Prediction Based on Betting Markets

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, 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 *efficient market hypothesis* by analyzing the market prices of contracts for Democratic and Republican nominees’ victories in each state.

The data files for 2008 and 2012 are available in CSV format as intrade08.csv and intrade12.csv, respectively. 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 Nominee’s market | Closing price (predicted vote share) of Democratic |
| PriceR Nominee’s market | Closing price (predicted vote share) of Republican |
| VolumeD market | Total session trades of Democratic Party Nominee’s |
| VolumeR market | Total session trades of Republican Party Nominee’s |

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. These data files are pres08.csv and pres12.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 and 2012 in the files polls08.csv and polls12.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 |

## Question 1

We will begin by using the market prices on the day before the election to predict the 2008 election outcome. To do this, subset the 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. We compare the closing prices for the two candidates in a given state and classify a candidate whose contract has a higher price as the predicted winner of that state. Which states were misclassified? How does this compare to the classification by polls presented earlier in this chapter? Repeat the same analysis for the 2012 election, which was held on November 6. How well did the prediction market do in 2012 compared to 2008? Note that in 2012 some less competitive states have missing data on the day before the election because there were no trades on the Republican and Democratic betting markets. Assume Intrade predictions would have been accurate for these states.

## Answer 1

#2008  
pres08<-read.csv("pres08.csv",header = T)  
intrade08<-read.csv("intrade08.csv",header = T)  
polls08<-read.csv("polls08.csv",header = T)  
  
#merge intraday and presidential data  
pred<-merge(intrade08,pres08,by="state")  
  
#Find the number of days to the election date  
pred$DaysToElection<-as.Date("2008-11-04")-as.Date(pred$day)  
  
#subset the data before the election  
data.before<-subset(pred,subset=(pred$DaysToElection==1))  
  
nframe<-data.frame(pres08$state)  
#Predicted intrade margin  
nframe$win.pred<-data.before$PriceD-data.before$PriceR  
# Actual election margin  
nframe$win.Elec<-data.before$Obama-data.before$McCain   
#Misclassified states  
nframe$pres08.state[sign(nframe$win.pred)!=sign(nframe$win.Elec)]

## [1] IA MS  
## 51 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA ... WY

#2012  
#Follow previous comment  
pres12<-read.csv("pres12.csv",header = T)  
intrade12<-read.csv("intrade12.csv",header = T)  
polls12<-read.csv("polls12.csv",header = T)  
  
pred12<-merge(intrade12,pres12,by="state")  
  
pred12$DaysToElection<-as.Date("2012-11-06")-as.Date(pred12$day)  
data.before.12<-na.omit(subset(pred12,subset=(pred12$DaysToElection==1)))#Omits states with missing data.  
  
nframe12<-data.frame(data.before.12$state)  
nframe12$win.pred<-data.before.12$PriceD-data.before.12$PriceR  
nframe12$win.Elec<-data.before.12$Obama-data.before.12$Romney   
  
nframe12$data.before.12.state[sign(nframe12$win.pred)!=sign(nframe12$win.Elec)]

## [1] FL  
## 50 Levels: AK AL AR AZ CA CO CT DE FL GA HI IA ID IL IN KS KY LA MA ... WY

## Question 2

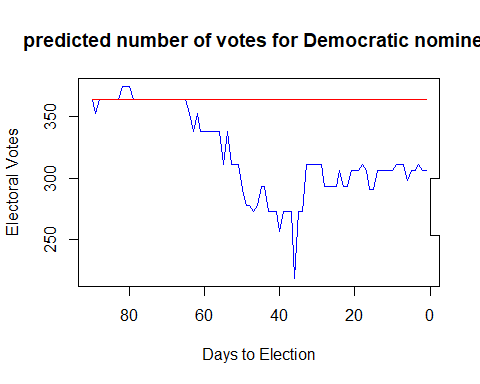
How do the predictions based on the betting markets change over time? Implement the same classification procedure as above on each of the last 90 days of the 2008 campaign rather than just the day before the election. Plot the predicted number of electoral votes for the Democratic party nominee over this 90-day period. The resulting plot should also indicate the actual election result. Note that in 2008, Obama won 365 electoral votes. Briefly comment on the plot.

## Answer 2

Days<-90:1  
sum.obama<-actual<-rep(NA,90)#Set up electoral votes for the predicted and actual for states Obama win.  
for(i in 90:1)  
{  
 data.before<-subset(pred,subset=(pred$DaysToElection==i))#Subset 2008 data on the ith day to the election.   
 data.before$obama.margin<-data.before$PriceD-data.before$PriceR#Calculates the margin  
 sum.obama[i]<-sum(data.before$EV[data.before$obama.margin>0])#Sums electoral votes for states Obama is predicted to win using postive margin.  
 actual[i]<-sum(pres08$EV[(pres08$Obama-pres08$McCain)>0])#Sums elctoral votes for obama in election  
}  
sum.obama

## [1] 364 353 364 364 364 364 364 364 375 375 375 364 364 364 364 364 364  
## [18] 364 364 364 364 364 364 364 364 364 353 338 353 338 338 338 338 338  
## [35] 338 311 338 311 311 311 291 278 278 273 278 293 293 273 273 273 256  
## [52] 273 273 273 218 273 273 311 311 311 311 311 293 293 293 293 306 293  
## [69] 293 306 306 306 311 306 291 291 306 306 306 306 306 311 311 311 298  
## [86] 306 306 311 306 306

#Plots the predicted number of votes for Democratic nominee in the last 90 days to election.  
plot(Days,sum.obama,type="l",col="blue",xlab="Days to Election",ylab="Electoral Votes",main="predicted number of votes for Democratic nominee",xlim=c(90,1))  
lines(Days,actual,col="red")  
legend(0,300,legend = c("Obama","Actual"),col=c("blue","red"),lwd = c(1,1))



## Question 3

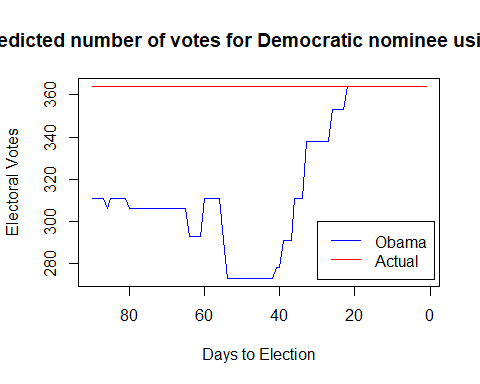
Repeat the previous exercise but this time use the seven-day *moving-average* price, instead of the daily price, for each candidate within a state. This can be done with a loop. For a given day, we take the average of the Session Close prices within the past seven days (including that day). To answer this question, we must first compute the seven-day average within each state. Next, we sum the electoral votes for the states Obama is predicted to win. Using the tapply function will allow us to efficiently compute the predicted winner for each state on a given day.

## Answer 3

#2008 intrade  
Days<-90:1  
w<-7  
sum.obama\_MA<-actual<-rep(NA,90)  
for(i in 90:1)  
{  
 data.before<-subset(pred,subset = ((pred$DaysToElection<(Days[i]+w)) & (pred$DaysToElection>=Days[i])))#subset data of 7 day moving avaerage before election   
 obama<-tapply(data.before$PriceD,data.before$state.name,mean)# calculate the moving average for Obama in each state  
 mccain<-tapply(data.before$PriceR,data.before$state.name,mean)# Calculate the moving average for McCain in each state  
 margin<-obama-mccain #Calculate the margin   
 sum.obama\_MA[i]<-sum(pres08$EV[margin>0])#Calculates the number of predicted electoral votes for Obama.  
 actual[i]<-sum(pres08$EV[(pres08$Obama-pres08$McCain)>0])#Calculates the actual electoral votes for Obama.  
}  
sum.obama\_MA

## [1] 311 311 311 311 306 311 311 311 311 311 306 306 306 306 306 306 306  
## [18] 306 306 306 306 306 306 306 306 306 293 293 293 293 311 311 311 311  
## [35] 311 291 273 273 273 273 273 273 273 273 273 273 273 273 273 278 278  
## [52] 291 291 291 311 311 311 338 338 338 338 338 338 338 353 353 353 353  
## [69] 364 364 364 364 364 364 364 364 364 364 364 364 364 364 364 364 364  
## [86] 364 364 364 364 364

plot(Days,sum.obama\_MA,type="l",col="blue",xlab="Days to Election",ylab="Electoral Votes",  
 main="predicted number of votes for Democratic nominee using MA",xlim=c(90,1))  
lines(Days,actual,col="red")  
legend(30,300,legend = c("Obama","Actual"),col=c("blue","red"),lwd = c(1,1))



## Question 4

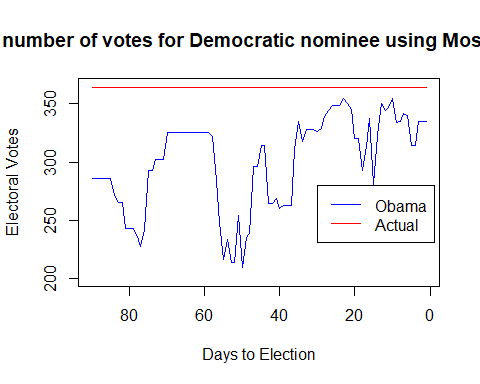
Create a similar plot for 2008 state-wide poll predictions using the data file polls08.csv. Notice that polls are not conducted daily within each state. Therefore, within a given state for each of the last 90 days of the campaign, we compute the average margin of victory from the most recent poll(s) conducted. If multiple polls occurred on the same day, average these polls. Based on the most recent predictions in each state,sum Obama’s total number of predicted electoral votes. One strategy to answer this question is to program two loops - an inner loop with 51 iterations for each state and an outer loop with 90 iterations for each day.

## Answer 4

#2008 State Polls  
  
polls<-merge(polls08,pres08,by="state")#Merge polls and election data  
polls$DaysToElection<-as.Date("2008-11-04")-as.Date(polls$middate)#Calculates the number of days to election.  
  
state.m<-unique(polls$state)#Create a vector of states.  
Days<-90:1  
polls.dat<-rep(NA,90)#Creates a vector of predicted number of electoral votes in states Obama won in the last 90 days.  
for (j in 90:1)  
{  
 margin<-rep(NA,51)  
 for (i in 1:51)   
 {  
 sub.state<-subset(polls,subset = (polls$state==state.m[i]))#subset polls data for the ith state.  
 x<-sub.state$DaysToElection #number of days to elction  
 sub.polls<-subset(sub.state,subset=(x==min(x[(x-Days[j])>=1])))#Subset data with most recent poll for the ith state  
 margin[i]<-mean(sub.polls$Obama.x)-mean(sub.polls$McCain.x) #Calculate the average margin  
 }  
 ndat<-cbind(pres08,margin)#Election data bind with average margin.   
 polls.dat[j]<-sum(ndat$EV[ndat$margin>0])# Sum the predicted number of electoral votes in the last 90 day  
   
}  
polls.dat

## [1] 286 286 286 286 286 286 271 265 265 243 243 243 236 227 242 293 293  
## [18] 302 302 302 325 325 325 325 325 325 325 325 325 325 325 325 322 284  
## [35] 248 216 233 214 214 254 209 235 239 296 296 314 314 264 264 269 260  
## [52] 263 263 263 311 335 318 328 328 328 326 329 339 344 349 349 349 355  
## [69] 350 345 320 320 293 313 337 281 323 350 344 347 355 334 335 342 340  
## [86] 314 314 335 335 335

plot(Days,polls.dat,type="l",col="blue",xlab="Days to Election",ylab="Electoral Votes",ylim=c(200,365),main="predicted number of votes for Democratic nominee using Most Recent Polls",xlim=c(90,1))  
lines(Days,actual,col="red")  
legend(30,280,legend = c("Obama","Actual"),col=c("blue","red"),lwd = c(1,1))



Q.5 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. Similarly, in a seperate analysis, regress Obama’s actual margin of victory on the Obama’s predicted margin from the latest polls within each state. Interpret the results of the regressions.

#2008 actual and intrade regression analysis  
#subset the data before the election  
data.before<-subset(pred,subset=(pred$DaysToElection==1))  
  
act.margin<-data.before$Obama-data.before$McCain #Actual margin  
intr.margin<-data.before$PriceD-data.before$PriceR #Iintrade margin  
fit.intrade<-lm(act.margin~intr.margin) #Fitted model of actual margin regressed on intrade margin  
summary(fit.intrade)

##   
## Call:  
## lm(formula = act.margin ~ intr.margin)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.455 -7.215 -0.623 4.006 61.704   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.30265 1.64616 0.791 0.433   
## intr.margin 0.22910 0.01995 11.481 1.69e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.61 on 49 degrees of freedom  
## Multiple R-squared: 0.729, Adjusted R-squared: 0.7235   
## F-statistic: 131.8 on 1 and 49 DF, p-value: 1.689e-15

#2008 actual and polls regression analysis  
#subset the data before the election  
data.before<-subset(pred,subset=(pred$DaysToElection==1))  
  
marginpolls<-c() # Creates a vector of margin using polls data before election  
for (i in 1:51)   
{  
 sub.state<-subset(polls,subset = (polls$state==state.m[i]))#Picks a state  
 x<-sub.state$DaysToElection # The number of days to election  
 sub.polls<-subset(sub.state,subset=(x==min(x[x>=1]))) # Subset of data with Latest poll in each state  
 marginpolls[i]<-mean(sub.polls$Obama.x)-mean(sub.polls$McCain.x) #Find the avaerage margin #obama.x and McCain.x are the polls prediction in the merged data  
}  
marginpolls<-round(marginpolls,1)  
act.margin<-data.before$Obama-data.before$McCain #Actual margin  
  
fit.polls<-lm(act.margin~marginpolls) # Fitted model of actual margin and polls margin  
summary(fit.polls)

##   
## Call:  
## lm(formula = act.margin ~ marginpolls)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.3842 -2.7214 0.3918 3.4218 13.0106   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.70986 0.78696 0.902 0.371   
## marginpolls 1.10855 0.04063 27.285 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.54 on 49 degrees of freedom  
## Multiple R-squared: 0.9382, Adjusted R-squared: 0.937   
## F-statistic: 744.5 on 1 and 49 DF, p-value: < 2.2e-16

# Q.6

Do the 2008 predictions of polls and intrade accurately predict each state’s 2012 elections results? Using the fitted regressions from the previous question,forecast Obama’s actual margin of victory for the 2012 election in two ways. First, use the 2012 intrade price margins from the day before the election as the predictor in each state. Recall that the 2012 intradedata do not contain market prices for all states. Ignore states without data. Second, use the 2012 poll predicted margins from the latest polls in each state as the predictor, found in “poll12.csv”

#Intrade predicting 2012 margin using 2008 margin regression analysis  
pred.2012<-data.before.12#Data of intrade before election omitting states with NA  
pred.2012$intr.margin.2012<-pred.2012$PriceD-pred.2012$PriceR#2012 intrade margin  
pred.2012$act.margin.2012<-pred.2012$Obama-pred.2012$Romney#2012 election margin  
  
#2012 predicted margin using 2008 regression  
pred.2012$pred.12.08<- 1.30265+0.22910 \*pred.2012$intr.margin.2012  
  
#Number of states not accurately predicted  
data.before.12$state[sign(pred.2012$pred.12.08)!=sign(pred.2012$act.margin.2012)]

## [1] FL  
## 50 Levels: AK AL AR AZ CA CO CT DE FL GA HI IA ID IL IN KS KY LA MA ... WY

#Polls predicting 2012 margin using 2008 margin regression analysis  
polls12<-read.csv("polls12.csv",header = T)  
polls.pred.12<-merge(polls12,pres12,by="state")#merge 2012 polls and election data  
  
polls.pred.12$DaysToElection<-as.Date("2012-11-06")-as.Date(polls.pred.12$middate)#Days to election  
  
state.12<-unique(polls.pred.12$state)#selects states which polls were conducted (there were 47 states in 2012)  
  
marginpolls12<-c() # Creates a vector of margin using latest polls data before election  
for (i in 1:length(state.12))   
{  
 sub.state<-subset(polls.pred.12,subset = (polls.pred.12$state==state.12[i]))#Picks a state  
 y<-sub.state$DaysToElection # The number of days to election  
 sub.polls.12<-subset(sub.state,subset=(y==min(y[y>=1]))) # Subset of data with Latest poll in each state  
 marginpolls12[i]<-mean(sub.polls.12$Obama.x)-mean(sub.polls.12$Romney.x) #Find the avaerage margin #obama.x and McCain.x are the polls prediction in the merged data  
}  
  
#Creates dataframe of 47 states  
data.p<-as.data.frame(state.12)  
colnames(data.p)<-"state"  
  
data.p$marginpolls12<-round(marginpolls12,1)#merge 2012 poll margin  
polls.12<-merge(polls.pred.12,data.p,by="state")#merge 2012 election data  
  
#2012 predicted polls margin using 2008 regression  
polls.12$marginpolls12<- 0.70986 +1.10855 \*polls.12$marginpolls12  
  
polls.12$act.margin.2012<-polls.12$Obama.y-polls.12$Romney.y#2012 actual margin  
  
#Number of states not accurately predicted   
polls.12$state[sign(polls.12$marginpolls12)!=sign(polls.12$act.margin.2012)]

## factor(0)  
## 47 Levels: AL AR AZ CA CO CT FL GA HI IA ID IL IN KS KY LA MA MD ME ... WV