Stat 666
HW #4
Due on TBA

I. Pollution Sources and Ambient Pollution Measurements. Formulation and evaluation of environmental policy depends upon receptor models that are used to assess the impact and nature of pollution sources affecting the air quality for a region of interest. Pollution source apportionment (PSA) involves the use of ambient air quality data to estimate the abundance of pollutants emanating from the various pollution sources. The source contributions are generally estimated via nonlinear modeling that ensures nonnegativity in estimates. In this study, however, we are interested in looking at linear relationships between the air quality measurements and the source contributions obtained from a PSA analysis. We are interested in evaluating the performance of linear regression in predicting the source contributions $(y_1, \ldots, y_8)$ from the abundance of different chemical species in $PM_{2.5}$ $(x_1, \ldots, x_{18})$. The observed data matrix $\mathbf{X}$ is found in the file `PSAData.txt`. This data set consists of a header row containing the 18 chemical names, followed by 200 observations of chemical abundances for 18 important chemical constituents. The response matrix ($\mathbf{Y}$) contains the 8 source contributions associated with each measurement and are found in the file `PSAContributions.txt`. This data set consists of a header row containing the names of the 8 assumed pollution sources, followed by 200 measurements of the 8 sources. You may treat the 200 observations as a random sample from the population of possible measurements. (That is, for this analysis, we choose to ignore the temporally-correlated nature of the air quality measurements.)

For each question, show your work and give a BRIEF response/interpretation.

1. Is there a significant relationship between the chemical abundances (xs) and the source contributions(ys)? What is $\hat{\boldsymbol{B}}$?

2. What is the essential dimensionality of the relationship between the xs and ys? Is that dimensionality directly evident from an inspection of $\hat{\boldsymbol{B}}$?

3. Use canonical correlation analysis to identify and interpret the important dimensions of the relationship between the xs and the ys.

4. Traditionally, experts have assumed that heavy metals are important to the understanding of some pollution sources. In terms of the linear regression considered here, is the heavy metal Pb an important factor in the overall prediction of pollution source emissions? That is, can we drop Pb from the list of measured predictors without losing significant predictive ability?

[continued on next page]

II. Body Fat and Physiological Measurements. A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. For example, one can estimate body fat from tables using age and various skin-fold measurements obtained with a caliper. Other texts give predictive equations for body fat using body circumference measurements (e.g. abdominal circumference) and/or skin-fold measurements. We are interested in a linear regression prediction of body density and percent body fat. The columns of the data set `bodyfat.txt` are as follows:

y_1 = Density determined from underwater weighing
y_2 = Percent body fat
x_1 = Age (years)
x_2 = Weight (lbs)
x_3 = Height (inches)
x_4 = Neck circumference (cm)
x_5 = Chest circumference (cm)
x_6 = Abdomen 2 circumference (cm)
x_7 = Hip circumference (cm)
x_8 = Thigh circumference (cm)
x_9 = Knee circumference (cm)
x_{10} = Ankle circumference (cm)
x_{11} = Biceps (extended) circumference (cm)
x_{12} = Forearm circumference (cm)
x_{13} = Wrist circumference (cm)

5. Although the two ys are very highly correlated, there is some concern that each of these two measures of body composition may have different relationships with the xs. Is there a reason to be concerned about the ys being different in some sense, or are they responding similarly to the xs?

6. For quick and inexpensive assessment of body fat and body density, it may not be practical to measure all 13 of the predictors. Alternatively, we would like to have a subset of predictors that adequately predicts the bivariate response. Use any approach you'd like to propose model using only a subset of the predictor variables that adequately predict the body fat related measurements (y_1 and y_2).

7. Using canonical correlation analysis, describe and interpret the multivariate relationship between the body fat related measurements and your reduced set of predictors from the previous problem.