STA 437/1005, Fall 2009 — Assignment #2

Due at the start of the lecture on November 2. Please hand it in on 8 1/2 by 11 inch paper, stapled in the upper-left corner, without any folder or other packaging around it. Note that I've decided that only the last assignment will have the two-part solution/critique form that I discussed earlier, so for this assignment you just hand in your solution.

This assignment is worth 10% of the course grade. It is to be done by each student individually. You may discuss this assignment in general terms with other students, but the work you hand in should be your own. In particular, you should not leave any discussion of this assignment with any written notes or other recordings, nor receive any written or other material from anyone else by other means such as email.

For this assignment, you will examine two real data sets. You should first make decisions about whether there are outliers or other data points that may not be reliable, or are not representative. For a real project, you would likely be able to consult the people who gathered the data about any possible problems, but for this assignment, you will have to judge how reliable data points are based only on the numbers you see, your common sense, and the descriptions provided with the data. You can use the various tools discussed in class and the textbook, such as scatterplots, histograms, QQ plots, and statistical distance.

You should also decide whether some variables should be transformed in some way, and whether the data can usefully be regarded as having a multivariate normal distribution (perhaps after transformations).

You should also try to draw some preliminary conclusions, without doing formal statistical tests, as described for each data set below. These conclusions should of course be based on the data as modified by deleting outliers or transforming variables, as you decide is desirable. You may decide that you should look at more than one version of the data (eg, with or without some transformation), but you should do this only if you have a good reason to think both versions provide information that may be useful.

You will need to use R for this assignment. Part of the purpose of the assignment is to get you started using R, and more generally to learn how to handle the preliminary data processing that is needed for most projects. You must hand in a listing of the R commands you used to draw your conclusions, and the output these commands produced (text output or plots). You don't have to (and shouldn't) hand in every R command you typed. You should instead hand in the R commands that someone wanting to replicate or critique your work would need to look at (or run) to see how you arrived at your conclusions.

You must also hand in a writeup of your conclusions, which refers to the output of the R commands to justify these conclusions.

This handout, the data sets, and some hints about useful R commands, are available from the course web page, at `http://www.utstat.utoronto.ca/~radford/sta437/`

Data set 1: This data set (from statlib) was gathered in order to access how well percentage of body fat for adult men can be determined from their height, weight, and circumference of various body parts. The course web site has the description taken from statlib, as well as the data from statlib, to which I have added a header line with the names of the variables (so you should use the `head=TRUE` option when reading this data using `read.table`).

There is data on 252 men. The first two variables are their body density (density) and the percent of fat (pcfat), which according to the statlib description was computed from the density by the formula

$$\text{pcfat} = 495/\text{density} - 450$$

The remaining variables are their age, weight, height, and 10 measurements of the circumference of their neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist. See the description from statlib for more details.

The data was gathered for the task of predicting body fat from easier measurements, but it could also be used to study how various body measurements relate and how they vary over the male population, so you should look at this data in general multivariate terms, not just in terms of a regression for pcfat on the other variables.

As described at the beginning of this assignment, you should look for measurements that appear to be unreliable, or even if correct, are so unusual that they are better ignored when trying to draw general conclusions. You should also consider whether the data appears to be multivariate normal, and if not whether it could be made closer to normal by some transformation.

Although not mentioned in the data set description, you should also use this data to assess how well the Body Mass Index (BMI) correlates with percent body fat. BMI is defined as the weight in kilograms divided by the square of the height in meters. You should look at other possible definitions for BMI, of the form: height to the power p divided by weight to the power q. The standard definition has p=1 and q=2. You should see whether other choices for p and q might produce a signficantly better correlation with percent body fat.

Data set 2: This data set (from the UCI repository of machine learning data sets) contains satellite image measurements in the vicinity of various locations in an agricultural region. Each observation has intensity values (in the range 0 to 255) for four spectral bands at each pixel in a 3x3 array. From ground observations, the actual type of ground cover at the location of the centre pixel was also determined.

I have selected a subset of 2512 observations in which the ground cover was either red soil (coded as class 1), cotton crop (coded as class 2), or grey soil (coded as class 3). I have also added a header line with names for the 37 variables. The last variable is "class", from 1 to 3. The first 36 variables are the image measurements, with names of the form Xn, where X is A, B, C, D, E, F, G, H, or I, for the nine pixels (E is the centre pixel), and n is 1, 2, 3, or 4. Since there is a header line, you should use the `head=TRUE` option when reading this data using `read.table`.

You should look at this data with a view to trying to determine the class of the centre pixel from the measurements at the nine pixel locations. As we will cover later in the class, one class of methods for classification relies on the distributions *within each class* being multivariate normal, so you should check whether this seems to be true, or whether it can be made close to being true by some transformation. You can also look for outliers, though since this data is automatically recorded, there is less possibility of outright errors in the numbers.

You should also use scatterplots with class identified by colour to judge how promising the idea of classifying the centre pixel based on these measurements is. Does there seem to be a lot or a little overlap of classes in scatterplots involving two variables?

You should also consider reducing the dimensionality of the data by taking the mean or the median of the values for each of the nine pixels (separately for each of the four spectral bands). This would reduce the number of variables (apart from the class) from 36 to four. Do you think the mean or the median would work better?