Midterm 02
Linear Models – Fall 2015
Due: 2015-12-07

This assignment should be treated similarly to a problem set, with the one caveat that you must work on it by yourself without collaborating with other students. You are welcome to ask Jason on I for help on understanding the questions but we will not help with the content of the questions. You are allowed to access whatever notes or texts you would like.

1. (15 points) Consider a 3-by-4 matrix X and the regression vector β equal to $(1, 1, 1)^t$. Define y as the following (i.e., no error terms of offsets):

$$y = X\beta.$$

Construct X such that that solving the least squares problem (that is, without a noise term, the same way I did in class) by inverting the matrix X^tX fails to reconstruct the correct β but using the pseudoinverse and QR trick both do. Then repeat this with an example for which all of these methods fail. Full credit for any working solution, bonus points if the matrix X looks 'reasonable' at first glance. For each of the two cases case print out the matrix and show the three error rates (direct solve, QR of X, and pseudoinverse of X). Hint: the inverse condition number need to be between `sqrt(.Machine$double.eps)` and `.Machine$double.eps` for the first matrix, and less than both in the second.

2. (10 points) Construct a simulated dataset X and response vector y such that the lasso solution path *returns* a vector from the active set back into the inactive set. That is, there exists a j such that $\widehat{\beta}_j^{\lambda_1}$ is non-zero but $\widehat{\beta}_j^{\lambda_2}$ for some $\lambda_2 < \lambda_1$. Do not give me the data, but rather supply R code which generates such a dataset. For full credit, explain why you would expect your simulated model to behave like this. Note: Do not just simulate a large set of data until this happens!

2. (75 points) This question concerns a dataset constructed from a corpus of texts. I have supplied a term frequency matrix (which gives the words counts for any word contained in at least 3 documents), and a set of two classification vectors. These indicate whether the author was older than 22 years of age, and whether the author self-identified as female. The data are contained here (you will need to load the Matrix package as the data matrix is stored in a sparse format):

```
http://www.stat.yale.edu/~tba3/class_data/metaData.Rds
http://www.stat.yale.edu/~tba3/class_data/mmLemma.Rds
http://www.stat.yale.edu/~tba3/class_data/lset.Rds
```

The rows of the meta data and data matrix line up together. The data *lset* gives the word corresponding to each column of the data matrix. Notice that approximately one third of the responses are set to `NA`. Your task is to fill in the missing predictions for both of these responses. You will need to supply a csv file formated like the following (spacing is not important):

id, female, age
0001, 1, 1
0002, 1, 0
0003, 1, 1
0004, 0, 1

This should be done for all of the samples, not just the missing ones! Notice that you need discrete predictions 0 or 1, not probabilities. Your goal is to minimize the raw prediction errors.

To answer this question, upload your predictions as a plain text csv file to the classesV2 site and write a short 1 page response describing how you approached this problem and how well you believe your prediction will do on the unknown data. This should contain absolutely no code, and be entirely in prose. You may include tables or figures, but the total response should be no longer than two standard sized pages long. Full credit will be given for any reasonably predictive model and approach; bonus points for particularly clever ideas and top performing estimates.