

ELECTION FORECASTING REVISITED

In the recitation from Week 3, we used logistic regression on polling data in order to construct US presidential election predictions. We separated our data into a training set, containing data from 2004 and 2008 polls, and a test set, containing the data from 2012 polls. We then proceeded to develop a logistic regression model to forecast the 2012 US presidential election.

In this homework problem, we'll revisit our logistic regression model from Week 3, and learn how to plot the output on a map of the United States. Unlike what we did in the Crime lecture, this time we'll be plotting predictions rather than data!

First, load the ggplot2, maps, and ggmap packages using the library function. All three packages should be installed on your computer from lecture, but if not, you may need to install them too using the install.packages function.

Then, load the US map and save it to the variable statesMap, like we did during the Crime lecture:

```
statesMap = map_data("state")
```

The maps package contains other built-in maps, including a US county map, a world map, and maps for France and Italy.

PROBLEM 1.1 - DRAWING A MAP OF THE US (1/1 point)

If you look at the structure of the statesMap data frame using the str function, you should see that there are 6 variables. One of the variables, group, defines the different shapes or polygons on the map. Sometimes a state may have multiple groups, for example, if it includes islands. How many different groups are there?

Answer: 63



EXPLANATION

You can count the number of different values of the group variable by using the command `table(statesMap$group)`. There are 63 different values.

Alternatively, you could use the command `length(table(statesMap$group))` as a shortcut to counting the number of groups in the table output.

The variable "order" defines the order to connect the points within each group, and the variable "region" gives the name of the state.

Hide Answer

You have used 1 of 3 submissions


PROBLEM 1.2 - DRAWING A MAP OF THE US (1/1 point)

You can draw a map of the United States by typing the following in your R console:

```
ggplot(statesMap, aes(x = long, y = lat, group = group)) + geom_polygon(fill = "white", color = "black") + coord_map("mercator")
```

As we saw in lecture, we used the option `coord_map`. This sets the type of map projection. The option "mercator" just creates a flat projection. Other options include "orthographic" and "polyconic". To see all options, type `?mapproject` in your R console.

We specified two colors in `geom_polygon` -- fill and color. Which one defined the color of the outline of the states?

- ☐ fill
- ☒ color 
- ☐ Neither

EXPLANATION

In our plot, the states are outlined in black, which is the color we specified for the option "color". To confirm that this is changing the outline color of the states, you can try re-running the command with a different color:

```
ggplot(statesMap, aes(x = long, y = lat, group = group)) + geom_polygon(fill = "white", color = "pink") + coord_map("mercator")
```

Hide Answer

You have used 1 of 1 submissions

PROBLEM 2.1 - COLORING THE STATES BY PREDICTIONS (2/2 points)

Now, let's color the map of the US according to our 2012 US presidential election predictions from the Week 3 Recitation. We'll rebuild the model here, using the dataset [PollingImputed.csv \(/c4x/MITx/15.071x/asset/PollingImputed.csv\)](/c4x/MITx/15.071x/asset/PollingImputed.csv). Be sure to use this file so that you don't have to redo the imputation to fill in the missing values, like we did in the Week 3 Recitation.

Load the data using the `read.csv` function, and call it "polling". Then split the data using the `subset` function into a training set called "Train" that has observations from 2004 and 2008, and a testing set called "Test" that has observations from 2012.

Note that we only have 45 states in our testing set, since we are missing observations for Alaska, Delaware, Alabama, Wyoming, and Vermont, so these states will not appear colored in our map.

Then, create a logistic regression model and make predictions on the test set using the following commands:

```
mod2 = glm(Republican~SurveyUSA+DiffCount, data=Train, family="binomial")
```

```
TestPrediction = predict(mod2, newdata=Test, type="response")
```

`TestPrediction` gives the predicted probabilities for each state, but let's also create a vector of Republican/Democrat predictions by using the following command:

```
TestPredictionBinary = as.numeric(TestPrediction > 0.5)
```

Now, put the predictions and state labels in a `data.frame` so that we can use `ggplot`:

```
predictionDataFrame = data.frame(TestPrediction, TestPredictionBinary, Test$State)
```

To make sure everything went smoothly, answer the following questions.

For how many states is our binary prediction 1, corresponding to Republican?

Answer: 22

What is the average predicted probability of our model?

Answer: 0.4852626

EXPLANATION

You can create the data frame predictionDataFrame by running the following lines of R code:

```
polling = read.csv("PollingImputed.csv")
```

```
Train = subset(polling, Year < 2012)
```

```
Test = subset(polling, Year == 2012)
```

```
mod2 = glm(Republican~SurveyUSA+DiffCount, data=Train, family="binomial")
```

```
TestPrediction = predict(mod2, newdata=Test, type="response")
```

```
TestPredictionBinary = as.numeric(TestPrediction > 0.5)
```

```
predictionDataFrame = data.frame(TestPrediction, TestPredictionBinary, Test$State)
```

You can answer the two questions with the functions table(TestPredictionBinary) and mean(TestPrediction).

Hide Answer

You have used 1 of 3 submissions

PROBLEM 2.2 - COLORING THE STATES BY PREDICTIONS (2/2 points)

Now, we need to merge "predictionDataFrame" with the map data "statesMap", like we did in lecture. Before doing so, we need to convert the Test.State variable to lowercase, so that it matches the region variable in statesMap. Do this by typing the following in your R console:

```
predictionDataFrame$region = tolower(predictionDataFrame$Test.State)
```

Now, merge the two data frames using the following command:

```
predictionMap = merge(statesMap, predictionDataFrame, by = "region")
```

Lastly, we need to make sure the observations are in order so that the map is drawn properly, by typing the following:

```
predictionMap = predictionMap[order(predictionMap$order),]
```

How many observations are there in predictionMap?

Answer: 15034

How many observations are there in statesMap?

Answer: 15537

EXPLANATION

If you type str(predictionMap), you should see that there are 15034 observations, and if you type str(statesMap) you should see that there are 15537 observations.

Hide Answer

You have used 2 of 3 submissions

PROBLEM 2.3 - COLORING THE STATES BY PREDICTIONS (1/1 point)

When we merged the data in the previous problem, it caused the number of observations to change. Why? Check out the help page for merge by typing ?merge to help you answer this question.

- ☐ Merging the data just combines the two data frames like it would if we used rbind, so the number of observations increased.
- ☐ We have more observations for each state now, because some observations have the statesMap data, and some observations have the prediction data.
- ☒ Because we only make predictions for 45 states, we no longer have observations for some of the states. These observations were removed in the merging process. ✓
- ☐ We merged the observations for which our predictions are identical.

EXPLANATION

When we merge data, it only merged the observations that exist in both data sets. So since we are merging based on the region variable, we will lose all observations that have a value of "region" that doesn't exist in both data frames. You can change this default behavior by using the all.x and all.y arguments of the merge function. For more information, look at the help page for the merge function by typing ?merge in your R console.

Hide Answer

You have used 1 of 1 submissions

PROBLEM 2.4 - COLORING THE STATES BY PREDICTIONS (1/1 point)

Now we are ready to color the US map with our predictions! You can color the states according to our binary predictions by typing the following in your R console:

```
ggplot(predictionMap, aes(x = long, y = lat, group = group, fill = TestPredictionBinary)) + geom_polygon(color = "black")
```

The states appear light blue and dark blue in this map. Which color represents a Republican prediction?

- ☒ Light blue ✓
- ☐ Dark blue

EXPLANATION

Our logistic regression model assigned 1 to Republican and 0 to Democrat. As we can see from the legend, 1 corresponds to a light blue color on the map and 0 corresponds to a dark blue color on the map.

Hide Answer

You have used 1 of 1 submissions

PROBLEM 2.5 - COLORING THE STATES BY PREDICTIONS (1/1 point)

We see that the legend displays a blue gradient for outcomes between 0 and 1. However, in our model there are only two possible outcomes: 0 or 1. Let's replot the map with discrete outcomes. We can also change the color scheme to blue and red, to match the blue color associated with the Democratic Party in the US and the red color associated with the Republican Party in the US. This can be done with the following command:

```
ggplot(predictionMap, aes(x = long, y = lat, group = group, fill = TestPredictionBinary)) + geom_polygon(color = "black") +  
scale_fill_gradient(low = "blue", high = "red", guide = "legend", breaks = c(0,1), labels = c("Democrat", "Republican"), name = "Prediction  
2012")
```

Alternatively, we could plot the probabilities instead of the binary predictions. Change the plot command above to instead color the states by the variable TestPrediction. You should see a gradient of colors ranging from red to blue. How many states look purple in the map?

- ☒ 1 ✓
- ☐ 2
- ☐ 3

EXPLANATION

This plot can be generated by using the command:

```
ggplot(predictionMap, aes(x = long, y = lat, group = group, fill = TestPrediction)) + geom_polygon(color = "black") +
scale_fill_gradient(low = "blue", high = "red", name = "Prediction 2012")
```



The only state that appears purple is the state of Iowa.

[Hide Answer](#)

You have used 1 of 2 submissions

PROBLEM 3.1 - UNDERSTANDING THE PREDICTIONS (1 point possible)

What does it mean when a state appears with a purple "in-between" color on the map?

- ☐ Our logistic regression classifier failed to classify all states for which we have data
- ☐ Our logistic regression classifier is not as confident in its prediction for this state 
- ☒ This state was a close race in the 2012 election 

EXPLANATION


Our logistic regression classifier did classify each state for which we have data. However, a state will only appear bright red if the logistic regression probability was close to 1, and will only appear bright blue if the logistic regression probability was close to 0. An in-between color signifies less confidence in the prediction. This is a good way to visualize uncertainty. Although Iowa, the state that appears purple here, was a hard state for us to predict, we don't know whether or not it was a close race in the 2012 election. The color only represents what our model thought.

[Hide Answer](#)

You have used 1 of 1 submissions

PROBLEM 3.2 - UNDERSTANDING THE PREDICTIONS (1/1 point)

In the 2012 election, the state of Florida ended up being a very close race. It was ultimately won by the Democratic party. Did we predict this state correctly or incorrectly? To see the names and locations of the different states, take a look at the World Atlas map [here](http://www.worldatlas.com/webimage/testmaps/usanames.htm) (<http://www.worldatlas.com/webimage/testmaps/usanames.htm>).

- ☐ We correctly predicted that this state would be won by the Democratic party.
- ☒ We incorrectly predicted this state by predicting that it would be won by the Republican party. 

EXPLANATION

In our prediction map, the state of Florida is colored red, meaning that we predicted Republican. So we incorrectly predicted this state.

[Hide Answer](#)

You have used 1 of 1 submissions

PROBLEM 3.3 - UNDERSTANDING THE PREDICTIONS (2/2 points)

What was our predicted probability for the state of Florida?

Answer: 0.9640395

EXPLANATION

You can find the predicted probability for Florida by typing `predictionDataFrame` in your R console, and finding that Florida is the 6th observation, and then finding the 6th probability in the column `TestPrediction`.

What does this imply?

- ☐ Our prediction model did a good job of correctly predicting the state of Florida, and we were very confident in our prediction.
- ☐ Our prediction model did a good job of correctly predicting the state of Florida, but we were not very confident in the prediction.
- ☐ Our prediction model did not do a very good job of correctly predicting the state of Florida, but we were not very confident in our prediction.
- ☒ Our prediction model did not do a very good job of correctly predicting the state of Florida, and we were very confident in our incorrect prediction. ✓

EXPLANATION

We predicted Republican for the state of Florida with high probability, meaning that we were very confident in our incorrect prediction! Historically, Florida is usually a close race, but our model doesn't know this. The model only uses polling results for the particular year. For Florida in 2012, Survey USA predicted a tie, but other polls predicted Republican, so our model predicted Republican.

Hide Answer

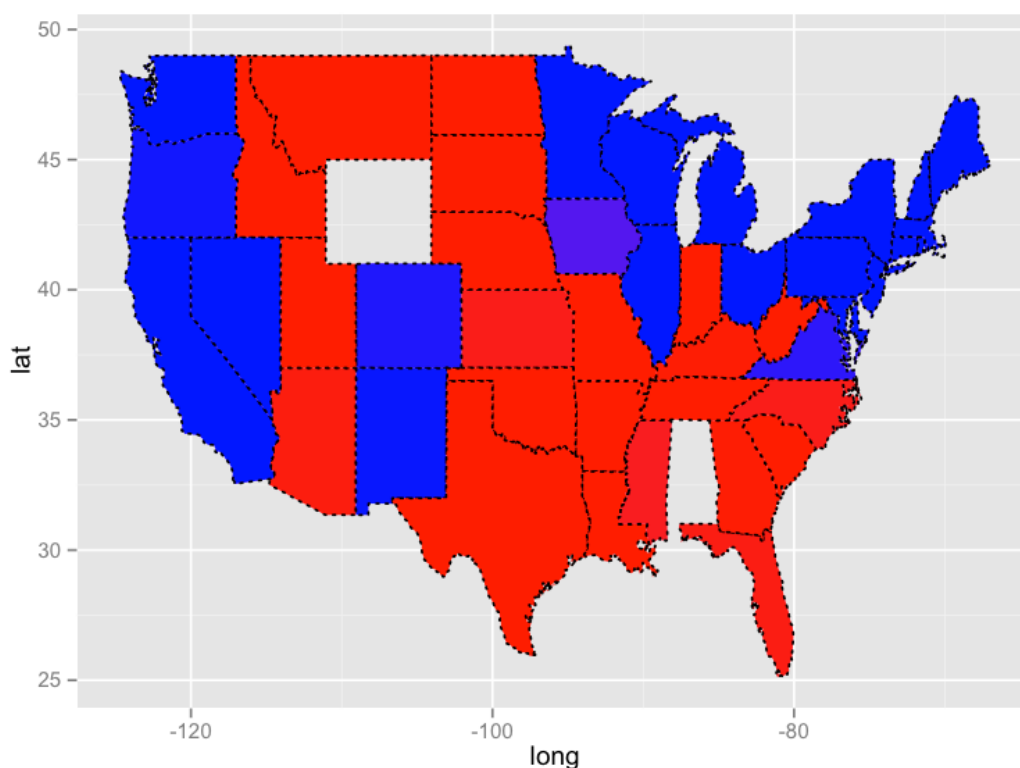
You have used 1 of 2 submissions

PROBLEM 4 - PARAMETER SETTINGS

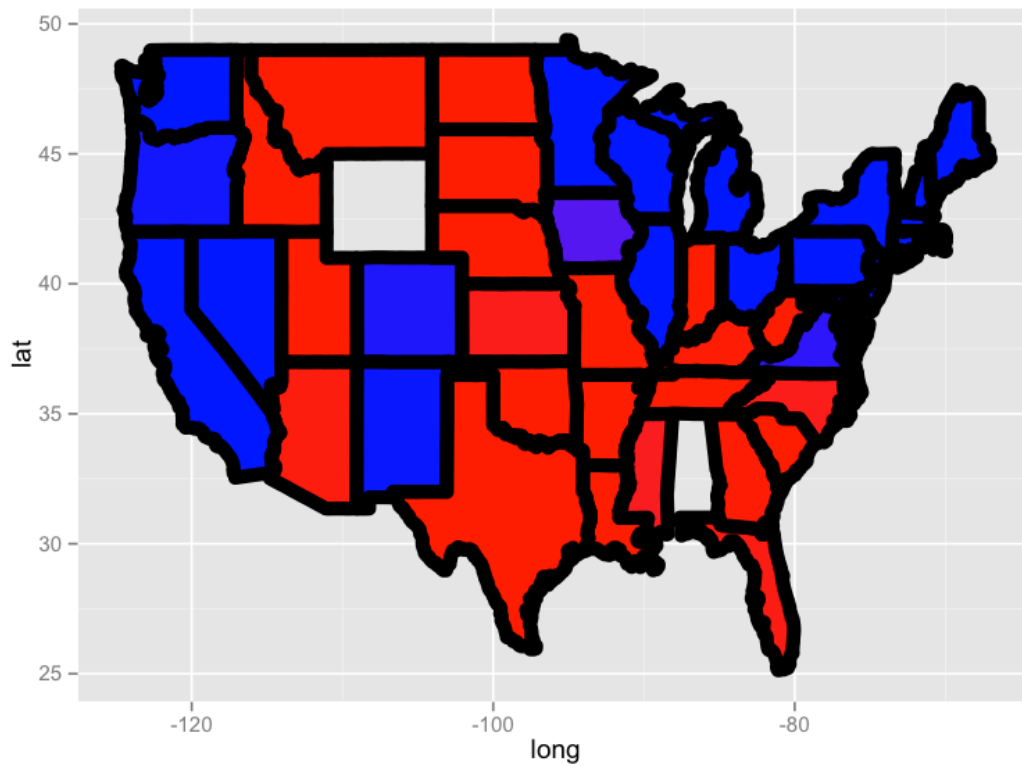
In this problem, we'll explore what the different parameter settings of `geom_polygon` do. Throughout the problem, use the help page for `geom_polygon`, which can be accessed by `?geom_polygon`. To see more information about a certain parameter, just type a question mark and then the parameter name to get the help page for that parameter. Experiment with different parameter settings to try and replicate the plots!

We'll be asking questions about the following three plots:

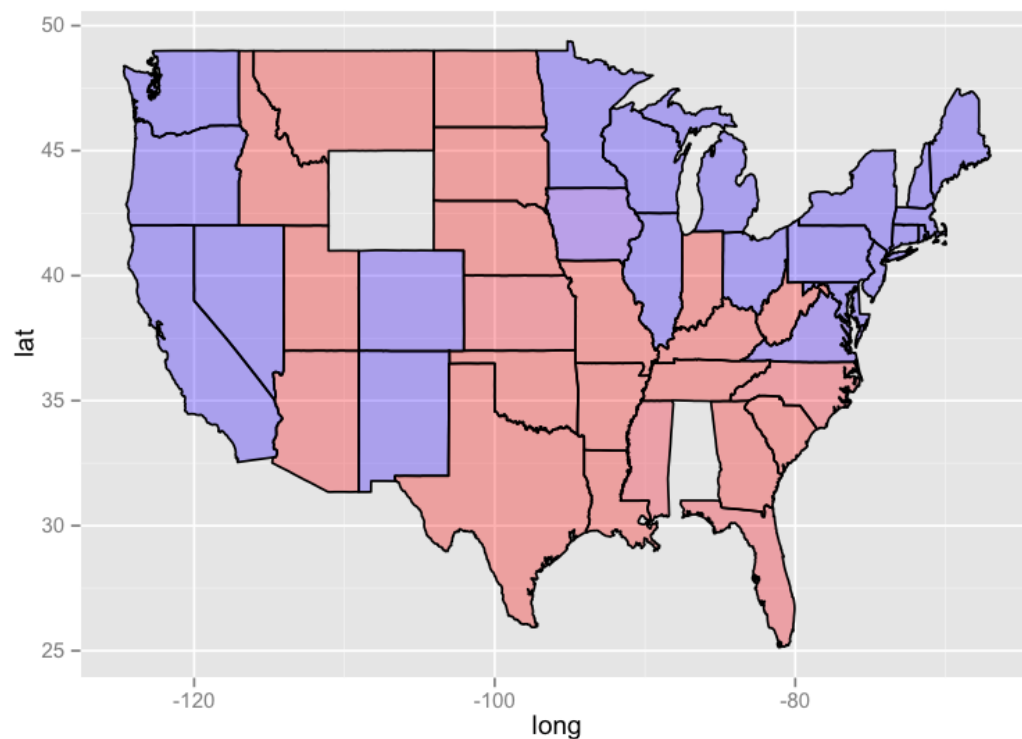
Plot (1)



Plot (2)



Plot (3)



PROBLEM 4.1 - PARAMETER SETTINGS (2/2 points)

Plots (1) and (2) were created by setting different parameters of `geom_polygon` to the value 3.

What is the name of the parameter we set to have value 3 to create plot (1)?

linetype

Answer: linetype

What is the name of the parameter we set to have value 3 to create plot (2)?

size

Answer: size

EXPLANATION

The first plot can be generated by setting the parameter `linetype = 3`:

```
ggplot(predictionMap, aes(x = long, y = lat, group = group, fill = TestPrediction))+ geom_polygon(color = "black", linetype=3) +  
scale_fill_gradient(low = "blue", high = "red", guide = "legend", breaks= c(0,1), labels = c("Democrat", "Republican"), name =  
"Prediction 2012")
```

The second plot can be generated by setting the parameter `size = 3`:

```
ggplot(predictionMap, aes(x = long, y = lat, group = group, fill = TestPrediction))+ geom_polygon(color = "black", size=3) +  
scale_fill_gradient(low = "blue", high = "red", guide = "legend", breaks= c(0,1), labels = c("Democrat", "Republican"), name =  
"Prediction 2012")
```

Hide Answer

You have used 1 of 3 submissions

PROBLEM 4.2 - PARAMETER SETTINGS (1/1 point)

Plot (3) was created by changing the value of a different `geom_polygon` parameter to have value 0.3. Which parameter did we use?

alpha

Answer: alpha

EXPLANATION

Plot (3) can be created by changing the `alpha` parameter:

```
ggplot(predictionMap, aes(x = long, y = lat, group = group, fill = TestPrediction))+ geom_polygon(color = "black", alpha=0.3) +  
scale_fill_gradient(low = "blue", high = "red", guide = "legend", breaks= c(0,1), labels = c("Democrat", "Republican"), name =  
"Prediction 2012")
```

The "`alpha`" parameter controls the transparency or darkness of the color. A smaller value of `alpha` will make the colors lighter.

Hide Answer

You have used 1 of 3 submissions

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion

 New Post



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy \(https://www.edx.org/edx-privacy-policy\)](https://www.edx.org/edx-privacy-policy)