Welcome to the Statistical Machine Learning midterm. For problems 1 and 2, you can use a variety of resources, including our text, your notes, and web resources. But you cannot ask anybody anything, and your use of resources should be to aide the technical parts of arriving at your answer and not for getting the answer itself. For question 3, you may only use the textbook, course slides, and your notes. Honor code applies as usual. Enjoy!

1. You have a sample of 55 measurements, and you are interested in the mean of this dataset. However, you believe that some (roughly 10%) of the measurements have been corrupted and are likely from a population with a larger mean and standard deviation than your population of interest. One way to handle this contamination (though maybe not the best way, but let's set that aside for now) is to compute the trimmed mean, which (in R) trims a percentage of the data from each of the tails of your sample. So for instance, if you were to call the function *mean(pop, trim=0.1)*, roughly 10% of the lowest values, and 10% of the highest values, will be discarded (I say roughly because 10% of 45 is 4.5 and one can't remove a half observation) (and while we are at it, the method by which R chooses how many points to trim, and the fact that it trims *symmetrically* may not be exactly what you want, but the writers of the function had to make a choice, and if you need a different behavior for your particular problem, well tough cheese! You'll have to write your own function. It's not hard.).
   So you settle on calculating the trimmed mean with the R function *mean(pop, trim=0.1)*.
   a. Do you expect the standard error of the trimmed mean to be smaller or larger than the standard error of the mean?
   b. Simulate such a dataset by drawing 50 data points from a normal distribution with a mean of 100 and a standard deviation of 10, and 5 data points from a normal distribution with a mean of 120 and a standard deviation of 50. Calculate the standard error of the trimmed mean (hint: an analytical solution is crazy messy)
   c. What are the properties of your estimation method? Run a simulation study to calculate the bias and variance of your estimate. This requires you to do 2 things:
      i. Draw multiple datasets (like 100,000) from your distribution, calculate their trimmed means, and calculate the standard error of those trimmed means.
      ii. Draw multiple datasets from your distribution, calculate the standard error of the trimmed means of each dataset using your method in b.
      iii. Use i. and ii. to calculate the bias and variance of your estimate.

2. You will find two .csv files attached to this midterm- 'Xvals.csv', and 'Yvals.csv'. It should be no surprise to anybody that 'Xvals' contains a matrix of covariates (as usual, observations on rows, covariates on columns) and 'Yvals' are the values being predicted. This is actually a real dataset, but I'm not going to tell you what it is. Please do not attempt to figure out which dataset it is. You don't need that info to complete this problem.

   Your task is to build a prediction model for this dataset. The deliverable here is a discussion of your model including the steps you took in building the model, the choices you made, and an analysis of the model. You don't need to submit a set of predictions like you did in your homework, though you should submit your code.

   Here are some important things to keep in mind:

   a. ML algorithms are often black box algorithms. But that doesn't mean we don't put significant effort into evaluating those models and trying to understand their performance. You should put that effort in here.
   b. ML algorithms are often black box algorithms. But that doesn't mean that there isn't a significant amount of work in preparing the data to use as input. You should put serious thought into these variables. There are some interesting transformations you should consider, and implications of those transformations. You may also want to create other variables that are functions of the variables you have (and by 'may', I might take that as a hint that you should) (and while I'm on generous hints, the transformations you consider might only require a single R function- you may have a little bit of work to do after the transformation. Again, try to carefully think through the implications of what you are doing).

3. We have learned about the lasso and ridge regression which constrain the coefficients in our regression. **For this question, you may only use the textbook "Introduction to Statistical Machine Learning" and your notes/ the slides.**

a. Describe the purpose of constraining the coefficients of the regression.

b. In some cases, the coefficients may fall into natural groupings, such as economic indicators that fall naturally into groups, problems in genomics, areas of the brain, stocks, etc.

In these case, we can use the group lasso, which solves the following problem

$$\underset{\beta}{\operatorname{argmin}} \left[\left(y - \beta_0 - \sum_{p} \sum_{j=1}^{J_p} \beta_j^p X_j^p \right)^2 + \lambda \sum_{p} \sqrt{q_p} \cdot \left|\left|\beta^p\right|\right|_2 \right]$$

where $\left|\left|\beta^p\right|\right|_2 = \sqrt{\beta_1^2 + \beta_2^2 + \cdots + \beta_{J_p}^2}$ and q_p accounts for the size of group p.

Use your knowledge of ridge and lasso to describe the effect on the coefficients of using this penalized regression. You should conclude by discussing why it is called the group lasso.