

# 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
<code>day</code>	Date of the session
<code>statename</code>	Full name of each state (including District of Columbia in 2008)
<code>state</code>	Abbreviation of each state (including District of Columbia in 2008)
<code>PriceD</code>	Closing price (predicted vote share) of Democratic Nominee’s market
<code>PriceR</code>	Closing price (predicted vote share) of Republican Nominee’s market
<code>VolumeD</code>	Total session trades of Democratic Party Nominee’s market
<code>VolumeR</code>	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
<code>state.name</code>	Full name of state (only in <code>pres2008</code> )
<code>state</code>	Two letter state abbreviation
<code>Obama</code>	Vote percentage for Obama
<code>McCain</code>	Vote percentage for McCain
<code>EV</code>	Number of electoral college votes for this state

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

Name	Description
<code>state</code>	Abbreviated name of state in which poll was conducted
<code>Obama</code>	Predicted support for Obama (percentage)
<code>Romney</code>	Predicted support for Romney (percentage)
<code>Pollster</code>	Name of organization conducting poll
<code>midddate</code>	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

```
election_2008 <- read.csv("pres08.csv")
intrade_2008 <- read.csv("intrade08.csv")

election_trade <- merge(election_2008, intrade_2008, by = "state")
election_trade$DaysToElection <- as.Date("2008-11-04") - as.Date(election_trade$day)

for (i in 1:nrow(election_trade)) {
  election_trade$margin[i] <- election_trade$Obama[i] - election_trade$McCain[i]
  election_trade$price_margin[i] <- election_trade$PriceD[i] - election_trade$PriceR[i]
}
```

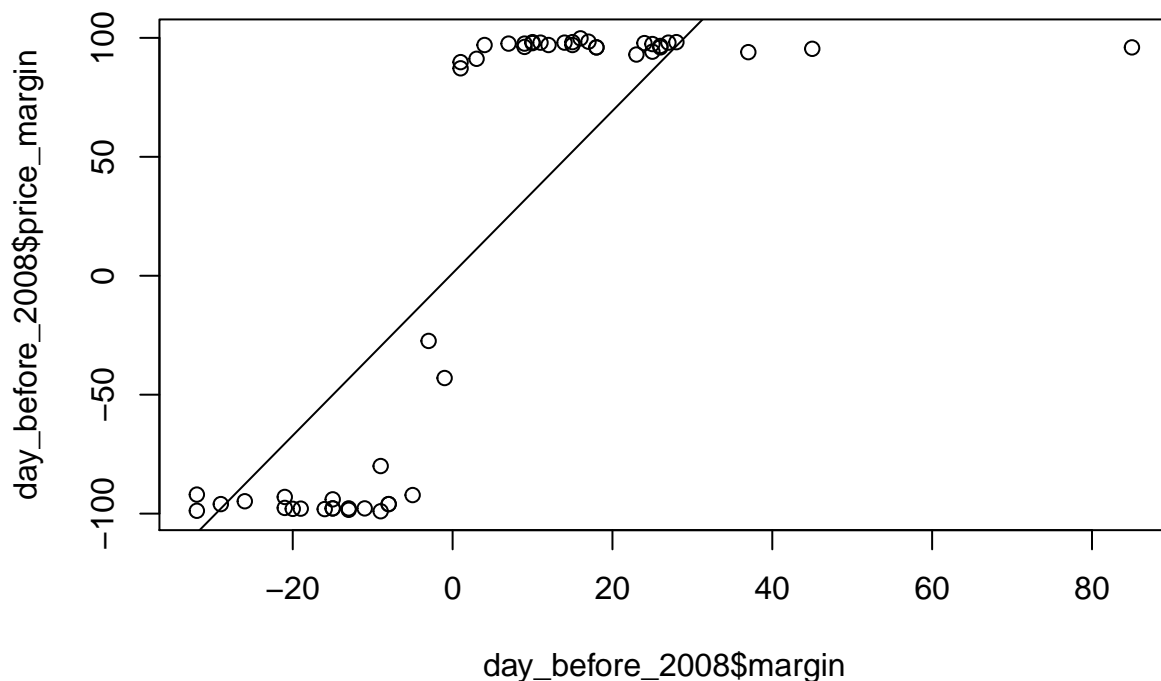
## Question 2

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

```
day_before_2008 <- subset(election_trade, subset = (day == "2008-11-04"))

plot(day_before_2008$margin, day_before_2008$price_margin)
abline(lm(day_before_2008$price_margin ~ day_before_2008$margin))
```



We know that positive numbers are associated with Obama predictions and negative numbers are associated with McCain predictions. We see that there is a positive prediction associated with Obama's margins of victory.

### 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 seems 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 tell 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

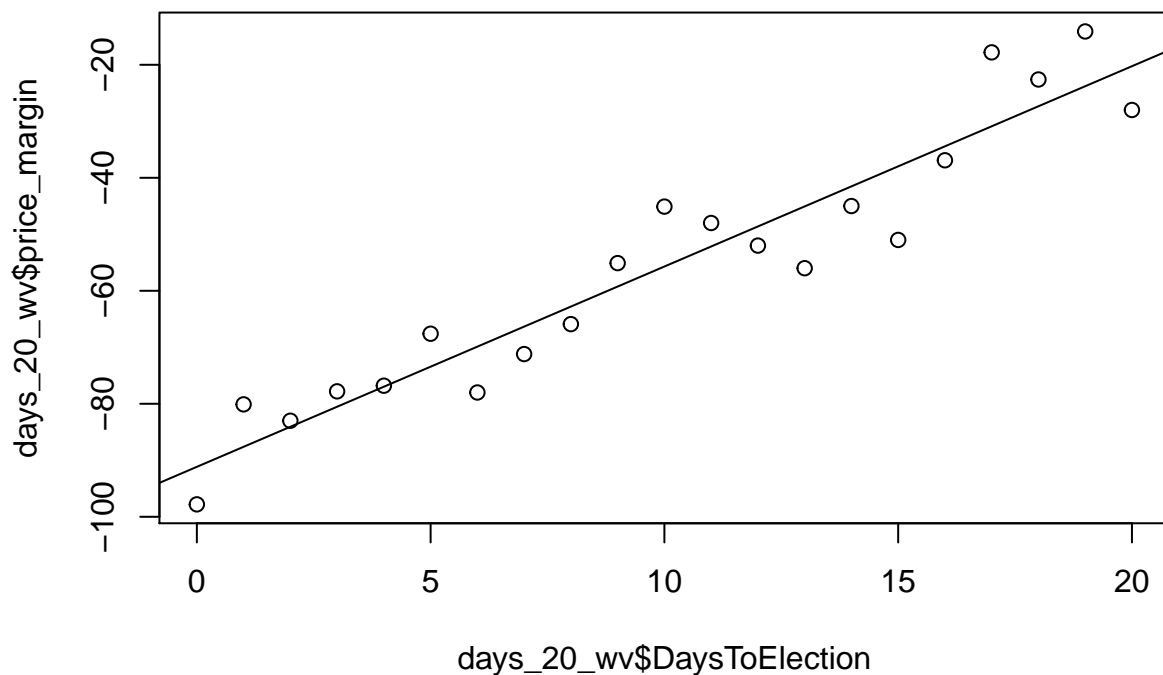
```

st.names <- unique(election_trade$state.name)

days_20_wv <- subset(election_trade, DaysToElection <= 20 & state.name == "West Virginia")

plot(days_20_wv$price_margin ~ days_20_wv$DaysToElection)
abline(lm(days_20_wv$price_margin ~ days_20_wv$DaysToElection))

```



```

#days_20_wv$fit <- coef(lm(days_20_wv$price_margin ~ days_20_wv$DaysToElection))

```

In this section we see that as the election gets closer, Obama's margin continues to decrease, into the negatives, showing an advantage for McCain. This shows that betters were increasingly assured that Obama would lose West Virginia.

## Question 4

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

## Answer 4

```

#days_20_all <- subset(election_trade, DaysToElection <= 20)

price_prediction <- rep(NA, 51)
#st.names <- unique(days_20_all$state.name)

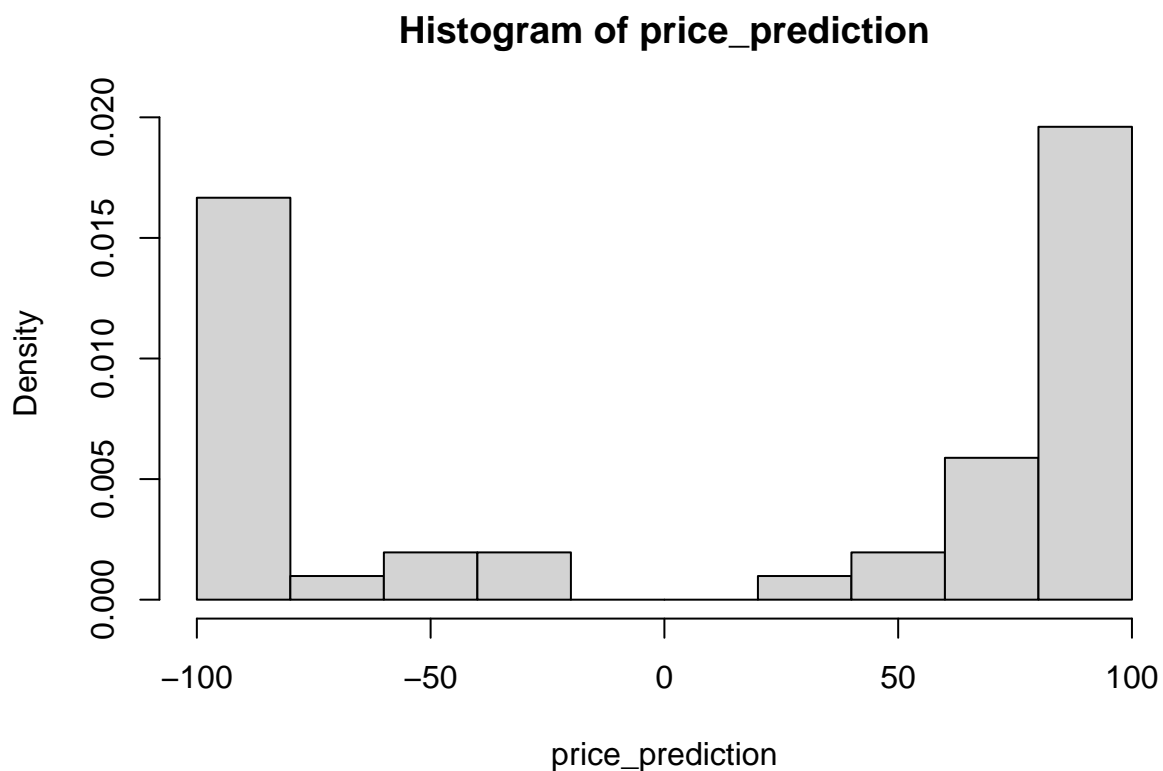
for (i in 1:51) {
  state.data <- subset(election_trade, state.name == st.names[i] & DaysToElection <= 20)

  #days_20_state <- state.data$DaysToElection <= 20
  lm.state <- lm(state.data$price_margin ~ state.data$DaysToElection)

  coeff.state <- coef(lm.state)
  price_prediction[i] <- coeff.state
}

hist(price_prediction, freq = FALSE)

```



This shows that the predictions are relatively set in the 20 days prior to the election and that the predictions are not all that volatile.

## 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

```
intrade_2008$DaysToElection <- as.Date("2008-11-04") - as.Date(intrade_2008$day)

two_weeks <- subset(intrade_2008, DaysToElection <= 14)

price_2 <- rep(NA, 51)

for (i in 1:nrow(two_weeks)) {
  two_weeks$price_margin[i] <- two_weeks$PriceD[i] - two_weeks$PriceR[i]
}

for (i in 1:51) {
  state.data <- subset(two_weeks, statename == st.names[i] & DaysToElection == 14)

  state.data$predict <- sign(state.data$price_margin)

  price_2[i] <- sign(state.data$price_margin[i])

  #lm.state <- lm(state.data$price_margin ~ state.data$DaysToElection)
  #coeff.state <- coef(lm.state)
  #price_2[i] <- lm.state
}

table(price_2)

## price_2
## -1
## 1
```