**Chapter 2: Data examination**

**“*A picture is worth a thousand words.*”**

* 1. **Importance of examining data**

Examining the data is a key step in virtually every statistical analysis. Try to do it as soon as possible since it can affect all other aspects of the analysis. Most datasets have errors or problems so it is important to plan for them by allowing extra time. It is also necessary to have a good picture of the main features of the data before beginning modeling.

The most important thing about this is probably to *remember to do it*. The methods are not highly technical, but sometimes in an unfamiliar situation it is easy to forget, so it should become one of the first things on your checklist.

**2.2 What to look for**

In examining the data you are looking for many things, including

* Weird data values (e.g. outliers or unusual points)
* Data points that are not possible (e.g. age > 120 years, negative weights, etc.)
* Changes over time that are not possible (e.g. 100 lb weight gains in 2 weeks)
* Data points that do not conform to theoretical relationships (e.g. weight of fat mass plus weight of non-fat mass should approximately equal body weight)
* Levels of categorical variables that shouldn’t exist (e.g. sex=‘K’)
* Levels of categorical variables with very few subjects
* Weird data patterns (e.g. skewness, curving, unequal variance)
* Missing data patterns (e.g. how many missing values on each variable)
* Some basic relationships among variables (e.g. which are most correlated)

**2.3 Techniques and software for examining data**

Methods for examining data are often simple, but applying them in effective ways can save lots of time and errors. Some examples of methods I routinely use are:

**SAS PROC UNIVARIATE *PLOT*; or R ‘summary’, ‘hist’, and ‘qqnorm’ functions**

* + The descriptive statistics printed out are useful to check reasonable ranges of variables (e.g. ages > 120 years or negative weights).
  + The descriptive statistics are also useful to see the number of missing values for each variable.
  + The stem-leaf plot produced by the PLOT option is really a histogram turned on its side. I find it very useful for quickly examining large numbers of variables for outliers and/or skewness. These can be done for subgroups using the BY or WHERE statements. The other two graphs (a boxplot and a terrible quality QQplot) are not so useful.
  + A CLASS statement in UNIVARIATE gives side by side comparisons, see example below.

**SAS DATA steps; R calculations**

To examine data for conformance to theoretical relationships I use DATA steps to calculate any relationships that should be approximately equal (e.g. CHECKWT = FATMASS + NONFATMASS) and the difference between the two values that should be similar (e.g. DIFFWT = WT – CHECKWT), and then use PROC UNIVARIATE as above to examine these differences. An understanding of the subject area is useful here. It is amazing how often errors can be caught this way.

**SAS PROC FREQ; R ‘table’ function**

To examine categorical variables I use SAS PROC FREQ. I look for impossible categories (e.g. sex=’K’), sparse categories (very few subjects), and missing values. TABLES X / MISSING creates a separate category for missing. Bivariate and higher tables can be made and can be very useful too.

I often collect the results from the above analyses and make a table showing missing data patterns, perhaps the number and percent missing for each variable.

**SAS PROC GPLOT; R ‘plot’ function, or ‘ggplot’ function**

To examine bivariate patterns in numerical data I use PROC GPLOT in SAS. Look for values that lie away from the bivariate pattern, curved patterns, unequal variances, clumps of points, etc.

I also use GPLOT to make “spaghetti plots” (individual subject time plots) for repeated measures data. Look for unusual values for individuals, unusual subjects, or skewness (most subjects low on the graph, with a few higher).

**R ‘pairs’ function, or many other similar functions**

The above suggestions assume you are using SAS. R is also excellent for data examination, and all of the things listed above can be done there. The pairwise scatterplot graphs in R are terrific and I use them whenever I have more than a couple of numerical variables to examine. See the section on handy computing.