ST810: Big Data – Challenge #1

Tentative due date: Thursday, Sept 24

Overview: The data consist of $n = 200$ subjects. For subject i we observe the response Y_i which measures their musical ability, and a three-dimensional image of brain activity $\mathrm{X}_i = X_{iuvw}$, where X_{iuvw} is the brain activity in voxel (u, v, w). Here the brain is partitioned into $p = 20^3 = 8,000$ voxels with $(u, v, w) \in \{1, ..., 20\}^3$. **Your objective it so build a statistical model that uses brain activity to predict the subjects' musical ability.** You should not collaborate with other students on this project.

Data: The data are available on the course webpage at

`http://www4.stat.ncsu.edu/~reich/BigData/assignments/C1.RData.`

The workspace contains the following objects:

1. `Y`: a $n = 200$ vector of musical ability scores
2. `X`: a $200 \times 20 \times 20 \times 20$ array of brain activity scores

The final 100 values of `Y` are missing. You will predict these values and be scored the accuracy of these predictions.

Final report: Your final report should be on two pieces of paper (front and back, single spaced, 11 font, 1 inch margins) and have the following section titles:

1. **Summary**: One paragraph overview of the objectives, methods, and results.
2. **Introduction**: Very briefly describe the problem to be addressed.
3. **Methods**: Describe your statistical methods. Do not include code. You should include a few equations. Provide sufficient detail that the analysis could be replicated by another student.
4. **Results**: Describe the important features of the statistical model, for example, which voxels are the most important, and how accurate are your predictions?
5. **Predictions**: For subjects 101-200, which ten do you predict have the highest musical ability?

Turn in three copies of your paper, one with your name on top and two without your name. All three should include an identifying code so they can be compiled.