Stat 133 Final Project: Part 2 (Due: 8/13 6pm)

July 28, 2014

For the second part of your project, you will explore the linguistic data that you cleaned and reformatted. Your next task is to answer the question of how strongly linguistic trends are tied to geography. This question is fairly open ended, and you are free to eplore it in a wide variety of ways. It is meant to mimic a real-world applied data analysis project, so we are not giving you specific tasks that you are required to complete. We expect that you will examine the data thoroughly and present your findings to us. To help get you started, we have compiled a list of several questions that you might address in your report. You will find many of these questions lead to others. Feel free to explore these new questions as they come up (even if that means not looking at some of the other questions we have provided).

- Pick several survey questions and examine their relationship with each other and geography. Can responses to one question be used to predict the responses to another?
- Experiment with dimension reduction using PCA. How does scaling the variables affect your results? Do certain questions seem to be given more weight in different principal components? Which ones? You will probably want to restrict this exploration to the first few PCs.
- Use clustering methods we discussed to try and group the observations based on their responses. How well do the groups correspond to geography as you change k? If the clustering corresponds to geography are there distinct boundaries or a continuum? Which questions do the groups seem to separate along?
- How stable are the clusters? What happens when you try to cluster using the original dataset? A dataset reduced via PCA? How does cluster stability change as you reduce the number of PCs? Try taking a subsample of the dataset. Does the link between clusters and geography change as you use different subsamples? Is the relationship stable for certain values of k?

Please present your findings in a writeup of no more than 2 pages (single spaced, excluding figures). The report should be written for someone with a background in statistics as well as with the linguistic data. This means that there is no need to introduce the data or describe any statistical methods. In addition to your report, you will give a 15 minute presentation of your findings to both instructors on 8/14.