HW2P3

September 21, 2016

0.0.1 Homework 2: More Exploratory Data Analysis

0.1 Gene Expression Data and Election Polls

Due: Thursday, September 29, 2016 11:59 PM

Submission Instructions To submit your homework, create a folder named **lastname_firstinitial_hw#** and place your IPython notebooks, data files, and any other files in this folder. Your IPython Notebooks should be completely executed with the results visible in the notebook. We should not have to run any code. Make sure to share the private repo with my github account (mdog) and submit the repo path through blackboard.

0.2 Introduction

John Tukey wrote in Exploratory Data Analysis, 1977: "The greatest value of a picture is when it forces us to notice what we never expected to see." In this assignment we will continue using our exploratory data analysis tools, but apply it to new sets of data: gene expression and polls from the 2012 Presidental Election and from the 2014 Senate Midterm Elections.

First: You will use exploratory data analysis and apply the singular value decomposition (SVD) to a gene expression data matrix to determine if the the date that the gene expression samples are processed has large effect on the variability seen in the data.

Second: You will use the polls from the 2012 Presidential Elections to determine (1) Is there a pollster bias in presidential election polls? and (2) Is the average of polls better than just one poll?

Finally: You will use the HuffPost Pollster API to extract the polls for the current 2014 Senate Midterm Elections and provide a preliminary prediction of the result of each state.

Data We will use the following data sets:

1. A gene expression data set called `exprs_GSE5859.csv` and sample annotation table called `sampleinfo_GSE5859.csv` which are both available on Github in the 2014_data repository: expression data set and sample annotation table.

2. Polls from the 2012 Presidential Election: Barack Obama vs Mitt Romney. The polls we will use are from the Huffington Post Pollster.

3. Polls from the 2014 Senate Midterm Elections from the HuffPost Pollster API.

0.3 Load Python modules

```
In [1]: # special IPython command to prepare the notebook for matplotlib
        %matplotlib inline

        import requests
        from io import StringIO
        import numpy as np
        import pandas as pd # pandas
        import matplotlib.pyplot as plt # module for plotting
        import datetime as dt # module for manipulating dates and times
        import numpy.linalg as lin # module for performing linear algebra operation
```

0.4 Problem 3: Is the average of polls better than just one poll?

Problem 3(a) Most undecided voters vote for one of the two candidates at the election. Therefore, the reported percentages underestimate the final value of both candidates. However, if we assume the undecided will split evenly, then the observed difference should be an unbiased estimate of the final difference.

Add a new column to the `election` DataFrame containg the difference between Obama and Romeny called `Diff`.

```
In [31]: #your code here
```

Problem 3(b) Make a plot of the differences for the week before the election (e.g. 5 days) where the days are on the x-axis and the differences are on the y-axis. Add a horizontal line showing 3.9%: the difference between Obama and Romney on election day.

```
In [32]: #your code here
```

Problem 3(c) Make a plot showing the differences by pollster where the pollsters are on the x-axis and the differences on the y-axis.

```
In [33]: #your code here
```

Is the *across poll* difference larger than the *between pollster* difference?
Answer:

Problem 3(d) Take the average for each pollster and then compute the average of that. Given this difference how confident would you have been of an Obama victory?

Hint: Compute an estimate of the SE of this average based exclusively on the observed data.

```
In [34]: #your code here
```

Answer:

Problem 3(e) Show the difference against time and see if you can detect a trend towards the end. Use this trend to see if it improves the final estimate.

```
In [35]: #your code here
```

Answer:

0.5 Discussion for Problem 3

Write a brief discussion of your conclusions to the questions and tasks above in 100 words or less.

---

1 Submission Instructions

To submit your homework, create a folder named **lastname_firstinitial_hw#** and place your IPython notebooks, data files, and any other files in this folder. Your IPython Notebooks should be completely executed with the results visible in the notebook. We should not have to run any code. Make sure to share the private repo with my github account (mdog) and submit the repo path through blackboard.

```
In [ ]:
```