Problem Set 3: Predicting Elections

Due 3/26, adapted from QSS 4.5.1

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

In this exercise, we study the prediction of election outcomes based on betting markets. In particular, we 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. Why might we expect betting markets like Intrade to accurately predict the outcomes of elections or of other events? Some argue that the market can aggregate available information efficiently. In this exercise, we will test this claim by analyzing the market prices of contracts for Democratic and Republican nominees' victories in each state. We will then compare how well betting markets predict election outcomes to how well opinion polls do.

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 in 2008)
`state`	Abbreviation of each state (including District of Columbia in 2008)
`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 (only in `pres2008`)
`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.csv` with variables:

Name	Description
`state`	Abbreviated name of state in which poll was conducted
`pollD`	Predicted support for Obama (percentage)
`pollR`	Predicted support for McCain (percentage)

The poll data for each state represents an average of polls from the last few weeks before the election.

Question 1

We will begin by using the market prices on the day before the election to predict the 2008 election outcome. First, subset the market price data such that it contains the market information for each state and candidate only on the day before the election. Note that in 2008 the election day was November 4. Then, merge the market information into the dataset with the 2008 election outcomes. Create a variable that is the difference in the closing price between the Democratic nominee (Obama) and the Republican nominee (McCain), as well as a variable that provides the vote margin between Obama and McCain. Provide a plot that shows the relation between the difference in closing price and the vote margin (be sure to label the axes). Also compute the correlation between the two variables.

Question 2

Estimate a linear regression where the dependent variable is the vote share margin and the independent variable is the difference in the closing prices. Interpret the intercept as well as the coefficient in substantive terms (that is, a sentence that someone who is not familiar with linear regression can understand). Can we reject the null hypothesis that there is no relationship between market price difference and vote share margin? What is the 95 percent confidence interval for the coefficient? What is the R^2 of this regression, and what does it mean? Then, plot the regression line along with the scatterplot of the two variables. Finally, what are the predicted vote share margins for each state according to the regression equation? In which state did Obama over-perform the predicted value the most? And in which did he under-perform the most?

Question 3

We will now examine how well opinion polls predicts vote outcomes. Repeat everything you did for Questions 1 and 2, but using the polling data as the independent variable.

Question 4

Based on the analyses you've done, were betting markets or opinion polls better at predicting the election outcome? Why?