Predicting Elections with Linear Models

Earlier in the chapter, we studied the prediction of election outcomes using polls. Here, we study the prediction of election outcomes based on betting markets. In particular, we analyze data for the 2008 and 2012 US presidential elections from the online betting company, 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 *efficient market hypothesis* by analyzing the market prices of contracts for Democratic and Republican nominees’ victories in each state.

The data files for 2008 is available in CSV format as intrade08.csv. The variables in these 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 | Total session trades of Republican Party Nominee’s market |

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. This data file is pres08.csv with 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 |

We’ll also use poll data from 2008 in the file polls08.csv, The variables in the polling data are:

|  |  |
| --- | --- |
| Name | Description |
| state | Abbreviated name of state in which poll was conducted |
| Obama | Predicted support for Obama (percentage) |
| Romney | Predicted support for Romney (percentage) |
| Pollster | Name of organization conducting poll |
| middate | Middle of the period when poll was conducted |

What is the relationship between the price margins of the Intrade market and the actual margin of victory? Using only the market data from the day before the election in 2008, regress Obama’s actual margin of victory in each state on Obama’s price margin from the Intrade markets. We’ll start by constructing Obama’s Intrade margin from 2008

## Question 1

Load the Intrade and Election outcome data for 2008 and merge them together.

Create a DaysToElection variable by subtracting the day of the election from each day in the dataset. Now create a state margin of victory variable to predict, and a betting market margin to predict it with.

## Answer 1

intrade08 <- read.csv("intrade08.csv")  
pres08 <- read.csv("pres08.csv")  
polls08 <- read.csv("polls08.csv")  
intresults08 <- merge(intrade08, pres08, by = "state")  
head(intresults08)

## state day statename PriceD VolumeD PriceR VolumeR state.name Obama  
## 1 AK 2008-02-28 Alaska 7.5 0 92.5 0 Alaska 38  
## 2 AK 2008-03-21 Alaska 7.5 0 92.5 0 Alaska 38  
## 3 AK 2007-05-04 Alaska 10.0 0 90.0 0 Alaska 38  
## 4 AK 2008-03-29 Alaska 8.0 0 92.0 0 Alaska 38  
## 5 AK 2008-03-22 Alaska 7.5 0 92.5 0 Alaska 38  
## 6 AK 2008-02-20 Alaska 7.5 0 92.5 0 Alaska 38  
## McCain EV  
## 1 59 3  
## 2 59 3  
## 3 59 3  
## 4 59 3  
## 5 59 3  
## 6 59 3

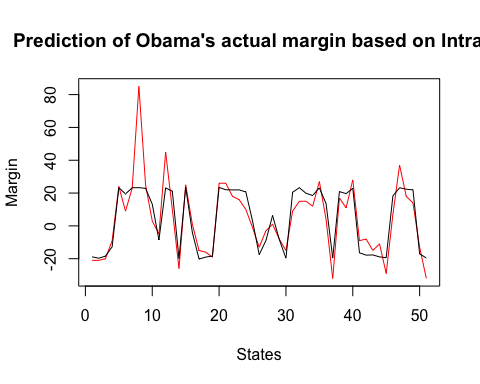
intresults08$DaysToElection <- as.Date("2008-11-04") - as.Date(intresults08$day)  
intresults08$obama.intmarg <- intresults08$PriceD -intresults08$PriceR  
intresults08$obama.actmarg <- intresults08$Obama - intresults08$McCain

## Question 2

Considering only the trading one day from the election, predict the actual electoral margins from the trading margins using a linear model. Does it predict well? How would you visualize the predictions and the outcomes together? Hint: because we only have one predictor you can use abline.

## Answer 2

latest08 <- intresults08[intresults08$DaysToElection == 1, ]  
int.fit08 <- lm(obama.actmarg ~ obama.intmarg, data = latest08)  
latest08$obama.fitmarg <- int.fit08$fitted.values  
plot(latest08$obama.actmarg,type = "l", col = "red",   
 xlab = "States",   
 ylab = "Margin",   
 main = "Prediction of Obama's actual margin based on Intrade")  
lines(latest08$obama.fitmarg)



## Question 3

Even efficient markets aren’t omniscient. Information comes in about the election every day and the market prices should reflect any change in information that seem to matter to the outcome.

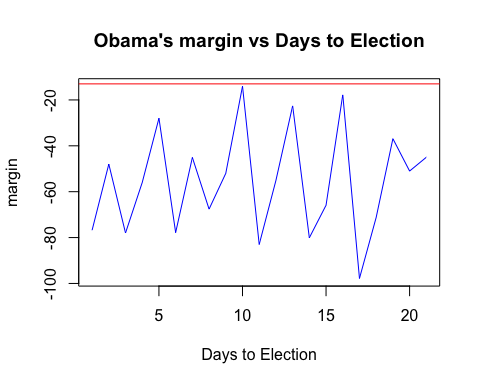
We can examine how and about what the markets change their minds by looking at which states they are confident about, and which they update their ‘opinions’ (i.e. their prices) about. Over the period before the election, let’s see how prices for each state are evolving. We can get a compact summary of price movement by fitting a linear model to Obama’s margin for each state over the 20 days before the election.

We will summarise price movement by the direction (up or down) and rate of change (large or small) of price over time. This is basically also what people in finance do, but they get paid more…

Start by plotting Obama’s margin in West Virginia against the number of days until the election and modeling the relationship with a linear model. [Hint: write your code using index notation. You then can copy it and paste it for the loop in the next question]. Use the last 20 days. Show the model’s predictions on each day and the data. What does this model’s slope coefficient tells us about which direction the margin is changing and also how fast it is changing? Hint: West Virginia is 50th [i<-50] on the alphabetical list of state names.

## Answer 3

state\_names <- sort(unique(latest08$state))  
last20<- intresults08[intresults08$DaysToElection <= 20, ]  
last20 <- subset(last20, subset = (state == state\_names[50]))  
plot(last20$obama.intmarg, type = "l", col = "blue",   
 main = "Obama's margin vs Days to Election",   
 xlab= "Days to Election",   
 ylab = "margin")  
abline(h= last20$obama.actmarg[1], col = "red")



reg.obama <- lm(data = last20, obama.actmarg ~ DaysToElection)  
summary(reg.obama)

## Warning in summary.lm(reg.obama): essentially perfect fit: summary may be  
## unreliable

##   
## Call:  
## lm(formula = obama.actmarg ~ DaysToElection, data = last20)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.022e-15 -6.414e-16 -3.242e-16 -7.000e-18 7.372e-15   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.300e+01 7.487e-16 -1.736e+16 <2e-16 \*\*\*  
## DaysToElection -6.343e-17 6.405e-17 -9.900e-01 0.334   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.777e-15 on 19 degrees of freedom  
## Multiple R-squared: 0.4998, Adjusted R-squared: 0.4734   
## F-statistic: 18.98 on 1 and 19 DF, p-value: 0.0003395

The direction is negative since slope is -6.343e-17. Hence, the closer to the election we were in West Virginia the less was Obama’s predicted slope. However, the rate of change is extremely small, very close to zero. So, we can disregard the slope in West Virginia since it is small and statistically insignificant.

## Question 4

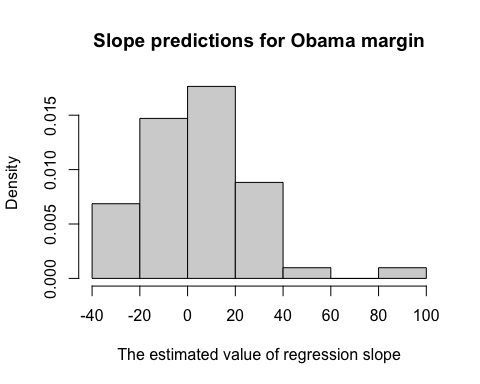
Now do it for all states and collect the slop coefficients to see [historgram] how volative the state estimates are.

## Answer 4

slopes.obama <- rep(NA, 51)  
for (i in 1:51) {  
 state\_names <- sort(unique(latest08$state))  
 last20days<-intresults08[intresults08$DaysToElection <= 20,]  
 last20days <- subset(last20days, subset = (state == state\_names[i]))  
 my.reg <- lm(obama.actmarg ~ DaysToElection, data = last20days)  
 slopes.obama[i] <- coef(my.reg)  
}

## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length  
  
## Warning in slopes.obama[i] <- coef(my.reg): number of items to replace is not a  
## multiple of replacement length

names(slopes.obama) <- state\_names  
hist(slopes.obama, freq = FALSE,  
 main = "Slope predictions for Obama margin",  
 xlab = "The estimated value of regression slope")



summary(slopes.obama)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -32.000 -13.000 4.000 4.314 17.500 85.000

The histogram shows that slopes are not that volatile for Obama’s margin, however there are some outliers - the very end of the left tail. The rest of the states are close to the mean 4.314

## Question 5

Now predict the winner of the election *one week before the election* using the Intrade data. To do so, first use the *two weeks before* that moment to fit state level linear models, then use those models to predict what will happen in each state. How well does the model do predicting the election outcome? [present your results vs. actual margins in a confusion table to show your classification results - p. 136 and command table()]

## Answer 5

intresults2week<- subset(intresults08, subset = (DaysToElection <= 21) & (DaysToElection > 7))  
obama7days <- rep(NA, 51)  
for( j in 1:51){  
   
}