**DESCRIPTION OF THE BODYFAT DATA**

For us in a practical sense, the bigger issue is how to use scatterplots, correlation and simple linear regression models to deepen our knowledge, understanding, and information about the data we are analyzing.  This is the domain of applied correlation and regression analysis.

Once you are done with data cleaning, which includes graphing and summary statistic computations on each variable separately, you should start to wonder and ask questions about which variables are related to one another.   In what way are they related and how could the observed patterns be modeled?

Here, in the world of correlation and regression, we are interested in looking for underlying patterns between two quantitative variables.  Depending on the task that is given to you, a) you may know ahead of time which variables should be related and you then set off to confirm the expectation; or, b) you may simply be given a data set and told to find something. In which case, you will have to figure out a strategy of what to do.

***The Data***

Whether your priority is looking good in a bathing suit or living to see 100, there’s one piece of information that might stand in the way. A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. Higher levels of body fat have been linked to an increased incidence of diabetes, high blood pressure and other heart risks. The problem is that accurate measurement of body fat is inconvenient and costly. The most accurate way to measure body fat is Hydrostatic weighing, commonly referred to as underwater weighing. This method compares a subject’s normal bodyweight (outside the water) to their bodyweight while completely submerged. Using these two numbers and the density and temperature of the water, operators can accurately nail down the subject’s density. This number is then used to estimate body composition, and ultimately percent body fat. But, unless you happen to have an underwater scale at home, you’re going to have to find a lab or a performance center that offers hydrostatic weighing. As a result, this method can be a bit inconvenient and more expensive compared to other techniques. Subjects also have to forcefully exhale as much air out of their lungs as possible (to reduce potential for error) and sit submerged completely underwater. So, it is desirable to have easy methods of estimating body fat that are not so inconvenient and costly.   But, the easy methods have differing degrees of unreliability.   Statistically, this translates into larger amounts of error and variability.    Keep in mind, when you think of variability - think standard deviation or sum of squared residuals (errors).

The data we are going to use has percentage body fat determined through Hydrostatic weighing for 252 men. In addition, these men are measured on a number of other, less intrusive body circumference measurements. The variables of our dataset are listed below and include:

* Density determined from underwater weighing
* Percent body fat - derived from Siri's equation which is a function of Density
* Age (years)
* Weight (lbs)
* Height (inches)
* Neck circumference (cm)
* Chest circumference (cm)
* Abdomen 2 circumference (cm)
* Hip circumference (cm)
* Thigh circumference (cm)
* Knee circumference (cm)
* Ankle circumference (cm)
* Biceps (extended) circumference (cm)
* Forearm circumference (cm)
* Wrist circumference (cm)

Our task in these presentations is to try and predict Percent Body Fat using regression and correlation with the other circumference variables to come up with a "best" but non-intrusive way to measure Percent Body Fat, as well as to try and judge the accuracy of our prediction.

Below, you can find the body fat data which is in a file called:    BODY\_FAT.CSV





