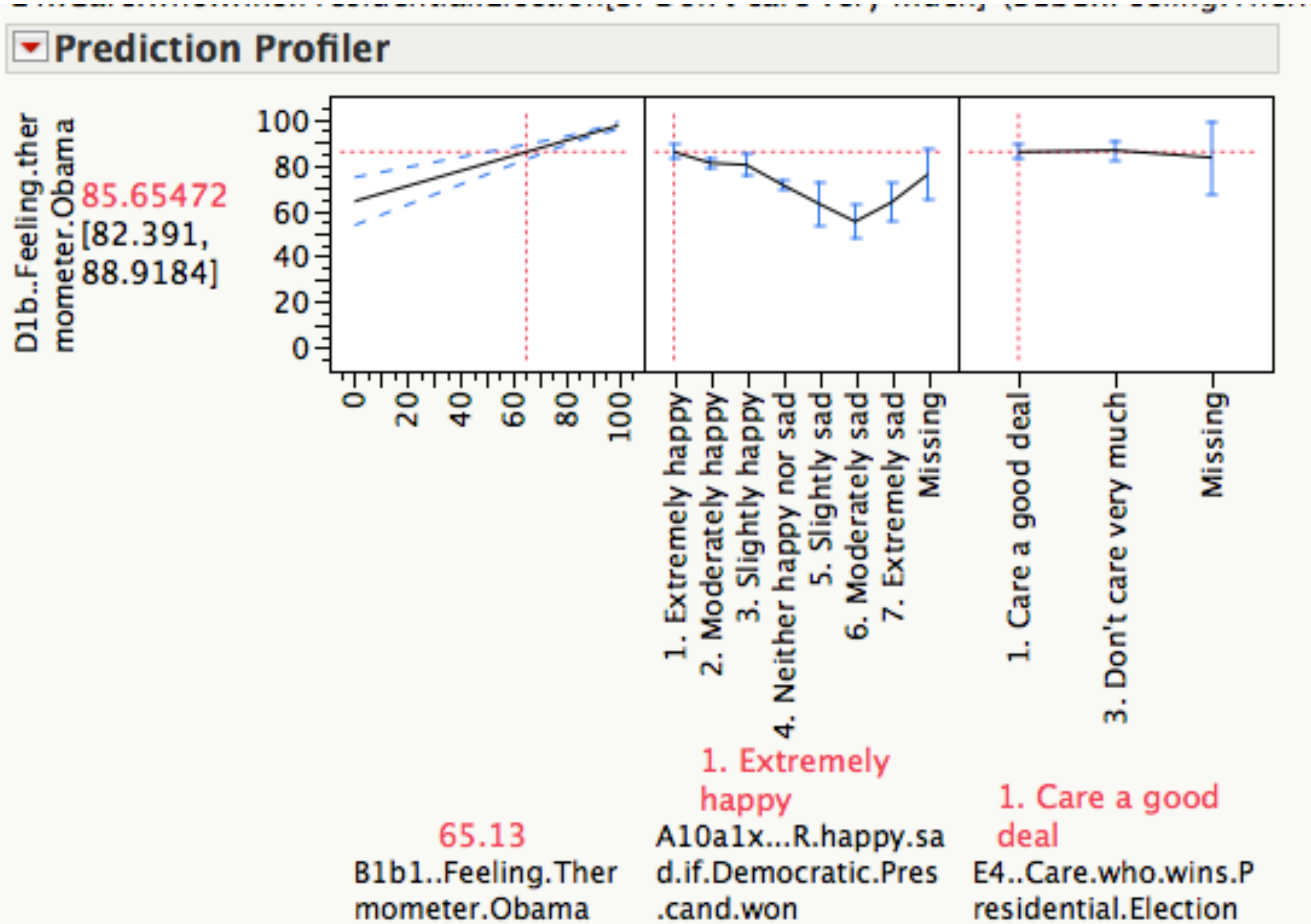


Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.278347	32.51196	0.32	0.7519
B1b1..Feeling.Thermometer.Obama	0.9681949	0.461061	2.10	0.0359*
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[1. Extremely happy]	9.8663472	5.835125	1.69	0.0911
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[2. Moderately happy]	5.0396631	5.729722	0.88	0.3792
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[3. Slightly happy]	4.2210901	6.144354	0.69	0.4922
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[4. Neither happy nor sad]	-4.538505	5.663989	-0.80	0.4231
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[5. Slightly sad]	-12.802	7.369606	-1.74	0.0826
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[6. Moderately sad]	-20.63233	6.695992	-3.08	0.0021*
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[7. Extremely sad]	-12.19253	6.97073	-1.75	0.0805
E4..Care.who.wins.Presidential.Election[1. Care a good deal]	2.4513298	7.926328	0.31	0.7572
E4..Care.who.wins.Presidential.Election[3. Don't care very much]	3.15552	8.00467	0.39	0.6935
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[1. Extremely happy]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.486502	0.150129	-3.24	0.0012*
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[2. Moderately happy]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.400236	0.156171	-2.56	0.0105*
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[3. Slightly happy]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.664516	0.198185	-3.35	0.0008*
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[4. Neither happy nor sad]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.182634	0.142532	-1.28	0.2003
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[5. Slightly sad]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.383674	0.214671	-1.79	0.0741
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[6. Moderately sad]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.309216	0.169471	-1.82	0.0683
A10a1x...R.happy.sad.if.Democratic.Pres.cand.won[7. Extremely sad]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.10462	0.158739	-0.66	0.5100
E4..Care.who.wins.Presidential.Election[1. Care a good deal]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.150387	0.439843	-0.34	0.7325
E4..Care.who.wins.Presidential.Election[3. Don't care very much]*(B1b1..Feeling.Thermometer.Obama-65.1302)	-0.249495	0.444175	-0.56	0.5744

Regression models often feel like a black box too...

Profile of Model

- Alternative way to look at a model
  - Visual presentation of effects vs tabular
  - What does the interaction do? (animated)


Prediction Profiler

D1b..Feeling.thermometer.Obama 85.65472 [82.391, 88.9184]

65.13 B1b1..Feeling.Thermometer.Obama

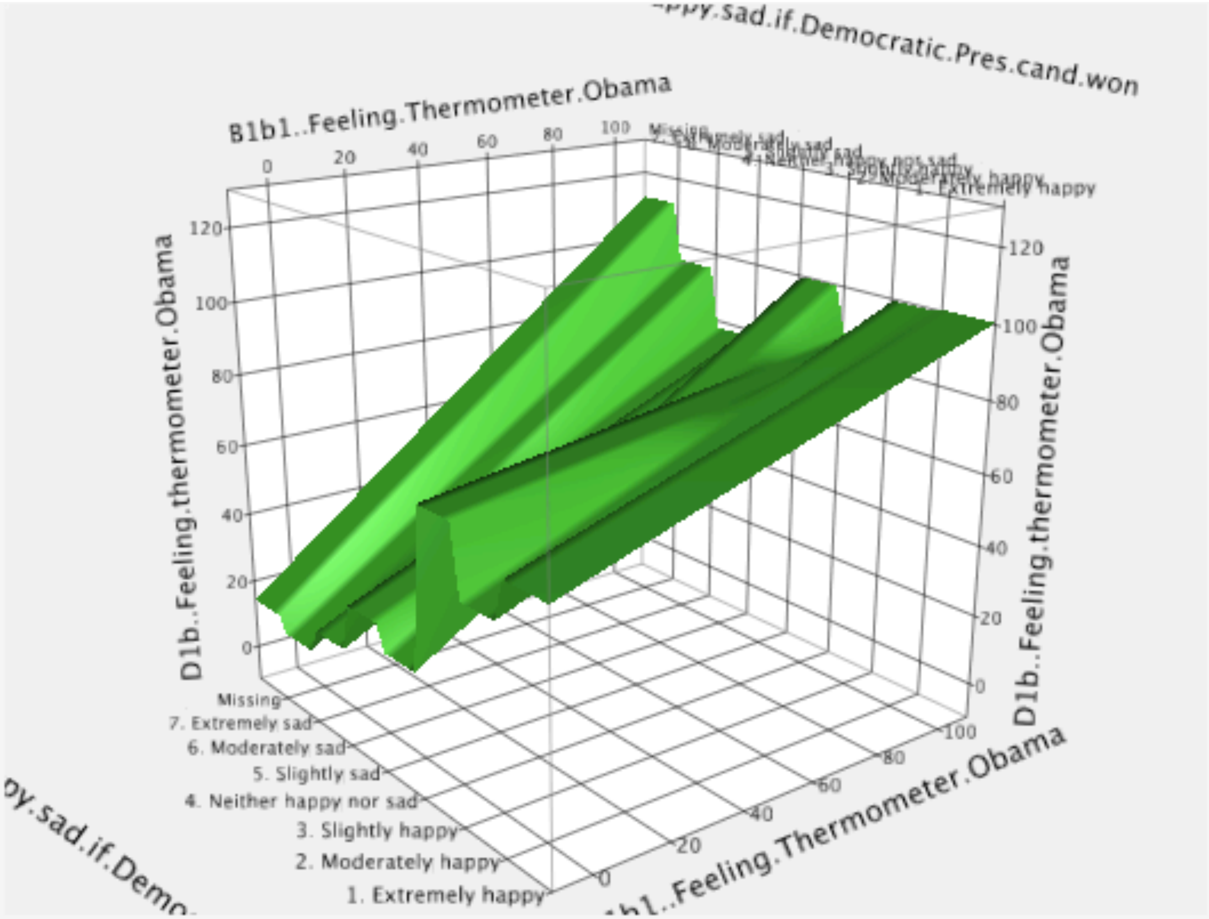
1. Extremely happy A10a1x...R.happy.sad.if.Democratic.Pres.cand.won

1. Care a good deal E4..Care.who.wins.Presidential.Election

Error bars indicate confidence

Looking at Fit

- Surface profile



How would it look were there no interaction?

Take-Aways

- Role for data mining in social sci research
    - Diagnostic
    - Better way to do what we do already
- Importance of models
    - Linking theory to data to allow inference
    - Standard error: bootstrap resampling
- Calibration
    - Check that a model is correct, on average
- Interactive visualization
    - Exploring data (plot linking, brushing)
    - Exploring models (profiling, surfaces)

Assignment

- Skim syllabus, bibliography
- Peek at the codebook for ANES
  - Will put on Newberry computers
- Think about modeling your own data
  - How did you decide on a model, hypotheses
- Come with questions...

Next Time

- Picking the features of a model.
- An often overlooked diagnostic.
- What to do about missing values?