Problem Set 5: Predicting Elections, Reprise

Due 4/23

Submit a short writeup of your answers as well as your R code (in a separate file) on Blackboard.

In Problem Set 3, we studied the prediction of election outcomes based on betting markets and state-level polls. Here, we do so again, but focusing on how well both precit the eventual outcome over the course of the campaign. We again analyze data for the 2008 US presidential elections from the online betting company, called Intrade. At Intrade, people trade contracts such as 'Obama to win the electoral votes of Florida.' Each contract's market price fluctuates based on its sales. The trading price data file is available in CSV format as `intrade08.csv`. The variables in this datasets are:

Name	Description
`day`	Date of the session
`statename`	Full name of each state (including District of Columbia)
`state`	Abbreviation of each state (including District of Columbia)
`PriceD`	Closing price (predicted vote share) of Democratic Nominee's market
`PriceR`	Closing price (predicted vote share) of Republican Nominee's market
`VolumeD`	Total session trades of Democratic Party Nominee's market
`VolumeR` m	Total session trades of Republican Party Nominee's arket

Each row represents daily trading information about the contracts for either the Democratic or Republican Party nominee's victory in a particular state. We will also use the election outcome data `pres08.csv` with the following variables:

Name	Description
`state.name`	Full name of state
`state`	Two letter state abbreviation
`Obama`	Vote percentage for Obama
`McCain`	Vote percentage for McCain
`EV`	Number of electoral college votes for this state

Finally, we'll use poll data in the file `polls08_ps5.csv`. This dataset tracks the latest poll numbers available in every state over the course of the election campaign. If polls were conducted in a given state on a given day, it provides the predicted support for Obama and McCain from those polls. If no poll was conducted in a state on a given day, the data provide the last available polling numbers for the two candidates. The variables are:

Name	Description
`state`	Abbreviated name of state in which poll was conducted
`middate`	Middle of the period when poll was conducted
`pollsD`	Predicted support for Obama (percentage)
`pollsR`	Predicted support for McCain (percentage)

Question 1

What is the predicted number of electoral college votes for Obama according to the *prediction markets* on each day in the 120 days before the election? To find out, first take the `intrade08` data and merge the `pres08` data into it. This will, among other things, provide you with information on the number of electoral votes that each state has. Also create a variable that indicates how many days before the election a given date was. Note that in 2008, the election took place on November 4.

Then, on a given day, Obama is predicted to receive all electoral college votes for a state if his closing price (predicted vote share) is higher than McCain's in that state. If his closing price is lower than McCain's, he is predicted to receive zero electoral college votes in that state. For each day from 120 days to one day before the election, compute the number of electoral college votes Obama is predicted to receive (the best way to do so is to use a loop). Plot the results, with time until the election on the horizontal axis and predicted number of electoral college votes on the vertical axis. Make sure to label the axes. Also plot a dashed horizontal line that indicates the number of electoral college votes Obama actually received (365). Briefly summarize the graph.

Question 2

What is the predicted number of electoral college votes for Obama according to the *state-level polls* on each day in the 120 days before the election? Follow the same steps as in Question 1, but using the `polls08_ps5` data instead. Provide a plot that shows both the predicted number of electoral college votes according to the opinion polls, as well as the predicted number according to the betting markets (from Question 1). Also plot a dashed horizontal line that indicates the number of electoral college votes Obama actually received (365). Briefly summarize the graph.

Question 3

Compute the prediction error for both indicators by subtracting the number of predicted electoral college votes from 365, the actual number of votes Obama received (note that you will *not* need a loop to do this). For each indicator, you should have a vector of length 120 that tells you how far off its prediction was from the eventual result on every day in the 120 days before the election. What is the average prediction error of each method? What is the average prediction error of each method in the last 30 days before the election? Summarize which method does better at predicting the eventual electoral college vote.