

Lecture 7. Logistic Models, Neural Networks

Questions from Previous Classes?

Web site for evolving software

- <http://gosset.wharton.upenn.edu/~foster/auction.html>

Topics

Logistic regression

Neural networks

Research applications

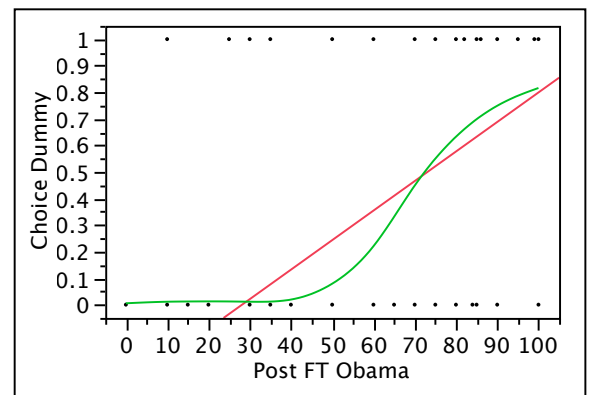
Logistic Regression

Regression with 0/1 response

- Interpret $E(Y|X)$ as $P(Y|X)$
- If calibrated, linear regression generates a very good estimate of a probability.

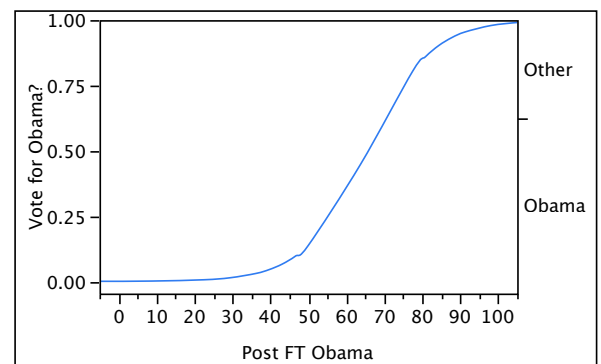
ANES example

- Limit data to those who actually voted in November presidential election (Exclusion)
- Response coded 1 if person voted for Obama and is 0 otherwise.
- Explanatory variable is post-election FT for Obama
- OLS fit is not so sensible as an estimate for the probability of voting for Obama. Not calibrated
- Spline shows evident lack of calibration, particularly at boundaries.



Logistic regression

- Fits sigmoidal curve that estimates the probability of an event.
- Response is categorical, limited to 2 choices: Obama/Other.
- Equation for curve is $\log \text{odds (Obama)} = -0.31 + 0.011 \text{ FT Obama}$
- Interpretation of intercept, slope



Logistic regression model

- Model for the probability of one category versus another (2 groups)
- Probability follows the S-shaped logistic curve $f(z) = 1/(1+e^{-z})$

$$P(Y|X) = 1/(1 + \exp(-a - bX))$$

Regression equation embedded in a transformation that keeps the probabilities in the range 0 to 1.

Slope captures multiplicative effect of an increase in the explanatory variable

- Latent variable interpretation
- Multiple logistic regression just adds more variables to the equation.

Inference in logistic regression

- Software finds the estimates that maximize the likelihood implied by the model, then computes a more elaborate standard error. $(L(y_1, \dots, y_n) = p^y (1-p)^{n-y})$
- Ratio of estimate to standard error is approximately normal(0,1) under usual assumptions of independent observations and correctly specified model.
- JMP shows the square of the t-statistics, and calls them chi-square. Just look at the p-values as usual. (-2 Log likelihood in OLS is the residual SS)

Example

- Similar hypotheses as used in linear regression.
- Just different test statistics.

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	252.1109	10	504.2217	<.0001*
Full	769.2543			
Reduced	1021.3651			
RSquare (U)		0.2468		
Observations (or Sum Wgts)		1568		
Converged by Objective				

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.019	0.242	69.83	<.0001*
Region[1. Northeast]	0.379	0.158	5.76	0.0164*
Region[2. North Central]	-0.014	0.129	0.01	0.9162
Region[3. South]	-0.539	0.101	28.61	<.0001*
R Age	-0.013	0.004	13.47	0.0002*
R Race[10. Black]	2.535	0.298	72.37	<.0001*
R Race[20. Asian]	-0.872	0.421	4.28	0.0386*
R Race[30. Native American]	-0.901	0.439	4.21	0.0401*
R Race[40. Hispanic or Latino]	-0.098	0.201	0.24	0.6270
R Race[50. White]	-1.499	0.169	78.98	<.0001*
R Gender[1. Male respondent selected]	-0.177	0.063	7.86	0.0051*

For log odds of Obama/Other

Stepwise

- JMP includes stepwise option for logistic, but don't try too many variables!

Neural Network

Structure

- Network of inputs (explanatory variables), hidden nodes (logistic regressions), and outputs (the response)

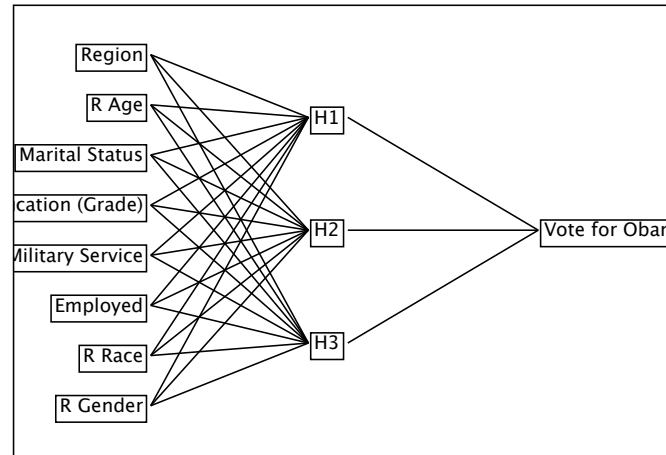
Analogy to neurons in brain tissue

- Hidden nodes combine input variables in a logistic regression

Key parameter to choose is the number of hidden nodes

Weights chosen to represent different combinations of the inputs

- Projection pursuit regression



Theoretical appeal

- Imagine splicing together lots of logistic curves, some positive, some negative.
- Can represent any function (if you have enough data to identify it)

Optimization

- Lots and lots of parameters, but nothing like standard errors and p-values
- Searching of the mountain top in a dense forest (Tours controls number of searches)
- Lack of p-values combined with many parameters leads to potential over-fitting and unknown (or suspect) accuracy.

Example

- Vote for Obama in 2008 election using anes_2008.jmp (JMP data with missing)
- JMP cross-validation option is essential for tuning the over-fit penalty so that does not grossly over-fit the observed sample. These use 50% of the data for cross-validation.
- Not much fit, not much over-fitting either

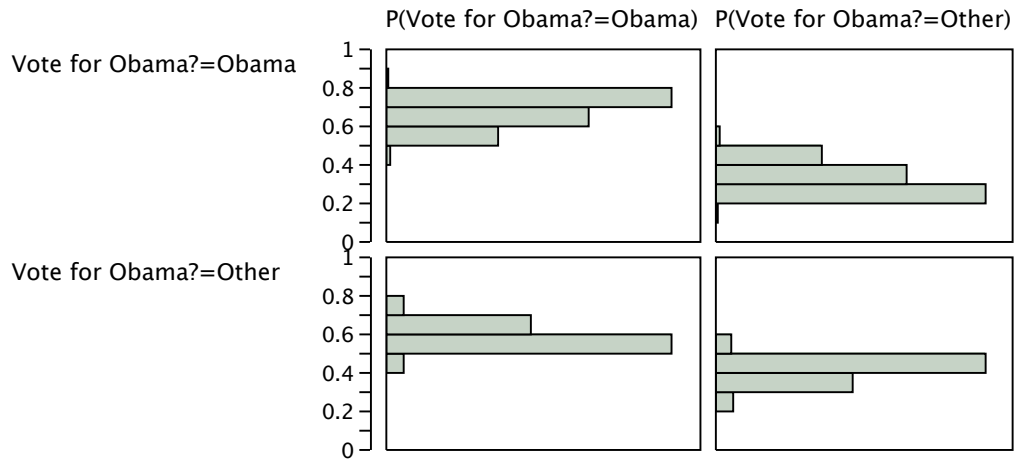
Fit History

Nodes	Penalty	RSquare	CV RSquare	.2 .4 .6 .8
3	1	0.34178	0.18106	
3	1.5	0.31487	0.18759	
3	3	0.27895	0.18370	
3	6	0.22578	0.15981	
3	12	0.12103	0.09078	

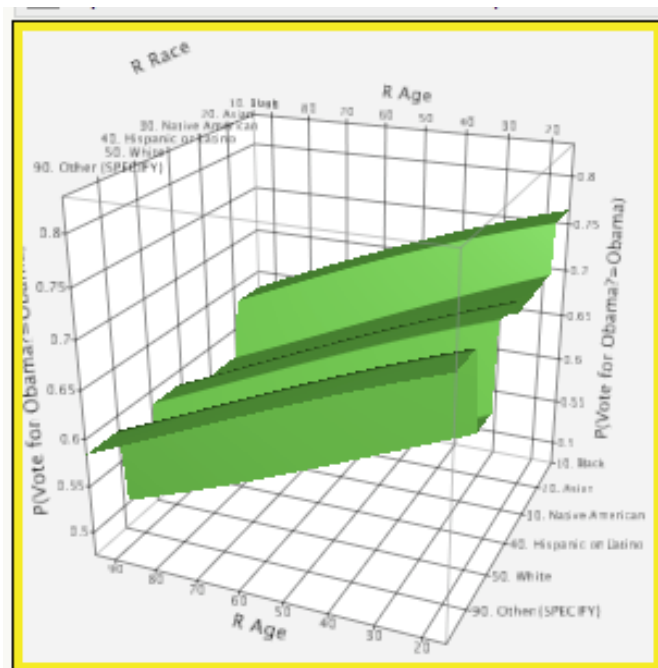
Eye candy

- Wish some of these tools were provided for regression

Predicted Value Histograms



- Surface view is particularly useful to understand the fit of the model. Shows
 - Interactions (saddle shapes as seen in regression)
 - Nonlinearity of some variables.

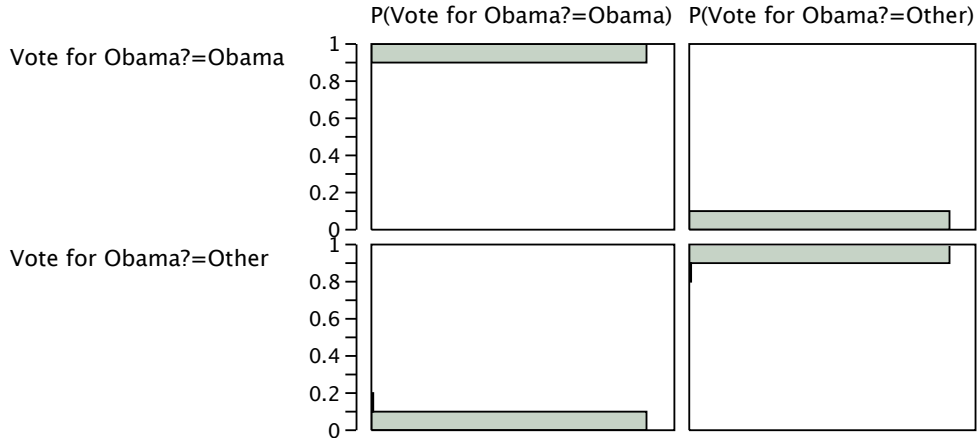


Bigger model

- Reach deeper into the FT data (through FT Atheists)
- Over-fitting becomes a real possibility

Fit History				
Nodes	Penalty	RSquare	CV RSquare	.2 .4 .6 .8
3	0.01	0.99683	0.22880	

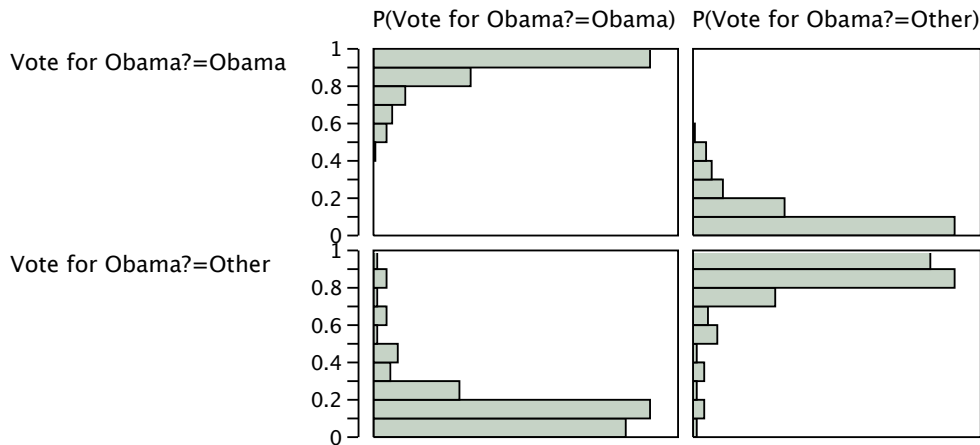
Predicted Value Histograms



- Increase the validation penalty to bring these into alignment

Fit History				
Nodes	Penalty	RSquare	CV RSquare	.2 .4 .6 .8
3	0.01	0.99683	0.22880	
3	1	0.87084	0.67660	
3	4	0.73574	0.65571	

Predicted Value Histograms



- Surface fit of this model is particularly interesting, but can be hard to figure out which to look at, so use the “Profiler” option to identify variables that impact the response.

FT Israel is particularly noticeable in the neural network fit.

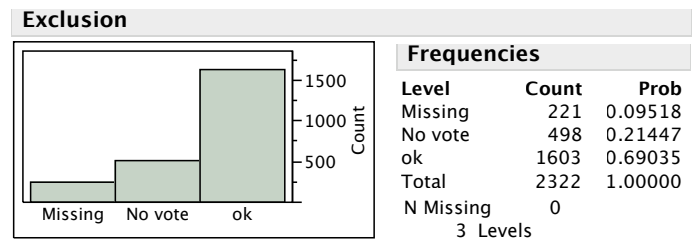
Role for neural network

- I personally find them interesting, particularly as implemented with visual aids in JMP. They are passing these days in computer science and machine learning (too hard to fit, too hard to validate, not much opportunities to improve them)
- Explore data for nonlinearity, missed features

Research Applications

Prediction

- How would those who didn't vote have gone in the 2008 election?
- The choice of predictions might be interesting, but the outcome even more so.



Data mining as a diagnostic

- Take your best shot, save the predictions
- Use your predictions as a single explanatory variable
- Does data mining find anything else? Start off a stepwise model with your predictions as the one explanatory variable in the initial model.

If your predictions are “sufficient”, then data mining won't be able to improve on them when predicting the response. They will have everything that's worth using.

If DM finds more, then you need to consider whether it's enough to be excited about and perhaps what it might mean.

Questions for Thought

Neural net vs logistic regression

- Does a neural net generate a better fit, one with more accurate predictions out of sample, than logistic regression? How would you decide?

Over-fitting

- Neural nets need cross-validation to control the fitting procedure and avoid over-fitting. Does logistic regression also need cross-validation?

Special case

- Suppose you fit a neural net with one hidden node. How does the fit of the neural network differ from the fit of a logistic regression built with the same variables?