Help

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 point possible)

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?

[   ]

[   ]

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

[Show Answer]   *You have used 0 of 3 submissions*

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

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

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

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). 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?

[   ]

[   ]

What is the average predicted probability of our model?

[   ]

[   ]

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

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?

How many observations are there in statesMap?

Show Answer   *You have used 0 of 3 submissions*

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

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.

Show Answer   *You have used 0 of 1 submissions*

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

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

Show Answer | *You have used 0 of 1 submissions*

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

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

○ 4

Show Answer | *You have used 0 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

Show Answer | *You have used 0 of 1 submissions*

PROBLEM 3.2 - UNDERSTANDING THE PREDICTIONS (1 point possible)

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).

○ 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.

Show Answer | *You have used 0 of 1 submissions*

PROBLEM 3.3 - UNDERSTANDING THE PREDICTIONS (2 points possible)

What was our predicted probability for the state of Florida?

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.
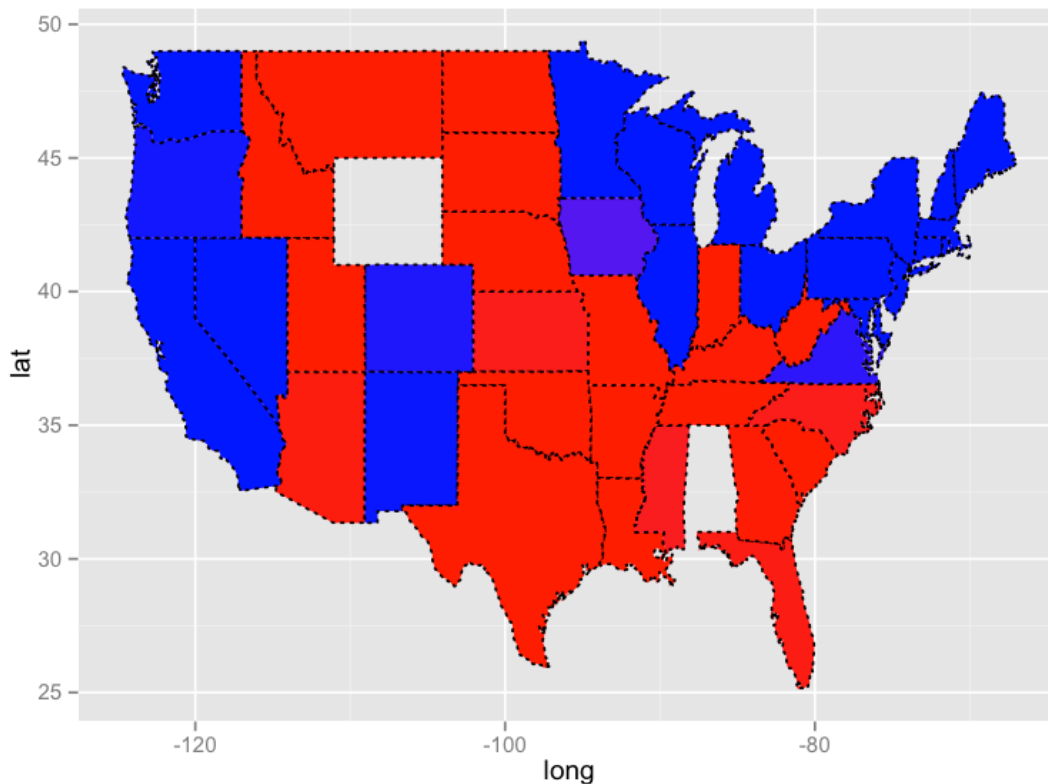
Show Answer   *You have used 0 of 2 submissions*

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 points possible)

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)?

[   ]

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

[   ]

Show Answer   *You have used 0 of 3 submissions*

PROBLEM 4.2 - PARAMETER SETTINGS (1 point possible)

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

[   ]

Show Answer   *You have used 0 of 3 submissions*

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

Show Discussion   ✏ **New Post**