Bonus Problem Set 4

Alin Ierima

Predicting Elections Using Betting Markets and Linear Models

Earlier in the fall term, 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* (for those interested, see an interview about Intrade here). 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 (*for more see here*). 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 this dataset 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

Question 1 (15 points)

In this problem set we are interested in the relationship between the price margins of the *Intrade* market and the actual margin of victory at the 2008 presidential elections. To study this relationship you will regress Obama's actual margin of victory (outcome variable) on Obama's price margin (predictor) from the *Intrade* markets in each state. You will do so using a loop in question 2. But first, we need to do some preparatory steps. Follow the below instructions:

- Load the *Intrade* (`intrade08.csv`) and the Election outcome data `pres08.csv` for 2008 and combine them together by `state` using the `merge` function. You can assigned the merged dataset to a new object called `intresults08`. *Hint: see a simple example how to use the merge function at the bottom of this document.*
- Create a `DaysToElection` variable by subtracting each *Intrade* session date in the dataset (variable `day`) from the day of the election (4th of November 2008). Below you can find guidelines how to achieve this. (9 points)
  - Hint: Notice that the `day` variable is a character variable in the form "2008-02-28", "2008-03-17" (general format `YYYY-MM-DD`). In order to be able to do any numeric operations with date variables (like calculate the number of days between two given dates), you first need to transform this character variable to a variable of date format. Otherwise R does not know that this variable contains dates information. You can transform a character variable to a date format using the function `as.Date(x, format)`. *See a simple example how to use the merge function at the bottom of this document.*
- Inside the `intresults08` object, create a *state margin of victory* variable (our outcome variable, call it `obama.actmarg`) - this is the difference in the vote percentage between Obama and McCain for each state. (3 points)
- Inside the `intresults08` object, create a *betting market margin* (our predictor, call it `obama.intmarg`) - this is the difference in the *closing price (predicted vote share)* for Obama (Democratic candidate) and McCain (Republican candidate). (3 points)

Answer 1

For your information - the first few observations in the merged dataset

```
intresults08 = merge (x=intrade08, y=pres08, by.x = "state", by.y = "state", all.x= T)

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
```

For your information - the first few observations in the `DaysToElection` variable

```
 intresults08$day = as.Date(intresults08$day, format="%Y-%m-%d")
 intrade08$day = as.Date(intrade08$day, format="%Y-%m-%d")
 election_date = "2008-11-04"
 election_date = as.Date (election_date, format="%Y-%m-%d")
```

```r
intresults08$DaysToElection = election_date - intresults08$day


head(intresults08$DaysToElection)
```

```
## Time differences in days
## [1] 250 228 550 220 227 258
```

For your information - the first few observations in the betting market margin variable

```r
intresults08$obama.intmarg =  intresults08$PriceD - intresults08$PriceR
head(intresults08$obama.intmarg)
```

```
## [1] -85 -85 -80 -84 -85 -85
```

For your information - the first few observations in the state margin of victory variable

```r
intresults08$obama.actmarg = intresults08$Obama - intresults08$McCain
head(intresults08$obama.actmarg)
```

```
## [1] -21 -21 -21 -21 -21 -21
```

Question 2 (15 points)

- Consider only the trading one day from the election (trading on the 3rd of November 2008) and regress Obama's actual margin of victory (`obama.actmarg`) on Obama's price margin (`obama.intmarg`). Hint*: Make a subset of the data where `DaysToElection` is equal to 1. Then predict Obama's actual electoral margins with Obama's trading margins using a linear regression model. (5 points)
- Let us visualize the fitted model. How would you visualize the predictions and the outcomes together? Follow the guidelines below: (4 points)
  - Make a scatterplot, where you plot the actual electoral margins (y-axis) against the trading margin.
  - Add the fitted regression line (line of best fit) to the plot. Hint: because we only have one predictor you can use `abline` for the line of best fit.
  - Color the solid points in *"steelblue2"* with transparency level 0.5, and the fitted line in *"orangered"*. Add dashed horizontal and vertical lines at 0 in gray (*"gray20"*)
- What do we find? Shortly discuss your findings in the text. (3 points)
- Does the linear regression model predict well? Elaborate your thinking. (Hint: think about what statistic we use to evaluate the model fit.) (3 points)

Answer 2

For your information - the first few observations in the subset one day before the election

```r
one_day = intresults08[intresults08$DaysToElection == 1, ]
head(one_day)
```

```
##      state        day  statename PriceD VolumeD PriceR VolumeR state.name Obama
## 717     AK 2008-11-03     Alaska    6.0      32   94.0      58     Alaska    38
## 1047    AL 2008-11-03    Alabama    3.3       0   95.0      90    Alabama    39
## 1789    AR 2008-11-03   Arkansas    3.9      11   90.0       5   Arkansas    39
## 2627    AZ 2008-11-03    Arizona   18.9    1662   80.2     139    Arizona    45
## 3242    CA 2008-11-03 California   97.9     100    2.1      10 California    61
## 4332    CO 2008-11-03   Colorado   91.5     149   12.0     103   Colorado    54
##      McCain EV DaysToElection obama.intmarg obama.actmarg
## 717      59  3         1 days         -88.0           -21
## 1047     60  9         1 days         -91.7           -21
## 1789     59  6         1 days         -86.1           -20
## 2627     54 10         1 days         -61.3            -9
## 3242     37 55         1 days          95.8            24
## 4332     45  9         1 days          79.5             9
```

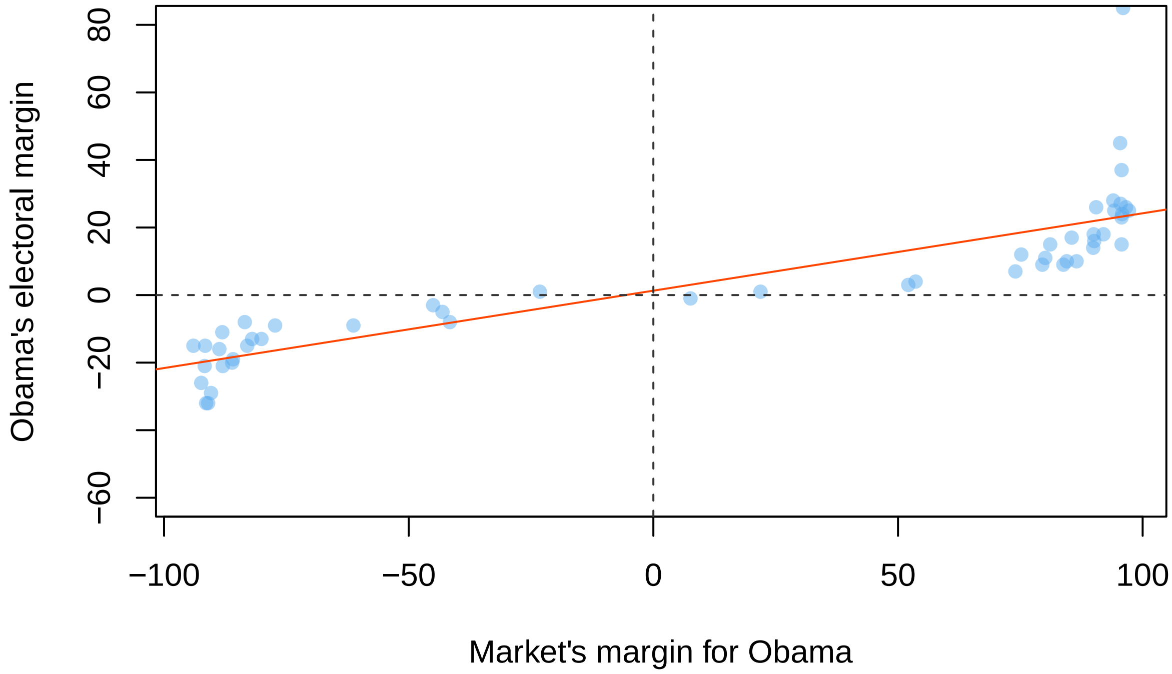
Linear Model Results

```
model = lm(obama.actmarg ~ obama.intmarg, data = one_day)
model
```

```
##
## Call:
## lm(formula = obama.actmarg ~ obama.intmarg, data = one_day)
##
## Coefficients:
##   (Intercept)  obama.intmarg
##        1.3027         0.2291
```

Scatterplot with the line of best fit

```
plot(one_day$obama.intmarg, one_day$obama.actmarg,pch=16, col = adjustcolor("steelblue2", alpha = 0.5),
abline(model, col = "orangered")
abline(h = 0, col = "gray20", lty = 2)
abline(v = 0, col = "gray20", lty = 2)
```



R^2 from the above model

```
summary(model)$r.squared
```

```
## [1] 0.7290092
```

Insert your answer here: Considering R squared goes from 0 to 1, a score of 0.73 means that the model fits the data quite nicely. Essentially, 73% of the variation can be explained by this model. This shows that the market is quite efficient at predicting electoral outcomes, or at least more than chance.

Question 3 (20 points)

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 market margin for *each state* over the 20 days before the election using days to election as a predictor.

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…

- Ultimately our goal is to fit linear models in each state using a loop, but here we will start with one state - West Virginia. For West Virginia:

  - Concentrate only on the last 20 days before the election. (Hint: Make a subset, where `DaysToElection` is smaller or equal to 20) (4 points)
  - Model the relationship between Obama's market margin and days to election with a linear regression model (4 points)
  - Plot Obama's market margin (y-axis) against the number of days until the election (x-axis) (4 points)
  - Show on the plot the model's predictions for each day starting from the 20th to the 0 day before elections. (4 points) (Hint: think about what approach we use to show the predictions of the outcome for different values of the predictor variable.)
- What does this model's slope coefficient tells us about which direction the margin is changing and also how fast it is changing? (4 points)

Answer 3

For your information: the first few observations in West Virginia 20 days before elections

```r
twenty_days =  intresults08[intresults08$DaysToElection <= 20 & intresults08$state== "WV" ,]
head(twenty_days)
```

```
##       state        day     statename PriceD VolumeD PriceR VolumeR
## 35803    WV 2008-10-31 West Virginia   13.0     191   89.8      98
## 35810    WV 2008-10-24 West Virginia   27.5      64   75.5      37
## 35822    WV 2008-10-29 West Virginia   11.0     133   89.0      51
## 35829    WV 2008-10-22 West Virginia   22.0     204   78.0     171
## 35836    WV 2008-10-15 West Virginia   36.0      48   64.0      30
## 35854    WV 2008-11-01 West Virginia   12.0     101   89.8      35
##          state.name Obama McCain EV DaysToElection obama.intmarg obama.actmarg
## 35803 West Virginia    43     56  5         4 days         -76.8           -13
## 35810 West Virginia    43     56  5        11 days         -48.0           -13
## 35822 West Virginia    43     56  5         6 days         -78.0           -13
## 35829 West Virginia    43     56  5        13 days         -56.0           -13
## 35836 West Virginia    43     56  5        20 days         -28.0           -13
## 35854 West Virginia    43     56  5         3 days         -77.8           -13
```

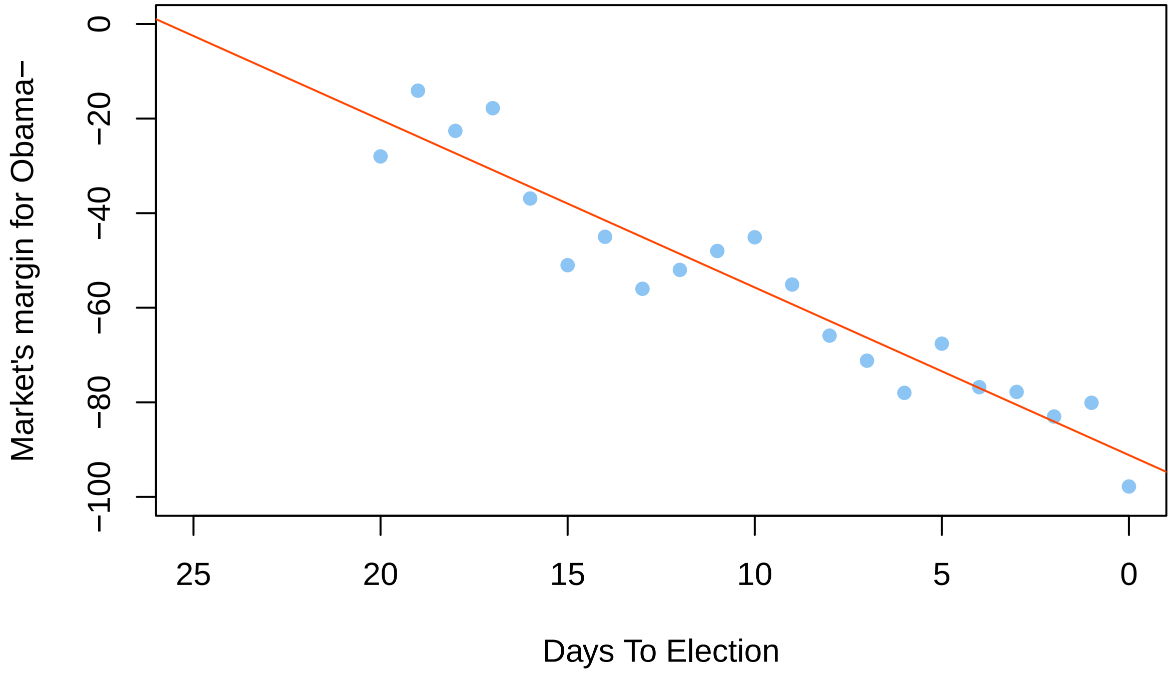
Linear Model Results for Obama's Martket Margit

```r
lm_twenty_days = lm(obama.intmarg ~ DaysToElection, data=twenty_days)
lm_twenty_days
```

```
##
## Call:
## lm(formula = obama.intmarg ~ DaysToElection, data = twenty_days)
##
## Coefficients:
##    (Intercept)  DaysToElection
##        -91.161           3.546
```

Plot Obama's market margin over time with a line of best fit

```r
plot(twenty_days$DaysToElection,twenty_days$obama.intmarg, xlim = c(25,0), ylim=c(-100,0), pch=16, col =
abline(lm_twenty_days, col= "orangered")
```



Insert your answer here:The slope shows that as the days progress and we get closer to the election day, the market's price margin for Obama increases by 3.546 for each day. This implies a fast change in market opinions towards Obama as we approach the elections.

Question 4 (25 points)

Now we would like to have a summary of the price movement for *each state* over the 20 days before the election. We will hence do the same regression model as in question 3, but this time for **all states using a loop**:

Specifically, use a loop which runs a linear model with Obama's market margin as the outcome variable and the`DaysToElection` as a predictor variable. Run a linear model for *each state* focusing on market's betting for the last 20 days. Follow the below steps:

- loop over each state. Inside the loop:
  - Make a subset for each state 20 days before elections (including the 20th day). (5 points)
  - Model the relationship between days to election and Obama's market margin with a linear regression model.(Hint: regress Obama's market margin on days to election for this subset.) (10 points)
  - Collect the slope coefficients (the coefficient for the effect of `DaysToElection`, not the intercept coefficient) from each state model and save these in an empty container for the corresponding

state. You will need this information later to see how volatile the state estimates are. (Hint: before you do the loop, create an empty container called `change`, where you will save the effect of days to election on Obama's betting market margin for each state. Name each observation with the state names.) (5 points)

- Plot the slope coefficients from all states (the values you saved in the empty container `change`) using a histogram. Shortly interpret the histogram and discuss your findings. (5 points)

Answer 4

For your information: the empty container where you will save the regression slopes for each state

```r
states = unique(intresults08$statename)
change = rep(NA, length(states))
names(change) = states
change
```

```
##               Alaska              Alabama             Arkansas
##                   NA                   NA                   NA
##              Arizona           California             Colorado
##                   NA                   NA                   NA
##          Connecticut District of Columbia             Delaware
##                   NA                   NA                   NA
##              Florida              Georgia               Hawaii
##                   NA                   NA                   NA
##                 Iowa                Idaho             Illinois
##                   NA                   NA                   NA
##              Indiana               Kansas             Kentucky
##                   NA                   NA                   NA
##            Louisiana        Massachusetts             Maryland
##                   NA                   NA                   NA
##                Maine             Michigan            Minnesota
##                   NA                   NA                   NA
##             Missouri          Mississippi              Montana
##                   NA                   NA                   NA
##       North Carolina         North Dakota             Nebraska
##                   NA                   NA                   NA
##        New Hampshire           New Jersey           New Mexico
##                   NA                   NA                   NA
##               Nevada             New York                 Ohio
##                   NA                   NA                   NA
##             Oklahoma               Oregon         Pennsylvania
##                   NA                   NA                   NA
##         Rhode Island       South Carolina         South Dakota
##                   NA                   NA                   NA
##            Tennessee                Texas                 Utah
##                   NA                   NA                   NA
##             Virginia              Vermont           Washington
##                   NA                   NA                   NA
##            Wisconsin        West Virginia              Wyoming
##                   NA                   NA                   NA
```

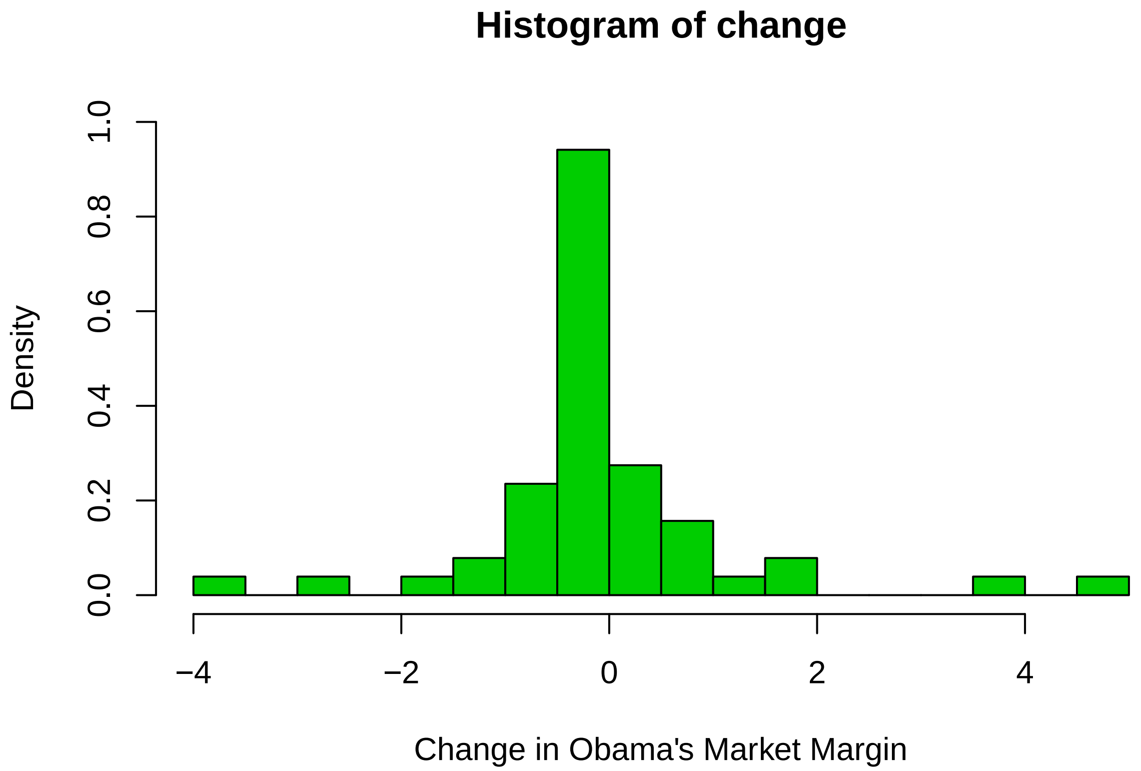
Filled empty container with the regression slopes (using a loop). Note: For presentation reasons we show below the regression slopes rounded to 4 digits

```r
for (i in states) {
  state_subset =intresults08[intresults08$statename == i & intresults08$DaysToElection <= 20, ]
 days_obama_model =lm(state_subset$obama.intmarg ~ DaysToElection, data = state_subset)
  slope_coefficient = coef(days_obama_model)["DaysToElection"]
  change[i] =slope_coefficient
}
round(change,4)
```

```
##               Alaska              Alabama             Arkansas
##               0.2164               0.0369               1.9410
##              Arizona           California             Colorado
##              -1.2206              -0.2875              -0.5126
##          Connecticut District of Columbia             Delaware
##              -0.4882              -0.0701              -0.1979
##              Florida              Georgia               Hawaii
##              -0.2777              -0.2116               0.0000
##                 Iowa                Idaho             Illinois
##              -0.6478              -0.0338              -0.0688
##              Indiana               Kansas             Kentucky
##              -3.8036               0.7883               0.3594
##            Louisiana        Massachusetts             Maryland
##               1.1121               0.0000              -0.2529
##                Maine             Michigan            Minnesota
##              -0.5534              -0.1716              -0.4708
##             Missouri          Mississippi              Montana
##               4.5612               0.2783              -0.6749
##       North Carolina         North Dakota             Nebraska
##              -2.5596               1.5582               0.2779
##        New Hampshire           New Jersey           New Mexico
##              -0.2344              -0.3391              -0.5294
##               Nevada             New York                 Ohio
##              -1.1973              -0.0406              -1.8116
##             Oklahoma               Oregon         Pennsylvania
##              -0.0919              -0.2644              -0.0665
##         Rhode Island       South Carolina         South Dakota
##              -0.3091               0.4457               0.6086
##            Tennessee                Texas                 Utah
##               0.7412               0.5810              -0.1423
##             Virginia              Vermont           Washington
##              -0.8231              -0.2344              -0.3000
##            Wisconsin        West Virginia              Wyoming
##              -0.2700               3.5456               0.2416
```

Distribution of the slope coefficients

```r
hist(change, freq=F, ylim = c(0,1), breaks = 15, col = "green3", xlab="Change in Obama's Market Margin"
```



Insert your answer here:In the last 20 days, most of the slope coefficients are around -0.5-0, indicating that as time went by, markets were changing their bets on who will win the election generally towards Obama.

Bonus Question 5 (20 points)

- Do the same plot as in question 3, but this time for all states **using a loop**. Specifically:
  - plot the relationship between Obama's market margin and days until election
  - add to the plot a line of best fit from the fitted linear model for this state
  - concentrate only on the observations 20 days before elections (including the 20th day)
- Make 1 graph with the first 25 states as separate plots (5 rows and 5 columns). Make sure to include the state name in each plot. You can do this inside the loop. Hint: the function `paste("some text")` might be useful here. You can find a similar example in one of our class sessions on loops and graphs (10 points)
- Make 1 graph with the remaining 26 plots (6 rows and 5 columns). Also here make sure to include the state name in each plot. (10 points)
- Hints:
  - you can use 2 separate loops for the 2 graphs.
  - first try to plot a few sates (say 4), to see whether your code works. To see the figure with all 25 states, knit the file. You will not be able to preview the figure with 25 plots in the console, as you will most probably get the error `"Error in plot.new() : figure margins too large"` - which means that the plot panel in RStudio is too small for the margins of the plot. You should see the plot in the knitted file. Do not hesitate to reach out if you have any questions here.

Answer 5

Plot with the first 25 states

```
# Make your plot in this R chunk, notice that this chunk includes some graph formatting
par(mfrow = c(5, 5))
for (i in states[1:25]) {

  twenty_days =  intresults08[intresults08$DaysToElection <= 20 & intresults08$statename== i ,]
lm_twenty_days = lm(obama.intmarg ~ DaysToElection, data=twenty_days)
plot(twenty_days$DaysToElection,twenty_days$obama.intmarg, xlim = c(25,0), ylim=c(-100,100), pch=16, col
abline(lm_twenty_days, col= "orangered")

}
```

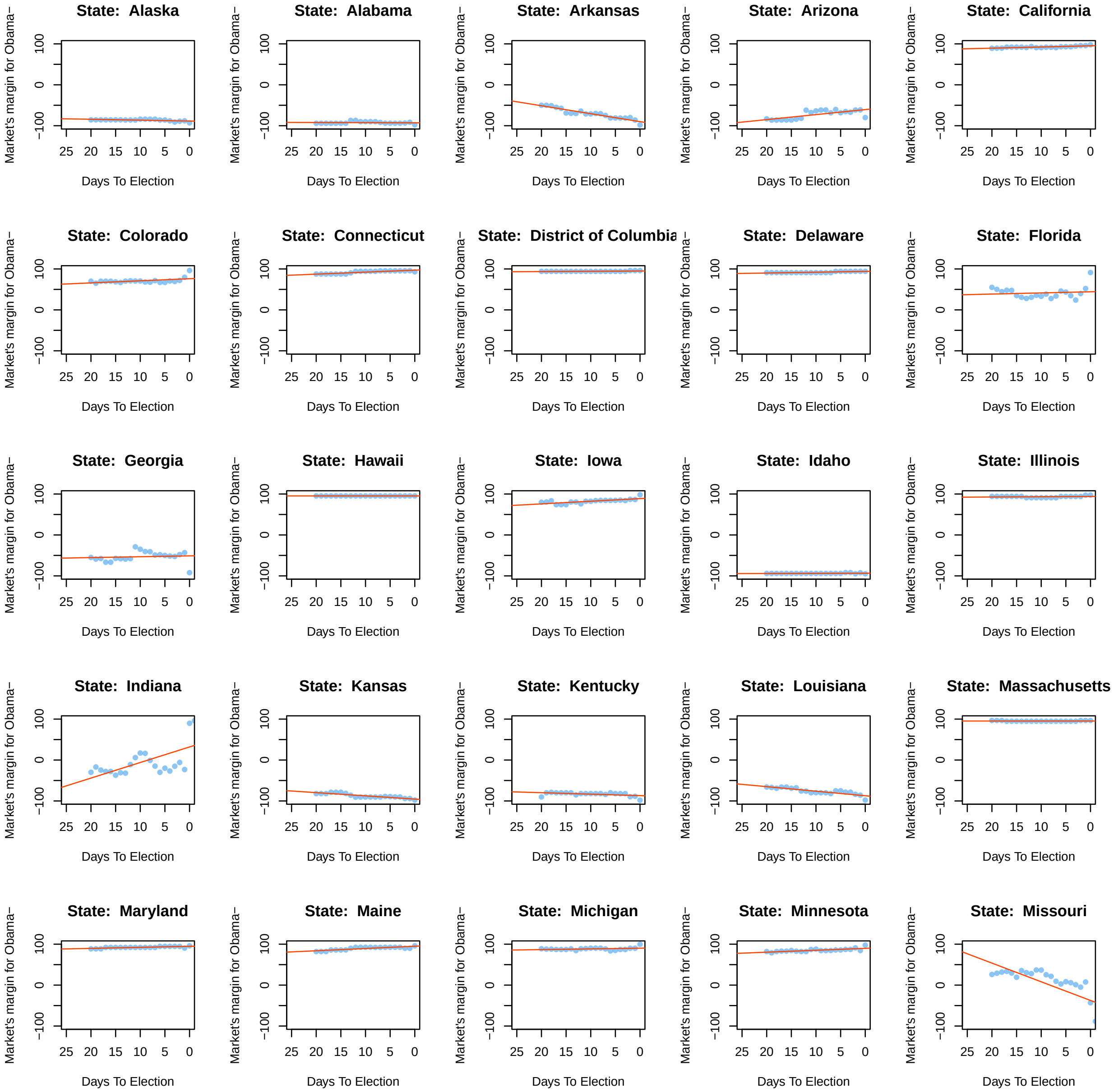
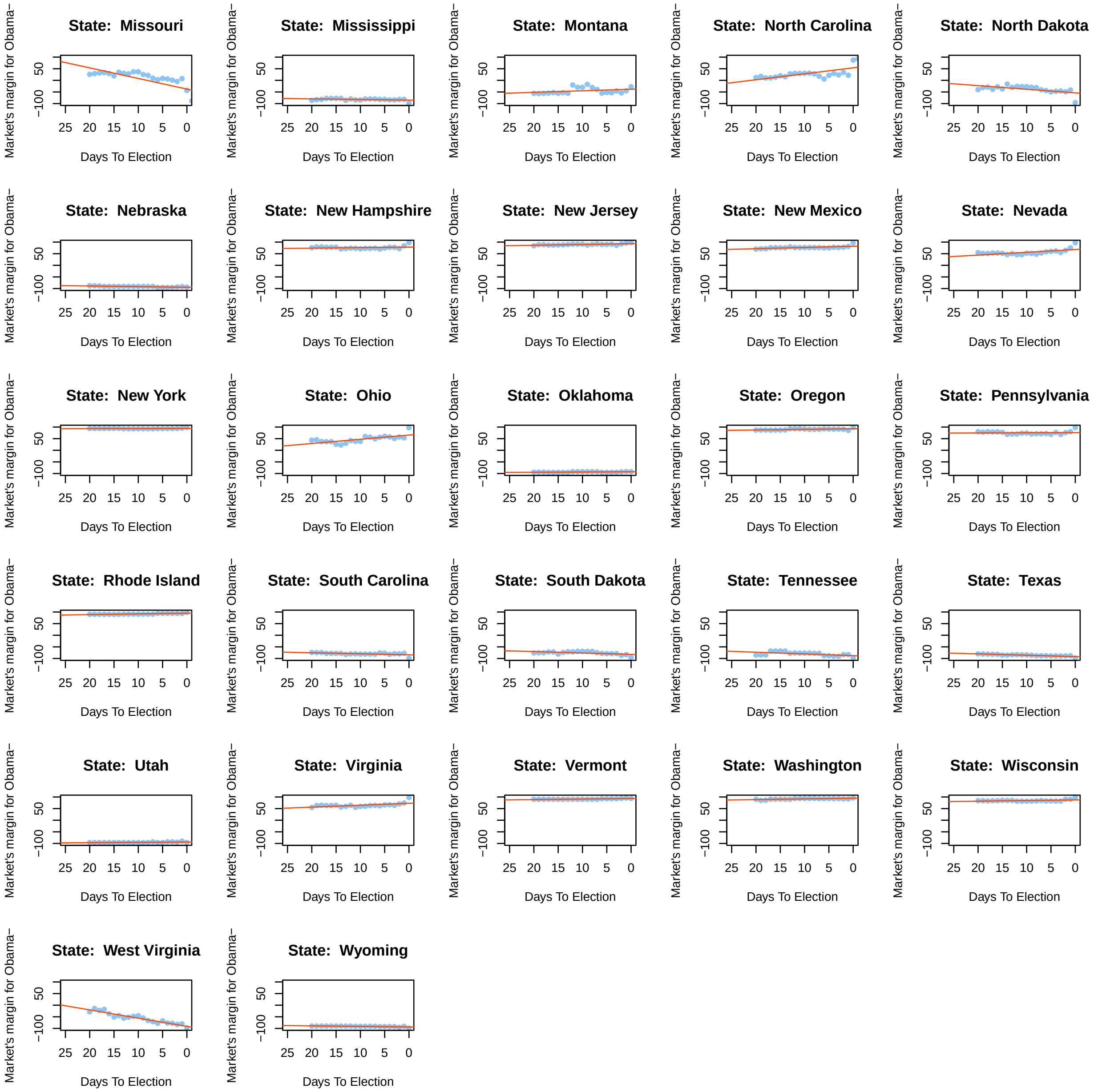


Figure with the remaining 26 plots

```
# Make your plot in this R chunk, notice that this chunk includes some graph formatting
par(mfrow = c(6, 5))
for (i in states[25:51]) {

  twenty_days =  intresults08[intresults08$DaysToElection <= 20 & intresults08$statename== i ,]
lm_twenty_days = lm(obama.intmarg ~ DaysToElection, data=twenty_days)
plot(twenty_days$DaysToElection,twenty_days$obama.intmarg, xlim = c(25,0), ylim=c(-100,100), pch=16, co
abline(lm_twenty_days, col= "orangered")

}
```



Question 6 (35 points)

In this question you are asked to predict the winner of the presidential election in USA for 2008 ***one week before the election*** using the *Intrade* data, where days to election is the predictor and the betting market margin is the outcome variable.

Use the *two weeks before* that moment (of one week prior to the election) to fit linear models for every state and make a prediction of the winner by state. You will need to use a loop to accomplish that.

Once you know who is the predicted winner in each state, you can calculate who is the predicted winner for the presidential election. Keep in mind that the winner of a given state receives all electoral college votes for this state (see variable `EV`). The presidential candidate who wins 270 out of 538 of the electoral college votes wins the presidential elections.

To accomplish this task follow the steps below:

- start by creating an empty container called `obama.predmarg`, where you will save your prediction about Obama's margin of victory in a given state. You will have such prediction for every state. Label the observations in the container with the state names. (5 points)
- Loop over each state:
  - Inside the loop make a subset of the data by selecting only observations with daily trading information done between 21 and 7 days before the election date for *one state.* (5 points)
  - Using this subset run a linear regression model of Obama's betting margin on days to election (5 points)
  - Use the model output to predict what will happen (who will be the winner) in this particular state **on election day**. In other words, use the model output for a given state to predict Obama's market margin of victory for the day of the election, hence when `DaysToElection` is equal to 0. Save these predictions in the empty container `obama.predmarg`. (Hint, you can use the `predict()` function.) (10 points)
    - Note: After the loop is done running, your container `obama.predmarg` will be filled with the predicted margin of victory for Obama for each state. If the predicted Obama margin is bigger than 0 for the election day, this means your model predicts that Obama will win. If it is below 0, this means that your model predicts that Obama will lose.

(Hint: be careful when you run the `predict()` function inside the loop. The `newdata` argument should contain your explanatory variables used in the model you pass to `predict()`. The `DaysToElection` variable is a `difftime` variable (you can see this by running `class(object_name$DaysToElection)`): recall that you created `DaysToElection` by subtracting the date of the betting session from the election date, where both dates were saves in *date format* using the function `as.Date()`. So, any value for `DaysToElection` in the `newdata` argument should also be of `difftime` class (difference in time). The easiest way to do so is to find the difference in time in your dataset (in your case 0 days difference), save this value in an object outside the loop, and then set the `DaysToElection` variable inside the `predict()` function to be equal to this object. Alternatively, you can create a difference in time equal to 0 by substracting the same date (in *date format*) from itself.

- Compare the prediction saved in `obama.predmarg` with the actual electoral results (`obama.actmarg`) using a plot. (Bonus 3 points)
- Make a contingency table summarizing the relationship between predicted and actual win of Obama across states. In particular, the table should show all 4 possible classifications: true positive (number of states for which Obama is predicted to win and actually won), false positive (number of states for which Obama is predicted to win, but lost in reality), true negative (number of states for which Obama is predicted to lose and actually lost), and false negative (number states for which Obama was predicted

to lose, but in reality won). (See the example in the answers file). Recall that margin of victory (in `obama.actmarg` and `obama.predmarg`) bigger than 0 indicates that Obama wins/is predicted to win. (Bonus 4 points)

- How well does the betting market model do predicting the election outcome? Calculate the number of electoral votes Obama was predicted to win. Compare these to the number of electoral votes Obama won in the 2008 presidential elections. (Hint: for this you need to consider the variable `EV`, which indicates the number of electoral college votes to be won for a given state. The candidate with most popular support wins all electoral college votes in a given state. (Bonus 3 points)

Answer 6

For your information: empty container for the predicted margin of victory for Obama for each state

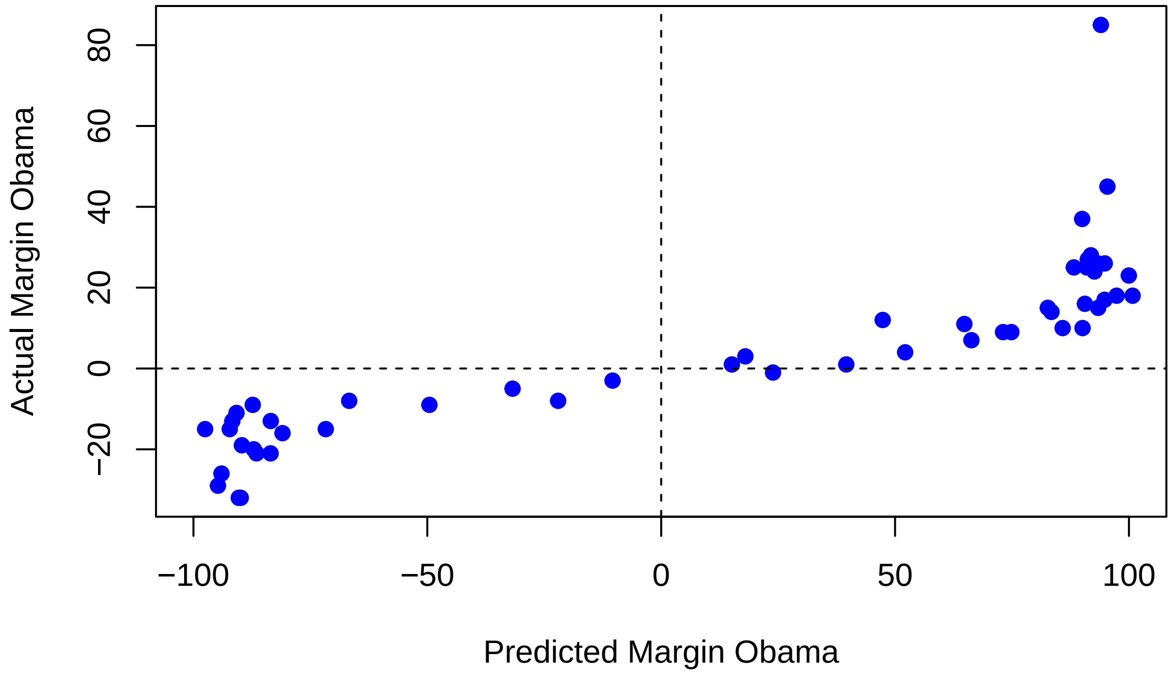
```
obama.predmarg = rep(NA, length(states))
names(obama.predmarg) = states
obama.predmarg
```

```
##               Alaska              Alabama             Arkansas 
##                   NA                   NA                   NA 
##              Arizona           California             Colorado 
##                   NA                   NA                   NA 
##          Connecticut District of Columbia             Delaware 
##                   NA                   NA                   NA 
##              Florida              Georgia               Hawaii 
##                   NA                   NA                   NA 
##                 Iowa                Idaho             Illinois 
##                   NA                   NA                   NA 
##              Indiana               Kansas             Kentucky 
##                   NA                   NA                   NA 
##            Louisiana        Massachusetts             Maryland 
##                   NA                   NA                   NA 
##                Maine             Michigan            Minnesota 
##                   NA                   NA                   NA 
##             Missouri          Mississippi              Montana 
##                   NA                   NA                   NA 
##       North Carolina         North Dakota             Nebraska 
##                   NA                   NA                   NA 
##        New Hampshire           New Jersey           New Mexico 
##                   NA                   NA                   NA 
##               Nevada             New York                 Ohio 
##                   NA                   NA                   NA 
##             Oklahoma               Oregon         Pennsylvania 
##                   NA                   NA                   NA 
##         Rhode Island       South Carolina         South Dakota 
##                   NA                   NA                   NA 
##            Tennessee                Texas                 Utah 
##                   NA                   NA                   NA 
##             Virginia              Vermont           Washington 
##                   NA                   NA                   NA
##            Wisconsin        West Virginia              Wyoming
##                   NA                   NA                   NA
```

Container filled with the linear model predictions for Obama's market margin of victory for the election day

```
 dte_0 = as.difftime (0, units = "days")
for (i in states) {
  subset_trading_info = intresults08[intresults08$DaysToElection <= 21 & intresults08$DaysToElection >=
  lm_subset_trading_info = lm(obama.intmarg ~ DaysToElection, data = subset_trading_info)
  predicted_margin = predict(lm_subset_trading_info, newdata = data.frame(DaysToElection =dte_0))



  obama.predmarg[i] = round(predicted_margin, 5)
}
```

```
obama_betteractmarg = tapply(intresults08$Obama - intresults08$McCain, intresults08$state, mean, na.rm =



plot(obama.predmarg, obama_betteractmarg, col = "blue", pch = 19,
     xlab = "Predicted Margin Obama", ylab = "Actual Margin Obama", xlim=c(-100,100))
abline(v=0, h=0, lty=2 )
```



Contingency Table

```
pred_win = obama.predmarg >=0
actual_win = obama_betteractmarg >=0
table (pred_win, actual_win)
```

```
##          actual_win
## pred_win FALSE TRUE
##    FALSE    21    0
##    TRUE      1   29
```

Insert your answer here: Not a lot of false positives! The predictions are generally true, whether Obama loses or not the market is quite successful in determining it.

Who is predicted to win: Obama or McCain? Number of electoral votes predicted to be won by Obama

```
pres08$predicted = ifelse(obama.predmarg > 0, 1, 0)
pres08$actual = ifelse(pres08$Obama - pres08$McCain >0,1,0)

sum(pres08$EV[pres08$predicted == 1])
```

```
## [1] 347
```

Number of electoral votes actually won by Obama

```
sum(pres08$EV[pres08$actual == 1])
```

```
## [1] 364
```

Insert your answer here: Of course, the model is not perfect. Even though the market is quite successful in determining the winner (the bigger picture), this does not mean that it would determine electoral votes accurately. Perhaps the difference between the predicted and actual EV is due to the false positive that we could also notice on the table above.

Evaluation

- 5 questions for a total of 100 points
- 2 bonus question for 30 points (See question 5 and parts of question

  6)

Code Usage examples

`merge` function

|:—— *Example for merge usage, you can skip this part if you know how to use the merge function*|:——|

```
- Let us illustrate the `merge function` using a simple hypothetical example
  - in this hypothetical example we will create two datasets
    - the first dataset has 4 individuals with a unique ID number and Name
    - the second dataset has 4 individuals with a unique ID number and age
    - identical id numbers in both dataset refer to the same individual
  - we would like to combine both datasets. Note that the datasets do not contain the entirley the same
  - we can combine the two datasets using the function merge and the unique ID variable
  - consider the below code
```

```r
data1 <- data.frame(ID = c(1, 2, 3, 4),
                    Name = c("Alice", "Bob", "Charlie", "David"))

data2 <- data.frame(ID.variable = c(2, 3, 4, 5),
                    Age = c(25, 30, 35, 22))

combined_data <- merge(x = data1, y = data2, by.x = "ID", by.y = "ID.variable", all.x = TRUE)
combined_data
```

```
##   ID    Name Age
## 1  1   Alice  NA
## 2  2     Bob  25
## 3  3 Charlie  30
## 4  4   David  35
```

```
  - the first argument in `merge` is x; pass to it the first dataset
  - the second argument in `merge` is y; pass to it the second dataset
  - by.x specified the variable in the first dataset (the data you passed to argument x) on which you w
  - by.y specified the merging variable in the second dataset (the data you passed to argument y). Here
  - all.x = TRUE, states that you want to keep all observations form the first dataset (the x dataset)
```

|:——————– End of merge example|:———————————————|

as.Date() function

|:—— *Example* ***as.Date()*** *function, you can skip this part if you know how to use this function*|:——|

```
- The first argument in the `as.Date()` function is called `x`,
    here you need to pass an object with the dates to be converted.
- The second argument `format`, describes the date format in your
    object with dates. In our case the dates come in the format
    `2008-02-28` or `YYYY-MM-DD`. To communicate this to R you need
    to set the argument `format = "%Y-%m-%d"`.
- If the format was `28.02.2008` or `DD.MM.YYYY` then you would set
    `format = "%d.%m.%Y"`. If the format was `28/02/08` or
    `DD/MM/YY` then you would set `format = "%d/%m/%y"`.
- Notice that when the year has 4 digits (e.g. YYYY or 2008) then %Y is
    written with a capital letter. When the year is denoted by 2
    digits (e.g. YY or 08), then we need to use lower case %y
    instead).
```

```r
# Example
date <- "28.02.2008" # date but saveds as a string\
data_transformed <- as.Date(x = date, format = "%d.%m.%Y")
class(data_transformed)
```

```
## [1] "Date"
```

```r
# The difference between two dates has a `difftime` format.
difference_between_two_dates <- data_transformed - data_transformed
difference_between_two_dates
```

```
## Time difference of 0 days
```

```r
class(difference_between_two_dates)
```

```
## [1] "difftime"
```

|:————————————- End of `as.Date` example|:——————————————————|