Math 242: Polling Project Part 2, Due Monday, October 3

In Part 1, we learned about visualizing the race and creating confidence intervals from polls. We also learned that there are a lot of ways bias can seep into our results. To mitigate the effects of these biases, it makes sense to combine results from several polls. This can be done via a simple regression model, where each poll acts as a different explanatory variable. Suppose that a poll is a number, telling how much Clinton is up over Trump (negative if Trump is up over Clinton). For example, if x1 = 3, this means the first poll found that Clinton was up by 3 points.

1. An extremely simplistic idea would be to plot pairs (t,x) where x is the poll value as above and t is the date of the poll (for simplicity, say t = 0 is Jan. 1, 2016 and t is measured in days), then to fit a simple linear regression to this line and predict the election based on the value of x when we plug t = 312 (the 312th day of the year is November 8, election day). Is there anything wrong with this approach?
2. A different approach is to ignore time and try to combine different polls to account for sampling variability and bias. Write out the general model form (both the version with the “~” and the version where you name all the coefficients) of a model for how much Clinton is up, using each of 3 different polls x1, x2, and x3, without interaction:
3. What would it mean, in real world terms, to allow the explanatory variables to interact for your model above? Do you think that’s a good modeling decision or a bad decision (i.e. which choice would give a better model)?
4. Suppose you created two models, one with interaction and one without. Suppose you also had a fourth poll you did not use to create the models. Describe, following the language in chapters 8 and 9, how you could use the fourth poll to see which model was better.
5. Perhaps we should consider weighting the different polls xi. What factors do you think should go into deciding which poll gets more weight and which less weight?
6. For now, let’s focus on trustworthiness of the polls and date the polls were conducted. Note that fivethirtyeight.com gives grades to the various polls:

<http://projects.fivethirtyeight.com/2016-election-forecast/national-polls/>

Describe how you could use these rankings to make a more complicated model that you think would fit better. Assume again that you have 3 polls x1, x2, x3 and they have trustworthiness scores w1, w2, w3 where 0 means totally untrustworthy and 1 means totally trustworthy. Write out the best model you can:

1. Now suppose you wanted to weight the polls by date instead of trustworthiness. Let t1, t2, t3 be the days of the year the polls were conducted (as in problem 1). Write out a model that factors in these times as weights for the 3 polls:
2. Senate races are a bit easier to model than the presidential race, because there isn’t an electoral college (the 17th amendment assures “direct election” by the voters). The methodology of fivethirtyeight.com for senate races is described at:

<http://fivethirtyeight.blogs.nytimes.com/methodology/>

Read that page and identify as many factors as you can that it seems they are using as weights for different polls.

1. What is the Partisan Voting Index?
2. What other terms go into the regression model? Which terms do you think are most important? Can you think of other terms that should go in?
3. How do you feel about the fact that election predictors and pollsters (whose words are often reported by the media) are allowed to make subjective decisions adjusting the polls? Is there any way to avoid that subjectivity? To see how much their modeling decisions matter, read the following article: <http://www.nytimes.com/interactive/2016/09/20/upshot/the-error-the-polling-world-rarely-talks-about.html?r=0>
4. In Step 6, they say they don’t rely on “any sort of theoretical calculation of the margin of error.” What theoretical calculations are they talking about? How does this relate to our discussion of parametric methods vs. bootstrapping? Note that some parametric methods are still being used, e.g. the gamma distribution in Step 7 (this is an example of a probability model, as in Chapter 11 of our book).
5. In Question (1), you were probably concerned about Extrapolation. In Step 5, Nate Silver suggests his models actually can predict the future, under certain circumstances. Comment on what you believe it would take in order to be able to predict an election 2 weeks ahead of time, i.e. under what circumstances would it be possible? How about 6 weeks ahead of time (i.e. our situation today).