

Chapter 13

PROBLEMS IN CORRELATION AND REGRESSION

Hayes, W. L. (1981). *Statistics*.
New York: Holt, Rinehart, & Winston

So far in applying the different variations of the general linear model we have confined our attention to situations where the independent variable X represents a set of essentially qualitative distinctions. These distinctions have sometimes represented preformed groups of subjects, and at other times they have stood for different levels of some treatment actually administered by the investigator. The model adopted in these instances has treated the variable X strictly as an indicator, taking on only the values 0 and 1 depending on which of a set of groups is being specified.

Now we are going to consider situations in which the independent variable X takes on *any real number value*. Such a variable may, for example, represent the amount of some treatment that a subject received in an experiment, or it may be the score that the subject earned on some test. In short, we are now going to deal with quantitative independent or predictor variables as well as a quantitative dependent variable Y .

One distinction that we will find useful concerns the sampling scheme employed for obtaining the values of the independent variable X . In some situations, the experimenter exercises no control whatsoever over the values of X that occur in the study, nor over the values of Y . Rather, in this sampling situation each individual included simply "brings along" a value of X and a value of Y . We will call this approach "a problem in correlation."

On the other hand, it may be that the investigator forms groups, and then arranges that a given group will receive a given, predetermined, value of X . This might be the situation in an experiment in which different amounts of medication are given to different groups of patients. Here, the values of X are not sampled, as in a problem in correlation, but rather are deliberately preselected. This sampling scheme will be called "a problem in regression."

444

In this chapter we will usually assume that the data are collected as for a problem in correlation, and defer consideration of the other sampling situation until Chapter 14. Although, both theoretically and mechanically, the methods employed in these two situations are virtually identical, the two approaches do differ in the actual sampling methods used, the breadth of inferences made from the results, and in some of the assumptions ordinarily made about the population sampled.

445
SIMPLE LINEAR
RELATIONS
Section 13.1

13.1 / SIMPLE LINEAR RELATIONS

Problems in correlation and in regression are both concerned with three main questions:

1. Does a statistical relation affording some predictability appear between the random variables X and Y ?
2. How strong is the apparent degree of the statistical relation, in the sense of possible predictive ability the relation affords?
3. Can a simple rule be formulated for predicting Y from X , and if so, how good is this rule?

The ordinary techniques we have studied heretofore apply to the first two of these questions, but the third is a new feature. In this chapter we are going to study the possibility of applying the *linear model as a rule for the prediction of Y from X* . We are going to act *as though* the true relation actually were a function, and, using a function rule, make predictions or "bets" about Y values from knowledge of X values. Then we are going to evaluate the *goodness* of this prediction rule in terms of how well one actually would do by predicting according to the rule. If the statistical relation actually is a function then some rule exists that affords perfect prediction; for the usual statistical relation, no rule permits perfect prediction, but some function rule may nevertheless provide a good "fit" to the relation under study. Proceeding in this fashion gives us two important advantages: quite often we are able to achieve a fair degree of predictive ability by adopting a particular function rule, even though the true relation itself is not really a precise function. Second, by studying our errors using this rule, we gain information about how the rule might be made better and how the general form of the relation might be specified more adequately.

The model that we will use to describe the relation between X and Y will be linear, just as in Chapters 10 through 12. However, the model we will first employ will be very simple:

$$y_i = a_0 + bx_i + e_i, \quad [13.1.1]$$

where, as before, a_0 is a constant that enters into the value of y_i for any individual i , b is a constant weight that applies to the value x_i , and e_i stands for error. Note that if there were no errors, then

$$y'_i = a_0 + bx_i, \quad [13.1.2]$$

which is strictly linear in the sense that a plot of all of the (x_i, y'_i) pairs of values would fall along a straight line.

Naturally, there is no law that says the relationship between values such as x_i and values such as y_i must be linear. The best description of how the x_i are related to y_i

values in a given set of data might call for a very different mathematical function. Why, then, do we emphasize such linear rules or models?

The reasons for starting with linear rules for prediction are several: linear functions are the simplest to discuss and understand; such rules are often good approximations to other, much more complicated, rules; and we will find that in certain circumstances the only prediction rule that *can* apply is linear. However, do not jump to the conclusion that just because we deal first with linear prediction this is the only important way to predict, or that all real relationships must be more or less linear functions. In the next chapter we will find that there are many other, nonlinear, function rules that might also be applied to a given problem.

Before we can turn to inferential methods appropriate to problems in correlation, we need some terminology for describing a linear relation between variables X and Y in any fixed set of N observations, or sample of N cases. Thus, the succeeding few sections will deal with the descriptive statistics of correlation and regression.

13.2 / THE DESCRIPTIVE STATISTICS OF CORRELATION AND REGRESSION

Just as we discussed how one could find and interpret the mean and variance as descriptions of particular aspects of a set of data, we will now turn to the problem of finding a linear rule that "fits" a given set of data as well as possible. For the moment, our interest is only in a specific set of data, the scores for some particular set of N observed individuals.

Imagine this kind of situation: a teacher of a large introductory college course is interested in the possible relationship between the high school preparation in mathematics that a student has and success in the course. In a particular semester the teacher has a class of 91 students, and at the outset each student is asked the number of mathematics courses taken in high school (four years). The teacher weights these courses in a routine way and assigns scores running from 2 through 8 to the students. Let us call these "mathematics scores" X .

The teacher, however, files these reports away and does not look at them until after the final examination in the course has been given. The actual raw scores on this examination will be called the variable Y . After both scores for each student are known, the teacher asks this question: "To what extent is there a linear relation between the X and the Y scores?" In other words, how well does a simple linear rule allow one to predict the Y score of a student drawn at random from this group, given the information about the X score? The problem is to find the best possible linear rule for predicting from these data, and then to evaluate the goodness of such a rule.

Actually, the teacher is not especially interested in predicting the raw Y score of a student so much as in the *relative* performance of the student in terms of Y . That is, the teacher would like to be able to predict the standard score z_y , given by

$$z_y = \frac{y - M_y}{S_y}$$

This prediction is to be based on the standard score z_x , where

$$z_x = \frac{x - M_x}{S_x}$$

Since a linear rule is to be used for this prediction, this means a rule of the form

$$z'_y = A + Bz_x \quad [13.2.1]$$

where B and A are constants. (Here, and in the following, we will use capital letters for the unknown constants in a rule involving z -scores, and lowercase letters for rules involving raw score values.) The predicted score is labeled as z'_y to indicate that it *need not* be the same as z_y , the true standard score for any given individual. Several individuals may have the same z_x standard score, but quite different z_y scores; by use of the rule, however, *only one* z'_y or predicted standard score will be given for each z_x value.

The problem is shown graphically in Figure 13.2.1. Here, the horizontal axis represents possible values for z_x , the vertical axis the possible values for z_y , and any point within the plane defined by these two axes represents a pair of z -scores (z_x, z_y) that might be associated with any individual observation. Points in the functional relation $z'_y = A + Bz_x$ lie along the straight line in the figure. For the particular value of z_x represented in the figure, the linear rule affords a *predicted* value z'_y ; this *need not* correspond to the actual value z_y corresponding to any individual showing the particular value of z_x shown in the figure. The extent of "miss" or error between the predicted value z'_y and z_y for an individual is represented by the vertical distance between the two points (z_x, z_y) and (z_x, z'_y). We would like our prediction rule to be such that, across all individuals, the fit between predicted and actual standard scores on Y is as good as possible. The reader may be puzzled by the use of the term "prediction" in this context; the teacher actually *has* the two scores for each of the 91 students. Why not merely look at the standardized Y score for any student? Actually the methods to be developed here will apply to situations where the user wants to go beyond the immediate data, and to forecast the Y or z_y score for an individual for which this information is not already available. However, the basis for these methods is best seen if one deals only with one intact group of N cases, each having two scores, X and Y . For the moment, "prediction" consists of drawing one case at random from this particular group, noting the z_x , and then finding a predicted value of z'_y by use of the linear rule.

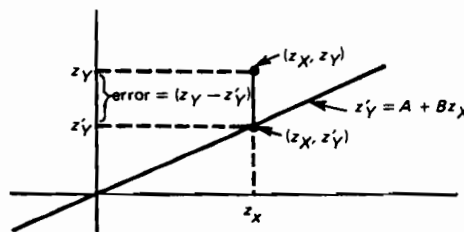


Figure 13.2.1

Plot of a linear regression equation for the prediction of the standard score on Y from the standard score on X .

The first problem is to find constants A and B that will make the linear rule give the "best possible" predictions. These constants are found by the method of least squares, which we have already encountered in Chapters 4 and 10. In this context, applying the criterion of least squares means that we want to minimize the sum of

squared errors in prediction. Thus, for any individual case i we will make some prediction, z'_i ; this need *not* be the same as the true value z_i for that individual, and so some error will exist.

$$e_i = (z'_i - z_i).$$

The least-squares criterion requires that we choose A and B in such a way that the average squared error over individual predictions be as small as possible. Thus, given N individuals i , we want to choose A and B so as to make

$$\frac{\sum_i (z'_i - z_i)^2}{N} \quad [13.2.2]$$

have its minimum possible value. (Please understand that in the following we are able to use elementary algebra to find the values of A and B only because the problem is a very simple one. In actual practice, the methods of the differential calculus would be used in this or any more complicated situation in order to find a least-squares solution.)

Now, first of all we will show that by the least-squares criterion, if we are predicting standard scores, the value of the constant A must be zero, so that the best linear rule is actually

$$z'_i = Bz_{xi}. \quad [13.2.3]$$

This can be shown as follows: Substituting 13.2.1 into 13.2.2 and rearranging terms we have

$$\frac{\sum_i (z'_i - z_i)^2}{N} = \frac{\sum_i [(Bz_{xi} - z_i) + A]^2}{N}. \quad [13.2.4]$$

On carrying out the square for each i and summing, we have

$$\begin{aligned} \frac{\sum_i [(Bz_{xi} - z_i) + A]^2}{N} &= \frac{\sum_i (Bz_{xi} - z_i)^2}{N} + 2A \frac{\sum_i (Bz_{xi} - z_i)}{N} + \frac{\sum_i A^2}{N} \\ &= \frac{\sum_i (Bz_{xi} - z_i)^2}{N} + A^2 \end{aligned} \quad [13.2.5]$$

since A and B are constants, and the mean of each set of z -scores must be zero.

Now, assuming B fixed, for what value of A can the expression on the right in 13.2.5 be at its smallest? The first term is a mean of squared numbers and hence must be positive, and A^2 must be positive as well; it follows that this entire expression can be at its smallest value *only* when A is zero. Thus, the value of A dictated by the least-squares criterion is zero.

Next, we will show that by the least-squares criterion the value of B for predicting z -scores must be

$$B = \frac{\sum_i z_{xi}z_{yi}}{N} = r_{xy}. \quad [13.2.6]$$

This value of B is actually the **correlation coefficient** (or Pearson product-mo-

ment correlation coefficient) r_{xy} , about which we will have much to say. By defini-
tion, [13.2.7*]

$$r_{xy} = \frac{\sum_i z_{xi}z_{yi}}{N}.$$

Thus, by the criterion of least squares, our prediction rule must be

$$z'_i = r_{xy}z_{xi}. \quad [13.2.8*]$$

We can show as follows how the least-squares criterion dictates that $B = r_{xy}$ for the prediction of standard scores. Since we know that A must be zero, this time we start by substituting expression 13.2.3 into 13.2.2.

$$\frac{\sum_i (z'_i - z_i)^2}{N} = \frac{\sum_i (Bz_{xi} - z_i)^2}{N}. \quad [13.2.9]$$

Expanding the square on the right hand of 13.2.9 and summing, we have

$$\begin{aligned} \frac{\sum_i (z'_i - z_i)^2}{N} &= \frac{B^2 \sum_i z_{xi}^2}{N} - \frac{2B \sum_i z_{xi}z_{yi}}{N} + \frac{\sum_i z_{yi}^2}{N} \\ &= B^2 - 2Br_{xy} + 1 \end{aligned} \quad [13.2.10]$$

since the variance of standard scores is always 1. Now suppose that B differed from r by some amount c , either a positive or negative number:

$$B = r + c.$$

Substituting $r + c$ for B in 13.2.10 above, we find

$$\begin{aligned} \frac{\sum_i (z'_i - z_i)^2}{N} &= (r + c)^2 - 2(r + c)r + 1 \\ &= r^2 + 2rc + c^2 - 2r^2 - 2rc + 1 \\ &= (1 - r^2) + c^2. \end{aligned} \quad [13.2.11]$$

When $B = r_{xy}$, so that c is zero, the mean squared error must be at its smallest,

$$\frac{\sum_i (z'_i - z_i)^2}{N} = (1 - r_{xy}^2), \quad [13.2.12*]$$

and for any $c \neq 0$, the value must be $(1 - r_{xy}^2)$ plus a positive number, c^2 . Hence taking $B = r_{xy}$ gives the least squared error in linear prediction, on the average, for standard scores.

Expression 13.2.8 is called the **prediction equation** (or **regression equation**) for z_y predicted from z_x . The constant B as found from 13.2.6 is called the **standardized regression coefficient** or **standardized regression weight** for predicting z_y from z_x . Thus in the special case where standard scores z_y are to be predicted from standard scores z_x , the value of $B = r_{xy}$. However, as we will see, when raw scores are to be predicted, a somewhat different form of regression weight will be needed.

The notion of the mean squared error (13.2.2) is an important one in its own right, and we will give it a special symbol and name. Let

$$S_{z'z}^2 = \frac{\sum (z'_i - z_{yi})^2}{N} = 1 - r_{xy}^2 \quad [13.2.13^*]$$

be called the **sample variance of estimate for standard scores**. This variance of estimate reflects the *poorness* of the linear rule for prediction of standard scores, the extent to which squared error is, on the average, large.

Most often, however, this index is discussed in terms of its positive square root

$$S_{z'z} = \sqrt{1 - r_{xy}^2} \quad [13.2.14^*]$$

which is called the **sample standard error of estimate for predicting standard scores**.

Obviously, there is a close connection between the size of the standard error of estimate in a sample and the value of r in the regression equation: **the larger the absolute value of r_{xy} , the smaller is the standard error of estimate**. Now we turn to some interpretations of the index r_{xy} .

13.3 / SOME PROPERTIES OF THE CORRELATION COEFFICIENT IN A SAMPLE

The connection between the variance of estimate and the correlation coefficient shows us at once that r_{xy} can take on values *only* between -1 and 1 . Notice that the variance of estimate, being a weighted sum of squares, can be only a positive number or zero. If r_{xy} were less than -1 , or greater than $+1$, then the variance of estimate could not be positive. Hence, $-1 \leq r_{xy} \leq 1$.

What does it mean when r_{xy} is exactly zero? When this is true, one predicts $z'_y = 0$, corresponding to the mean Y , *regardless* of the value of X : for any X , the mean of Y is the best linear prediction when the correlation is zero. Furthermore,

$$S_{z'z}^2 = 1$$

for $r_{xy} = 0$. This means that when the correlation coefficient is zero the variance of estimate for standard scores is exactly the same as the variance of the standard scores z_y with X *unspecified*. Thus, when r_{xy} is zero, predicting by the linear rule *does not* reduce the variability of z_y below the variability present when z_x is unknown. In short, the fact that r_{xy} is zero implies that if a predictive statistical relation exists for the set of data, it is not linear, and the linear rule gives no predictive power.

On the other hand, when r_{xy} is either $+1$ or -1 , the variance of error in prediction is zero, so that each prediction is exactly right. These ultimate limits for r_{xy} can occur only when X and Y are functionally related, and follow a linear rule.

All intermediate values of r_{xy} indicate that some prediction is possible using the linear rule, but that this prediction is not perfect, and some error in prediction exists. Any value of r_{xy} between 0 and 1 in absolute magnitude indicates either that the relationship is not functional or that if it is a function, the rule is not exactly linear, although a linear rule does afford some predictability.

In the regression equation for standard scores the correlation coefficient plays the role of converting a standard score in X into a predicted standard score in Y . Rather loosely, the correlation coefficient can be said to be "the rate of exchange," the value of a "standard deviation's worth" of X in terms of *predicted* standard deviation units of Y .

Finally, the fact that we are able to define the correlation coefficient in terms of standard scores shows that r_{xy} is a *dimensionless index* of linear relationship. This means that r_{xy} does not depend on the units of measurement for either X or Y , nor on what value is called "zero" for either variable. Either X or Y or both can be given a linear transformation (multiplying by a positive constant and adding another constant), and the correlation between the transformed variables will be the same as r_{xy} .

13.4 / THE PROPORTION OF VARIANCE ACCOUNTED FOR BY A LINEAR RELATIONSHIP

Just how good is a linear rule for predicting values of Y from values of X in a given sample? In order to answer this question, we need an index of the *strength of linear relationship* between X and Y in the data. We can approach this problem as follows:

We already know that the variance of any set of standard values z_y has to be 1.00 . However, what is the variance of the *predicted* values z'_y ? We can think of this as *variance accounted for, variability among the z_{yi} values for different observations i directly attributable to the fact that they have different z_x scores*. When we take the variance of the predicted z'_y values we have

$$S_{z'_y}^2 = \frac{\sum (z'_i)^2}{N} - \left(\frac{\sum z'_i}{N} \right)^2 \quad [13.4.1]$$

First of all notice that

$$\sum \frac{z'_i}{N} = \sum \frac{r_{xy} z_{xi}}{N} = 0,$$

since the sum of all of the z_{xi} values must be zero. Thus, the second term on the right, the squared mean of all the predicted z'_y values, is itself zero. Then we find

$$\sum \frac{(z'_i)^2}{N} = \sum \frac{r_{xy}^2 (z_{xi})^2}{N} = r_{xy}^2 \quad [13.4.2^*]$$

The variance of the predicted z'_y values, or the variance explained by the z_{xi} values, is thus r_{xy}^2 .

Since the variance of the original z_y values has to be 1.00 , then if we take the ratio of the variance of the predicted values to the total variance of the z_y values, we find that

$$\left(\begin{array}{l} \text{proportion of variance explained} \\ \text{by linear regression of } Y \text{ on } X \end{array} \right) = \frac{S_{z'_y}^2}{S_{z_y}^2} = r_{xy}^2 \quad [13.4.3^*]$$

the proportion of variance in Y accounted for, or explained, by X is given by r_{xy}^2 . Thus, an index of the "goodness" of the linear rule for predicting X from Y is given by r_{xy}^2 , the proportion of variance in X accounted for by Y under the linear rule. The

index r_{xy}^2 is usually termed the coefficient of determination. You can always think of r_{xy}^2 as representing the strength of linear relationship in a given set of data. Furthermore, although it was convenient to discuss the proportion of variance accounted for initially in terms of z-values, r_{xy}^2 is the same either for z-scores or for raw values so that these interpretations are valid for r_{xy}^2 whether we are discussing standardized or raw scores.

Thus, if the value of a correlation coefficient is .50 (positive or negative in sign) then some .25 of the variability in Y is accounted for by specifying the linear rule and X . If the correlation is .80, then 64 percent of the variance in Y is accounted for in this way. A correlation of positive or negative 1.00 means that 100 percent of variability in Y can be accounted for by the linear rule and X , but if $r_{xy} = 0$, none of the variability is thereby accounted for. All in all, not the correlation coefficient per se but the square of the correlation coefficient informs us of the "goodness" of the linear rule for prediction.

13.5 / THE IDEA OF REGRESSION TOWARD THE MEAN

The term *regression* has come to be applied to the general problem of prediction by use of a wide variety of rules, although the original application of this term had a very specific meaning. The term "regression" is a shortened form of **regression toward the mean in prediction**. The general idea is that **given any standard score z_x , the best linear prediction of the standard score z_y is one relatively nearer the mean of zero than is z_x** .

This can be illustrated quite simply from our regression equation 13.2.8. Suppose that an individual has a standard score z_x of 2. Also suppose that the regression equation we have found for the group to which the individual belongs is

$$z'_y = .5z_x.$$

Then we *predict* this individual to have a z_y score of 1, since

$$z'_y = .5(2) = 1.$$

Notice that we predict the individual to fall relatively *nearer* the mean on Y than on X . That is, we predict in accordance with *regression toward the mean*. For another set of data, the regression equation might be

$$z'_y = -.75z_x.$$

Now in this instance, suppose that the z_x for some randomly selected individual were 1.5. Then

$$z'_y = (-.75)(1.5) = -1.125.$$

Since the correlation coefficient is negative, the prediction is that this individual falls *below* the mean of Y , given that the X value is above the mean.

However, in absolute terms, we again predict a standing relatively *closer* to the mean on Y than on X .

This principle of predicting relatively closer to the mean, or regression toward the mean, is a feature of any linear prediction rule that is best in the "least-squares" sense of Section 13.2. The idea is that if we are going to use such a linear rule for prediction, then it is always a good bet that an individual will fall *relatively closer*

to the group mean on the thing predicted than on the thing actually known. This does not imply that the actual Y value *must* fall relatively closer to the mean than does the value of X , however, but only that our best bet is that it will do so. Regression toward the mean is not some immutable law of nature, but rather a statistical consequence of our choosing to predict in this linear way, using the criterion of least squares in the choice of a rule.

13.6 / THE REGRESSION OF z_x ON z_y

In a true correlation problem, nothing makes it necessary to think of X as the independent variable, or the value somehow known first or predicted from. It is entirely possible to consider a situation where one might want to predict z_x from knowing z_y . What does this do to the linear prediction rule, the correlation coefficient, and so on?

In the first place, the same argument used in Section 13.2 shows that for predicting z_x from z_y ,

$$z'_x = r_{xy}z_y, \quad [13.6.1]$$

where the correlation coefficient is, just as before, given by 13.2.7.

It is tempting to ask why we do not just solve the original regression equation for z'_y in terms of z_x in order to get

$$z_{xi} = z'_y / r_{xy}.$$

However, recall what the symbols z'_y and z'_x actually represent. These are *predicted* values and do not necessarily symbolize the actual values of z_y and z_x at all. Solving the expression 13.2.8 for z_x might be useful if one wanted to know the value of z_x known, given that z'_y were the predicted value, although it is hard to see why anyone would ordinarily want this information. The form of the regression equation used (13.2.8 or 13.6.1) depends strictly on which variable, X or Y , is designated as the independent variable, or the thing known first in a prediction situation.

In prediction of X from Y , the sample variance of estimate for standard scores is

$$S_{z_x:z_y}^2 = 1 - r_{xy}^2 \quad [13.6.2]$$

(notice the reversal in subscripts from 13.2.13 when z_x is predicted from z_y). The proportional variance in X accounted for by Y is, once again, r_{xy}^2 .

This brings up the point that the correlation coefficient is a *symmetric* measure of linear relationship. So long as we are talking about the correlation coefficient alone, it is immaterial which we designate as the independent and which the dependent variable; the measure of possible linear prediction is the same. However, when we deal with the actual regression equations themselves, this symmetry is not usually present. As we shall see in the next section, it does make a difference whether you are predicting Y from X or X from Y when it comes to finding the regression equations and errors of estimate for raw scores.

13.7 / THE REGRESSION EQUATIONS FOR RAW SCORES

Up to this point, we have considered only the problem of predicting standard scores from standard scores. Introducing regression and correlation in terms of standard

scores makes the algebra somewhat easier, and the essential ideas somewhat simpler. Nevertheless, each feature of correlation and regression shown for standard score prediction is also valid for the prediction of raw scores. For any given set of data, each standard score corresponds uniquely to some raw score, and vice versa, so that linear prediction which is optimal in standard score terms is optimal in raw score terms as well.

It is quite simple to put the regression equation for prediction z_y from z_x into raw score form. We start with

$$z'_{yi} = r_{xy} z_{xi}$$

which is exactly the same as

$$\frac{(y'_i - M_y)}{S_y} = r_{xy} \frac{(x_i - M_x)}{S_x}$$

where y'_i is the predicted raw score for the individual, and x_i is the known raw score on the independent variable. A little algebraic manipulation gives

$$y'_i = M_y + \frac{r_{xy} S_y}{S_x} (x_i - M_x) \tag{13.7.1*}$$

This is the raw score form of the regression equation for prediction of Y from X .

It will be convenient to write this regression equation as

$$y'_i = M_y + b_{y,x}(x_i - M_x) \tag{13.7.2*}$$

where

$$b_{y,x} = \frac{r_{xy} S_y}{S_x} \tag{13.7.3*}$$

The value $b_{y,x}$ is called the **unstandardized or raw score regression coefficient** of Y on X . This value $b_{y,x}$ gives the best (least-squares) prediction of raw Y scores from raw X scores.

In an identical way, we can turn the regression equation for z_x predicted from z_y into

$$x'_i = M_x + b_{x,y}(y_i - M_y) \tag{13.7.4}$$

This is the raw score form of the regression equation for predicting X from Y . Here

$$b_{x,y} = \frac{r_{xy} S_x}{S_y} \tag{13.7.5}$$

is the unstandardized or raw score regression coefficient for predicting X from Y .

Notice that when no specification is put on which is to be regarded as the independent or predictor variable, there are two possible regression coefficients, $b_{y,x}$ and $b_{x,y}$, and that

$$\sqrt{b_{y,x} b_{x,y}} = r_{xy} \tag{13.7.6}$$

the square root of the product of the two regression coefficients (i.e. their geometric mean) is the correlation coefficient.

Figure 13.7.1 shows the two raw score regression lines that might apply to a given set of data.

C. FC. b. Section.

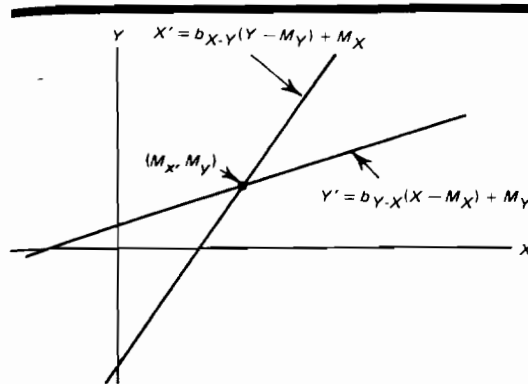


Figure 13.7.1

Plot of the two regression lines for predicting Y from X and X from Y .

When prediction of raw scores is to be carried out, the **sample variance of estimate** for predicting Y from X is

$$S^2_{y,x} = S^2_y(1 - r^2_{xy}) \tag{13.7.7}$$

and the **sample standard error of estimate** is

$$S_{y,x} = S_y \sqrt{1 - r^2_{xy}} \tag{13.7.8*}$$

Similarly, the sample variance of estimate for predicting X from Y is

$$S^2_{x,y} = S^2_x(1 - r^2_{xy}) \tag{13.7.9}$$

and the sample standard error of estimate is

$$S_{x,y} = S_x \sqrt{1 - r^2_{xy}} \tag{13.7.10*}$$

Finally, remember that the proportion of variance accounted for by linear relationship (either in predicting Y from X or X from Y) is r^2_{xy} , just as for standardized scores.

13.8 / COMPUTATIONAL FORMS FOR r_{xy} AND $b_{y,x}$

Although the sample correlation coefficient was actually defined in Section 13.2 as a summed product of standard scores (13.2.7), in practice the index is seldom computed in this way. An equivalent raw score computational form will now be found. We will show that the raw score form of the correlation coefficient is

$$r_{xy} = \frac{(\sum x_y / N) - M_x M_y}{S_x S_y} \tag{13.8.1*}$$

Starting with the definition, 13.2.7, and substituting the raw score equivalents of z_{xi} and z_{yi} , we have

$$\begin{aligned} r_{xy} &= \frac{\sum_i (x_i - M_x)(y_i - M_y)}{NS_xS_y} \\ &= \frac{1}{NS_xS_y} \left[\sum_i x_i y_i - \sum_i x_i M_y - \sum_i y_i M_x + NM_x M_y \right] \\ &= \frac{1}{NS_xS_y} \left[\sum_i x_i y_i - 2NM_x M_y + NM_x M_y \right] \\ &= \frac{\left(\sum_i x_i y_i / N \right) - M_x M_y}{S_x S_y} \end{aligned}$$

Still another computing form for r_{xy} that is often useful when work is being done on a pocket or desk calculator is

$$r_{xy} = \frac{N \sum_i x_i y_i - \left(\sum_i x_i \right) \left(\sum_i y_i \right)}{\sqrt{\left[N \sum_i x_i^2 - \left(\sum_i x_i \right)^2 \right] \left[N \sum_i y_i^2 - \left(\sum_i y_i \right)^2 \right]}} \quad [13.8.2]$$

Given the correlation coefficient, it is then possible to find the sample raw score regression coefficient $b_{y,x}$ directly from

$$b_{y,x} = r_{xy} \left(\frac{S_y}{S_x} \right) \quad [13.8.3]$$

However, for many problems it is desirable to calculate $b_{y,x}$ without first finding r_{xy} . This may be done most simply by taking

$$b_{y,x} = \frac{\sum_i x_i y_i - NM_x M_y}{\sum_i x_i^2 - NM_x^2} \quad [13.8.4^*]$$

or

$$b_{y,x} = \frac{N \sum_i x_i y_i - \left(\sum_i x_i \right) \left(\sum_i y_i \right)}{N \sum_i x_i^2 - \left(\sum_i x_i \right)^2} \quad [13.8.5]$$

A great many hand-held or pocket calculators are preprogrammed to provide the value of the correlation coefficient, along with b and other values. All one has to do is to enter each of the pairs of (x_i, y_i) values, and the calculator produces r_{xy} , in addition to the two means and variances, with $b_{y,x}$ if desired. Many also include the feature of giving the predicted y' value in terms of some entered x value, and x' from y . The ease of use and wide availability of these inexpensive calculators has taken much of the former labor from correlation computations.

If a number of correlations, along with their associated regression weights and

equations must be found, the job will likely be done by computer. Statistical computing packages such as SPSS contain a variety of methods for computing correlations and related statistics. Various other methods for computing r_{xy} are available. One method especially popular in past years is designed for data grouped into a joint frequency distribution or "scatter plot," and this version is to be found in many elementary statistics texts. However, for the reasons just mentioned this is outdated as a practical method, and we will not consider it here.

13.9 / AN EXAMPLE OF CORRELATION COMPUTATIONS FOR A SAMPLE

Consider once again the example of Section 13.2. The teacher collected data for a class of 91 students, obtaining for each a score X , based on the number of courses taken in high school mathematics, and a score Y , the actual score on the final examination for the course. For each of the 91 individuals, the x_i and the y_i values are shown in Table 13.9.1.

x	y	x	y	x	y	x	y
4	36	4	25	3.5	25	7.5	41
3.5	19	4	19	3.5	22	6.5	44
6	38	3	24	3	22	7.5	35
7	52	3	9	2.5	6	5	32
4	20	2	7	2	5	2	3
3.5	12	2	2	3	29	2.5	12
2	10	3.5	34	2.5	26	4.5	16
8	53	3	23	2	17	4	27
3	16	3.5	26	5	41	3.5	17
3	26	6	33	4	24	5.5	25
4.5	27	3.5	17	2.5	7	3.5	17
3	8	4	18	5	19	2.5	16
3	24	2.5	13	2	9	8	40
2.5	23	2.5	10	4.5	28	7.5	38
5.5	32	5	27	6.5	28	6	27
6.5	37	7.5	42	6	34	3.5	23
8	40	8	48	4	18		
4	19	3	26	3.5	23		
3.5	22	2.5	10	3.5	8		
4	35	2	22	6	46		
2.5	18	2	16	6.5	32		
4	25	4	30	7	37		
4	12	4	3	2	14		
2.5	18	6	20	5.5	25		
3.5	18	7	46	4	21		

Table 13.9.1
High school mathematics scores (x) and final examination scores (y).

The mean of the Y values is

$$\frac{\sum_i y_i}{N} = \frac{2169}{91} = 23.84$$

$$\frac{\sum x_i}{N} = \frac{381.5}{91} = 4.19.$$

The two standard deviations are

$$S_x = \sqrt{\frac{\sum x_i^2}{N} - M_x^2} = \sqrt{\frac{1874.75}{91} - 17.56} = 1.74$$

$$S_y = \sqrt{\frac{\sum y_i^2}{N} - M_y^2} = \sqrt{\frac{64043}{91} - 568.34} = 11.64.$$

The correlation coefficient will be computed next:

$$\frac{\sum x_i y_i}{N} = \frac{(4)(36) + (3.5)(19) + \dots + (3.5)(23)}{91} = 116.26$$

so that

$$r_{xy} = \frac{(\sum x_i y_i / N) - M_x M_y}{S_x S_y} \\ = \frac{116.26 - (4.19)(23.84)}{(1.74)(11.64)} \\ = .81.$$

Thus, the regression equation for predicting z_{y_i} from z_{x_i} for these data is

$$z'_{y_i} = .81z_{x_i}.$$

The raw score regression coefficient $b_{y,x}$ is

$$b_{y,x} = r_{xy} \frac{S_y}{S_x} \\ = (.81) \frac{11.64}{1.74} \\ = 5.42.$$

Using this regression coefficient we find that the raw score regression equation for predicting Y from X is

$$y'_i = (5.42)(x_i - 4.19) + 23.84.$$

For instance, given that an individual has a high school mathematics score of 5, the teacher can predict that the score on the course examination is

$$y'_i = (5.42)(5 - 4.19) + 23.84 \\ = 28.23.$$

Figure 13.9.1 shows a plot of the regression equation, together with the actual (x , y) pairs for these data. Notice that although the actual pairs of scores do tend to cluster about the predicted (x , y') pairs, there is nevertheless some "scatter" of the actual Y scores about the predicted value for each X . This, of course, is reflected in the fact that the obtained r_{xy} is not 1.00.

459
ASSUMPTIONS
MADE IN
COMPUTING
CORRELATION AND
REGRESSION
COEFFICIENTS FOR
SAMPLE DATA
Section 13.10

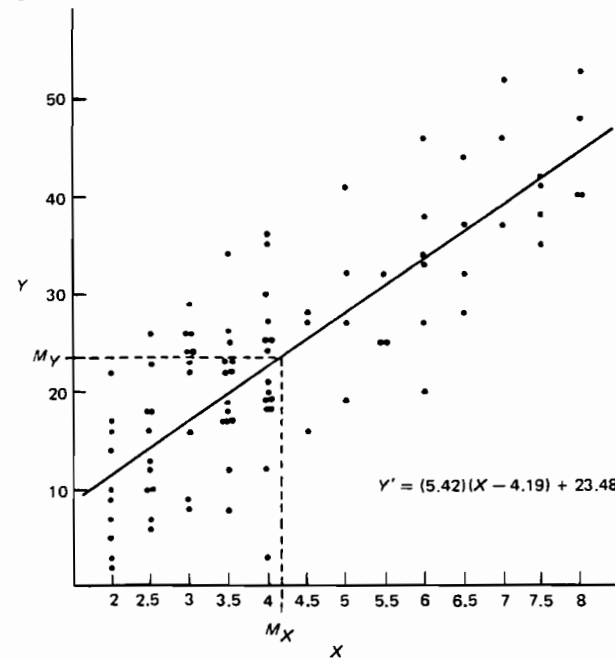


Figure 13.9.1

Plot of the data in Table 13.9.1, showing the regression line for predicting Y from X .

For these data, the sample standard error of estimate is

$$S_{y-x} = S_y \sqrt{1 - r_{xy}^2} \\ = (11.64)(.586) \\ = 6.82.$$

13.10 / ASSUMPTIONS MADE IN COMPUTING CORRELATION AND REGRESSION COEFFICIENTS FOR SAMPLE DATA

A few comments are in order about the propriety of computing correlations and regression equations for sample data. In some of the older literature in social and behavioral research several misleading ideas appear about when it is proper to compute these indices and equations as *descriptive* statistics. We need to be clear about

these matters. It is *not* necessary to make any assumptions at all about the form of the distribution, the variability of Y scores within X columns or "arrays," or the true level of measurement represented by the scores in order to employ linear regression and correlation indices to describe a given set of data. So long as there are N distinct cases, each having two numerical scores, X and Y , then the *descriptive* statistics of correlation and regression may be used. In so doing, we describe the data *as though* a linear rule were to be used for prediction, and this is a perfectly adequate way to talk about the tendency for *these* numerical scores to associate or "go together" in a linear way *in these data*.

The confusion has arisen because in inference about true linear relationships in populations, and in some applications of regression equations to predictions beyond the sample, assumptions do become necessary, as we shall see presently. However, one may apply correlation techniques to any set of paired-score data, and the results are valid descriptions of two things: the particular linear rule that best applies, and the goodness of the linear prediction rule as a summarization of the tendency of Y scores to differ systematically with differences in X *in these data*.

It is true, however, that the possible values that r_{xy} may assume depend to some extent upon the forms of *marginal* distributions of both X and Y in the joint data table. Unless the distributions for X and Y are similar in form, it is not necessarily true that the obtained value of r_{xy} can range between -1 and $+1$. In fact, it is possible to produce examples where the forms of the distributions of X and Y are very different, and the maximum possible absolute value of r_{xy} is only .3 or less. The fact that the value of the correlation coefficient can, in principle, range from -1.00 to $+1.00$ does not imply that the opportunity for a linear predictive relation to appear in a sample has nothing to do with the marginal distributions of the X and Y scores. In the same way, the actual possible range of the correlation in a population depends on the marginal distributions. This fact has very important implications for those who study the patterns of correlations in multivariate studies, and particularly for those who must employ some variant on, or approximation to, the correlation coefficient in such studies. An informative discussion of these issues is given in Carroll (1961).

A related problem with the value of r_{xy} in a sample has to do with selection of cases to appear in the sample, and in particular with systematic restrictions on the range of X (or of Y) values that appear. This introduces a bias into the obtained value of r_{xy} as an estimate of the actual correlation had this selection not been exercised. In particular, the absolute value of the correlation coefficient tends to be lowered by the introduction of such systematic selection reducing the possible range of one or both variables. These matters are also discussed by Carroll.

There is, incidentally, a large literature devoted to special-purpose correlations of various kinds. Among these are the *point-biserial correlation*, designed for the situation in which there are only two independent variable or X values, each accompanied by an array of Y values. Another historically important formulation of the correlation coefficient is the *tetrachoric correlation*, computed when both X and Y are arranged into widespread classes. Still another, the "phi" coefficient, is mentioned in Chapter 15. In addition, various corrections to the correlation coefficient may be made, in order to allow for various kinds of groupings or curtailments of the X and Y values. We will not devote further space to these matters here. A modern

summary of some of these ideas may be found in Guilford and Fruchter (1978), however.

13.11 / POPULATION CORRELATION AND REGRESSION

Imagine a population where each distinct elementary event qualifies for one and only one joint (x,y) event, where X and Y are random variables having some known bivariate distribution. (The general features of bivariate distributions are outlined in Appendix C.) That is, each individual i is assumed to be associated with a pair of values, (x_i, y_i) , and, under random sampling of individuals, there is a probability density to be assigned to each conceivable (x_i, y_i) pair.

Now using the population rather than the sample as a reference group, the linear model for a correlation problem is still

$$y_i = a + bx_i + e_i.$$

However, for the population, if we apply the same least-squares argument as for any other group, we find that the required value of the constant a for predicting raw values is

$$a = \mu_y - \beta_{yx} \mu_x \quad [13.11.1]$$

For the raw score regression coefficient b we have, for the population,

$$\beta_{yx} = \rho_{xy} \frac{\sigma_y}{\sigma_x} \quad [13.11.2^*]$$

The population correlation coefficient ρ_{xy} (lowercase Greek "rho") is defined by

$$\rho_{xy} = \frac{\text{cov.}(X,Y)}{\sigma_x \sigma_y} \quad [13.11.3]$$

where, as defined in Section B.3,

$$E[(X - \mu_x)(Y - \mu_y)] = \text{cov.}(X,Y).$$

That is, the population correlation coefficient is the covariance of X and Y , divided by the product of the population standard deviations for X and Y . (c.f. Section B.3)

Putting these facts together, we obtain the raw score prediction or regression equation for the population as follows:

$$y'_i = \mu_y + \beta_{yx} (x_i - \mu_x). \quad [13.11.4^*]$$

For obvious reasons one is usually more interested in the regression equation as it applies to a population than in the corresponding equation for a sample. Thus, the most useful feature of such an equation is the ability it provides for predicting the y' score of any individual, regardless of whether that individual was a member of some sample. In the same way, in order to make a general statement about the linear relationship that exists between X and Y , we would like to have the value of ρ_{xy} rather than the value of some sample r_{xy} . On the other hand, the sample usually provides all of the information we have about the population. Fortunately, we can make inferences about these population values from the sample, especially if the population has a particular form, as discussed below.

For any such population, we can discuss the *true* variance of estimate for Y predicted from X ,

$$\sigma_{Y.X}^2 = \sigma_Y^2(1 - \rho_{XY}^2), \quad [13.11.5^*]$$

as well as the true variance of estimate for X predicted from Y

$$\sigma_{X.Y}^2 = \sigma_X^2(1 - \rho_{XY}^2). \quad [13.11.6^*]$$

The two **true standard errors of estimate** are thus

$$\sigma_{Y.X} = \sigma_Y \sqrt{1 - \rho_{XY}^2} \quad [13.11.7^*]$$

and

$$\sigma_{X.Y} = \sigma_X \sqrt{1 - \rho_{XY}^2}. \quad [13.11.8^*]$$

Notice that just as for a sample, one can interpret the square of the correlation coefficient as **the proportion of variance accounted for by linear regression**. That is, $\sigma_{Y.X}^2$ is "error" variance in the use of the linear rule, the variability *not* accounted for by linear regression. Thus,

$$\sigma_Y^2 - \sigma_{Y.X}^2 \quad [13.11.9^*]$$

is *the true variance accounted for by linear regression*, or the reduction in variance accomplished by using a linear prediction rule. It follows that

$$\rho_{XY}^2 = \frac{\sigma_Y^2 - \sigma_{Y.X}^2}{\sigma_Y^2},$$

the square of the population correlation coefficient is *the true proportion of variance accounted for* by the use of a linear prediction rule.

Given a random sample, the value of the sample regression coefficient $b_{Y.X}$ is our best available estimate of $\beta_{Y.X}$, the population regression coefficient. Moreover, the best estimate of $\beta_{Y.X}(\mu_X)$ is given by $b_{Y.X}(M_X)$. As usual, our best estimate of μ_Y is simply M_Y .

Since each of these estimates corresponds to a term in the sample regression equation, we can use the sample equation itself as our best estimate of the population regression equation; that is, our estimate of the population regression equation is given by

$$y'_i = b_{Y.X}(x_i - M_X) + M_Y.$$

In the example of Section 13.9.1 the teacher *is* justified in using the sample regression equation to predict for new students drawn from the same population as the original class, *provided* that the class used to find this regression equation is actually a random sample from a population of such students.

In addition, an unbiased estimate of the *true* variance of estimate for predicting Y from X is given by

$$\text{est. } \sigma_{Y.X}^2 = \frac{N}{N-2} S_{Y.X}^2. \quad [13.11.10^*]$$

13.12 / CORRELATION IN BIVARIATE AND MULTIVARIATE NORMAL POPULATIONS

In a problem in correlation our main interest often focuses on the value of r_{xy} itself, especially as an estimate of ρ_{XY} for the population. Given a particular assumption about the population distribution of joint (x,y) events we can test not only if X and Y are linearly related, but also if any systematic relationship at all exists between the two variables.

As shown in Appendix C, in a joint distribution of discrete random variables a probability is associated with each possible X and Y pair, (x,y) . A similar conception holds when X and Y are continuous variables, and a probability is associated with any joint interval of values. Such joint distributions of two random variables are known as bivariate distributions. Although any number of theoretical bivariate distributions are possible in principle, by far the most studied is the **bivariate normal distribution**. The density function for this joint distribution has a rather elaborate-looking rule. However, for standardized variables, this can be condensed to

$$f(z_X, z_Y) = \frac{e^{-G}}{K}$$

where

$$G = \frac{(z_X^2 + z_Y^2 - 2\rho_{XY}z_Xz_Y)}{2(1 - \rho_{XY}^2)}$$

and

$$K = 2\pi\sqrt{(1 - \rho_{XY}^2)}.$$

Notice that in a bivariate normal distribution, the population correlation coefficient ρ_{XY} appears as a parameter in the rule for the density function. Thus, even though z_X and z_Y are both standardized variables, the particular bivariate distribution cannot be specified unless the value of the correlation ρ_{XY} is known.

In a bivariate normal distribution, the marginal distribution of X over all observations is itself a normal distribution, and the marginal distribution of Y is also normal. Furthermore, given any X value, the *conditional* distribution of Y is normal; given any Y , the conditional distribution of X is normal. In other words, if a bivariate normal distribution is conceived in terms of a table of joint events where the number of possible X values, of Y values, and of possible joint (x,y) events is infinite, then within any possible row of the table one would find a normal distribution; a normal distribution also exists within any possible column, and the marginals of the table also exhibit normal distributions.

For our purposes, however, the feature of most importance in a bivariate normal distribution is this: **given that densities for joint (x,y) events follow the bivariate normal rule, then X and Y are independent if and only if $\rho_{XY} = 0$** . For any joint distribution of (x,y) the independence of X and Y implies that $\rho_{XY} = 0$, but it may happen that $\rho_{XY} = 0$ even though X and Y are *not* independent. However, *for this special joint distribution, the bivariate normal, $\rho_{XY} = 0$ both implies and is implied by the statistical independence of X and Y . The only predictability possible in a bivariate normal distribution is that based on a linear rule.*

On the other hand, just because the distribution of X and the distribution of Y both happen to be normal, when considered as marginal distributions, this does *not* necessarily mean that the joint distribution of (x,y) values is bivariate normal. Hence, it is entirely possible for a nonlinear statistical relation to exist even though both X and Y are normally distributed when considered separately. It is not, however, possible for any but a linear relation to exist when X and Y jointly follow the bivariate normal law.

Most of the classical theory of inference about correlation and regression was developed in terms of the bivariate normal distribution. **If one can assume such a joint population distribution, inferences about correlation are equivalent to inferences about independence or dependence between two random variables.** For the kinds of problems here called **correlation problems**, the assumption of a bivariate normal distribution is usually made. When this assumption is valid, any inference about the value of ρ_{xy} is equivalent to an inference about the independence, or degree of dependence between two variables; this is not, however, a feature of regression problems where the bivariate normal assumption need not be made. As always, by adopting more stringent assumptions about the form of the population distribution, one is able to make much more positive statements from sample results.

The generalization of the idea of a bivariate normal distribution to more than two variables is the *multivariate normal distribution*. Here, too, the parameters of the multivariate normal distribution are the mean and variance of each variable, as well as the correlation between each pair of variables. The conditional distribution of any variable holding the others constant is normal, the conditional distribution of any pair of variables holding the others constant is bivariate normal, and so on. As in the case of the bivariate distribution, the only predictability possible in a multivariate normal distribution is linear; $\rho_{xy} = 0$ for any pair of variables X and Y implies and is implied by the independence of X and Y . Further discussion of multivariate normal distributions will be found in any text on multivariate statistics, such as that by Timm (1975).

13.13 / TESTS IN CORRELATION PROBLEMS

For many correlation problems the only hypothesis of interest is

$$H_0: \rho_{xy} = 0.$$

where the alternative hypothesis may be either directional or nondirectional.

When this hypothesis is true and the population can be assumed to be bivariate normal in form, the distribution of the sample correlation coefficient tends, rather slowly, toward a normal distribution for increasing N . The standard error of this distribution of sample correlations r_{xy} is approximately

$$\sigma_{r_{xy}} = 1/\sqrt{N} \quad [13.13.1]$$

When sample size is reasonably large (say, $N \geq 50$) then it is possible to test the significance of r_{xy} in this way, by forming the usual z -statistic and referring it to the normal distribution.

On the other hand, it can be shown that for any sample size N of 3 or more a test of this hypothesis in a bivariate normal population is given by the t ratio:

$$t = \frac{r_{xy} \sqrt{N-2}}{\sqrt{1-r_{xy}^2}} \quad [13.13.2^*]$$

with $N - 2$ degrees of freedom.

Under the assumptions made in a problem in correlation, the value of r_{xy} may be used directly as an estimator of ρ_{xy} for the population. Although it is a sufficient and consistent estimator for ρ_{xy} , the sample correlation is slightly biased; however, the amount of bias involves terms of the order of $1/N$, and for most practical purposes can be ignored.

As mentioned earlier, for very large samples the distribution of the sample correlation coefficient may be regarded as approximately normal when $\rho_{xy} = 0$. Even for relatively small samples ($N > 4$) this sampling distribution is unimodal and symmetric. However, when ρ_{xy} is other than zero, the distribution of r_{xy} tends to be very skewed. The more that ρ differs from zero, the greater is the skewness. When ρ_{xy} is greater than zero, the skewness tends to be toward the left, with intervals of high values of r_{xy} relatively more probable than similar intervals of negative values. When ρ_{xy} is negative, this situation is just reversed, and the distribution is skewed in the opposite direction. The fact that the particular form of the sampling distribution depends upon the value of ρ_{xy} makes it impossible to use the t test for other hypotheses about the value of the population correlation, or to set up confidence intervals for this value in some direct elementary way. Although the sampling distribution of r_{xy} for $\rho_{xy} \neq 0$ has been fairly extensively tabled, it is much simpler to employ the following method.

R. A. Fisher showed that tests of hypotheses about ρ_{xy} , as well as confidence intervals, can be made from moderately large samples (about $N \geq 10$) from a bivariate normal population if one uses a particular *function* of r_{xy} , rather than the sample correlation coefficient itself. The function used is known as the Fisher r to Z transformation, given by the rule

$$Z = \frac{1}{2} \log_e \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right). \quad [13.13.3]$$

Fisher showed that for virtually any value of ρ_{xy} , for samples of moderate size the sampling distribution of Z -values is *approximately normal*, with an expectation given approximately by

$$E(Z) = \zeta = \frac{1}{2} \log_e \left(\frac{1 + \rho_{xy}}{1 - \rho_{xy}} \right). \quad [13.13.4]$$

(The population value of Z , corresponding to ρ , is denoted by ζ , small Greek zeta.) The sampling variance of Z is approximately

$$\text{var}(Z) = \frac{1}{N-3}. \quad [13.13.5]$$

The goodness of these approximations increases the *smaller* the absolute value of ρ_{xy} and the *larger* the sample size. For moderately large samples the hypothesis that ρ_{xy} is equal to any value ρ_0 (not too close to 1 or -1) can be tested. This is done in terms of the test statistic

referred to a normal distribution. The value taken for $E(Z)$ or ζ depends on the value given for ρ_0 by the null hypothesis:

$$\zeta = E(Z) = \frac{1}{2} \log_e \left(\frac{1 + \rho_0}{1 - \rho_0} \right) \quad [13.13.7^*]$$

and the sample value of Z is taken from the sample correlation,

$$Z = \frac{1}{2} \log_e \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right) \quad [13.13.8^*]$$

It should be emphasized that the use of this r to Z transformation *does* require the assumption that the (x,y) events have a bivariate normal distribution in the population. On the surface, this assumption seems to be a very stringent one, which may not be reasonable in some situations, though there is some evidence that the assumption may be relatively innocuous in others. However, the consequences of this assumption's not being met seem largely to be unknown. Perhaps the safest course is to require rather larger samples in uses of this test when the assumption of a bivariate normal population is very questionable.

Table VI in Appendix D gives the Z -values corresponding to various values of r . This table is quite easy to use, and makes carrying out the test itself extremely simple. Only positive r and Z values are shown, since if r is negative, the sign of the Z -value is taken as negative also.

For example, suppose that we wanted to test the hypothesis that $\rho_{xy} = .50$ in some bivariate normal population. A sample of 100 cases drawn at random gives a correlation r_{xy} of $.35$. The hypothesis is to be tested with $\alpha = .05$, two-tailed.

Then, from the Table VI, we find that for $r_{xy} = .35$,

$$Z = .3654.$$

For $\rho_{xy} = .50$, we find

$$\zeta = E(Z) = .5493.$$

The test statistic is then

$$\frac{.3654 - .5493}{\sqrt{1/97}} = -1.81.$$

In a normal sampling distribution, a standard score of 1.96 in absolute value is required for rejecting the hypothesis at the .05 level, two-tailed. Thus, we do not reject the hypothesis that $\rho_{xy} = .50$ on the basis of this evidence. Observe that the test made in terms of Z leads to an inference in terms of ρ .

Occasionally one has two *independent* samples of N_1 and N_2 cases respectively, where each is regarded as drawn from a bivariate normal distribution, and computes correlation coefficient for each. The question to be asked is, "Do both of these correlation coefficients represent populations having the *same* true value of ρ_{xy} ?" When a test of the hypothesis that the two populations show equal correlation is provided by the ratio

$$\frac{z_1 - z_2}{\sigma_{(z_1 - z_2)}}$$

where Z_1 represents the transformed value of the correlation coefficient for the first sample, Z_2 the transformed value for the second, and

$$\sigma_{(z_1 - z_2)} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \quad [13.13.10]$$

For reasonably large samples (say, 10 in each) this ratio can be referred to the normal distribution. Remember, however, that the two samples must be independent (in particular, not involving the same or matched subjects) and the population represented by each must be bivariate normal in form.

More generally, suppose that there are J independent samples, each drawn from a bivariate normal distribution of (x,y) pairs. Each sample j yields a sample correlation r_j between X and Y . Then the hypothesis that the true value ρ_{xy} is the same for all of the populations can be tested by the statistic

$$V = \sum_j (n_j - 3)(Z_j - U)^2 \quad [13.13.11]$$

which is distributed as chi-square with $J - 1$ degrees of freedom when the null hypothesis that $\rho_1 = \rho_2 = \dots = \rho_j$ is true. Here, n_j is the number of observations in the sample j , and

$$U = \frac{\sum_j (n_j - 3)Z_j}{\sum_j (n_j - 3)} \quad [13.13.12]$$

13.14 / CONFIDENCE INTERVALS FOR ρ_{xy}

If the population has a bivariate normal distribution of (x,y) events, then the r to Z transformation can be used to find confidence intervals, very much as for a mean of a large sample. It is approximately true that for random samples of size N , an interval such as

$$Z - z_{(\alpha/2)} \sqrt{\frac{1}{N - 3}} \leq \zeta \leq Z + z_{(\alpha/2)} \sqrt{\frac{1}{N - 3}} \quad [13.14.1^*]$$

will cover the true value of ζ with probability $1 - \alpha$. Here, Z is the sample value corresponding to r_{xy} , ζ is the Z -value corresponding to ρ_{xy} , and $z_{(\alpha/2)}$ (*definitely* to be distinguished from Z) is the value cutting off the upper $\alpha/2$ proportion in a normal distribution. Thus, the expression 13.14.1 above gives the $100(1 - \alpha)$ percent confidence interval for ζ . On changing the limiting values of Z back into correlation values, we have a confidence interval for ρ_{xy} .

In the example given in the preceding section,

$$Z = .3654$$

$$N = 100,$$

so that

$$\sqrt{\frac{1}{N-3}} = .1.$$

For $\alpha = .05$, $z_{(\alpha/2)} = 1.96$, so that the 95 percent confidence interval for ζ is given approximately by

$$.3654 - (1.96)(.1) \leq \zeta \leq .3654 + (1.96)(.1)$$

or

$$.1694 \leq \zeta \leq .5614.$$

The corresponding interval for ρ_{xy} is then approximately

$$.168 \leq \rho_{xy} \leq .510$$

(the correlation values here are taken to correspond to the nearest tabled Z values). We can assert that the probability is about .95 that sample intervals such as this cover the true value of ρ_{xy} .

13.15 / ANOTHER EXAMPLE OF A CORRELATION PROBLEM

A study was made of the tendency of the height of a wife to be linearly related to that of her husband, and it was desired to find a sample correlation between husband and wife's heights, and to use this to test the hypothesis of no linear relationship.

A sample of 15 American couples was drawn at random, and the data are shown in Table 13.15.1

Couple	Heights in inches	
	X(Wife's height)	Y(Husband's height)
1	70	75
2	67	72
3	70	75
4	71	76
5	67	70
6	64	68
7	71	72
8	63	67
9	65	67
10	65	68
11	65	68
12	65	71
13	66	68
14	65	71
15	61	62

Table 13.15.1

Heights for a sample of American married couples.

Couple	Transformed scores				
	u	v	uv	u^2	v^2
1	10	5	50	100	25
2	7	2	14	49	4
3	10	5	50	100	25
4	11	6	66	121	36
5	7	0	0	49	0
6	4	-2	-8	16	4
7	11	2	22	121	4
8	3	-3	-9	9	9
9	5	-3	-15	25	9
10	4	-2	-8	16	4
11	5	-2	-10	25	4
12	5	1	5	25	1
13	6	-2	-12	36	4
14	5	1	5	25	1
15	1	-8	-8	1	64
	94	0	142	718	194

Table 13.15.2

Transformed data from Table 13.15.1.

For these data the computations for the correlation coefficient can be simplified by subtracting 60 from the height of each wife, and 70 from the height of each husband; this does not alter the value of r_{xy} obtained (Section 13.3). Then the new scores are as shown in Table 13.15.2 for u and v , respectively.

The correlation coefficient computed by the formula 13.8.2 turns out to be

$$\begin{aligned} r_{xy} &= \frac{(15)(142) - (94)(0)}{\sqrt{[(15)(718) - (94)^2][(15)(194) - (0)^2]}} \\ &= \frac{2130}{\sqrt{(1934)(2910)}} \\ &= .90. \end{aligned}$$

On the evidence of the sample, we conclude that there is a very strong linear relation between the heights of wives and husbands.

If we wished only to test the hypothesis that the true correlation is zero, we would employ the t test given by 13.13.2.

$$\begin{aligned} t &= \frac{(.90)\sqrt{15-2}}{\sqrt{1-(.90)^2}} \\ &= \frac{3.24}{.44} \\ &= 7.36. \end{aligned}$$

This greatly exceeds both the values required for $\alpha = .05$ and for $\alpha = .01$ for a t with 13 degrees of freedom (two-tailed).

Suppose, however, that the question had been, "Does the height of the wife and

the linear relation account for more than 50 percent of the variance in the observed heights of husbands?" That is, we actually want to test the hypothesis

$$H_0: \rho_{XY} \leq .50$$

against the alternative

$$H_1: \rho_{XY} > .50.$$

Given the assumption that ρ is positive, this is equivalent to the test of

$$H_0: \rho_{XY} \leq \sqrt{.50} \text{ or } H_0: \rho_{XY} \leq .707$$

against

$$H_1: \rho_{XY} > .707.$$

Here, the Fisher r to Z transformation and expressions 13.13.6 through 13.13.8 are used to find the test statistic

$$\begin{aligned} \frac{Z - E(Z)}{\sqrt{1/(N - 3)}} &= \frac{1.47 - .88}{\sqrt{1/12}} \\ &= 2.04. \end{aligned}$$

In a normal sampling distribution, this exceeds the value required for the 5 percent significance level, one-tailed. Thus we may safely conclude from this sample that more than 50 percent of the variance in Y is accounted for by the apparent linear relation with Y .

This example is made up, of course, and correlations this large are not usually found in social or behavioral work. It does illustrate one thing, however. Even though the correlation found is sizable, it makes no sense at all to think of the height of the wife as "causing" the height of the husband, or that of the husband the height of the wife. These are simply two numerical measurements that happen to occur together in a more or less linear way, according to the evidence of this sample. The reason *why* this linear relation exists is completely out of the realm of statistics, and the correlation coefficient and tests shed absolutely no light on this problem. In this example, it is perfectly obvious that personal preferences and current standards of society cause *some* selection to occur in the process of mating, and these factors in turn underlie our observations that (x, y) pairs do occur in a particular kind of relationship. As a description of a population situation, our inferences may very well be valid, but this fact alone gives us no license to talk about the cause of the apparent linear relation.

13.16 / OTHER INTERVAL ESTIMATES IN BIVARIATE NORMAL CORRELATION PROBLEMS

Under the bivariate normal correlation model, it is quite possible to form confidence intervals for β_{YX} , the true regression coefficient. The $100(1 - \alpha)$ percent confidence interval is found from

$$b_{YX} - \frac{\text{est. } \sigma_{YX}^{t(\alpha/2)}}{S_X \sqrt{N}} \leq \beta_{YX} \leq b_{YX} + \frac{\text{est. } \sigma_{YX}^{t(\alpha/2)}}{S_X \sqrt{N}} \quad [13.16.1^*]$$

where

$$\text{est. } \sigma_{YX} = \sqrt{\frac{NS_Y^2 - Nb_{YX}^2 S_X^2}{N - 2}} = \sqrt{\frac{NS_Y^2(1 - r^2)}{N - 2}} \quad [13.16.2]$$

The $t_{(\alpha/2)}$ value is found for $N - 2$ degrees of freedom.

Occasionally, interval estimates are desired for the predicted values y' using the population regression rule, and some specific x value. Remember that the predicted value y' found using a *sample* regression equation does not necessarily agree with the y' that would be found using the population regression equation. There are two possible sources of disagreement between a sample y' and the true value of y' as found from the population rule: the sample mean M_Y may be in error, and the sample estimate of β_{YX} may be wrong to some extent. Considering both of these sources of error, we have for a given score x the following confidence interval for *predicted* y' :

$$\begin{aligned} y' - t_{(\alpha/2)} \text{ est. } \sigma_{YX} \sqrt{\frac{1}{N} + \frac{(x - M_X)^2}{NS_X^2}} &\leq \text{true } y' \\ &\leq y' + t_{(\alpha/2)} \text{ est. } \sigma_{YX} \sqrt{\frac{1}{N} + \frac{(x - M_X)^2}{NS_X^2}} \end{aligned} \quad [13.16.2^*]$$

The number of degrees of freedom is again $N - 2$.

In other words, here, where $y' = M_Y + b_{YX}(x - M_X)$, and where true

$$y' = \mu_Y + \beta_{YX}(x - \mu_X),$$

there must be two kinds of variability in the distribution of Y' values over samples, given a constant value for $x - \mu_X$. The first source of variability is the difference between a given value of the mean M_Y and the true mean μ_Y . The second source of variability is the difference between the sample regression coefficient b_{YX} and the true coefficient β_{YX} . The two terms under the radical sign in [13.16.2] reflect these two sources of variability.

Be sure to notice the interesting fact that **the regression equation found for a sample is not equally good as an approximation to the population rule over all the different values of X** . The sample rule is at its best as a substitute for the population rule when $X = M_X$, the mean of the X values, since the confidence interval is smallest at this point. However, as X values grow increasingly deviant from M_X in either direction, the confidence intervals grow wider.

13.17 / PARTIAL AND PART CORRELATIONS

We now turn to a group of methods which extend the basic ideas of bivariate correlation and regression to any situation where more than two variables are involved. Certainly, in practical situations it is seldom that only two items of information about an individual will be known. Thus, in the employment setting several ratings of background, education and experience may be used, and several preemployment tests may be administered. Each of these things will be involved to some extent in the decision to hire or not to hire. The admission officer in a university may have not only college entrance examination scores, but also other information such as the high school record, the applicant's employment record, and the educational level of