A Probability Model for Betting Market Election Prediction

Earlier in this chapter, we used pre-election polls with a probability model to predict Obama's electoral vote share in the 2008 US election. In this exercise, we will apply a similar procedure to the Intrade betting market data analyzed in an exercise in Chapter 4 (see Section 4.6.1). This exercise is based on David Rothschild (2009). "Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases." *Public Opinion Quarterly* vol. 73, no. 5, pp. 895–916. The 2008 Intrade data is available as `intrade08.csv`. The variable names and descriptions of this data set are available in table 4.9. Recall that each row of the data set represents daily trading information about the contracts for either the Democratic or Republican Party nominee's victory in a particular state. The 2008 election results data are available as `pres08.csv`, whose variable names and descriptions appear in table 4.1.

Question 1

We analyze the contract of the Democratic Party nominee winning a given state j. Recall from Section 4.5 that the data set contains the contract price of the market for each state on each day i leading up to the election. We will interpret the `PriceD` as the probability p_{ij} that the Democrat would win state j if the election were held on day i. To treat `PriceD` as a probability, divide it by 100 so it ranges from 0 to 1. How accurate is this probability? Using only the data from the day before Election Day (November 4, 2008) within each state, compute the expected number of electoral votes Obama is predicted to win and compare it with the actual number of electoral votes Obama won. Briefly interpret the result. Recall that the actual total number of electoral votes for Obama is 365, not 364, which is the sum of electoral votes for Obama based on the results data. The 365-total includes a single electoral vote that Obama garnered from Nebraska's 2nd Congressional District. McCain won Nebraska's four other electoral votes because he won the state overall.

Question 2

Next, using the same set of probabilities used in the previous question, simulate the total number of electoral votes Obama is predicted to win. Assume that the election in each state is a Bernoulli trial where the probability of success (Obama winning) is p_{ij}. Display the results using a histogram. Add the actual number of electoral votes Obama won as a solid line. Briefly interpret the result.

Question 3

In prediction markets, people tend to exaggerate the likelihood that the trailing or "long shot" candidate will win. This means that candidates with a low (high) p_{ij} have a true probability that is lower (higher) than their predicted p_{ij}. Such a discrepancy could introduce bias into our predictions, so we want to adjust our probabilities to account for it. We do so by reducing the probability for candidates who have a less than 0.5 chance of winning, and increasing the probability for those with a greater than 0.5 chance. We will calculate a new probability p^*_{ij} using the following formula proposed by a researcher: $p^*_{ij} = \Phi(1.64 \times \Phi^{-1}(p_{ij}))$ where $\Phi(\cdot)$ is the CDF of the standard Normal random variable and $\Phi^{-1}(\cdot)$ is its inverse, the quantile function. The `R` functions `pnorm` and `qnorm` can be used to compute $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$, respectively. Plot p_{ij}, used in the previous questions, against p^*_{ij}. In addition, plot this function itself as a line. Explain the nature of the transformation.

Question 4

Using the new probabilities p^*_{ij}, repeat Questions 1 and 2. Do the new probabilities improve predictive performance? **Hint**: you can compute root mean squared error to help determine performance.

Question 5

Compute the expected number of Obama's electoral votes using the new probabilities p_{ij}^* for each of the last 120 days of the campaign. Display the result as a time series plot. Briefly interpret the plot.

Question 6

For each of the last 120 days of the campaign, conduct a simulation as done in Question 2 using the new probabilities p_{ij}^*. Compute the quantiles of Obama's electoral votes at 2.5% and 97.5% for each day. Represent the range from 2.5% to 97.5% for each day as a vertical line, using a loop. Also, add the estimated total number of Obama's electoral votes across simulations. Briefly interpret the result.