



“r”:

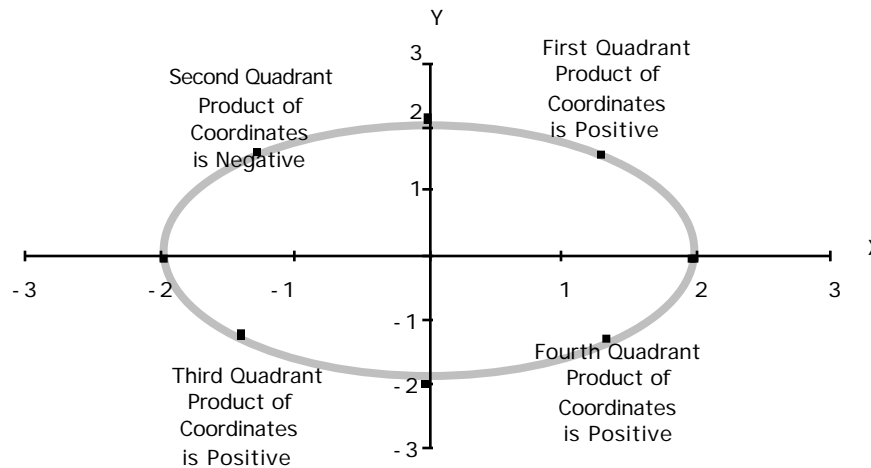
The Measure of Correlation

When a data analysis has been accomplished, and when the result is worth communicating, it becomes necessary to write a report. Reports are not as detailed as the work. And your reader may not be particularly fascinated by the details of the real work of hypothesis construction, examination of residuals for their pattern, revisions of hypotheses, and so forth that led to the result that merits a report. Creativity is a wonderful and peculiar process, but ultimately work will be judged by the result not than the process. At this point convention has greater value than it does in the creative process itself. And, where it is appropriate this is the time to use a conventional measure of the strength of the correlation between two variables. It is not as useful, as powerful, or as subtle as the examination of residuals, but it is conventional to cummarize linear correlation between two variables with a number “r”.

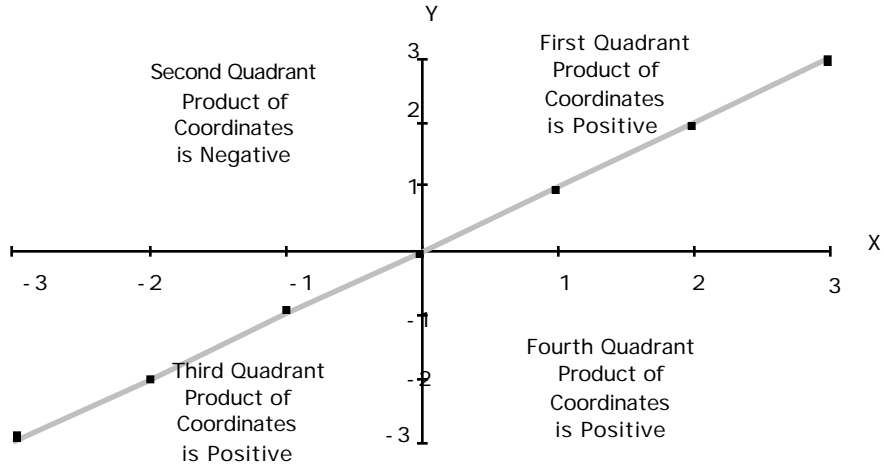
“r” is something of a magic number in statistics because r shows up in at least three contexts, three contexts where the same number makes sense for three different purposes. Here, I will introduce r as a measure of correlation, taking the pedagogical path of introducing r as if for the first time — as an answer to the question “How strong is the correlation?”

Just about the only kind of correlation that statistical technique is well prepared to talk about is a straight line correlation. And there are three things to be said about any straight line. In statistics, as in geometry, a line has an intercept and it has a slope. What's left, in data analysis, is the strength of the correlation or, to use one of the conventional terms, the "goodness of fit". A single number representing correlation is a clumsy weapon as compared to the human eyeball inspecting a pattern of residuals, but it is conventional, and no conventional journal article that uses linear correlation will be without conventional numbers. (With luck, you can use the residuals, and show the residuals, especially if there is a pattern — the serious reader will appreciate the fact that you have to satisfy two audiences, the reader who looks for convention as well as the serious reader.)

The basic intuition is that this collection of data points

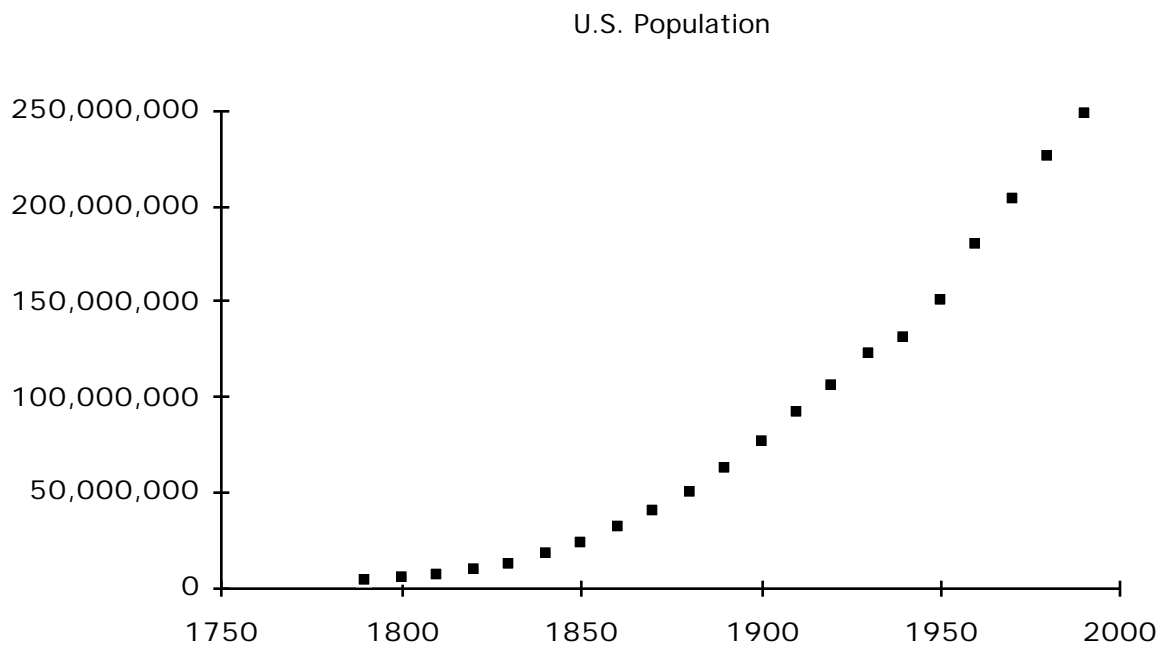


is not a line. While this collection of data points

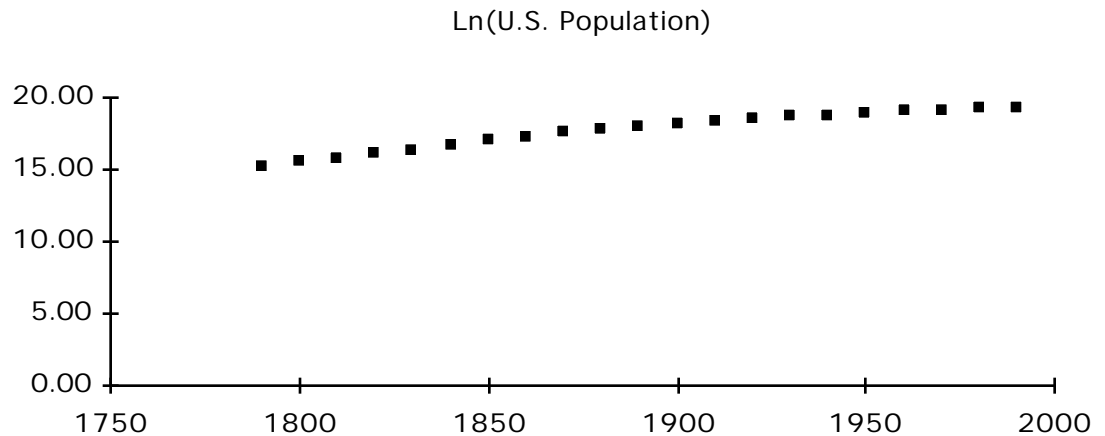


is a line (shown with six data points and a line. The data points are supposed to represent reality, while the line is supposed to represent what our fertile human imaginations would to “see” in those data).

That is what I am thinking about as some sort of ideal of linear data. But this is the kind of stuff we actually look at:



The first thing you do with such data is pay some attention to their behavior. A reasonable guess for a transformation of population to a well behaved variables is the logarithm, getting us something initially more linear looking, like



Now, after you've done all the real work of hypothesis construction, examining the residuals for pattern, revising hypotheses, and so forth, it comes time to write a report. Reports are not nearly as detailed as the work and, for the report, it is useful to put forth a summary statistic that answers the question "How linear?"

To invent a number that will answer that question, "How strong is the linear correlation?", I think, "What property of Figure ___ (the simple line) can I summarize in a number. When I look at something truly linear, the number should say "good"; when I look at something truly non-linear (such as a circle) the computed number should say bad (bad meaning -- not at all like the number I get for a line).

The property I am going to work with, the basic intuition, is that a positively-sloped line will have lots of data points in quadrants I and III, while it will have very few data points in Quadrants II and IV. Fortunately, quadrants I and III have a common property that allows me to detect dominance of these two quadrants. Specifically, the product of two coordinates in these quadrants is positive while, in the opposite quadrants, in II and IV, the product of two coordinates is negative. So I can invent a number that adds up the dominance of I and

III as compared to II and IV by adding up the products of coordinates. . This is not yet the right answer, not yet, but it is the right start.

$$\text{Rough intuition for correlation: } \sum_{i=1}^n X_i Y_i$$

Standardizing for number of data points:

That is the basic intuition. But this would-be index of correlation has serious flaws that need cleaning up to make it useful. For example, data sets come in different sizes, they have different “n’s”. This index would misjudge them: Using this index twelve data points with exactly the same pattern as six data points would, nevertheless, get twice the value for correlation. So, the index is better if it is modified to present an average instead of a sum.

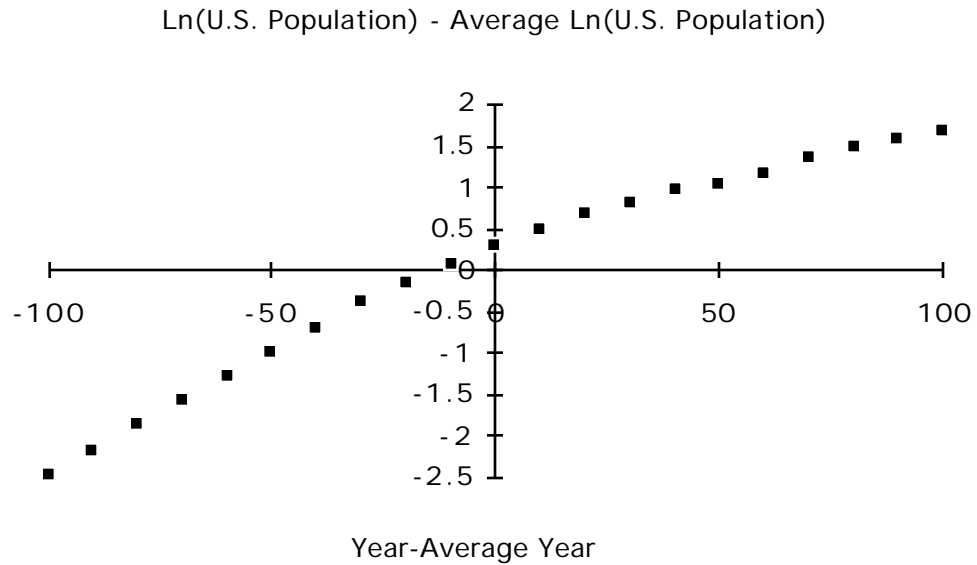
$$\text{Rough intuition for correlation: } \frac{1}{n} \sum_{i=1}^n X_i Y_i$$

Standardizing for the “origin”:

Another problem with this index is that data and even perfectly linear without ever passing through the origin. Like the population data above, the data could all lie in the first quadrant and still show correlation. Surely that’s a problem — unless we care to conclude that all positive numbers are correlated. So the index of correlation can be improved by revising the data — moving the origin to the center of the data. Within the “least squares” framework, that center should be at the mean of x and the mean of y, leading to a modified index

$$\text{Better intuition for goodness of fit: } \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Example with log population, translating the origin to the mean puts most of the data into proper quadrants (for the translated data).



Standardizing for Scale of the Variables:

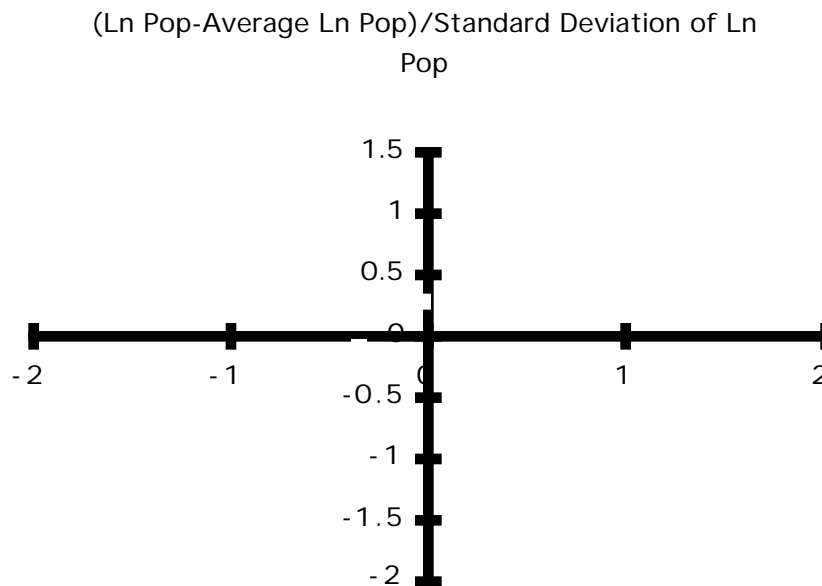
Finally there is a problem of scale. When you simply look at that graph you see a strong relation. But when you examine the numbers themselves, there is an enormous difference in scale between these two variables, ± 2.5 for one, ± 100 for the other. That doesn't affect the picture but it strongly affects this candidate for the index of correlation.

The index can be improved again by standardizing the scale: We standardize the "scale" of x by computing its standard deviation and then dividing by its standard deviation. The standard deviation for these numbers from 1790 to 1990 is 62. The standard deviation for these numbers from 15 to 19 is 1.309. But in standardized form, *subtracting the mean and dividing by the standard deviation*, both standardized variables have mean 0 and both standardized variables have

standard deviation 1 — standardizing both variables to approximately the same scale. So, rewriting the numbers by subtracting the mean and dividing by the standard deviation I get

Better intuition for correlation: $\frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$

In graphical form for the population data:



(fix graph, It is right, but make the point be making both scales have the same ticks and numbers on them -- Excel shrunk one scale ***)

There you see the relation (re-expressed) in what is called “standard form”. This example is typical of standard form graphs in that the graph is centered on zero, the ranges are much the same and, in

fact, they will usually run in a range of plus or minus two standard deviations of the mean. The “imperfection” of this particular correlation (which is not linear) has expressed itself by a little bit of “leakage” into the second and fourth quadrants — affecting the xy products (these will be negative) and numerically diminishing the overall measure of correlation.

Conventionally we call

$$X_i = \frac{x_i - \bar{x}}{s_x} \quad \text{and} \quad Y_i = \frac{y_i - \bar{y}}{s_y}$$

standardized variables

With these standardizations, that’s it: We use the average cross product of these standardized forms of the original variables as the measure of linear correlation, naming it “ r ”:

$$r = \frac{1}{n} \sum_{i=1}^n X_i Y_i$$

This “ r ” is very very important in the standard least squares approach. Enjoy the simplicity of the concept and the simplicity of the equation (in standard form) — but remember the thread that got us through the data, in this case through years and log populations, to the simple “ X ” and “ Y ” and r . Remember the thread because when we get to the next step in the standard “least squares” approach you have to do two things: You have to use these standardized forms, and you also have to remember how to follow the thread back to the data. The number “ r ” is an abstraction. And, as always, the numbers must come back to the data.

Exercises

Most statistical programs will compute "r" for you. Don't do that. For now, compute r by working through the detail: Compute the mean for each variable. Compute the standard deviation for each variable. Compute the standardized form for each variable. And compute the mean cross product for the pair of variables, that's the correlation coefficient, r.

Consider the correlation between U.S. population and year. Compute the correlation coefficient, r, with and without the use of the logarithm of population. What do the correlation coefficients report?

Consider the correlation between body weight and brain weight. Compute the correlation coefficient, r, with and without the use of the logarithms of each variable. What do the correlation coefficients report?