Logistic Regression & Classification

Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania

Wharton
Department of Statistics

Questions

- Did you see the parade? Watch fireworks?
- Do you need to do model selection?
  - What's a big model?
  - Size of n relative to p
- How to cut and paste figures in JMP?
  - Selection tool in JMP
- Other questions?
  - Review cross-validation and lasso, in R

Classification

- Response is categorical
  - Predict group membership rather than value
  - Several ways to measure goodness of fit
- Confusion matrix
  - Label "good" if estimated P(good) > ξ

    How should you pick the threshold ξ?

Want both large

  - Sensitivity $n_{11}/(n_{11}+n_{12})$
    Specificity $n_{22}/(n_{21}+n_{22})$
  - Role for economics and calibration

		Claim	
		Good	Bad
Actual	Good	n_{11}	n_{12}
	Bad	n_{21}	n_{22}

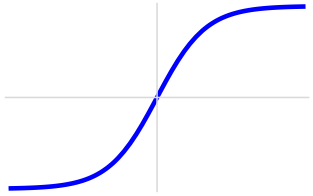
Sensitivity a.k.a. recall
Precision = $n_{11}/(n_{11}+n_{21})$

- ROC Curve
  - Graphs sensitivity and specificity over a range of decision boundaries (whether you care about them or not)

Logistic Regression

- Model
  - Assumes latent factor $\theta = x_1\beta_1 + \ldots + x_k\beta_k$ for which the log of the odds ratio is θ

$$\log \frac{P(good)}{1-P(good)} = \theta$$



  - Logistic curve resembles normal CDF
- Estimation uses maximum likelihood
  - Compute by iteratively reweighted LS regression
  - Summary analogous to linear regression

-2 log likelihood ≈ residual SS
chi-square overall ≈ overall F
chi-square estimates ≈ t^2

Example

- Voter choice
  - Fit a linear regression
  - Calibrate
  - Compare to logistic regression
- Data
  - 4,404 voters in ANES 2012
  - Response is Presidential Vote
    - Categorical for logistic
    - Limit to Obama vs Romney (just two groups, n=4,188)
    - Dummy variable for regression (aka, discriminant analysis)
  - Explanatory variables
    - Simple start: Romney-Obama sum comparison (higher favors Obama)
    - Multiple: add more via stepwise

anes_2012

Level	Count	Prob
Did not vote	1109	0.18746
Missing	403	0.06812
Voted	4404	0.74442
Total	5916	1.00000

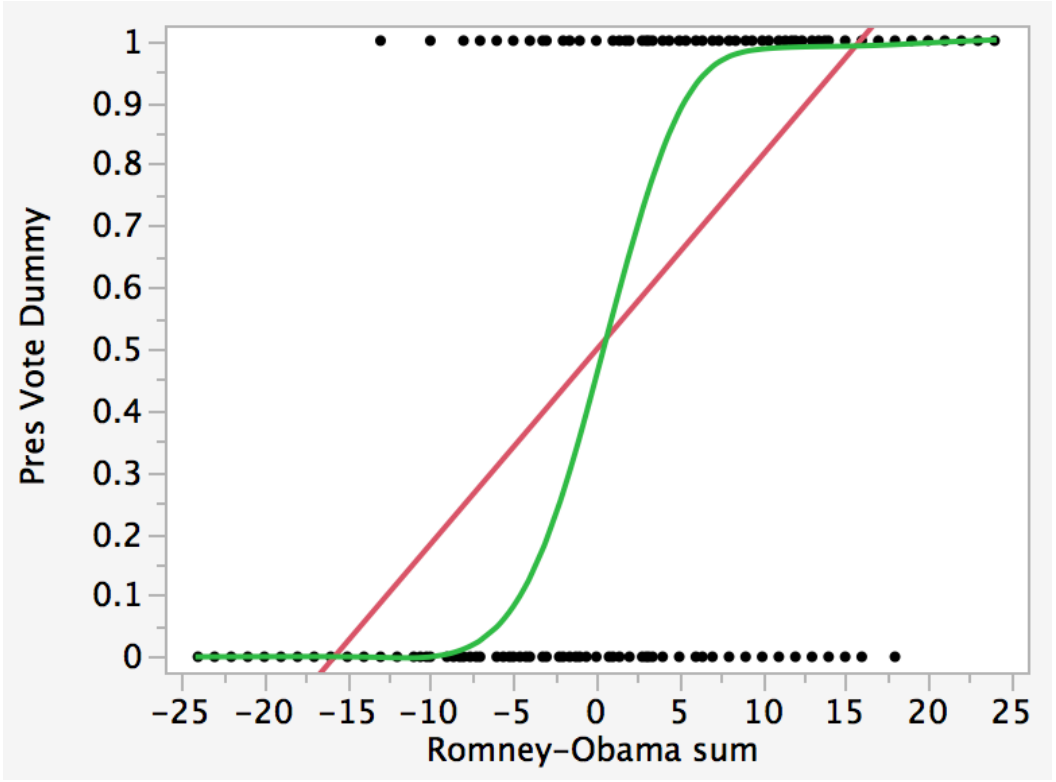
anes_2012_voters

Level	Count	Prob
Obama	2496	0.59599
Romney	1692	0.40401
Total	4188	1.00000

note over-sampling

Linear Regression

- Highly significant, but problematic



Uncalibrated!

Spline shows how to fix the fit

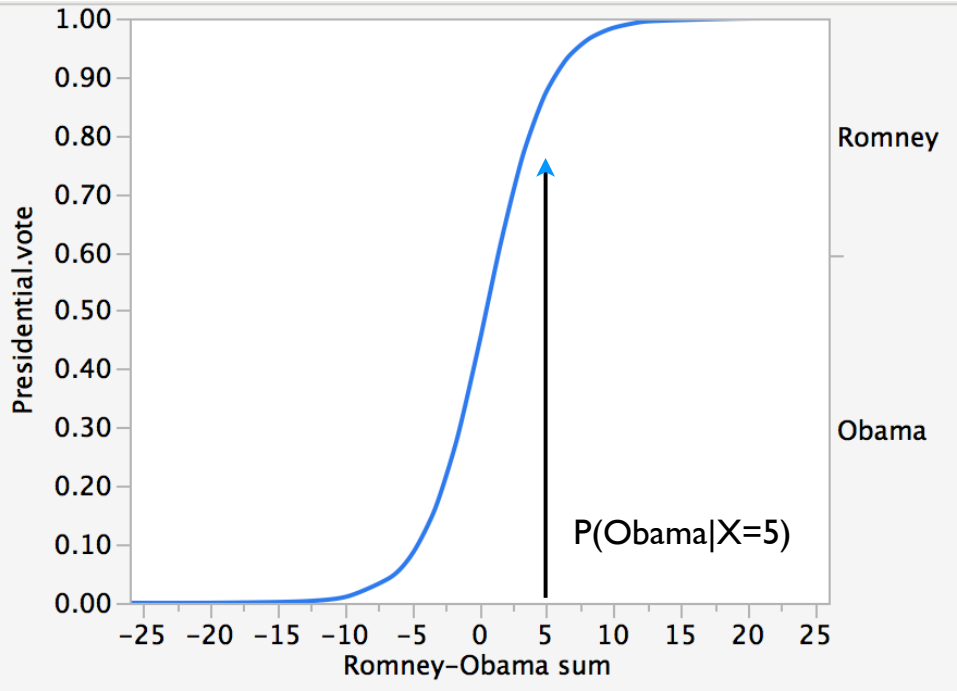
Smoothing Spline Fit, lambda=1	
R-Square	0.822578
Sum of Squares Error	178.9146
Change Lambda:	4487.454

Save predictions from spline*

*Fancy name: nonparametric single index model.

Logistic Regression

- Fitted model describes log of odds of vote



-2 Log Likelihood = Residual SS

Model	−LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	2180.0803	1	4360.161	<.0001*
Full	645.1642			
Reduced	2825.2444			

RSquare (U)	0.7716
AICc	1294.33
BIC	1307.01
Observations (or Sum Wgts)	4188

ChiSquare = t^2

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	−0.229923	0.074323	9.57	0.0020*
Romney−Obama sum	0.4340748	0.0166403	680.47	<.0001*

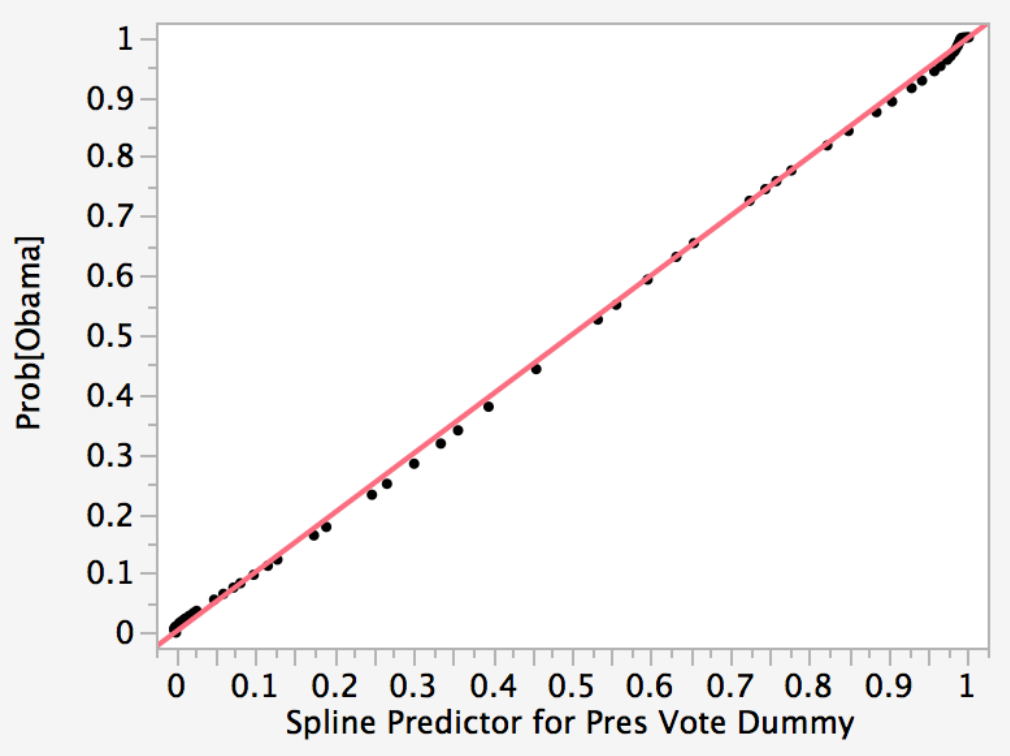
For log odds of Obama/Romney

save estimated probabilities...

Interpretation of slope, intercept?

Logistic ≈ Calibrated LS

- Compare predictions from the two models
  - Spline fit to dummy variable
  - Logistic predicted probabilities

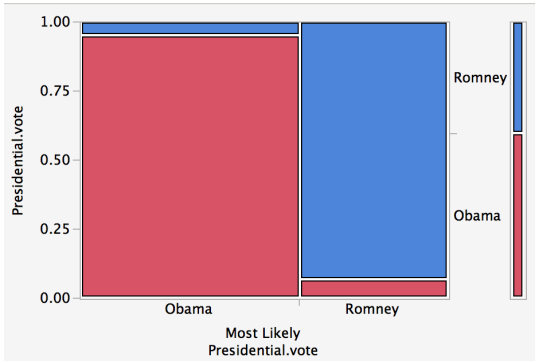


Moral

Calibrating a simple linear regression can reproduce the fit from a logistic regression

Goodness of Fit

- Confusion matrix counts classification errors
  - What threshold ξ should we use? ½ ?



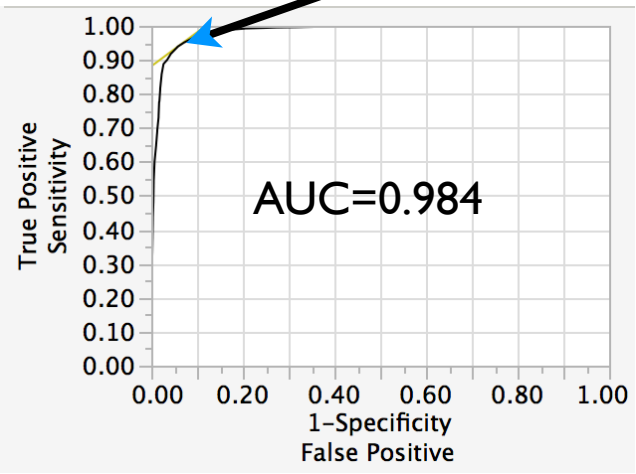
Most Likely Presidential.vote

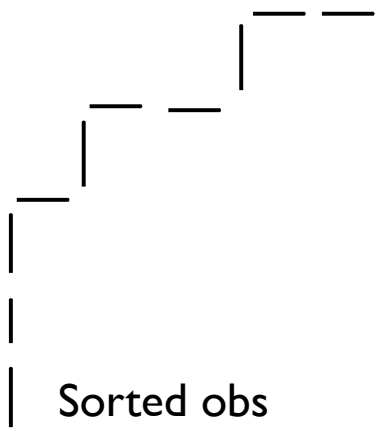
Presidential.vote	Count Row %	Obama	Romney	
	Obama	2377 95.23	119 4.77	2496
	Romney	125 7.39	1567 92.61	1692
		2502	1686	4188

sensitivity

specificity

- ROC Curve evaluates all thresholds





AUC=Area under the curve

Adding Variables

- Substantive model
  - Add party identification to the model. Better fit?

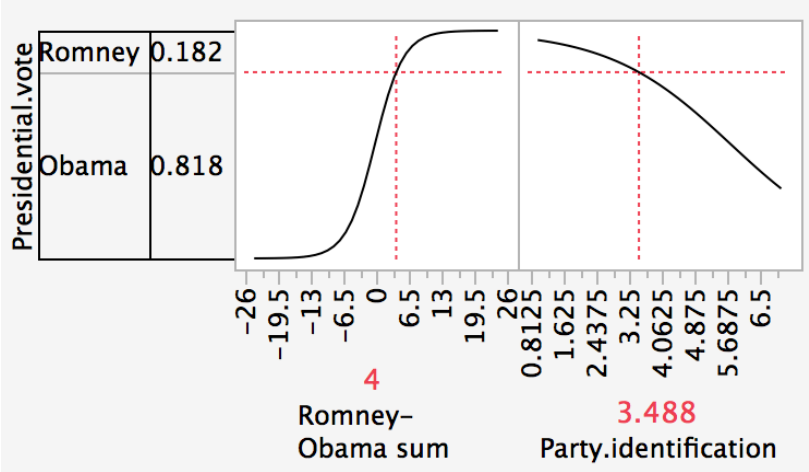
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.37277731	0.2084883	129.52	<.0001*
Romney-Obama sum	0.35354487	0.0168809	438.63	<.0001*
Party.identification	-0.6555776	0.0491522	177.89	<.0001*

Actual / Predicted

Training	Obama	Romney
Obama	2405	91
Romney	107	1585

- Profiler helps interpret effect sizes
  - Clear view of nonlinear effects

Dragging levels shows that model is nonlinear in probabilities.



Note that the interaction between these is not stat significant in logistic, but it is if modeled as linear regr.

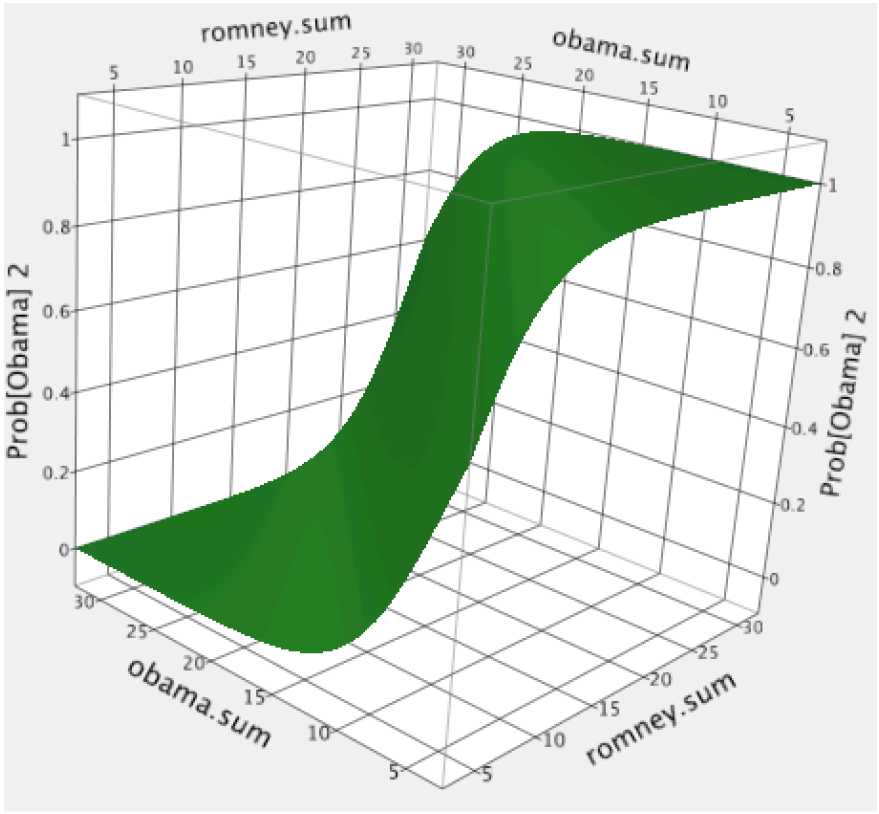
More Plots

- Surface plots are also interesting
  - Will be useful in comparison to neural network

Procedure:

Save prediction formula

Graph>Surface plot



Software is too clever… recognized Obama-Romney

Defeat by removing formula & converting tovalues (Cols>Column info…)

Stepwise Logistic

- Logistic calculations
  - Slower than OLS

    Each logistic fit requires an iterative sequence of weighted LS fits.
  - Add more variables, stepwise

    With categorical response, it takes a while to happen!

    Plus no interactions, missing indicators yet.
  - Cheat

    Swap in a numerical response, and get instant stepwise dialog
- Try some interactions!
  - Gender with other factors

    Gender interactions alone doubles number of effects

    Stepwise dialog takes a bit more time!
  - Best predictors are not surprising!

    Stop at rough Bonferroni threshold

    Useful confirmation of simpler model

Step	Parameter
1	Feeling.thermometer..Obama
2	Feeling.thermometer..Romney
3	Obama..like.dislike.scale
4	Romney..like.dislike.scale
5	Party.identification
6	Presidential.job.approval
7	(obama.handling.sum-9.42792)*Gender{Female-Male}

Refit Model

- Build logistic model
  - Use OLS to select features

    Not ideal, but better than not being able to do it at all!

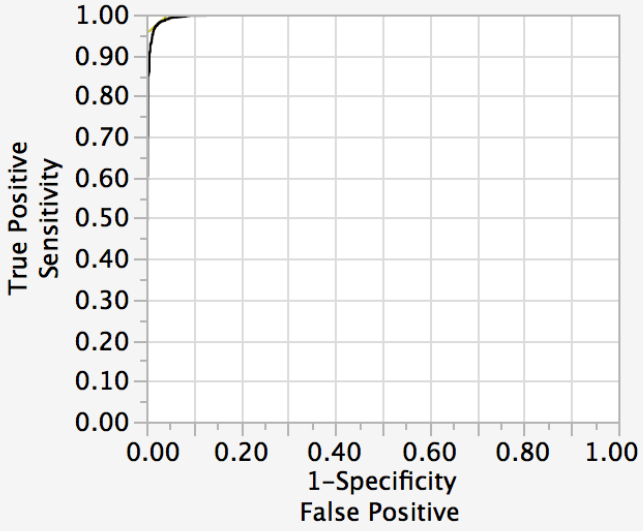
    Remove 'unstable' terms
  - Stepwise logistic on fewer columns

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	2535.9273	9	5071.855	<.0001*
Full	289.3171			
Reduced	2825.2444			

RSquare (U)	0.8976
AICc	598.687
BIC	662.034
Observations (or Sum Wgts)	4188

About ½ the errors of simple model

Actual		Predicted
Training	**Obama**	**Romney**
Obama	2452	44
Romney	51	1641

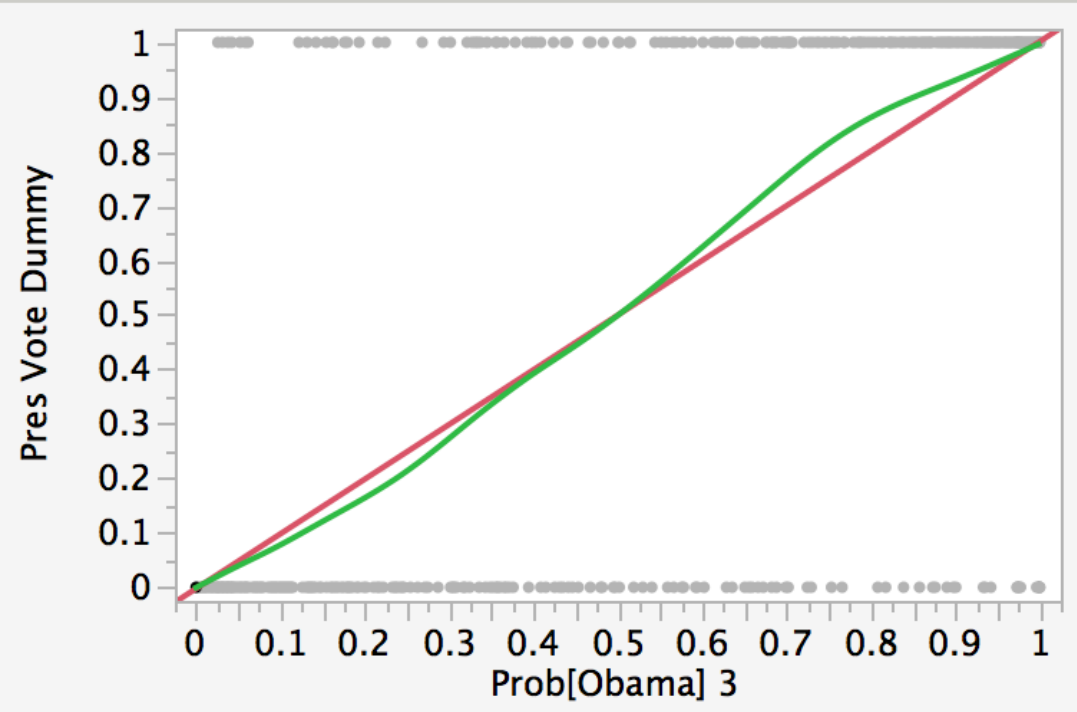


Using Presidential.vote='Obama' to be the positive level

AUC
0.99593

Calibrating the Logistic

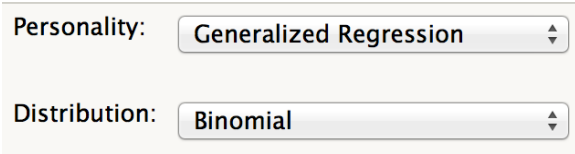
- Logistic fit may not be calibrated either!
  - Probabilities need not tend to 0/1 at boundary
  - Latent effect not necessarily logistic
  - Hosmer-Lemeshow test



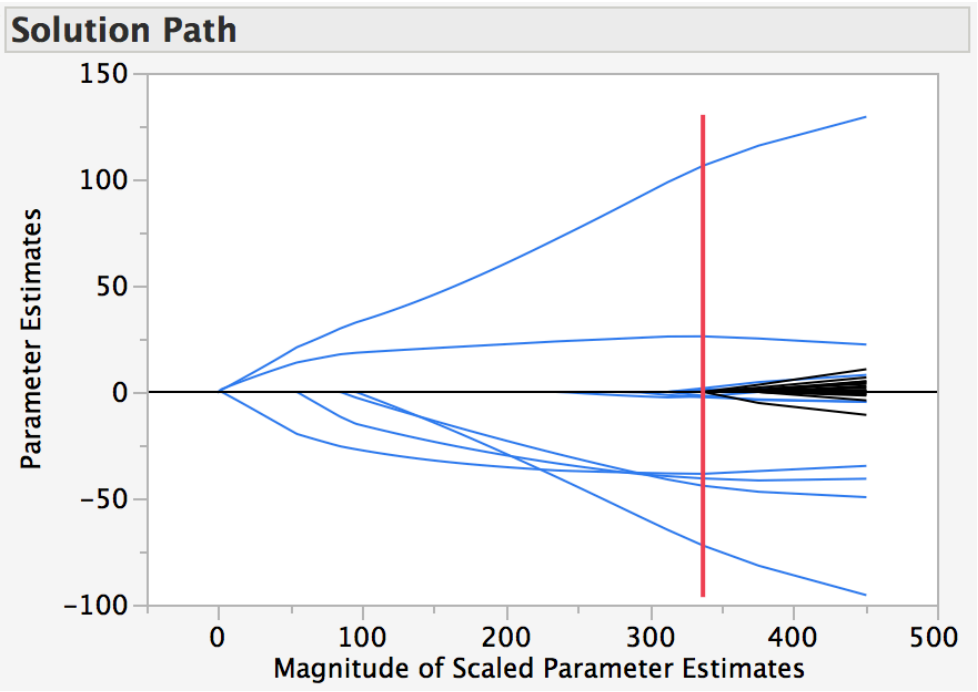
Very nearly linear

Lasso Alternative

- Convert prior stepwise dialog to 'generalized regression'



- Use BIC in JMP for faster calculation
  - generally similar terms



Term	Estimate
Intercept	0.9845193
Romney-Obama sum	0
Party.identification	-44.11835
Feeling.thermometer..Democratic.Party	0
Feeling.thermometer..Republican.Party	0
Media Frequency	0
Media Attention	0
Feeling.thermometer..Obama	26.127024
Feeling.thermometer..Romney	-40.67963
Feeling.thermometer..Biden	1.0302577
Feeling.thermometer..Ryan	-2.470447
Obama..like.dislike.scale	106.3254
Romney..like.dislike.scale	-72.21049
Presidential.job.approval	-38.51804
obama.handling.sum	-2.141782
Financial.situation.past.year	0
Economy.next.year	0
Unemployment.past.year	0
Economic.blame..Obama	0
Ideology	0
Obama..Ideological.placement	0
reduce.deficit.sum	0
Obama..Health.insurance.plan.scale	0
Support.for.Obamacare	-1.786785
Offshore.drilling	1.7338066

Which is better?

- Stepwise or BIC version of Lasso
  - What do you mean by better?
    - If talking squared error, then LS fit will look better
    - Not so clear about which is the better classifier
- Comparison
  - Exclude random subset of 1,000 cases
    - Exclude more to test than to fit (ought to repeat several times)
    - Need enough to be able to judge how well models do
  - Repeat procedure
    - Select model using stepwise and lasso
    - Calibrate (need formula for that spline)
    - Save predictions
    - Fit logistic using same predictors
  - Apply both models to the held-back data

Level	Count	Prob
Test	1000	0.23878
Train	3188	0.76122
Total	4188	1.00000

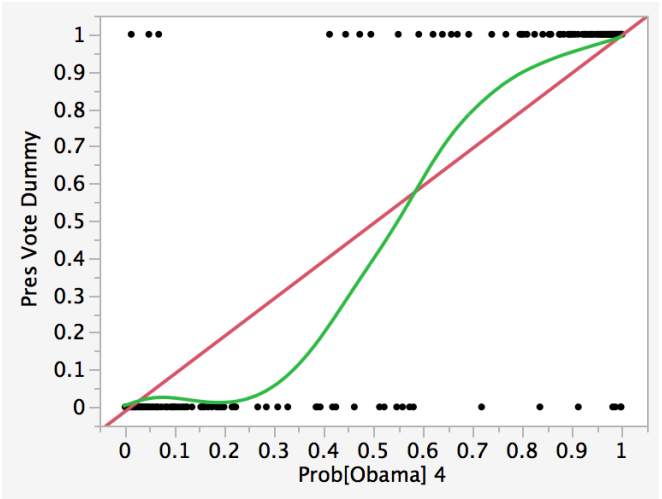
Easier to do in R than in JMP, unless you learn to program JMP (it has a language too)

Results of Comparison

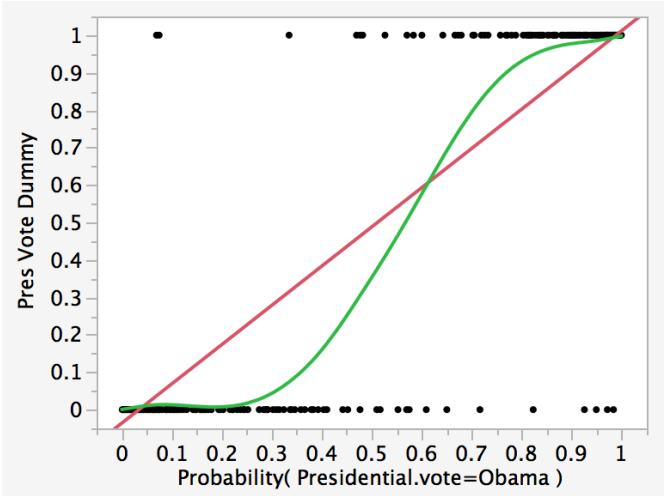
- Repeat procedure
  - Stepwise with region and gender interactions
  - Lasso fit over same variables
- Calibration plots, test samples
  - Both appear slightly uncalibrated

same errors?
brush plots

logit



Smoothing Spline Fit, lam	
R-Square	0.94160
Sum of Squares Error	14.3053



lasso

Smoothing Spline Fit, lamb	
R-Square	0.943333
Sum of Squares Error	13.88118

Results of Comparison

- Cross-validation of confusion matrix
  - Sensitivity and specificity
  - Very, very similar fits, with no sign of overfitting

Train

Most Likely Presidential.vote Logit

Presidential.vote	Obama	Romney	
Obama	1889 98.13	36 1.87	1925
Romney	41 3.25	1222 96.75	1263
	1930	1258	3188

Most Likely Presidential.vote Lasso

Presidential.vote	Obama	Romney	
Obama	1891 98.23	34 1.77	1925
Romney	46 3.64	1217 96.36	1263
	1937	1251	3188

Test

Most Likely Presidential.vote Logit

Presidential.vote	Obama	Romney	
Obama	564 98.77	7 1.23	571
Romney	13 3.03	416 96.97	429
	577	423	1000

Most Likely Presidential.vote Lasso

Presidential.vote	Obama	Romney	
Obama	565 98.95	6 1.05	571
Romney	14 3.26	415 96.74	429
	579	421	1000

Logit + Stepwise

Lasso + BIC

Take-Aways

- Logistic regression
  - Model gives probablities of group membership
  - Iterative (slower) fitting process
  - Borrow tools from OLS to get faster selection

    Not ideal, but workable
- Goodness of fit
  - Confusion matrix, sensitivity, specificity

    Need to pick the decision rule, threshold ξ
  - ROC curve

    Do you care about all of the decision boundaries?
- Comparison using cross-validation
  - Painful to hold back enough for a test
  - Need to repeat to avoid variation of C-V

    Easier with command-line software like R.

Some questions to ponder...

- What does it mean for a logistic regression to be uncalibrated?
  - Hint: Most often a logistic regression lacks calibration at the left/right boundaries.
- How is it possible for a calibrated linear regression to have smaller squared error but worse classification results?
- Might other interactions might improve either regression model?
- What happens if we apply sampling weights?

Next Time

- Enjoy Ann Arbor area
  - Canoeing on the Huron
    Whitmore Lake to Delhi
  - Detroit Institute of Art

- Tuesday
  - No more equations!
  - Neural networks combine several logistic regr
  - Ensemble methods, boosting