STAT 100 Notes April 1, 2010

Almost all the conceptual material of the course has been introduced already. In the seven hours of lectures left, we will review the concepts, contexts and techniques, especially those that require repetition. Today I will hand out a page listing the concepts, contexts, and techniques and ask for feedback on which ones require the most revisiting. But before I start that, I want to say a bit more about **Standard Units, Correlation,** and **Simple Linear Regression.** These are not strictly new topics but the coverage of them will be a bit more detailed than previously.

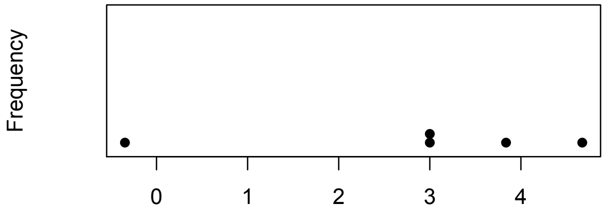
Standard Units:

Any data set can be converted in to a standard unit form. For some purposes, the data set in standard units is equivalent to the original data set, and it is much simpler to use.
Standard units simply re-express a data value in terms of its number of standard deviations from its mean.

x -> z=(x - Mean)/SD

For example, if I have a sample of 5 students who report number of cell phone uses so far today as

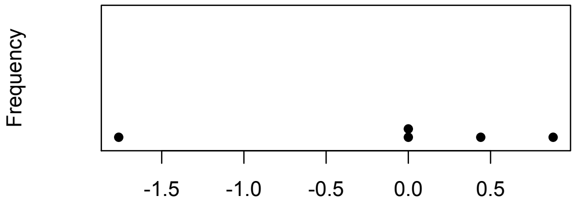
3, 0, 5, 3, 4



I can compute Mean = 3.0 and SD = 1.9 and so the data in standard units is

0/1.9, -3/1.9, 2/1.9, 0/1.9, 1/1.9 which simplifies to

0, -1.58, 0, 1.05, 0.53



Same Graph, different units.

Expressing units as "number of SDs from mean" instead of "uses" keeps all of the information about the configuration of the data but ends up with data that has mean 0 and SD = 1.

Expressed in standard units, data always has mean 0 and SD=1.

Now it will be easy to describe a measure called the **Correlation Coefficient, r.**

For the same 5 students, suppose we have data on the number of Facebook friends:

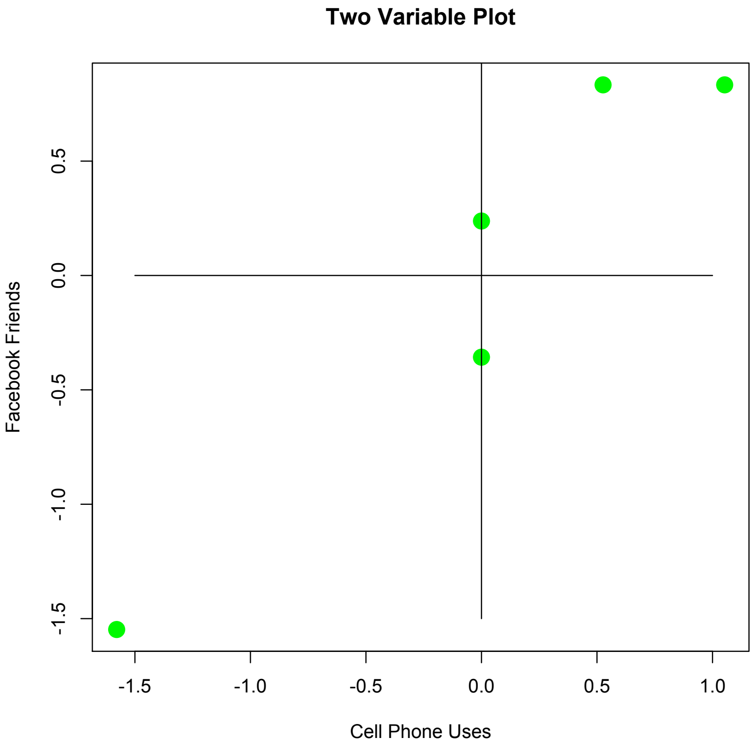
10, 0, 15, 20, 20

The mean of these is 13 and the SD is 8.4. Converting to standard units,

-0.36, -1.55, 0.24, 0.83, 0.83

which again have mean 0 and SD = 1.

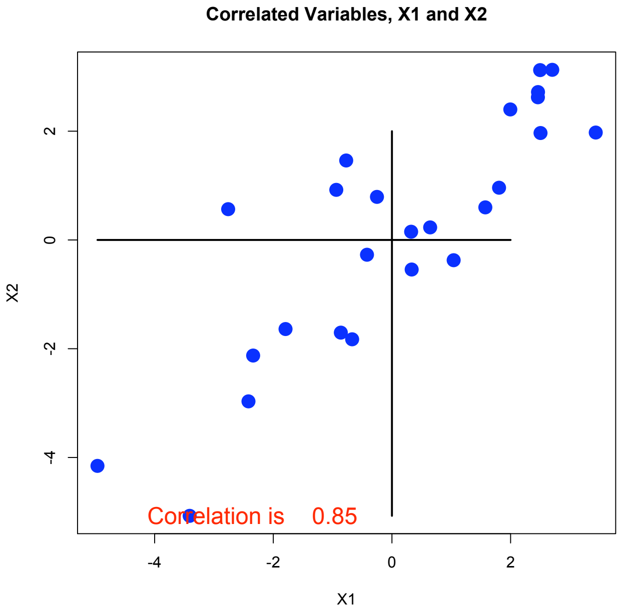
Now lets plot both variables on a graph.

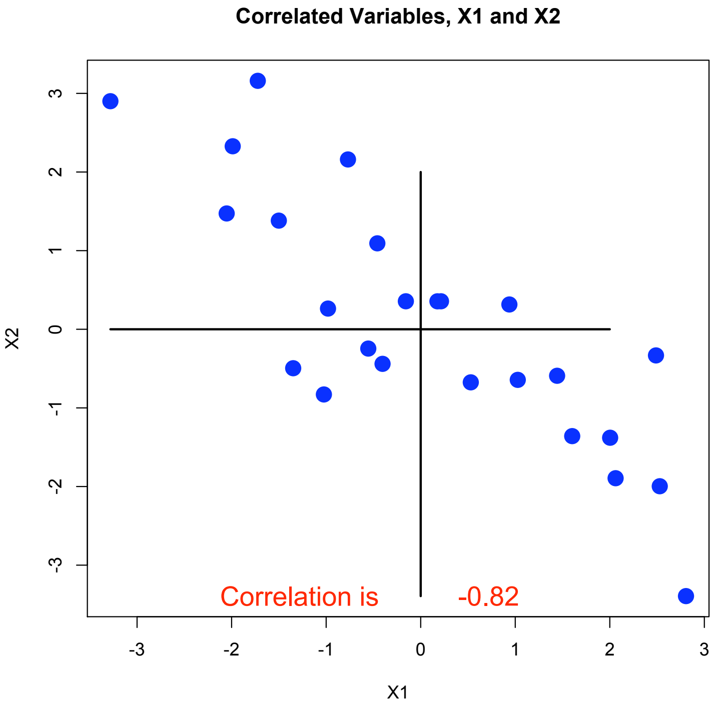


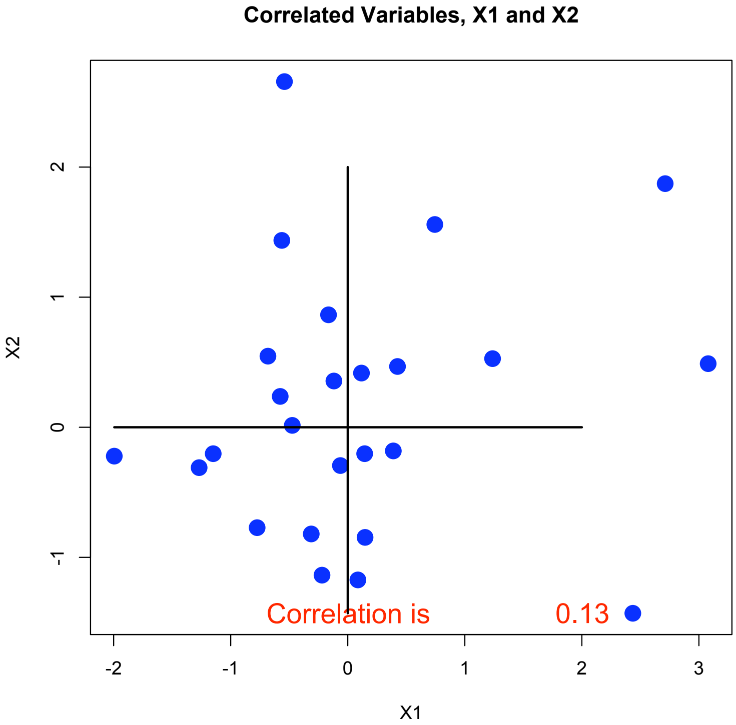
Note the quadrants where the points tend to be – upper right and lower left. Think of the products of the two coordinates – they will be positive, right?

It turns out the average product is 0.75, and is a measure of the "correlation" of the two variables. (If you get .96 on your calculator, it is actually correct but uses an n-1 in the denominator instead of the n that an ordinary mean would use. For reasonable size samples, n or n-1 does not make any important difference, and the ordinary mean is easier to remember.)

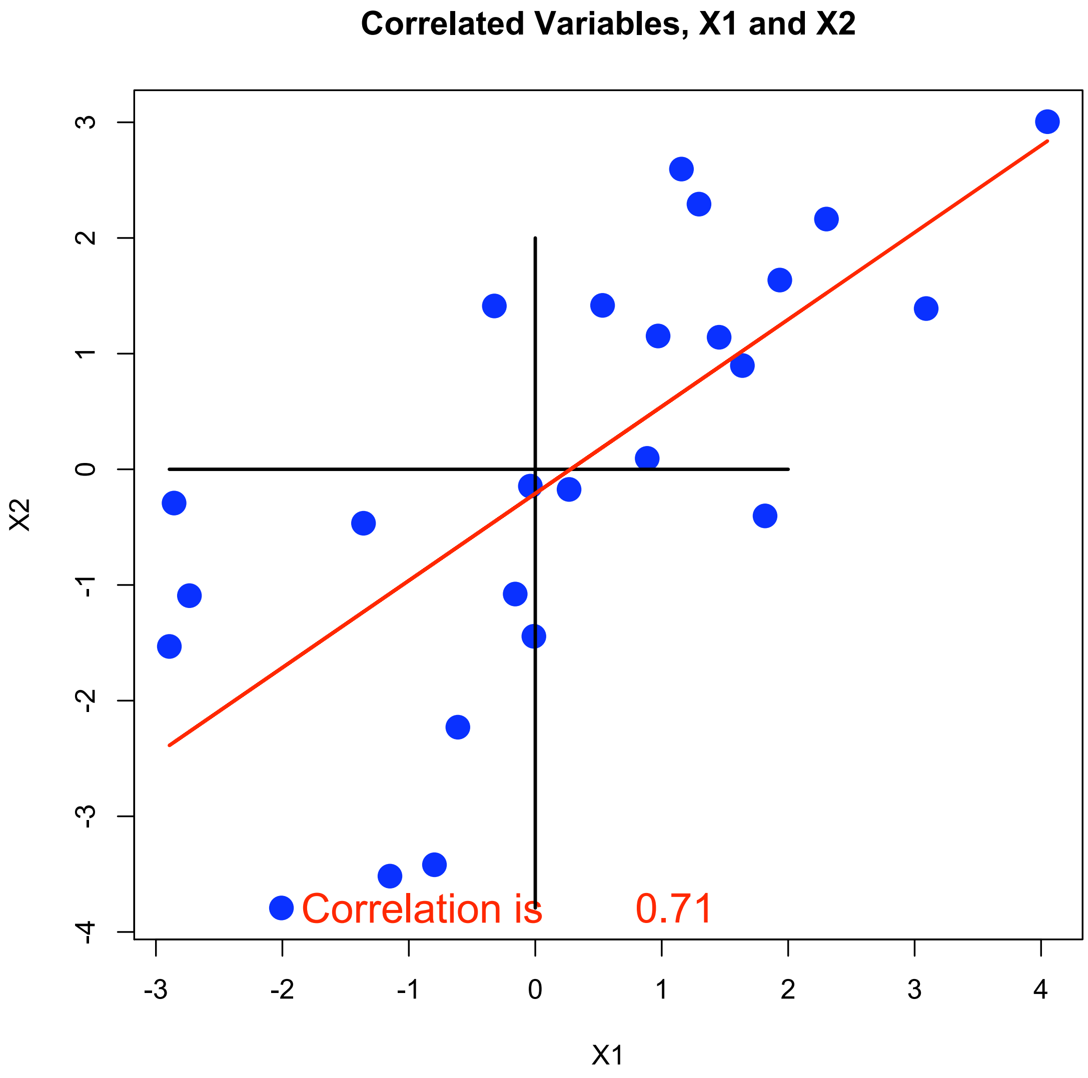
Correlation is useful for summarizing the extent to which one variable can predict another variable. A correlation of 1 suggests perfect prediction. In the following we will explore some data patterns and their correlations:







It turns out that in standard units, the correlation coefficient is also the slope of the regression line:



So this summarizes the connection between standard units, the correlation coefficient, and simple linear regression.

This topic has three components:

Concept: The predictive relationship between two variables ("correlation")
Example: Cell phone use and Facebook Friends
Technique: Standard Units, Correlation Coefficient, Linear Regression

Next we want to review all the Concepts, Examples and Techniques in the Course.