Predicting the United States Presidential Election

Nicholas Archambault

24 March 2021

Introduction

The *efficient market hypothesis* of economics states that markets reflect all information, which all market participants share equally. For the stock market, this implies that it is inherently impossible to "beat the market" consistently over the long run since future performance of a stock is dictated only by pertinent, time-relevant information, not by past behavior. But what about other markets?

This project tests the ability of the market to efficiently aggregate available information by predicting the results of the United States presidential election based on betting and polling data. The online company Intrade allows bettors to trade contracts like, "Obama to win the electoral votes of Florida." The market prices of the contracts fluctuate based on their sales in much the same way stock prices move based on activity surrounding them. We will analyze the market prices of contracts for Democratic and Republican nominees' victories in each U.S. state during the 2008 and 2012 presidential elections.

Data

Variables and descriptions for the six different data sets utilized in this analysis can be found below.

Intrade Prediction Market Data from 2008 (`intrade08.csv`) and 2012 (`intrade12.csv`)

Variable	Description
day	date of the session
statename	full name of state (including Washington D.C. in 2008)
state	abbreviated name of state (including Washington D.C. 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 nominee's market
VolumeR	total session trades of Republican nominee's market

2008 U.S. Presidential Election Data (`pres08.csv`)

Variable	Description
state	abbreviated name of state
state.name	full name of state
Obama	Obama's vote share (percentage)
McCain	McCain's vote share (percentage)
EV	number of Electoral College votes for each state

2008 U.S. Presidential Election Polling Data (`polls08.csv`)

Variable	Description
state	abbreviated name of state where poll was conducted
Obama	predicted support for Obama (percentage)
McCain	predicted support for McCain (percentage)
Pollster	name of organization conducting the poll
middate	middate of the period when the poll was conducted

2012 U.S. Presidential Election Data (`pres12.csv`)

Variable	Description
state	abbreviated name of state
Obama	Obama's vote share (percentage)
Romney	Romney's vote share (percentage)
EV	number of Electoral College votes for each state

2012 U.S. Presidential Election Polling Data (`polls12.csv`)

Variable	Description
state	abbreviated name of state where poll was conducted
Obama	predicted support for Obama (percentage)
Romney	predicted support for Romeny (percentage)
Pollster	name of organization conducting the poll
middate	middate of the period when the poll was conducted

1. Predicting State Electoral Results with Betting Data

We begin by predicting the state-by-state outcome of the 2008 and 2012 elections using market prices from the day before Election Day. The candidate whose contract for a given state has the higher closing price on that date is predicted to win that state.

```
# Load data
options(stringsAsFactors = FALSE)
intrade08 <- read.csv("intrade08.csv")
intrade12 <- read.csv("intrade12.csv")
polls08 <- read.csv("polls08.csv")
polls12 <- read.csv("polls12.csv")
pres08 <- read.csv("pres08.csv")
pres12 <- read.csv("pres12.csv")


# Make election results variables
pres08 <- within(pres08, vote_margin <- Obama - McCain)
pres08$dem_victory <- with(pres08, vote_margin >= 0)
pres12 <- within(pres12, vote_margin <- Obama - Romney)
pres12$dem_victory <- with(pres12, vote_margin >= 0)

# Market data predictions
```

```r
intrade08 <- within(intrade08, price_margin <- PriceD - PriceR)
intrade08$pred_dem_victory <- with(intrade08, price_margin >= 0)
intrade12 <- within(intrade12, price_margin <- PriceD - PriceR)
intrade12$pred_dem_victory <- with(intrade12, price_margin >= 0)

# Poll predictions
polls08 <- within(polls08, poll_margin <- Obama - McCain)
polls08$pred_dem_victory <- with(polls08, poll_margin >= 0)
polls12 <- within(polls12, poll_margin <- Obama - Romney)
polls12$pred_dem_victory <- with(polls12, poll_margin >= 0)

# Transform dates
intrade08$day <- as.Date(intrade08$day)
intrade12$day <- as.Date(intrade12$day)
polls08$middate <- as.Date(polls08$middate)
polls12$middate <- as.Date(polls12$middate)

# Merge election data with intrade and poll data
intrade08 <- merge(intrade08, pres08, by = "state")
intrade12 <- merge(intrade12, pres12, by = "state")
polls08 <- merge(polls08, pres08, by = "state", suffixes = c("_poll", "_elec"))
polls12 <- merge(polls12, pres12, by = "state", suffixes = c("_poll", "_elec"))

# Sort electoral votes by state name
electoral_votes08 <- with(pres08, tapply(EV, state, head, n = 1))

# Sort days 1-90 before election
dates08 <- as.Date("2008-11-04") - 90:1

# Function to identify incorrectly predicted states
incorrect <- function(dataset, date) {
  subset(dataset, (day == date) &
           !is.na(pred_dem_victory) &
           (pred_dem_victory != dem_victory))
}

# Get incorrectly predicted states for 2008 and 2012 elections
incorrect(intrade08, as.Date("2008-11-03"))$statename
```

```
## [1] "Indiana"  "Missouri"
```

```r
incorrect(intrade12, as.Date("2012-11-05"))$statename
```

```
## [1] "Florida"
```

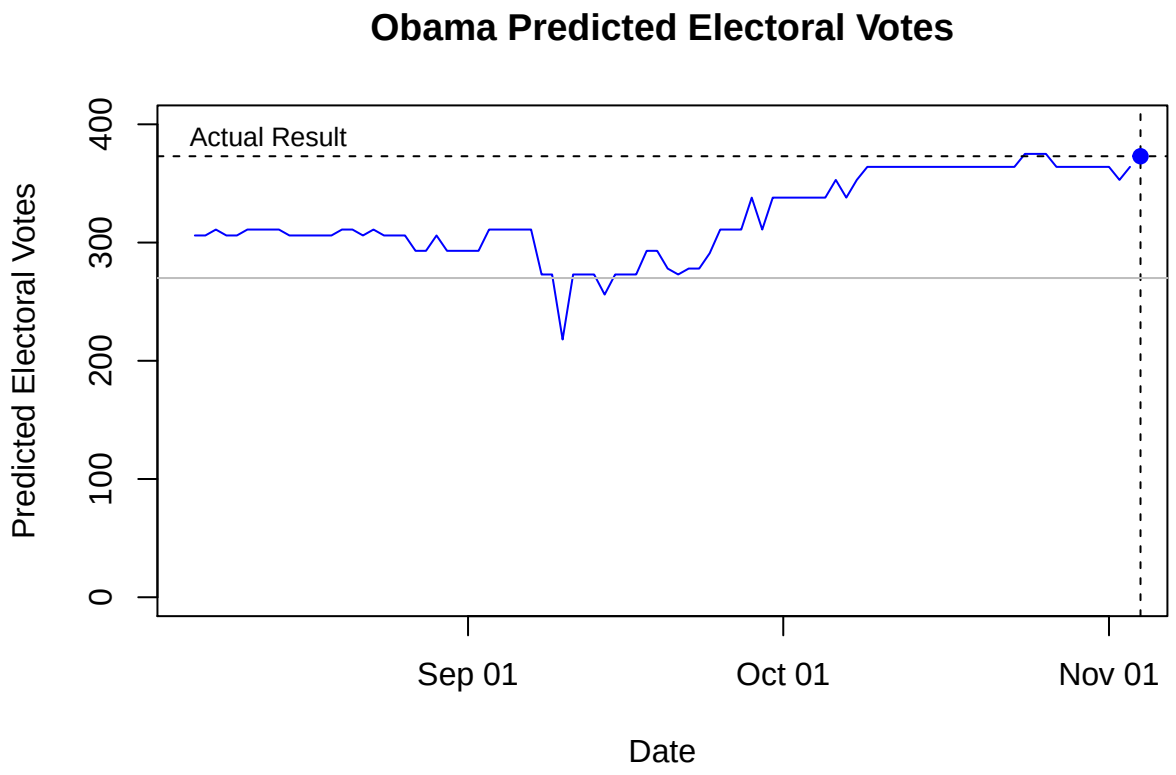
The betting market appears to predict election results quite well. For the 2008 election, the market only misclassified Indiana and Missouri, just two out of the 51 voting states, including Washington D.C. With these results, the overall accuracy rate for the 2008 election was approximately 96%, which is better than the accuracy rate of listed polling data from the day prior to the 2008 election. Besides Indiana and Florida, polls in 2008 also misclassified North Carolina.

In 2012, the results were even better, correctly predicting 98% of state outcomes. Only Florida was misclassified. In general, these results indicate that the betting market is efficient, serving as an accurate mechanism for forecasting election results.

2. Change in Betting Markets Over Time

In the previous section, we used the final day of betting market data to predict election outcomes. It would be useful to understand how the betting market reaches its final-day status by examining how predictions derived from the betting market evolve over time. Using the same method as above, whereby a state's predicted winner is the candidate whose closing contract price is higher for a given day, we can plot the predicted election results for the 90 days preceding the election.

```
# Function that predicts EV totals using market data from given date
predicted_EVs <- function(date) {
  sum(subset(intrade08, (day == date) & pred_dem_victory, EV))
}

# Call function for each date
electoral_votes <- sapply(dates08, predicted_EVs)

# Plot
plot(dates08, electoral_votes,
     main = "Obama Predicted Electoral Votes",
     xlab = "Date",
     ylab = "Predicted Electoral Votes",
     ylim = c(0, 400), type = "l", col = "blue")
abline(v = as.Date("2008-11-04"), lty = 2)
abline(h = 373, lty = 2)
abline(h = 270, col = "gray")
points(x = as.Date("2008-11-04"),
       y = 373,
       col = "blue", pch = 19)
text(as.Date("2008-08-13"), 390, "Actual Result", cex = 0.8)
```



The intersection of the dashed lines represents Barack Obama's true electoral vote count on election day

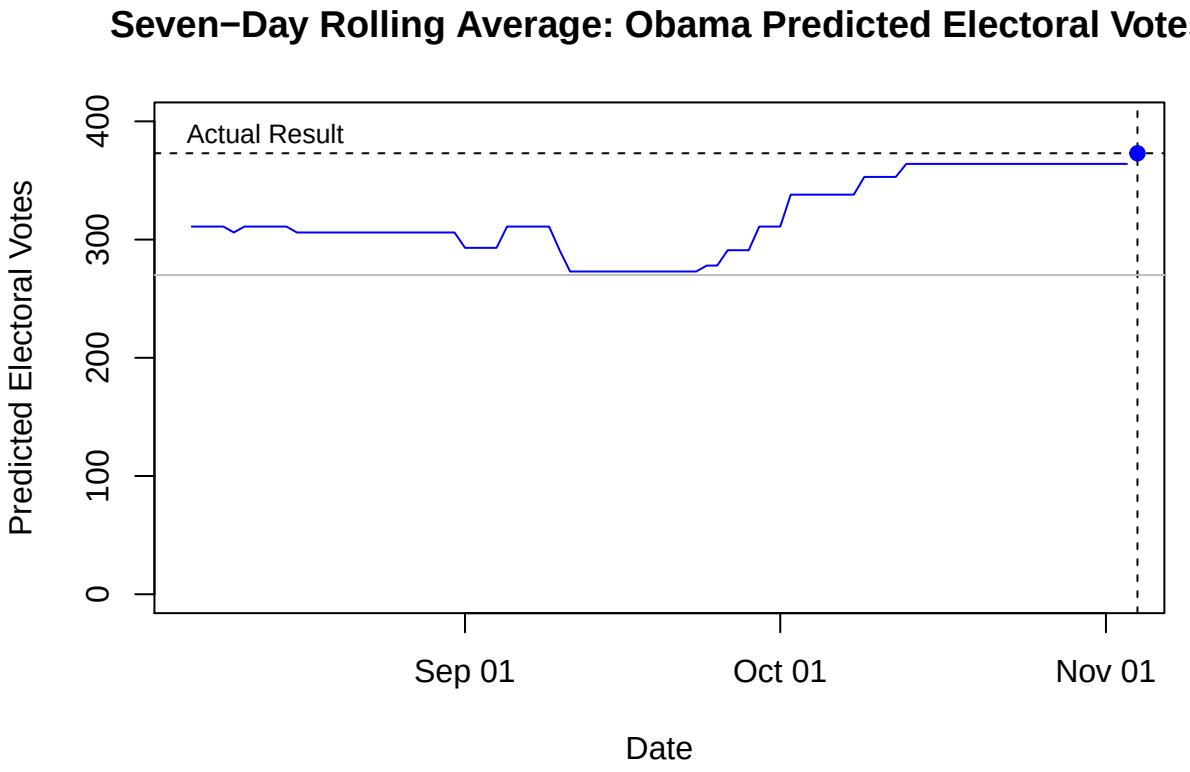
2008, and the solid gray line denotes the majority threshold for electoral votes. Trends in the betting market reveal that Obama maintained a comfortable position for most of the 90 days preceding the election. In August 2008, Obama held a modest predicted lead of about 35 electoral votes. Despite some fluctuation, that lead remained intact. John McCain briefly surged ahead before Obama regained the lead, which he widened during the final months before the election. There is little change in predicted electoral totals within roughly the last 25 days prior to the election.

Market data predictions are rather poor for much of August and September, likely the result of voters' unfamiliarity with the candidates and their policies. Predictions improve substantially by October and settle upon a faithful prediction of Obama's final electoral vote total. Despite the prediction's fluctuation in the early run-up to the election, this analysis reinforces the idea that the efficiency of the betting market renders it a solid predictor of electoral results.

3. Weekly Moving Average Price

To visualize higher-level patterns in the data, we can compute the weekly rolling average of Obama's closing price in each state.

```
rolling_average <- function(date) {
  # Data from zero to six days before date
  intrade08_prior <- subset(intrade08, subset = day %in% (date - 0:6))
  # State average over seven days
  average <- with(intrade08_prior, tapply(price_margin, state, mean))
  # Get predicted wins multipled by electoral votes
  sum(electoral_votes08[average >= 0])
}

# Function for each date
electoral_avg <- sapply(dates08, rolling_average)

# Plot
plot(dates08, electoral_avg,
     main = "Seven-Day Rolling Average: Obama Predicted Electoral Votes",
     xlab = "Date",
     ylab = "Predicted Electoral Votes",
     ylim = c(0, 400), type = "l", col = "blue")
abline(v = as.Date("2008-11-04"), lty = 2)
abline(h = 373, lty = 2)
abline(h = 270, col = "gray")
points(x = as.Date("2008-11-04"),
       y = 373,
       col = "blue", pch = 19)
text(as.Date("2008-08-13"), 390, "Actual Result", cex = 0.8)
```



This analysis largely replicates the results from the previous section. The main difference is that the plot shows less variation since prices are averaged over seven days. The rolling average yields greater stability and consistency over the course of the election cycle, correctly predicting over the entire 90-day period that Obama would win the election. A downside of this approach, however, is that it obfuscates day-to-day information, like how breaking news might have affected perceived chances of victory on individual dates.

4. Prediction Based on Recent Polls

Until now, we have examined only betting market data and found it to be a reasonably stable predictor of electoral results. A more traditional method of gauging sentiment prior to an election is polling. A number of polls were conducted in each state at various, irregular points in the run-up to the 2008 election. Information on those polls is collected in the `polls08.csv` file; a randomized sample of this file is provided below.

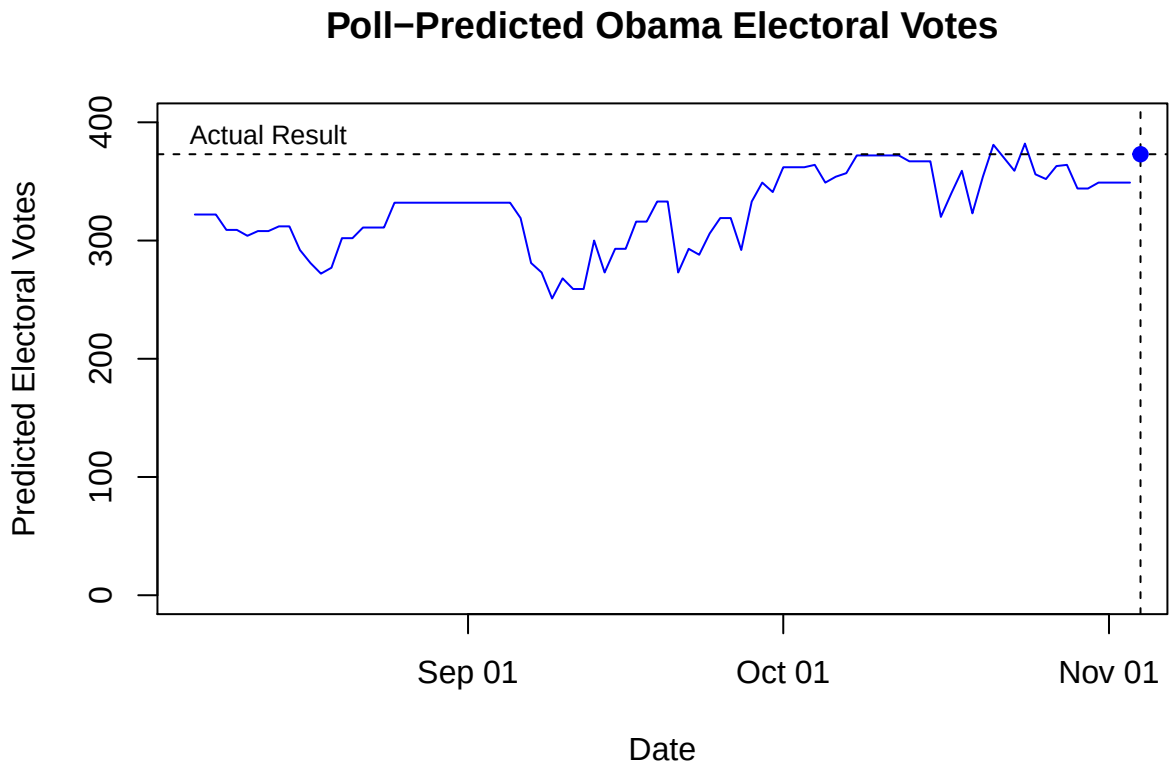
```
set.seed(1234)
sampler <- polls08[sample(nrow(polls08), 6), 1:7]
knitr::kable(sampler[order(sampler$middate), ], row.names = F)
```

state	Pollster	Obama_poll	McCain_poll	middate	poll_margin	pred_dem_victory
OR	SurveyUSA-3	50	41	2008-03-15	9	TRUE
TN	ARG-4	36	59	2008-09-17	-23	FALSE
OH	Suffolk U.-4	51	42	2008-10-17	9	TRUE
WV	Research 2000-3	43	49	2008-10-23	-6	FALSE
PA	Rasmussen-1	53	46	2008-10-27	7	TRUE
MT	Research 2000-3	44	48	2008-10-29	-4	FALSE

We can create a plot similar to those above, this time using daily predicted margin of victory within a state as determined by the most recent poll conducted. If multiple polls were administered on the same day, their results will be averaged. Collecting the poll predictions, we can track Obama's total number of predicted

electoral votes through time.

```
poll_prediction <- function(date) {
  # Polls on or before given date
  polls08_date <- subset(polls08, middate <= date)
  # Max date within state
  polls08_date$max_date <- with(polls08_date, ave(middate, state, FUN = max))
  # Keep data from max date
  polls08_date <- subset(polls08_date, middate == max_date)
  # Average poll margin for each state
  state_avg_poll <- with(polls08_date, tapply(poll_margin, state, mean))
  # Get predicted electoral votes
  sum(electoral_votes08[state_avg_poll >= 0])
}

# Function for each date
poll_electoral_votes <- sapply(dates08, poll_prediction)

# Plot
plot(dates08, poll_electoral_votes,
     main = "Poll-Predicted Obama Electoral Votes",
     xlab = "Date",
     ylab = "Predicted Electoral Votes",
     ylim = c(0, 400), type = "l", col = "blue")
abline(v = as.Date("2008-11-04"), lty = 2)
abline(h = 373, lty = 2)
points(x = as.Date("2008-11-04"),
       y = 373,
       col = "blue", pch = 19)
text(as.Date("2008-08-13"), 390, "Actual Result", cex = 0.8)
```



It appears that polls are less stable and accurate at predicting electoral results than either daily or rolling weekly average betting market data. Electoral vote predictions from betting market data hovered around their true values in the final days. The rolling weekly averages remained constant at the true values for both candidates during the final three weeks. In contrast, the polling data — though it consistently and correctly projects Obama as the winner — reveals more substantial fluctuation and discrepancy from the true electoral vote total in the days leading up to the election.

5. Betting Market Price Margins and Actual Victory Margins

Using betting market data from the day before the 2008 election, we can understand the relationship between prediction and reality by regressing Obama's actual margin of victory in each state on Obama's price margin within the state from the Intrade market.

Similarly, we can regress Obama's actual margin of victory on his predicted margin from recent polling within each state.

```
# Fit prediction model for market data
intrade08_elec_day <- subset(intrade08, day = as.Date("2008-11-03"))
intrade08_lm <- lm(vote_margin ~ price_margin, data = intrade08_elec_day)
summary(intrade08_lm)
```

```
##
## Call:
## lm(formula = vote_margin ~ price_margin, data = intrade08_elec_day)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.367  -6.572  -1.963   5.727  79.432
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.5681423  0.0608285   91.54   <2e-16 ***
## price_margin 0.2799326  0.0009194  304.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.66 on 36889 degrees of freedom
## Multiple R-squared:  0.7154, Adjusted R-squared:  0.7154
## F-statistic: 9.271e+04 on 1 and 36889 DF,  p-value: < 2.2e-16
```

```
# Average poll data for day closest to given date
poll_margin_by_date <- function(poll_data, date) {
  # Find all polls on or before date
  poll_data <- subset(poll_data, middate <= as.Date(date))
  # Find polls closest to given date
  poll_data <- subset(poll_data, middate == ave(middate, state, FUN = max))
  # Avg. within state among all polls close to date
  poll_data$poll_margin_avg <- with(poll_data, ave(poll_margin, state))
  # Remove duplicates
  subset(poll_data, !duplicated(state))
}

# Fit prediction model for polls
polls_elec_day <- poll_margin_by_date(polls08, "2008-11-03")
polls08_lm <- lm(vote_margin ~ poll_margin_avg, data = polls_elec_day)
```

```r
summary(polls08_lm)
```

```
##
## Call:
## lm(formula = vote_margin ~ poll_margin_avg, data = polls_elec_day)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3834  -2.7205   0.3556   3.4224  13.0111
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.70908    0.78695   0.901    0.372
## poll_margin_avg  1.10856    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
```

The coefficients are positive in each model, indicating that both market and poll data are predictive of the actual election outcome. We knew this from previous sections. The poll data appears to perform better than the market data, with an R^2 of 0.9382 vs 0.7154.

It is hard to interpret differences in the models beyond that. There is no reason why we ought to believe that the market price margins are linearly related to true vote margins. This market concerns the winner of the election, not the vote margin. For example, if the market judged Obama to have a solid 10 percentage point lead in some state, the betting prices would likely be 100% in his favor since Obama's victory in that state would be all but certain. The same would be true if Obama held a 50 point lead — that is to say, the market prices reflect winning probabilities, not victory margins. There is a relationship between the two, but it is hardly linear. As a result, market prices are better at predicting state winners than victory margins.

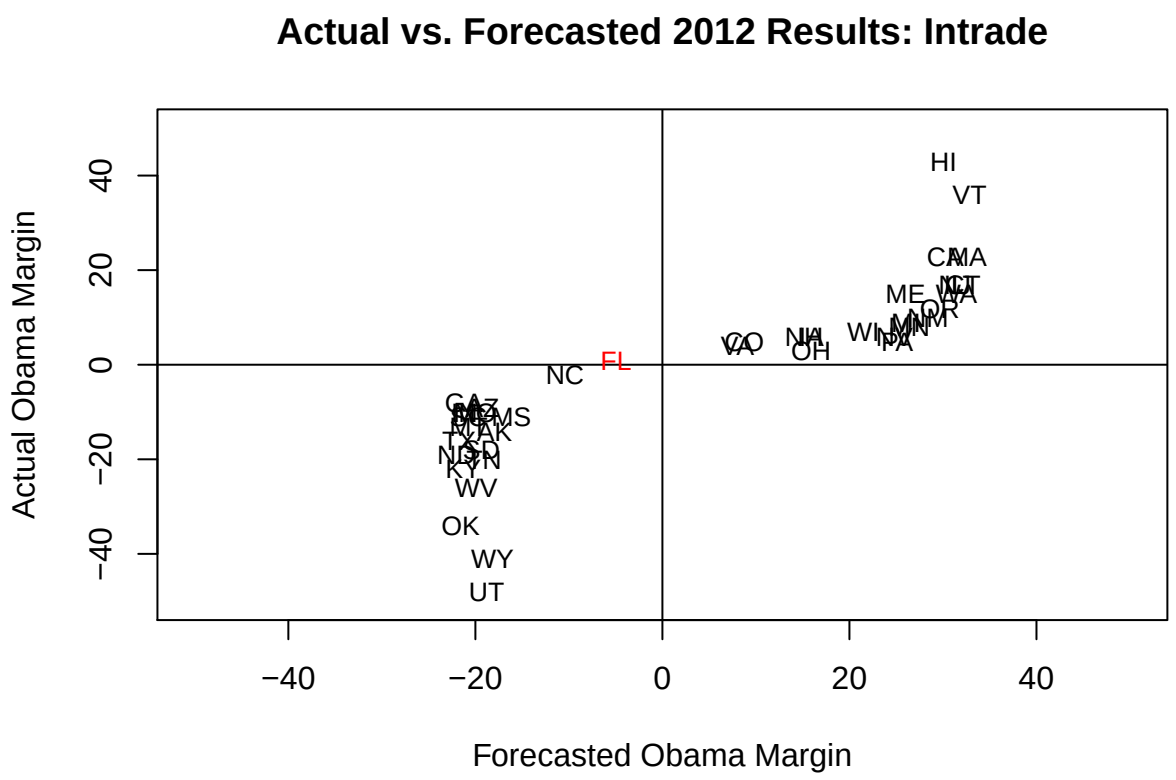
The opposite is true for poll data; this explains why the market performs better in previous sections, where we predicted winners, while the polls perform better in this exercise where we predict margin of victory.

6. Using 2008 to Predict 2012

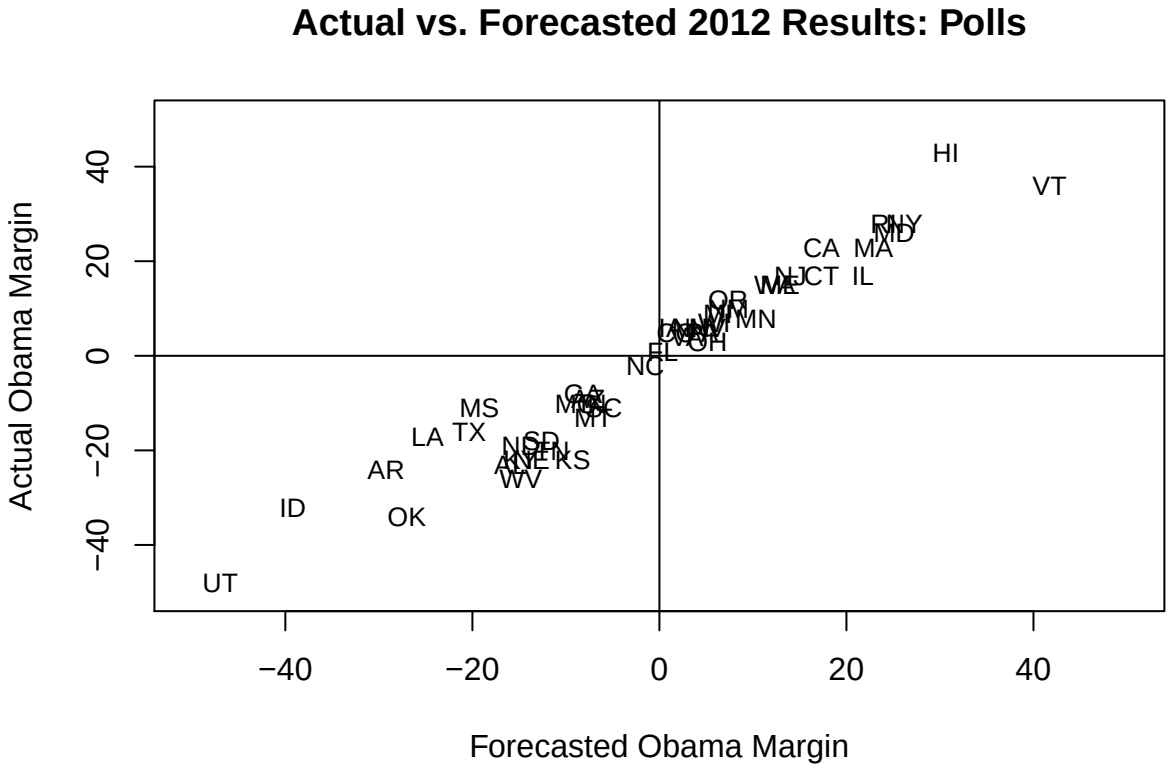
Having gleaned an understanding of the predictive power of both market data and polls, we attempt to forecast Obama's true margin of victory in the 2012 election using similar methods. This is done first by using 2008 Intrade prices from the day before the election as the predictor in each state, then by using 2008 poll-predicted margins from the most recent polls in each state.

```r
# Predict 2012 election using 2008 model and market data
intrade12_elec_day <- subset(intrade12, !is.na(price_margin) &
                               day == as.Date("2012-11-5"))
intrade12_elec_day$pred_vote_margin <- predict(intrade08_lm,
                                               newdata = intrade12_elec_day)
intrade12_elec_day$misclassified <- with(intrade12_elec_day,
                                         (pred_vote_margin >= 0) != dem_victory)

plot(intrade12_elec_day$pred_vote_margin, intrade12_elec_day$vote_margin,
     type = "n",
     xlim = c(-50, 50), ylim = c(-50, 50),
     xlab = "Forecasted Obama Margin", ylab = "Actual Obama Margin",
```

```
     main = "Actual vs. Forecasted 2012 Results: Intrade")
abline(h = 0)
abline(v = 0)
text(intrade12_elec_day$pred_vote_margin,
     intrade12_elec_day$vote_margin, intrade12_elec_day$state,
     cex = 0.8, col = ifelse(intrade12_elec_day$misclassified, "red", "black"))
```




```
# Predict 2012 election using 2008 model and polls
polls12_elec_day <- poll_margin_by_date(polls12, "2012-11-5")
polls12_elec_day$pred_vote_margin <- predict(polls08_lm,
                                             newdata = polls12_elec_day)
polls12_elec_day$misclassified <- with(polls12_elec_day,
                                       (pred_vote_margin >= 0) != dem_victory)

plot(polls12_elec_day$pred_vote_margin, polls12_elec_day$vote_margin,
     type = "n",
     xlim = c(-50, 50), ylim = c(-50, 50),
     xlab = "Forecasted Obama Margin", ylab = "Actual Obama Margin",
     main = "Actual vs. Forecasted 2012 Results: Polls")
abline(h = 0)
abline(v = 0)
text(polls12_elec_day$pred_vote_margin,
     polls12_elec_day$vote_margin, polls12_elec_day$state,
     cex = 0.8, col = ifelse(polls12_elec_day$misclassified, "red", "black"))
```



The inputs of the prediction model act as predictions themselves, but they might be biased one way or the other. Since the coefficient for the poll data in 2008 is greater than 1, it seems that polls tend to systematically underestimate support for Obama. It appears that both the market and poll data are solid predictors of 2012 election outcomes. The market data misclassify one state (Florida) and the poll data correctly classify all states. Interestingly, if we were to take only the raw poll data in 2012 without using the 2008 prediction model refinement, it would misclassify Florida as well.

Conclusion

In this project we have explored the efficiency of the market through two proxies for predicting election results: online betting data and polling. Both sources of data were found to be relatively stable predictors of the total number of electoral votes earned by Barack Obama in the 2008 U.S. presidential election. The betting market data appeared to be more accurate and less susceptible to fluctuation, especially in the days and weeks immediately preceding the election.

Poll data is better at predicting margin of victory than betting data, as shown in the regression analyses of Section 5. Betting data is more useful in choosing a winning candidate — a type of classification problem — since its odds and closing price will likely be the same no matter if a candidate's lead is a solid 10 percentage points or 50. We find that both market and poll data from 2008 are solid predictors of the 2012 electoral results, misclassifying just a single state between them.