David Youngberg
BSAD 210—Montgomery College

LECTURE 20: CORRELATION AND CAUSATION

I. Of Correlation and Causation

a. The scatter diagram graphically illustrates if two variables are correlated, or that if one value changes the other will change in a predictable way.
   i. *Positively correlated* means the values change in the same direction, such as "time studied" and "grade earned."
   ii. *Negatively correlated* means the values change in the opposite direction, such as "time partying" and "grade earned."

b. Correlation is actually a spectrum. Sometimes, there is weak correlation and other times the correlation is strong. Consider the three dot plots below. Each represents positive correlation but note that the data sets are very different:

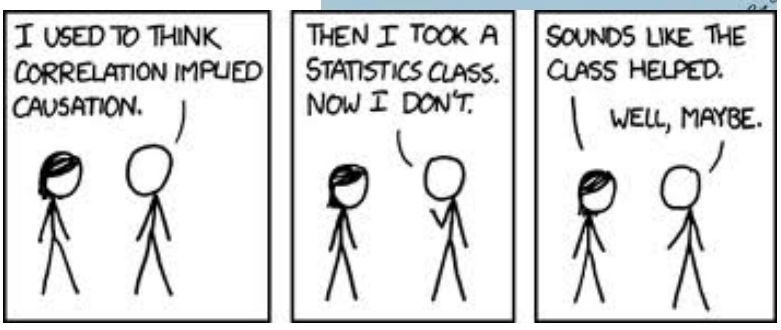
c. We can sum up those differences with the *correlation coefficient*—a single number which captures the strength and kind of the correlation.
   i. A positive value indicates positive correlation and a negative value indicates negative correlation.
   ii. The closer the value is to 0, the weaker the correlation.
   iii. The value cannot be greater than 1 or less than -1.

d. Of course, *correlation does not mean causation*. Just because it looks like two variables run together doesn't mean they do. Two other things could be going on:
   i. *Reverse causation*—when the dependent and independent are confused (Greater CO_2 emissions cause people to earn more?s)
   ii. *Confounding variable*—variable that causes both independent and dependent variables (Does a greater portion of agricultural workers lower infant mortality? No. Income causes both.)

e. At the same time, *correlation is evidence for causation*. The internet is full of people, upon seeing a strongly correlated pair of variables, dismissing any possibility that the two are connected because it doesn't mean causation. But it <u>should</u> give you reason to pause. And if you can find an explanation *why* one would cause another, you're in good standing.

   i. Hedge fund manager and blogger James Altucher dismisses evidence of higher earning potential thanks to a college degree by invoking the tired mantra.[1] Of course, there are good reasons to think college causes higher earnings such as credentials, signaling, and skill building.

f. You need a narrative—some sort of reason—why one thing can cause another. In the comic on the left, it makes sense the male character knows "correlation doesn't mean causation" because the statistics course would emphasize such thinking. If he learned, since the class, that North Korea is an oppressive dictatorship which puts disgruntled citizens into death camps, then that's probably a coincidence. North Korea politics aren't covered in (most) statistic courses.



[1] http://www.jamesaltucher.com/2011/01/10-more-reasons-why-parents-should-not-send-their-kids-to-college/

II. Correlation Coefficient

a. Open the Data Set 4; you'll find an expanded version of the country dataset we first encountered in Data Set 2.
b. In Data Set 2, we used a scatterplot to visually see if there was a correlation between two variables. Now we can do something much faster and more precise: the correlation coefficient.
c. Are more populated countries wealthier?
    i. Perhaps; we can imagine that more populated countries might allow for greater specialization of the labor force. Or people tend to be attracted to wealthier countries.
    ii. But we could also point out that as people get wealthier, they tend to have fewer kids.
d. Select any blank cell—there are a lot of missing data in D through G so I'd select one of those blank cells—and type "=CORREL(B2:B237,C2:C237)" and press ENTER.
    i. You should get a correlation coefficient of -0.07267. Note that just because it's negative, doesn't mean there's a negative correlation in any real sense; this is so close to zero, there's functionally no correlation.
    ii. Why is this? That's a big question. Perhaps none of our theories are true. Perhaps they are *all* true and together create no net effect. Perhaps some other factor we didn't consider and *that* makes no net effect.
    iii. There also could be a pattern we can't see; perhaps the data is making a U-shaped pattern. That would read as a zero correlation.
    iv. Figuring out exactly what's going on requires more data and greater investigation.
e. Double click the cell you calculated the correlation coefficient in. You'll notice two boxes around the data we used. Place your cursor over the population one until the box becomes thicker and the cursor becomes a cross of arrows. Left-click the border of the box and drag it over it Column I: popdensity, or population density. Let go.
    i. You're now measuring average income (GDP/capita) with population density; and the correlation coefficient is now about 0.23404.
    ii. This is a clear positive correlation, but it's also pretty noisy.
    iii. Still, this is interesting. Again, we don't know the causation but it forms the basis of a story to be told.

III. Correlation Table
   a. Suppose we wanted to know all the correlations between each pair of variables. With 28 variables, it would take a while to input the CORREL command for every possible pair. It's some 378 pairs.
   b. Excel has a built-in function which lets you make a table of all combinations of correlation coefficients.
   c. Go to Data >>> Data Analysis >>> Correlation
      i. If you don't see Data Analysis, follow these steps to enable it:
         1. File >>> Options >>> Add-ins >>> Go… (next to Manage: Excel Add-ins).
         2. Click the Analysis ToolPak box and click OK.
         3. For Mac users, here's a helpful video.
   d. For the Input Range, select all the data, save the countries. Be sure to include the name of each variable.
      i. The Input Range field should say: "$B$1:$AC$237"
   e. Make sure it's grouped by columns and select "Labels in first row."
   f. Select where the output will appear; I placed mine on the same sheet: A240.
   g. Press ENTER.

IV. Examining the table
   a. In A240, you'll see a big table with variables in the first row, along the top, and those variables in the first column.
   b. Any cell displays the correlation coefficient for the variable in that column with the variable in that row.
      i. Note that the cell which has the same variable in the column and the row displays a 1. A variable is perfectly positively correlated with itself.
      ii. Thus there is a diagonal line of 1s in any correlation table.
      iii. Note as well that the top half is empty; that's because it would be redundant. GDP/cap in the column with population in the row would have the same result as population in the column with GDP/cap in the row.
   c. Looking at the table, something should stand out: three pairs are perfectly correlated: one positive and two negative.
      i. This is suspicious; perfect correlation doesn't happen unless something's unusual is going on.
   d. Take a look at the actual data: we don't have many observations for the aid received variable nor for unemployment data. Indeed, we only have

data for both aid received and the unemployment rate for two countries: the Philippines and Turkey.

e. Recall that perfect correlation results in a scatterplot of the observations forms a perfectly straight line: you can draw a straight line that intersects with every data point.
f. If only one have two observations, *of course* you can draw a line that intersects with every data point; a line is defined by two points!
    i. In other words, we don't have enough information to determine if there's any kind of correlation between these pairs.
    ii. A word of caution: if there were three pairs of data, rather than two, the correlation coefficient wouldn't stand out as so unusual. Three pairs isn't much better than two, though, when claiming a correlation so it's also a good idea and check to see how many observations you actually have.