In your writeups, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project.

1. **Principal component regression (PCR).** Suppose we have a response $y$ and $p$ predictors $X_1, \ldots, X_p$. In PCR, one transforms the variables into principal components $Z_1, \ldots, Z_p$ and regress $y$ onto the principal components. One typically only uses a subset of the principal components in the regression, making a kind of regularized estimation. Often the principal components with the highest variance are selected (see Section 11.5.4 in the textbook for more information).

The data set *wheat.dat.txt* contains measurements about 79 wheat grain samples. The first column is the percentage of water content and the columns X1–X100 are Near Infrared (NIR) reflectance spectra. (If you are interested in spectrometry and wish to learn more, check the spectrometry wikipedia entry.)

(a) Plot the response (water) and the NIR data. The NIR spectra are probably best viewed by each spectrum (row) as a curve. Are there any special structures, or extreme observations that should perhaps be removed?

(b) Perform a principal component regression (PCR) with ten components. Compute the regression coefficients, the predicted values, and the residuals. Do you have comments?

(c) Estimate the prediction error from the model above which uses ten components, and explain your method.

(d) The number ten is a little arbitrary at this point. Should we use fewer or more? Use cross-validation to estimate the accuracy of models with 0, 1, 2, \ldots components. Plot the estimated prediction error versus the number of components, and decide how many components you want to include in the model. How large is the estimated prediction error for the chosen model size? Which components seem to be the most important for predicting the response?

(e) **Bonus question.** Give an estimate of the prediction error that someone—completely unaware of the good stuff we learn about in ACM 118—would get by using all the variables X1–X100 to predict the response.