Chapter Five

Correlation and regression

1. *Introduction*

Up to this point we have been concerned with single variables. The present chapter will discuss relationships between two variables, and the measurement of this relationship by means of the product-moment correlation coefficient. This coefficient varies in value between 0 and 1.0. It can be positive or negative in sign. If scores on one variable rise as scores on the other one rise then the correlation is positive, while if scores on one variable fall as the other scores rise the correlation is negative. For example height and weight are positively correlated because taller people tend to be heavier than smaller people, while mental speed and age are negatively correlated in adults, as mental speed drops with increasing age. If there is no relationship at all between two variables the correlation is zero. If the relationship is perfect, i.e. if there is complete correspondence between the two variables, the correlation will be 1.0. (Complete correspondence in this case is indicated by individuals obtaining exactly the same *Z* score on both variables).

An important point to bear in mind is that the product-moment correlation coefficient measures the strength of a linear relationship between two variables. If the relationship between two variables is not linear, then the correlation coefficient will be of little use. A linear relationship exists when the graph showing the relationships between two variables is a straight line, or near enough to a straight line, for a straight line to be a reasonable approximation to it. Figure 5.1 shows some linear and non-linear relationships.

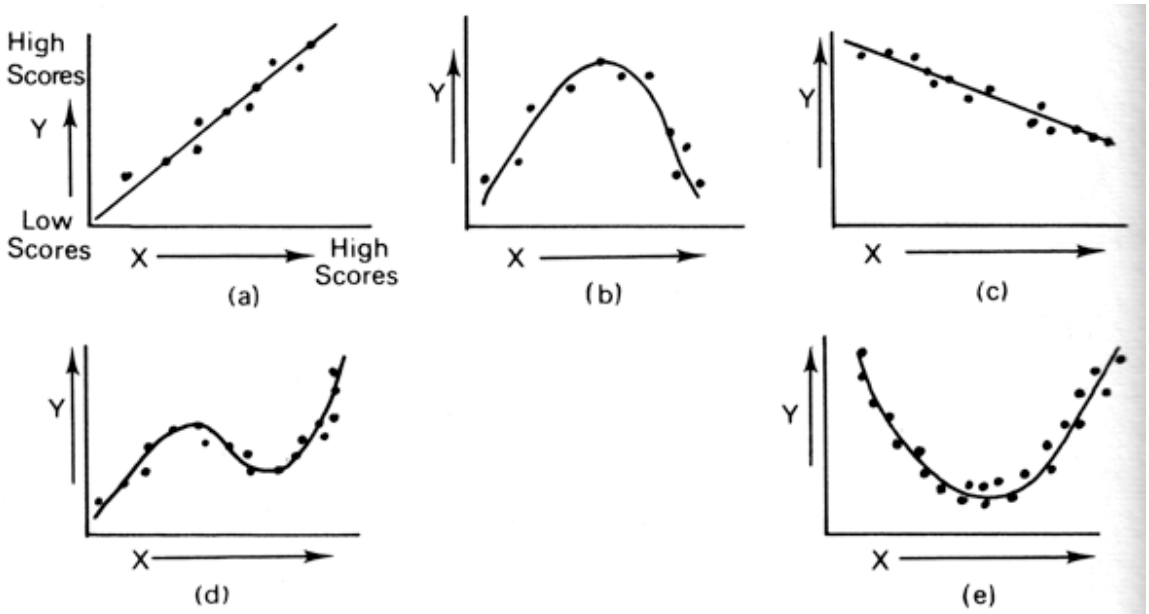

Figure 5.1 Some linear and non-linear relationships between two variables *X* and *Y*

Problems

A. Which of the relationships in Figure 5.1 can be adequately described by a product-moment correlation coefficient?

B. Which of the relationships represents a negative correlation?

C. Which relationships are curvilinear?

Answers

A. a; c.

B. c.

C. b; d; e.

2. *The Basic Formulae for the Product-Moment Correlation Coefficient.*

Suppose that we have two variables X and Y, the product moment correlation coefficient between them is symbolised r_{xy} and its formula is:

$$r_{xy} = \frac{\sum Z_X Z_Y}{N} \qquad (5:1)$$

From the formula it can be seen that the correlation coefficient is the mean of the products of the *Z* scores. To obtain the coefficient by this formula we need to take the two *Z* scores obtained by an individual, i.e. his *Z* score on *X* and his *Z* score on *Y*, and multiply them together. This is repeated for all individuals and the products so obtained are summed. This sum is then divided by *N*, and the result is the value of r_{xy}.

As an example suppose that seven individuals complete tests *X* and *Y* and obtain the following scores:

	Scores		
Individuals	*Test X*	*Test Y*	
A	1	12	
B	2	14	Mean *X* = 4
C	3	10	σ_x= 2.0
D	4	6	
E	5	8	Mean *Y* = 8.0
F	6	2	σ_y= 4.0
G	7	4	

Converting these scores into *Z* scores and finding the products of each pair of *Z* scores gives the following:

Individuals	Z_X	Z_Y	$Z_X Z_Y$
A	-1.5	+1.0	-1.5
B	-1.0	+1.5	-1.5
C	-0.5	+0.5	-0.25
D	0	-0.5	0
E	+0.5	0	0
F	+1.0	-1.5	-1.5
G	+1.5	-1.0	-1.5

$\sum Z_x Z_y = -6.25$ $N = 7$ $r_{xy} = -0.89$

This version of the formula for the product moment correlation coefficient is not the one most commonly found in basic text books, but it will be an extremely useful one for our purposes.

By simple manipulation this formula can be converted into a more common one, which will also be useful later.

$$r_{xy} = \frac{\sum xy}{N\sigma_x\sigma_y} \qquad (5:2)$$

Proof

(1) $r_{xy} = \dfrac{\sum Z_x Z_y}{N}$

(2) $Z_x = \dfrac{x}{\sigma_x}$; *and* $Z_Y = \dfrac{y}{\sigma_y}$.

(3) So $r_{xy} = \dfrac{\sum\left(\dfrac{x}{\sigma_x}.\dfrac{y}{\sigma_y}\right)}{N}$

(4) Multiplying numerator and denominator by $\sigma_x . \sigma_y$ gives

$$\frac{\sum xy}{N\sigma_x\sigma_y}$$

Neither of these formulae is very convenient for computational purposes so further operations would be necessary to derive easily computable formulae, but in the forms given they will be ideal for our purposes. At this stage it should be noted that by multiplying both sides of 5:2 by $.\sigma_x .\sigma_y$ we obtain:

$$\frac{\sum xy}{N\sigma_x\sigma_y} \qquad (5:3)$$

The term on the right is called the covariance of *X* and *Y*. Covariances and formula (5:3) will be used frequently in later sections.

3. *The Scatter Diagram*

A scatter diagram is a graphical device for showing the distribution of scores on two variables. The diagram is constructed by taking all subjects with a given score, X_1, on one variable and plotting the distribution of *Y* scores for these individuals. This will be the first column of the scatter diagram. Next the *Y* scores for all obtaining

score X_2 are plotted, forming the second column and so on. As an example suppose that the following scores are obtained on two tests.

	Tests			*Tests*	
Individuals	*X*	*Y*	*Individuals*	*X*	*Y*
A	1	1	F	3	4
B	1	2	G	3	3
C	2	3	H	4	4
D	2	2	I	4	5
E	2	3	J	5	4

These scores can be plotted on a scatter diagram as shown in Figure 5.2.

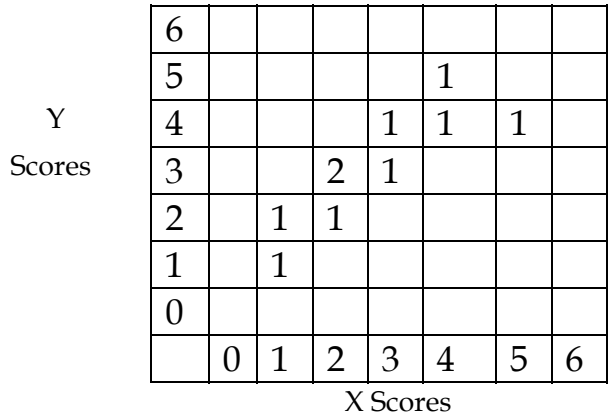

Figure 5.22 A scatter diagram of the data in table 5.3

Of the two subjects scoring 1 on test *X*, one obtained a score of 1 on test *Y*, and the other a score of 2. Of the three obtaining a score of 2 on test *X*, two scored 3 on *Y* and one scores 2, and so on. The scatter diagram can of course be read both ways. It is easy to see what *X* scores were obtained by people scoring 3 on *Y* for example.

Usually there is a much larger number of scores involved in a scatter diagram. If we draw a scatter diagram without grid lines and represent individuals by dots, we might obtain something by dots, we might obtain something like that shown in Figure 5.3.

Figure 5.3 Another scatter diagram

In this scatter diagram there is a tendency for those scoring high on X to also score high on Y. If the correlation was 1.0 there would be no scatter of scores within columns, and the points would all fall on a straight line. When the correlation is zero the scores will fall in a circular pattern, (if the variables are normally distributed). As the correlation rises from 0 to 1.0, the shape of the scatter becomes more elongated and ellipsoid until it becomes a single straight line. Figure 5.4 shows these changes visually.

Figure 5.4 Changes in the outline of the scatter plot as the correlation between two variables increases

Changes in the range of scores on one variable will (a) affect the range of scores on the other, and (b) the size of the correlation coefficient obtained. Suppose that the relationship between *X* and *Y* for the full range of scores is as shown in Figure 5.5. This would represent a reasonably high degree of correlation. If, however, we had a sample of subjects whose range of ability on *X* fell in the range *A* to *B*, this would curtail the range of *Y* scores to the range *C* to *D*. The shape of the scatter diagram obtained for this group would be as depicted in the smaller figure on the right in Figure 5.5.

Figure 5.5 The effects of restriction of range on the shape of the scatter plot

This smaller scatter diagram is more like the pattern of zero correlation, than is the scatter diagram for the full range, and indeed a restriction in range generally reduces the value of the correlation coefficient, (see Chapter 12, Section 4).

Problem

An investigator interested in the general relationship between creativity and intelligence, after finding a low relationship between these variables in a sample of university students, concludes that creativity is largely independent of intelligence. Should he have done so?

4. *Correlation and Predicition:*

(1) Guessing

The main use of correlation coefficients is prediction. The existence of a significant correlation coefficient means that X scores can be predicted from *Y* scores with better than chance accuracy. If there were no correlation between *X* and *Y* then knowing the individual's *X* score would tell us nothing about the likely *Y* score. If we know nothing about an individual's *Y* score and we have to guess it, then our best bet is that the score obtained will be the mode. If we know what outcomes are possible, and do not know which outcome will occur, our best bet is that the most common outcome will occur. This will lead to fewer mistakes in the long run than any other bet. If you know that someone has a set of cards consisting of two hearts and 11 spades in their hand, and someone chooses one at random, and you have to guess what it is, your best bet is that it is a spade. Following exactly the same principle the mode is the best bet in the case of test scores. With normal distributions the mode is the same as the mean. So the best strategy in attempting to predict Y from X or *X* from *Y* in the absence of any correlation between them is to choose the mode which in the case of test scores will usually be the mean. By choosing the mode we will be absolutely right more frequently than by choosing any other value.

However, being absolutely right is not the only criterion we might choose. In predicting test scores we might be more concerned with the average distance of our predictions from the true value. We might decide that we want our *average error* to be as near to zero as possible. Suppose in the absence of other information we guess the mean as the most likely score, i.e. for each individual and then compute the differences between the obtained score *X* and the predicted score, the mean, we will have a distribution of *(X - M)*'s summing across individuals, $\Sigma(X - M)$ is obtained, and this we have seen earlier (2:3) is equal to 0. So if we choose the mean the average error of our predictions will be zero. It is possible to show that this

will be true of no other value. Suppose that a different value *D* is chosen, where *D* is a score other than the mean.

(1) *D* = Score other than the mean.

(2) $D = M - A$ (where *A* is a positive or negative number).

(3) $X - D = X - (M - A)$.
(4) $X - (M - A) = X - M + A$.

(5) Therefore $\Sigma(X - D) = \Sigma (X - M + A)$.

(6) So $\Sigma(X - D) = \Sigma X - NM + NA$.

(7) But $M = \Sigma X/N$, so $\Sigma X = NM$.

(8) So (6) becomes $NM - NM + NA$.

(9) So $\Sigma (X - D) = NA$.

(10) And $\sum \frac{(X - D)}{N} = \frac{NA}{N} = A$.

The value of *A* differs from zero, therefore, choosing a point other than the mean leads to a greater average error than would have ensued from the choice of the mean. Thus if our interest is in obtaining the smallest average deviation between guessed and actual score we choose the mean.

To summarize this section:

(1) If it is important to be absolutely right, guess the mode.
(2) If it is desirable that average error in prediction should be zero, then guess the mean.

If we are interested in the smallest absolute average deviation, the best measure of central tendency to use, and the best guess to make, would be the median. The sum of absolute deviations, i.e. deviations disregarding their sign, is smaller when the median is used than when any other point is chosen.

Fortunately most test scores are normally distributed and thus the mean = the mode = the median. So there is no problem in deciding which to use as the best bet.

5. *Correlation and prediction: (2) Linear Regression*

Suppose that two tests have been administered to a group of subjects and the following scores obtained.

	Tests			*Tests*	
Subject	*X*	*Y*	*Subject*	*X*	*Y*
A	0	0	E	4	20
B	1	5	F	5	25
C	2	10	G	6	30
D	3	15			

If these scores are plotted on a graph with *X* as the horizontal axis and *Y* as the vertical one, it will be seen that a straight line fits the points exactly. This is shown in Figure 5.6.

Figure 5.6

We can also construct an equation $Y = bX$ for predicting Y scores from X scores. In this case $Y - 5X$, each Y value is five times the corresponding X value.

The value b tells us the slope of the line. By definition the slope of a line is given by taking two Y values, say Y_1 and Y_2 and their corresponding X values, X_1 and X_2 , and finding the value of:

$$\frac{Y_2 - Y_1}{X_2 - X_1}$$

This slope is equal to the ratio of the change in the Y variable to the change in the X variable. If Y goes down as X goes up, slope is negative, while if both rise together, slope is positive.

The next concept to be introduced is the intercept. The Y intercept is the point where the line crosses the Y axis. Consider the following sets of scores.

	Tests			*Tests*	
Subject	*X*	*Y*	*Subject*	*X*	*Y*
A	0	5	E	4	25
B	1	10	F	5	30
C	2	15	G	6	35
D	3	20			

If we plot these again the relationship is linear as in Figure 5.7.

Figure 5.7

But this time the line cuts through the *Y* axis at the value of 5. Therefore, a simple formula of the type $Y = bX$ will no longer suffice. The formula has to be modified by taking the intercept into account. The symbol for an intercept is *'a'*.

$$Y = bX + a \qquad (5.4)$$

This is the general formula for a linear relationship.

For the data presented above *b* can be found to be 5 and *'a'* can be seen to be 5, thus $Y = 5X + 5$.

In psychology, data, seldom, if ever, falls exactly on a straight line. A group of individuals obtaining a given *X* score will not all get the same *Y* score. Even if we compute the means for each group with a given *Y* score, the means are not likely to lie on a straight line. For example, let us plot the following data:

	Tests			*Tests*	
Individuals	*X*	*Y*	*Individuals*	*X*	*Y*
A	1	2	G	3	4
B	1	3	H	3	5

C	1	4	I	3	6
D	2	3	J	4	5
E	2	5	K	4	6
F	2	5	L	4	8

It can be seen in Figure 5.8 that although the points do not lie upon a straight line it is obvious that a linear prediction rule might have some value here.

Figure 5.8

The problem is how do we find a straight line to fit the data, when the points do not lie in a straight line. We need some criterion by which to choose amongst possible straight lines which might be fitted to the data. The criterion used is the least squares criterion.

The line of best fit is defined to be that line which minimizes the squared deviations between predicted and obtained scores. A line so chosen is known as a regression line.

If we decide that we want a linear equation for predicting *Z* scores on test *Y*, symbolized $\left(\hat{Z}_y\right)$, from *Z* scores on test *X*, (Z_x), then will need a formula of the following type:

$$\hat{Z}_y = bZ_x + a$$

$\hat{Z}_y$ is used rather than Z_y to indicate that it is an estimate of Z_y which ill differ from Z_y by the quantity $\hat{Z}_y - Z_y$. The least squares principle states that we must choose the values in the equation to make $\sum\left(\hat{Z}_y - Z_y\right)^2 / N$, as small as possible. It can be demonstrated that for this to be true:

$$a = 0 \qquad (5.5)$$

Proof

(1) $\hat{Z}_Y = bZ_X + a$

(2) So $Z_Y - \hat{Z}_Y = \left(Z_Y - bZ_X\right) - a$

(3) and $\left(Z_Y - \hat{Z}_Y\right)^2 = \left(Z_Y - bZ_X\right)^2 + a^2 - 2a\left(Z_Y - bZ_X\right)$

(4) so $\sum\left(Z_Y - \hat{Z}_Y\right)^2 = \sum\left(Z_Y - bZ_X\right)^2 + Na^2 - 2a\sum\left(Z_Y - bZ_X\right)$

(5) and $\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = \frac{\sum\left(Z_Y - bZ_X\right)^2}{N} + a^2 - 2a\left(\frac{\sum Z_Y}{N} - b\frac{\sum Z_X}{N}\right)$

(6) as $\frac{\sum Z_Y}{N} = \overline{Z}_Y$ *and* $\frac{\sum Z_X}{N} = \overline{Z}_X$

these both = 0, and therefore $2a\left(\frac{\sum Z_Y}{N} - b\frac{\sum Z_X}{N}\right) = 0$

(7) thus $\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = \frac{\sum\left(Z_Y - bZ_X\right)^2}{N} + a^2$

(8) as a^2 must be positive, (all squared numbers are), for $\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N}$ to be at its lowest '*a*' must equal zero

Having demonstrated that '*a*' must be zero, we can simplify the equation thus:

$$\hat{Z} = bZ_X \qquad (5:6)$$

It can also be shown that bZ_X must equal $r_{xy}Z_X$ if the least squares criterion is to be met.

Proof

(1) $Z_Y - \hat{Z}_Y = Z_Y - bZ_X$

(2) $\left(Z_Y - \hat{Z}_Y\right)^2 = Z_Y^2 + b^2 Z_X^2 - 2bZ_X Z_Y$

(3) $\sum\left(Z_Y - \hat{Z}_Y\right)^2 = \sum Z_Y^2 + b^2 Z_X^2 - 2b\sum Z_X Z_Y$

(4) $\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = \frac{\sum Z_Y^2}{N} + b^2\frac{\sum Z_X^2}{N} - 2b\frac{\sum Z_X Z_Y}{N}$

(5) $\frac{\sum Z_Y^2}{N} = \sigma_Z^2 = 1; \frac{\sum Z_X^2}{N} = \sigma_Z^2 = 1;$

and $\frac{\sum Z_X Z_Y}{N} = r_{xy}$

So we obtain

$$\frac{\sum\left(Z_Y - \hat{Z}_Y\right)^2}{N} = 1 + b^2 - 2br_{xy}$$

(6) It will now be shown that if b is any value other than r_{xy} then $1 + b^2 - 2br_{xy}$ will be larger in value than if b is equal to r_{xy}

(a) if $b = r_{xy}$ then (5) becomes $1 + r^2_{xy} - 2r^2_{xy} = 1 - r^2_{xy}$

(b) if b was other than r_{xy}, say $r_{xy} - C$ then (5) becomes $1 + (r_{xy} - C)^2 - 2(r_{xy} - C)r_{xy}$

(c) this equals: $1 + r^2_{xy} + C^2 - 2Cr_{xy} - 2r^2_{xy} + 2Cr_{xy}$ $= 1 - r^2_{xy} + C^2$. Which is 6(a) + C^2

(d) C^2 must be positive as it is a squared value, so 6(c) must be larger than 6(a). Therfore, r_{xy} is the value which gives the smallest value of $\sum\left(Z_Y - \hat{Z}_Y\right)^2$

Equation (5:6) is called the linear regression equation for predicting $\hat{Z}_Y$ from Z_X. For raw scores the linear regression equation will be:

$$\hat{Y} = r_{xy}\frac{\sigma_y}{\sigma_x}\left(X - M_x\right) + M_y \qquad (5:7)$$

Problems

A. What will be the value of $\hat{Z}_Y$ from Z_X. For raw scores the linear regression equation will be:

$$\hat{Y} = r_{xy} \frac{\sigma_y}{\sigma_x}(X - M_x) + M_y \qquad (5:7)$$

Problems

A. What will be the value of $\hat{Z}_Y$ when Z_X equals $\overline{Z}_X$?
B. What will be the value of $\hat{X}$ when Y equals M_Y?

Answers

A. $\hat{Z}_Y = r_{xy} Z_x$; therefore when $Z_X = Z_X = 0$.
$\hat{Z}_Y = 0 \times r_{xy} = 0 = \overline{Z}_Y$.

B. $\hat{X} = r_{xy} \frac{\sigma_x}{\sigma_y}(Y - M_y) + M_x$; therefore when $Y = M_y$,

$$X = r_{xy} \frac{\sigma_x}{\sigma_y}(M_y - M_y) + M_x = M_x$$

In both cases, when the predictor variable assumes its mean value, the predicted value becomes the mean of the criterion variable.

Thus a regression line passes through the point of intersection of M_x and M_y. It is also true that:-

$$\hat{M}_{y=} M_y \qquad (5:8)$$

Proof

(1) $\hat{M}_y = \frac{\sum \hat{Y}}{N}$

(2) $$\frac{\sum \hat{Y}}{N} = \frac{\sum \left(M_y + r_{xy} \left(\sigma_y / \sigma_x \right) \left[X - M_x \right] \right)}{}$$

(This is obtained by use of Formula (5:7)

(3) Therefore;

$$\frac{\sum \hat{Y}}{N} = \frac{NM_y + r_{xy} \left(\sigma_y / \sigma_x \right) \left[\sum X - NM_x \right]}{N}$$

$$= M_y = r_{xy} \frac{\sigma_y}{\sigma_x} \left[M_x - M_x \right]$$

(4) Therefore: $\hat{M}_y = M_y$

We can now prove Formula (5:7)

Proof

(1) $$\hat{Z}_Y = \frac{\hat{Y} - \hat{M}_y}{\sigma_y} = \frac{\hat{Y} - M_y}{\sigma_y}$$

(2) and $\hat{Z}_Y = r_{xy} Z_X$

(3) but $Z_X = \frac{X - M_x}{\sigma_x}$

(4) So $\hat{Z}_Y = \dfrac{\hat{Y} - M_y}{\sigma_y} = r_{xy} \dfrac{(X - M_x)}{\sigma_x}$

(5) Multiplying the last two terms of (4) by σ_y gives

$$\hat{Y} - M_y = \sigma_y r_{xy} \frac{(X - M_x)}{\sigma_x}$$

(6) A little rearrangement gives:

$$\hat{Y} - M_y = r_{xy} \frac{\sigma_y}{\sigma_x} (X - M_x)$$

(7) Adding M_y to both sides we obtain:

$$\hat{Y} = r_{xy} \frac{\sigma_y}{\sigma_x} (X - M_x) + M_y$$

This, as stated above, is the raw score linear regression equation for predicting *Y* from *X*.

Problems

Given two tests *X* and *Y* with $M_x = 50$; $\sigma_x = 10$, and $M_y = 100$, $\sigma_y = 20$, and $r_{xy} = +0.80$:

A. Find the predicted $\hat{Z}_Y$ for someone who scores 30 on test *X*.
B. Find the predicted raw score $(\hat{Y})$ for someone who scores 90 on text *X*.

Answers

A. $\hat{Z}_Y = r_{xy} Z_X$, *and* $Z_X = \dfrac{30-50}{10} = -2.0$

So $r_{xy}\, Z_X = 0.80 \times (-0.20) = -1.60$

So $\hat{Z}_Y = -1.60$

B. $\hat{Y} = r_{xy} \dfrac{\sigma_y}{\sigma_x}(X - M_x) + M_y$ so

$$\hat{Y} = 0.80 \frac{20}{10}(90-50) + 100 = 164$$

The slope of the regression line of the *Y* scores on the *X* scores, i.e. the best fit line for predicting *Y* from *X*, has been shown to be $r_{xy}(\sigma_y / \sigma_x)$. If in (5:7) we had been concerned with predicting *X* from *Y* instead of *Y* from *X* we would have found that instead of $r_{xy}(\sigma_y / \sigma_x)$ we would have obtained $r_{xy}(\sigma_x / \sigma_y)$.

This would be the slope of the regression line for predicting *X* from *Y*. In both cases the slope is the product of the correlation coefficient and the ratio of the standard deviations, and in both cases the standard deviation of the predicted variable is the numerator of the ratio. The moral of this tale is that except in the case where $r_{xy} = 1.0$ there will be two regression lines in the scatter diagram one with slope $b_{y.x}$ and one with slope $b_{x.y}$. The subscript y.x means *Y* predicted from *X*, and x.y means *X* predicted from *Y*. The slopes of the regression lines as correlation increases are shown in Figure 5.9, where it can be seen that as r_{xy} increases, the regression lines get closer together until at a correlation of 1.0 they become one line.

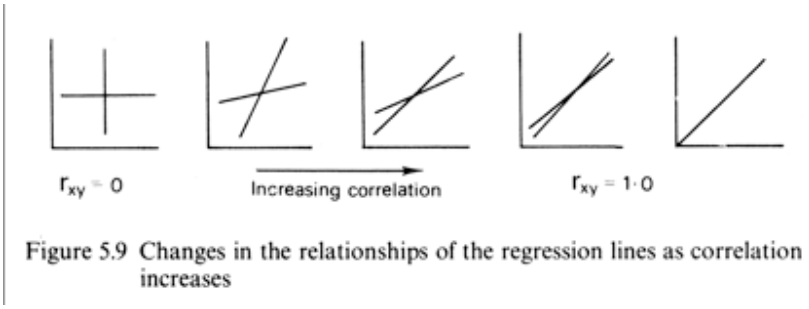

Figure 5.9 Changes in the relationships of the regression lines as correlation increases

Figure 5.9 Changes in the relationships of the regression lines as correlation increases