Data Science Math – Final

Andrew Goldberg

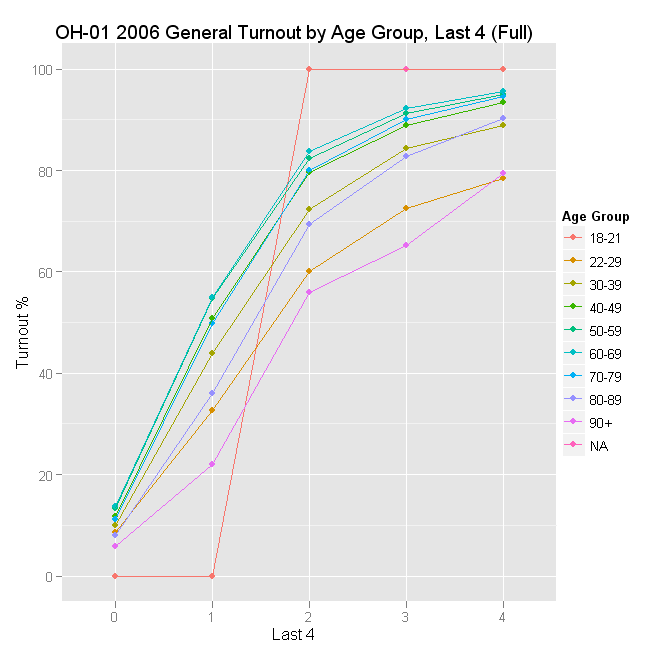
07/26/15

Logistic Regression for political turnout targeting

Large scale political campaigns typically cost tens of millions of dollars or more, yet are decided by very small margins. In 2012 Obama beat Romney by 3 percent (51% to 48%) and both campaigns spent roughly 1 billion dollars each. These circumstances make every vote matter, and political targeters are now hired by campaigns to, among many things, find voters who can be persuaded to vote for their candidate or need additional motivation to turn out to vote.

Voter persuasion is a very complex process and is difficult to measure because there is no surefire way of knowing who a voter ultimately supported. Voter turn-out is easier to study because states provide voterfiles detailing in which elections voters cast a ballot, giving the targeter clear observations that can be used to model future turnout. For instance, a targeter can look at the general demographics and vote history of those who voted in the 2012 presidential election as a model of those who will vote in the upcoming 2016 presidential election.

Commonly accessible free data (offensivepolitics.net):

* **Gender**: Using birth data from the Social Security Administration, matching each voter’s first name to a probable gender. About 9% of names were unable to be matched and coded as an empty string.
* **Age Group (2010)**: Using the birth year, calculating age as of 2010, and then assigning each voter to an age group: 18-21,22-29,30-39, 40-49, 50-59, 60-69, 70-79, 80-89, and 90+.
* **Age Group (2006)**: Using the birth year, calculating age as of 2006, and then assigned each voter to an age group: 18-21,22-29,30-39, 40-49, 50-59, 60-69, 70-79, 80-89, and 90+.
* **Household**: voters grouped discrete households using the full street address and zip code.
* **Marriage status**: Using the household variable, performing a very simple marriage determination: people living in the same household with a difference in age < 15 years were flagged as married.
* **Last4 (2006)**: Measures participation in the last 4 major elections prior to 2006: 2004 Primary and General, and 2002 Primary and General. Range 0-4.
* **Last4 (2008)**: Measures participation in the last 4 major elections prior to 2008: 2006 Primary and General, and 2004 Primary and General. Range 0-4.
* **Last4 (2010)**: Measures participation in the last 4 major elections prior to 2010: 2008 Primary and General, 2006 Primary and General. Range 0-4.

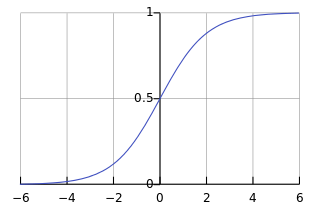
Age and past voting history are typically the strongest predictors of future turnout.

Linear regression works for many modeling tasks in politics, but because turnout has only two results (0 = didn’t vote, 1 = voted), logistic regression is typically used. Logistic regression uses variables like past voting behavior, age, gender and party to model the probability (0 to 1%) that a voter will turn out to vote. Campaigns are only interested in spending resources on voters who they motivate to get out to the polling booth, so voters with a score around .5 are targeted, while unlikely voters or very likely voters are typically ignored.

Logistic regression function. t is the explanatory variable

\sigma (t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}},

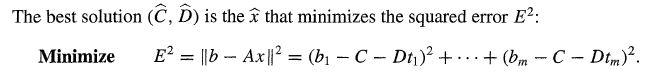
Logistic regression curve from Wikipedia

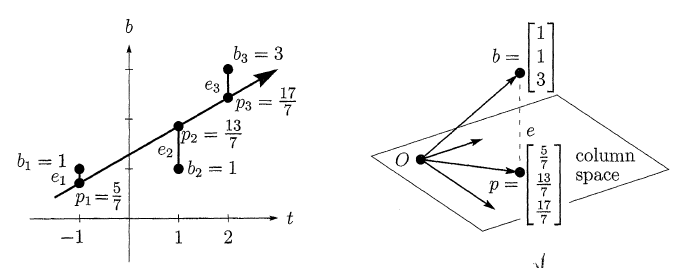


Targeters often use the ordinary least squares method to fit a logistic regression model, which is similar to many linear regression models. Ordinary least squares uses linear algebra to plot variables on a multi-dimensional matrix and finds the best line that fits the data. As data points are scattered around the line, the best fitting line minimizes the squared error between each data point and the line.

We want to find the line that minimizes the squared vertical distances (e in the following graph) between each data point and the model line, C + Dt = b.

From Professor Strang’s linear algebra text:





R Code using the glm package to fit the data (offensivepolitics.net)

# create temporary variables inside the data frame for 2006 values

vfs$last4 <- vfs$last4.g2006

vfs$age <- vfs$age.2006

# create a model for voters with at least 4 years of voting history

full.lr <- glm(turnout.g06 ~ last4 + age + gender+party+married,data=vfs[ele.full,],family=binomial)

# run ANOVA against the full table to test for term significance

anova(full.lr,test="Chisq")

# create a model for voters with less than 4 years of voting history

partial.lr <- glm(turnout.g06 ~ last4 + age + gender+party+married,data=vfs[ele.partial,],family=binomial)

# run ANOVA against the partial table to test for term significance

anova(partial.lr,test="Chisq")

Once the models are fit, targeters will run the voterfile through the model to predict and score each voter’s likelihood to turn out. The scores of turnout probability can be compared to actual past turnout to determine the effectiveness of the model.

Below is a chart of projected 2010 turnout in Ohio’s first congressional district. Many explained that the higher turnout in 2010 was due to the Obama campaign’s continual organizing activity and voter registration, especially among younger voters.

