Neuendorf

Covariation/Cross-Product, Covariance, and Correlation

Attached is one good introduction to the Pearson correlation coefficient that examines the background of its calculation. It is from:
Blalock, H. M. (1979). *Social statistics* (revised 2nd ed.). New York: McGraw-Hill.

But first, an introduction to some basic formulas that show how there is a progression from Covariation to Covariance to Correlation:

Covariation is a sum total of how much two variables co-vary, in terms of their original units, using cross-products of how far the scores are from their means. It is sometimes (as in SPSS) called the Cross-product or Cross-product deviations:

$$\text{Covariation}_{XY}\text{ (Cross-product)} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Covariance is an averaged indicator of how much two variables co-vary, in terms of their original units. It is the Covariation divided by the number of cases (pairs of X, Y scores):

$$\text{Covariance}_{XY}\ (\text{Cov}_{XY}) = \frac{\text{Covariation}_{XY}}{n} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n}$$

Correlation is a sort of "standardized" version of the Covariance, that is, the Covariance divided by the product of the two variables' standard deviations:

$$\text{Correlation}_{XY} = \frac{\text{Cov}_{XY}}{sd_x \cdot sd_y} = \left[\frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n}\right] \Big/ sd_x \cdot sd_y$$

So, remember when you see SPSS output that includes Covariations/Cross-products or Covariances, and the values are quite large, that these are in the variables' original units. The Correlation will always be "standardized" such that it ranges from -1.0 to 1.0. Thus, you might find, for example, that measures of "I like sick humor" and "I like humor about sex" (both measured with a 0 to 10 scale) have the following:

Covariation/Cross-product = 934.763
Covariance = 3.527
Correlation = .404

Only the last, the correlation, has much meaning without knowing about (a) the scale of measurement, and (b) the number of cases in the sample. . .

prediction equation, if we knew that there were 8 per cent blacks in a given city, the estimated median income differential would be

$$Y_p = a + b(8) = 561.83 + (19.931)(8) = \$721.28$$

We can see from the figure that approximately the same result would have been obtained graphically. Incidentally, by setting $X = 8$ and solving for Y, we have located a second point on the line which can then be used for the purpose of drawing in the line on the scatter diagram.

17.2 Correlation

Henceforth let us suppose that X is stochastic and therefore not subject to the control of the investigator. Not only do we want to know the *form* or nature of the relationship between X and Y so that one variable can be predicted from the other, but also it is necessary to know the *degree* or strength of the relationship. Obviously, if the relationship is very weak, there is no point in trying to predict Y from X. Sociologists are often primarily interested in discovering *which* of a very large number of variables are most closely related to a given dependent variable. In exploratory studies of this sort, regression analysis is of secondary importance. As a science matures and as important variables become identified, attention can then be focused on methods of exact prediction. Some statisticians are of the opinion that entirely too much attention has been given to correlation and too little to regression analysis. Whether or not this is true depends, of course, on the state of knowledge in the science concerned.

The correlation coefficient r to be discussed in this section was introduced by Karl Pearson and is often referred to as product-moment correlation in order to distinguish it from other measures of association. This coefficient measures the amount of spread about the *linear* least-squares equation. There is a comparable population coefficient rho (ρ) that measures the goodness of fit to the true regression equation. We obtain an estimate r of this parameter by measuring deviations from the line computed by least squares.

Since the regression equation represents the path of the means of Y's for given X's, it would also be possible to measure spread about this line by taking a standard deviation from the line.[6] Researchers in most

[6] The exact nature of such a measure will be discussed below. For the present we can simply point out that it represents an extension of the notion of a standard deviation, where the mean of the Y's is no longer taken as fixed but is considered to be a function of X.

applied fields have become accustomed to the correlation coefficient, however, and the correlation coefficient is probably here to stay. It has the advantage of being easily interpreted, and its range is from -1.0 to 1.0, a fact that is appealing to most practitioners. As we shall see, the relationship between the correlation coefficient and the standard deviation about the least-squares line is a very simple one, and this fact can be used to provide an interpretation for r.

It has been mentioned that r has an upper limit of 1.0. If all points are exactly on the straight line, r will be either 1.0 or -1.0 depending on whether the relationship is positive or negative. If the dots are randomly scattered, r will be zero. The better the fit, the larger the magnitude of r. This is indicated in Fig. 17.7.

Notice that r is a measure of *linear* relationship, being a measure of the goodness of fit of the least-squares straight line. You should not jump to the conclusion that if $r = 0$ (or if $\rho = 0$), there is no relationship whatsoever. If there is no relationship, it follows that r will be approxi-

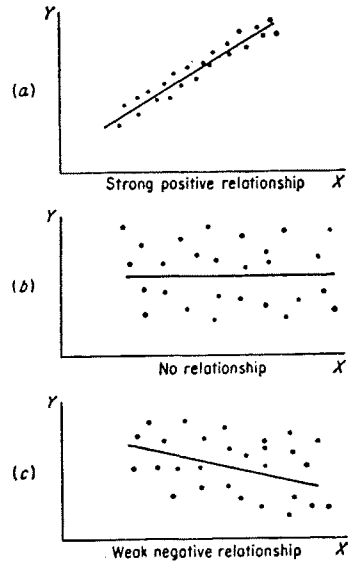

Figure 17.7 Scattergrams showing different strengths and directions of relationships between X and Y.

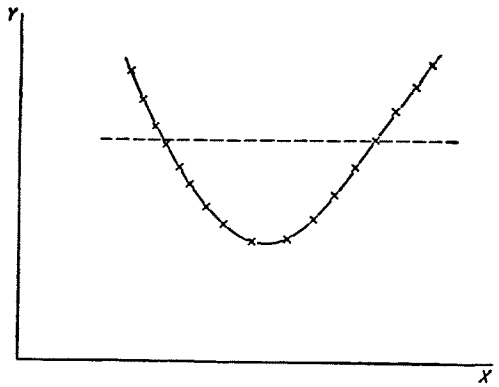

Figure 17.8 Scattergram for perfect nonlinear relationship for which $r = 0$.

mately zero and there will be a random scatter of points. There may, however, be a perfect curvilinear relationship and yet r can be zero, indicating that there is no *straight line* that can fit the data. In Fig. 17.8 this is actually the case. Therefore, if a researcher finds a correlation of zero, he or she should be careful not to infer that the two variables are unrelated. Usually, inspection of the scattergram will indicate whether there is in fact no relationship or whether the relationship is sufficiently nonlinear to produce a zero correlation. In most sociological problems relationships can be approximated reasonably well with straight lines. This does not mean that one should not be alert for possible exceptions, however.

We have not as yet defined the correlation coefficient, but we can easily do so in terms of the formula

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{[\Sigma(X - \bar{X})^2][\Sigma(Y - \bar{Y})^2]}} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} \qquad (17.6)$$

In words, the correlation coefficient is the ratio of the covariation to the square root of the product of the variation in X and the variation in Y. Dividing numerator and denominator by N, writing this quantity as N^2 within the radical, we see that r can also be defined as the ratio of the covariance to the product of the standard deviations of X and Y. The

covariance is a measure of the joint variation in X and Y, but its magnitude depends on the total amount of variability in both variables. Since the numerical value of the covariance can be considerably greater than unity, it is inconvenient to use it directly as a measure of association. Instead, we standardize by dividing by the product of the two standard deviations, thereby obtaining a measure that varies between -1.0 and 1.0.

We have already seen that the covariance will be zero whenever X and Y are unrelated. It can also easily be shown that the upper limit of r is unity. Let us take the case where b is positive and where all points lie exactly on the line. Then for every Y we can write $Y = a + bX$. Since $(\bar{X},\bar{Y})$ also lies on the line, we have $\bar{Y} = a + b\bar{X}$. Therefore for all points on the line

$$Y - \bar{Y} = (a + bX) - (a + b\bar{X}) = b(X - \bar{X})$$

Hence $$\Sigma(X - \bar{X})(Y - \bar{Y}) = b\Sigma(X - \bar{X})^2$$

and $$\Sigma(Y - \bar{Y})^2 = b^2\Sigma(X - \bar{X})^2$$

Inspection of the numerator and denominator of r now indicates that under these conditions $r = 1.0$. Similarly, it can be shown that if all points lie exactly on a line with negative slope, the resulting r will be -1.0.

The relationship between the correlation coefficient and the slopes of the two least-squares equations should also be noted. If we let b_{yx} be the slope of the least-squares equation estimating the regression of Y on X, and if we let b_{xy} indicate the slope of the estimate of the regression of X on Y, we have by symmetry that

$$b_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2}$$

where $$X = a_{xy} + b_{xy}Y$$

Thus, r has the same numerator as both b's. If the b's are zero, it follows that r must also be zero and vice versa.

This might seem to lead to the conclusion that the strength of relationship is proportional to the slope of the least-squares line. This will be true

only if the denominator remains fixed, however. As will be discussed below, the numerical values of the b's depend on the size of the units of measurement.

The value of r has been standardized so that it is independent of the relative sizes of the standard deviations in X and Y. It would indeed be unfortunate if this were not the case, since we would hardly want a measure that varied according to whether we selected dollars or pennies as our monetary unit. It will be noted from the formulas for r and the b's that r^2 can be expressed in terms of the b's. Thus

$$r^2 = b_{yx}b_{xy} = \frac{[\Sigma xy]^2}{\Sigma x^2 \Sigma y^2} \tag{17.7}$$

You should verify that when r is 1.0 (or -1.0), $b_{yx} = 1/b_{xy}$ and that this means that the two least-squares equations coincide. Generally, as r approaches zero, the angle between the two lines becomes larger and larger until when $r = 0$, the lines have become perpendicular or orthogonal.

Finally, we can introduce a computing formula for r that involves the five sums previously obtained in connection with the computations of a and b.[1] The formula is

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \tag{17.8}$$

The numerator has, of course, already been computed, and so has part of the denominator. Thus the correlation between per cent black and the index of discrimination is

$$r = \frac{13(43{,}943.32) - (62.88)(8{,}557)}{\sqrt{[13(432.2768) - (62.88)^2][13(6{,}192{,}505) - (8{,}557)^2]}}$$
$$= \frac{33{,}199}{110{,}120} = .301$$

It should be noted that one can add or subtract values from either X or Y without affecting the value of the correlation coefficient. Likewise, r will be unaffected by a change of scale in either variable. This says in effect that the correlation between income and education is the same regardless of whether income is measured in dollars or pennies. But

[1] Except in cases where confusion might arise, we shall continue to make use of b without the subscripts to represent b_{yx}.

although the correlation coefficient is invariant under transformations of this sort, the least-squares equation is not. Adding or subtracting values will affect the numerical value of a. A change of scale will affect the slope of the line. For example, if every X is divided by 10 while Y is kept fixed, the resulting b will be multiplied by 10. You should verify that these properties hold by examining the formulas for r, a, and b. These facts may be used in order to simplify computations. For example, if X involves either a very large number or a very small decimal, a change of scale may reduce the risk of computing errors. Or if the X variable consists of values such as 1,207, 1,409, 1,949, and 1,568, it would probably be advisable to subtract 1,000 from each score. Certain computing routines require that all values be positive. In computing r, therefore, it may be necessary to add to each value a number that is slightly larger than the largest negative score.

Another fact about the correlation coefficient should be noted at this time. Since this measure involves both variances and covariances, it is highly affected by a few extreme values of either variable. Furthermore, the magnitude of r depends on the degree of general variability in the independent variable. Figure 17.9 illustrates these points. In Fig. 17.9*a*, the effect of one or two extreme values is to produce a moderately high correlation where none exists among the remaining cases. In Fig. 17.9*b*, we have a moderately high linear relationship except for the fact that extreme cases are out of line with the rest. In this latter instance we probably have an example of a nonlinear relationship. A scattergram will always be helpful in indicating the nature of the situation in any given problem. Let us now discuss what can be done if either of these situations should occur.

Figure 17.9*a* illustrates the point made above that the magnitude of the correlation coefficient depends on the range of variability in both variables. Had there been a larger number of extreme cases, the resulting distribution might have been as in Fig. 17.10. In this instance the overall correlation may be high, but within any limited range of X's the correlation may be close to zero. In effect, this indicates that there is insufficient variability in X within this limited range to counteract the effects of numerous uncontrolled variables. In reality, X is almost being held constant. Therefore, if a scattergram turns out to be similar to the one in Fig. 17.9*a*, one should always attempt to extend the range of variability in X by finding more extreme cases.

If extending the range of variability is not feasible empirically or if the researcher's interest is focused primarily on less extreme cases, it may be more sensible to exclude the extreme cases from the analysis

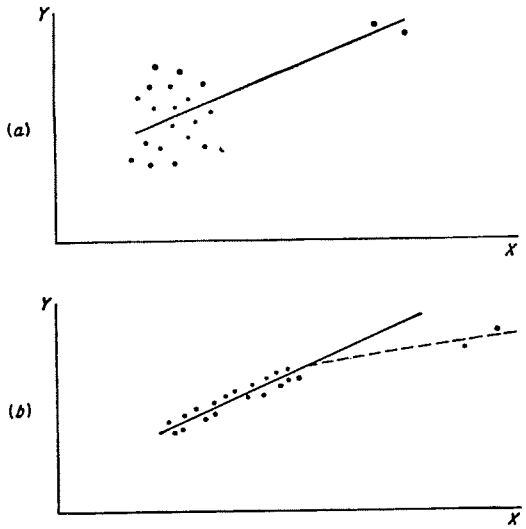

Figure 17.9 Scattergrams showing possible effects of extreme X values.

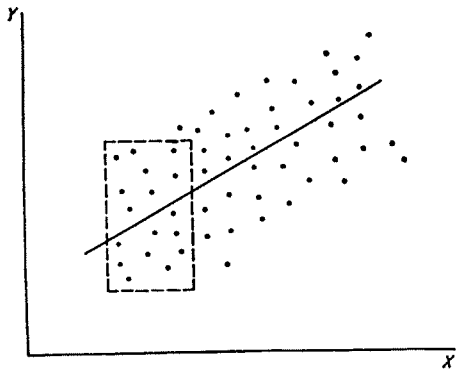

Figure 17.10 Scattergram showing no relationship within a limited range of variation in X but a positive relationship over the total range.

altogether. For example, suppose X is size of city and New York City appears in the sample. Unless there are a large number of cities of comparable size, and there are not, it may become necessary to confine one's attention to cities of less than 500,000. In some instances it would seem advisable to compute r both with and without the extreme cases. Obviously the decision made will depend upon the nature of the problem and the research interests of the social scientist. You should be cautioned that one or two extreme scores can have a very pronounced effect on the magnitude of r, and you should always take this into consideration in some manner. The range of variability should therefore be reported with correlation coefficients. This is another illustration of the important point that a single summarizing measure, no matter how superior it may be to other measures, can often be misleading.

If the data turn out to be as in Fig. 17.9*b*, we would obviously suspect nonlinearity. Again, if possible, additional extreme cases should be obtained. If there are only one or two extremes, it may be advisable to exclude these from the analysis. Situations of this sort illustrate the fact that within a limited range of variation a relationship may be approximately linear, but when extended, the linear model may be inappropriate. You should therefore be careful not to generalize beyond the limits of the data. A cautious statement such as, "Within the limits of _____ and _____ the relationship appears to be approximately linear" would be most appropriate.