6 Measuring Association: The Regression Coefficient

A sample has a distribution, a central tendency, and a variability. Each of these characteristics helps the researcher make sense out of the numbers. Although each reveals some important aspect of the data, none of them measures the relationship or association between two sets of numbers.

Examples of questions of association are: If you receive an A on the midterm test, will you tend to get a high mark on the final examination? If you buy a car that is supposed to get 32 miles per gallon, what kind of mileage will your car actually get? If your roommate likes you, will you like your roommate? These are all questions of association or relationship.

To speak of relationship or association there are two separate samples of numbers that are linked together. For instance, consider the two samples of midterm grades and final grades. The two samples are linked together by persons to whom the grades refer. This linkage is illustrated in Table 6.1. For instance, John R. obtained a 36 on the midterm and a 48 on the final. The unit that links together the two samples need not necessarily be a person. For instance, to relate husband's height to wife's height, the unit that links together the scores is married couple and not person. One potential source of confusion in interpreting a measure of association is determining the object that links the pair of scores together. For instance, consider the statement that more discipline leads to more academic achievement. The relationship between these two variables can be measured for students: Do students who receive more discipline achieve more? For classrooms: In classrooms where there is more discipline, do the students achieve more? And for schools: In schools where there is more discipline, do the students achieve more? So the object can be the student, the classroom, or the school.

TABLE 6.1 **Linking Together of Two Samples**

Midterm	Object	Final
36 ⟷	John R.	⟷ 48
89 ⟷	Mary P.	⟷ 78
93 ⟷	Paul T.	⟷ 81
78 ⟷	Jane A.	⟷ 95
90 ⟷	James S.	⟷ 82
81 ⟷	Jean M.	⟷ 89

If two variables are associated, then as the numbers in one sample vary, their partner numbers in the second sample vary in some related fashion. So association implies that the numbers vary together or, as stated in data analysis, the numbers *covary*.

The simplest way in which the numbers can covary is in a linear fashion. A relationship is said to be *linear* if a difference in one variable of a fixed amount results in a constant difference in the second variable. The term linear is used because when you plot a linear association on a graph, the result is a straight line.

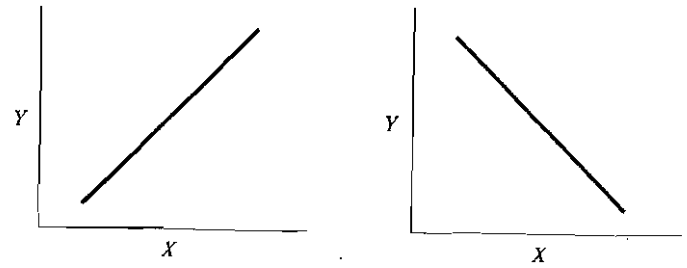
Linearity requires that a one-unit increase in variable X produces the same change in variable Y regardless of where that one-unit increase in X comes. Say, for instance, it is known that for every year of schooling a person earns on average $3000 more per year. If the relationship is linear, then an extra year of schooling provides the same amount of money ($3000) regardless of whether the extra year of schooling is one more year of college or one more year of high school.

The effect of a variable on another may be more complex than a linear one. Changes in a drug's dosage may be less potent at smaller concentrations than at larger ones. In this case the strength of a relationship between the variables increases as one variable increases. Any pattern of association between two variables that is not linear is referred to as *nonlinear* association. The important issue of nonlinear relationships is addressed in the next chapter.

There are two directions of linear association. The first type is positive association. Most often the expectation is that higher numbers in one sample are associated with higher numbers in the other sample. A student who does well on the midterm tends to do well on the final. The expectation is that the high numbers on the midterm are paired with the high numbers on the final and low numbers on the midterm go with low numbers on the final. Such a pattern of relationship is called a *positive* association. A *positive association* implies that as the numbers increase for one variable, they tend to increase for the other variable.

Sometimes as the numbers go up in one sample they go down in the other.

FIGURE 6.1 **Illustrations of positive (on the left) and negative (on the right) linear relationships.**



For instance, the more an adult weighs presumably the slower the person can run. This is an example of negative association: more weight, less speed. High numbers in one sample are associated with low numbers in the other. *Negative association* between two variables means that as the numbers increase in one sample, they decrease in the other.

The difference between a positive and a negative relationship is shown graphically in Figure 6.1. The positive relationship on the left of Figure 6.1 shows an ascending relationship, whereas on the right, the negative relationship is descending.

In Table 6.2 is a data set that will be used throughout this chapter. The data are memory scores from 16 men of various ages in the Boston area. It is then "person" that links the age and memory scores together. All men were given a test of short-term memory, which will be referred to as STM. Higher scores on STM indicate better short-term memory. The lowest possible score is zero and the maximum possible score is 24. All 16 men were being treated for alcoholism. The question considered in this chapter is the extent to which age and STM covary.

TABLE 6.2 **Ages and Short-Term Memory Scores (STM) of 16 Alcoholic Men**

Person	Age	STM	Person	Age	STM
1	48	14	9	56	2
2	46	7	10	54	12
3	44	12	11	65	12
4	52	10	12	35	18
5	22	24	13	63	5
6	43	11	14	39	18
7	51	9	15	30	14
8	54	19	16	47	8

Data were gathered by Dennis Ilchisin.

To describe the relationship between two variables the scores can be plotted on a graph. One variable is represented on the X axis and the other on the Y axis. Then each person's score is placed on the graph. A diagram in which the axes are the variables and the points are the data is called a *scatterplot*. Figure 6.2 is a plot of the score from person 1 in Table 6.2. The X axis is age and the Y axis is memory score. The person's age (48) and memory score (14) are located on the X and Y axes, respectively. Lines are drawn that are perpendicular to each axis from each score. The point at which the two lines intersect is a point in the scatterplot. When constructing a scatterplot, the perpendicular lines are not actually drawn. The dashed lines were drawn in Figure 6.2 only to show how to determine a point in a scatterplot.

Figure 6.3 illustrates the complete scatterplot for the age and memory data. The scatterplot itself tends to reveal whether there is any relationship between the two variables. Here, a negative relationship is suggested. Older persons tend to have lower memory scores.

The Regression Coefficient

Although a scatterplot describes relationship, most relationships are too weak to be clearly discerned from an examination of the scatterplot. Some method of capturing the strength of a linear relationship in a scatterplot is needed. One strategy for measuring association is to draw a straight line through the set of points in the scatterplot.

The slope of a line is the standard measure of linear association. For the slope, the variable on the X axis is called the predictor and the other is the

FIGURE 6.2 **How to determine a point in a scatterplot.**

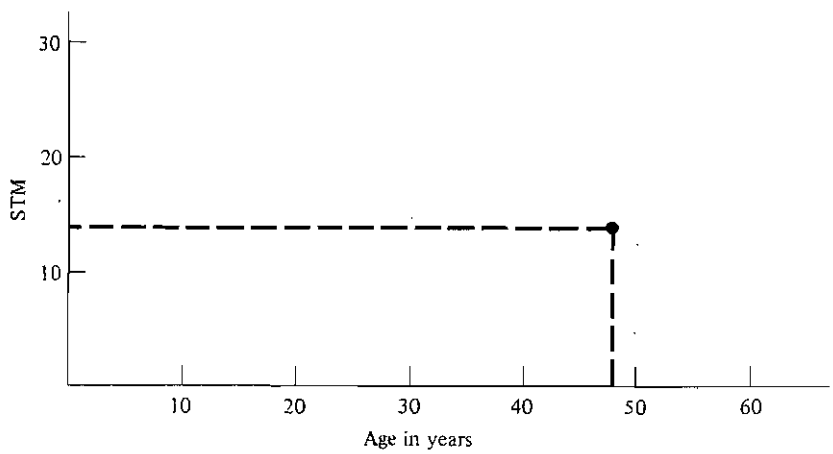
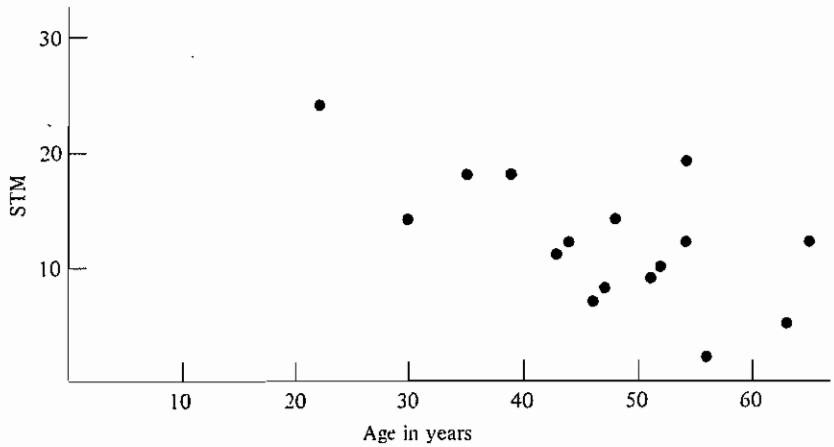


FIGURE 6.3 **Scatterplot for the age-memory study.**



criterion. The *slope* measures the effect of a change of one unit in the predictor on the criterion. Consider the following examples.

An industrial psychologist is interested in predicting who is more productive at a given factory. She devises a test that she thinks can predict productivity. The test would be the predictor and productivity would be the criterion.

A sociologist believes that the number of dollars that a community spends on schools per pupil will depend on the percentage of persons over 65 in the community. He believes that the relationship is negative: the larger the percentage of elderly, the less the amount spent on education. The predictor is the percentage over age 65, and the criterion is money spent per pupil.

A clinical psychologist believes that depression is related to diet. She believes that sugar in the diet leads to depression. The predictor is sugar consumption, and the criterion is depression.

The major measure of slope between a predictor and a criterion is a measure called the *regression coefficient*. Two different rationales for the regression coefficient are developed in this chapter. One is based on the notion that the regression coefficient is an average slope. The other is based on the notion of the regression coefficient as the best fitting line.

The Average Slope

Anyone who has decided to devote four years to obtaining a college degree must have wondered about the relationship between years of education and dollars earned. Does education predict income? So years of education is the predictor and income is the criterion. Imagine twin brothers Bob and Ray.

Ray graduated from college at State U. (16 years of education), and at age 30 he earns $35,000 a year. Bob, after finishing high school, elected not to go to college (twelve years of education), and at age 30 he earns $25,000 a year. Ray has four more years of education than Bob and he earns $10,000 more. Each year of education has brought Ray another $2,500 in income. The number 2,500 is a slope. Implicitly the following expression has been used.

$$\frac{\text{Ray's income} - \text{Bob's income}}{\text{Ray's education} - \text{Bob's education}}$$

The numerator is the difference between the brothers' incomes, which is $10,000. The denominator is the difference between the number of years of education. The *slope* states the amount of change in the criterion variable as a function of a one unit change in the predictor variable.

Suppose that there is a third brother, Mike, who went to college for only one year and so has 13 years of education. Mike's annual income is $26,000. There are now three different slopes: using Bob and Ray, Bob and Mike, and Ray and Mike. These three slopes are:

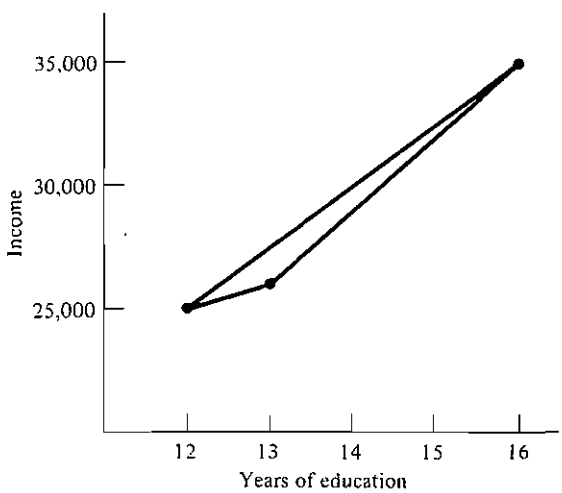
Bob and Ray: $2,500

Ray and Mike: $3,000

Bob and Mike: $1,000

These slopes measure how much money is earned for every year of education. In Figure 6.4 are the scores of Bob, Ray, and Mike in a scatterplot. The

FIGURE 6.4 **Scatterplot for the education and income example.**



horizontal or X axis is the predictor variable: number of years of education. On the vertical or Y axis is the criterion: dollars earned per year. There are three points in the graph for the three persons. These three points can be connected to form three different measures of slope.

A little notation can help. Education is denoted as X and income as Y. With three data points the scores are denoted X_1, X_2, and X_3, and Y_1, Y_2, and Y_3. The three measures of slope between the three pairs of lines are

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

$$\frac{Y_1 - Y_3}{X_1 - X_3}$$

$$\frac{Y_2 - Y_3}{X_2 - X_3}$$

These three measures of slope will not be equal unless the three points fall on a single straight line. In the example in Figure 6.4, the three pairs of points create three different measures of slope. To arrive at a single measure of slope, the three measures need to be averaged. The three measures could simply be averaged: (2500 + 3000 + 1000)/3 = 2167 for the example. Alternatively, the numerators and denominators of the three estimates of slope could be separately summed; that is,

$$\frac{(Y_1 - Y_2) + (Y_1 - Y_3) + (Y_2 - Y_3)}{(X_1 - X_2) + (X_1 - X_3) + (X_2 - X_3)}$$

However, this seemingly sensible solution results in the loss of the X_2, Y_2 data pair and yields, as an estimate of slope,

$$\frac{Y_1 - Y_3}{X_1 - X_3}$$

and so this average results in throwing away the measures of slope that involve X_2 and Y_2. To remedy this problem, the estimates of slope must be weighted in some fashion. The estimate of slope using two persons whose education differs markedly should be more reliable than using two persons whose education is quite similar. Therefore, one strategy is to weight by the difference between the scores on the predictor variable. Thus the more different two persons are on the predictor variable, the more their estimate of slope should be weighted.

Weighting by differences in the predictor variable makes intuitive sense. When persons do not differ at all on the predictor, the slope becomes impossible to measure. Weighting by differences in the predictor, the estimate of the slope becomes

$$\frac{(Y_1 - Y_2)(X_1 - X_2) + (Y_1 - Y_3)(X_1 - X_3) + (Y_2 - Y_3)(X_2 - X_3)}{(X_1 - X_2)^2 + (X_1 - X_3)^2 + (X_2 - X_3)^2}$$

or, for the example,

$$\frac{(10000)(4) + (9000)(3) + (1000)(1)}{16 + 9 + 1} = 2615$$

This is the measure of slope that is commonly used in social research and it is called the *regression coefficient*. The regression coefficient can be viewed as the average slope across all possible pairs of observations and weighted by difference on the predictor variable. Researchers never actually compute the slope for all possible pairs, but the regression coefficient does equal an average slope and can be interpreted as such.

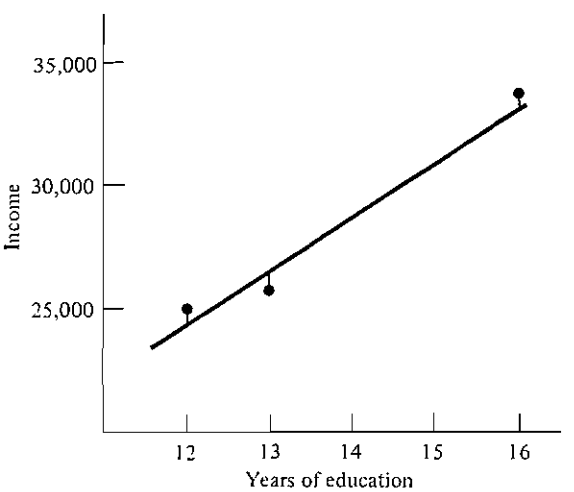
The Best-Fitting Line

There is a second and more commonly known rationale for the regression coefficient. It is the least-squares line, which is now described.

To measure the relationship between education and income, a line is fitted in the scatterplot between education and income. Because there are many such possible lines, some way is needed to determine what line is the "best" line. The best-fitting line is one that minimizes the errors.

Before proceeding any further, an error in prediction must be defined. In Figure 6.5 is the scatterplot of the three brothers' education and income. In the scatterplot there is a prediction line drawn. Also drawn are vertical lines from the line to the points in the scatterplot. The vertical length of the line can be viewed as an error in prediction. An *error* is defined as the vertical distance from the line to the score.

FIGURE 6.5 **Errors in a regression equation.**



These errors in prediction can be squared. The regression coefficient is defined as the slope of the line that minimizes the sum of squared errors. It is for this reason that the regression line is called a *least-squares* estimate; that is, the line that is chosen has the least sum of squared errors. For instance, for the education and income example, a line with a slope of 2615 has the lowest sum of squared errors. A line with any other slope has a greater sum of squared errors.

There are then two rationales for the regression coefficient. One is that it is a weighted average of all possible slope measures. The other is that the regression coefficient is the slope of the best-fitting line.

The Regression Equation

The regression line can be represented graphically as in Figure 6.5 or it can be represented by an equation. The equation is

$$Y = a + bX + e$$

The term a is the intercept, b is the slope or regression coefficient, and e is the error. The *intercept* is the predicted score for Y given that X is zero. The intercept is the point at which the regression line intersects the Y axis.

The predicted value Y given X, or $\hat{Y}$, equals

$$\hat{Y} = a + bX$$

The term $\hat{Y}$ is the predicted Y given a particular value of X. The error in prediction is then defined by $Y - \hat{Y}$. The error in prediction is the vertical distance of the regression line from the point in the scatterplot.

The regression coefficient has two important properties. First, the line always passes through the point $\bar{X}$, $\bar{Y}$. If the line did not pass through this point, it would no longer be the least-squares line. Second, the mean of the errors always equals zero; that is $\Sigma(Y - \hat{Y})/n = 0$.

The value of a regression coefficient depends on the unit of measurement. Adding or subtracting a constant to either the predictor or the criterion does not affect the regression coefficient; however, if the scores are multiplied or divided by a constant, the regression coefficient does change. If the predictor is multiplied by a constant, the regression coefficient is divided by the constant. If the criterion is multiplied by a constant, the regression coefficient is also multiplied by the constant.

Computation

The standard formula for the regression coefficient b, where X is the predictor and Y is the criterion, is

$$b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

The denominator of the slope formula is the numerator of the formula for the variance of the predictor. The numerator of the formula for slope is called the *sum of cross-products*. The denominator of the formula is called the *sum of squares* of the predictor variable. In general, the regression coefficient equals the sum of cross-products between the predictor and the criterion divided by the sum of squares of the predictor.

There is a computationally more efficient formula for the regression coefficient:

$$b = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum X^2 - (\sum X)^2/n}$$

This is the formula that is generally used to compute the regression coefficient.

The formula for the intercept, which is symbolized by a, can be obtained by solving for the predicted value of Y when X is at the mean. Because $\bar{X}$, $\bar{Y}$ is a point on the regression line, the predicted value of Y for $\bar{X}$ is $\bar{Y}$. The resulting prediction equation is $\bar{Y} = a + b\bar{X}$. Solving for a, the solution is

$$a = \bar{Y} - b\bar{X}$$

The intercept equals the mean of the criterion minus the product of the mean of the predictor and the regression coefficient.

The variance of the errors is symbolized by $s_{Y \cdot X}^2$ and is sometimes referred to as the mean square error or MSE. The formula for the variance of the errors is

$$s_{Y \cdot X}^2 = \frac{\sum(Y - \hat{Y})^2}{n - 2}$$

This formula does not look like a variance but it is. The numerator is the sum of squared errors. There is no need to subtract the mean of the errors because that mean is always zero. It is correct to divide by $n - 2$ instead of $n - 1$ because both the regression coefficient and the intercept have been estimated.

Actually the errors need not be individually computed to determine their variance. The following formula is often much simpler to compute:

$$s_{Y \cdot X}^2 = \frac{n - 1}{n - 2}(s_Y^2 - b^2 s_X^2)$$

To illustrate the use of the formulas consider the data previously presented in Table 6.2. To compute the regression coefficient, the intercept, and the variance of the errors, the following quantities are computed: ΣX, ΣY, ΣXY, ΣX^2, and ΣY^2. Because age is the predictor, it is denoted as X. And because STM is the criterion, it is denoted as Y. In Table 6.3, these computations for the age-memory study are illustrated. Laying out the numbers, as in Table 6.3, can simplify the computations. The slope for the example is

TABLE 6.3 **Computational Table for Age and STM Study**

Person	Age	STM	Age^2	STM^2	Age × STM	$\hat{STM}$	STM − $\hat{STM}$
1	48	14	2304	196	672	11.81	2.19
2	46	7	2116	49	322	12.44	−5.44
3	44	12	193	144	528	13.08	−1.08
4	52	10	2704	100	520	10.54	−.54
5	22	24	484	576	528	20.05	3.95
6	43	11	1849	121	473	13.40	−2.40
7	51	9	2601	81	459	10.86	−1.86
8	54	19	2916	361	1026	9.91	9.09
9	56	2	3136	4	112	9.28	−7.28
10	54	12	2916	144	648	9.91	2.09
11	65	12	4225	144	780	6.42	5.58
12	35	18	1225	324	630	15.93	2.07
13	63	5	3969	25	315	7.06	−2.06
14	39	18	1521	324	702	14.66	3.34
15	30	14	900	196	420	17.52	−3.52
16	47	8	2209	64	376	12.13	−4.13
Total	749	195	37011	2853	8511	195.00	0.00

$$b = \frac{8511 - (749)(195)/16}{37011 - 749^2/16} = -.3169$$

As the scatterplot shows, the slope is negative. As these men age a year, their short-term memory declines by about three-tenths of a unit, or for a decade the men lose about three points of memory score. Because the mean of the predictor is 749/16 or 46.8125 and the mean of the criterion is 195/16 or 12.1875, the intercept is

$$a = 12.1875 - (-.3169)(46.8125) = 27.0224$$

This is the predicted score for a person whose age is zero—that is, newborns. The resulting regression equation is

$$\text{STM} = 27.0224 - .3169(\text{Age}) + e$$

The variance of the errors requires the computation of the variances for age and STM. For age the variance is

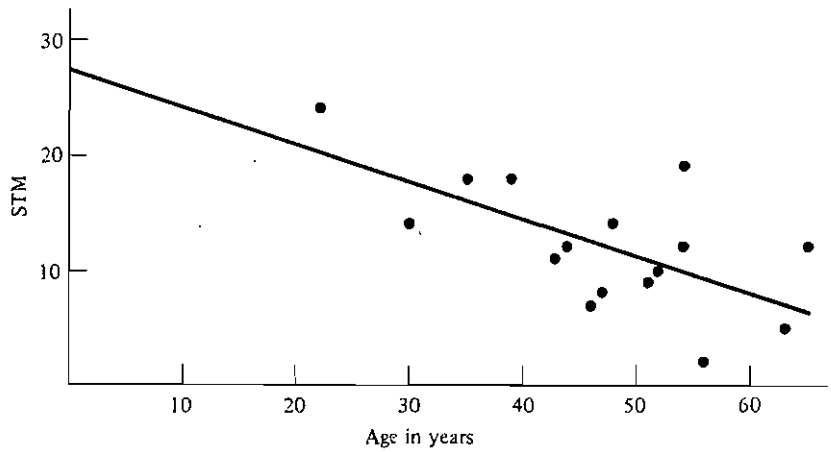
$$\frac{37011 - 749^2/16}{15} = 129.8958$$

And for STM the variance is

$$\frac{2853 - 195^2/16}{15} = 31.7625$$

The error variance is then

FIGURE 6.6 **Scatterplot and regression line for age-memory study.**



$$s_{Y \cdot X}^2 = \frac{15}{14}[31.7625 - (-.3169)^2(129.8958)] = 20.0546$$

Note that the error variation is considerably smaller than the variance of the criterion variable, which is 31.8292.

The predicted scores are also presented in Table 6.3. For instance, for person 1 whose age is 48, the predicted score is

$$27.0224 + (-.3169)(48) = 11.81$$

The error in prediction for person 1 equals the actual memory score of 14 minus the predicted score of 11.81, which is 2.19. The fact that the sum of the errors is zero is a mathematical necessity.

Finally, Figure 6.6 shows the scatterplot with the regression line plotted. To plot a line two points are needed. The two points that are used to plot the regression line are $X = 0$, $Y = 27.0224$ (the intercept) and $\bar{X} = 46.8125$, $\bar{Y} = 12.1875$ (the means). The line very clearly shows the declining memory scores with increasing age.

Interpretation of a Regression Coefficient

To compute a regression line, one variable must be treated as the predictor and the other as the criterion. The choice of which variable to designate as the predictor and which to use as a criterion should not be arbitrary. That is, one

should have either a practical or conceptual basis for making the designation. If one is not certain which variable to treat as the predictor, it may be more appropriate to use another measure of association such as the correlation coefficient, which is described in the next chapter.

If X is used to predict Y, the line obtained is different from the one obtained when Y is used to predict X. Hence there are two regression lines. To distinguish the two lines, the regression coefficient is often subscripted first by the criterion and then by the predictor. So b_{YX} implies that X is the predictor and Y is the criterion, and b_{XY} implies that Y is the predictor and X is the criterion. If Y is used to predict X, the formula for b_{XY} is as follows:

$$b_{XY} = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum Y^2 - (\sum Y)^2/n}$$

The relationship between b_{YX} and b_{XY} is straightforward. To convert b_{YX} into b_{XY}, the following formula is used.

$$b_{XY} = b_{YX}\left(\frac{s_X^2}{s_Y^2}\right)$$

Conversely,

$$b_{YX} = b_{XY}\left(\frac{s_Y^2}{s_X^2}\right)$$

There are two major purposes for the regression coefficient: prediction and explanation. In prediction, the following question is asked: If one knew someone's standing on a variable, how well would one be able to predict the person's standing on another variable? The purpose in prediction is not to change or alter reality, but merely to make good guesses about the future. The predictive use of a regression equation is valid within only the range of the predictor variable. Using a regression equation to predict scores of persons who do not score within the range of the predictor variable can be quite misleading. For the age-memory example the youngest person is 22 the oldest is 65. So any prediction for subjects younger than 22 or older than 65 involves an extending or extrapolating of the regression line beyond the sample used to estimate it. For the age-memory example, the intercept can be viewed as an extrapolation because no subjects are zero years of age. The intercept is 27.0224 and it predicts newborns would remember more than 27 items. Because 24 is the maximum possible score, it is logically impossible for anyone to score so high. This illustrates the dangers of extrapolation.

The second major use of a regression equation is for explanation. Here the researcher wants to claim that the regression equation indicates a causal effect. A causal interpretation states what would happen when reality is changed, whereas a predictive relation describes reality as it is. For instance, attitudes and behavior are generally strongly related. As an example, individual attitudes toward the use of seat belts predicts fairly well who will use

seat belts. Given this association, one can also attempt to change persons' attitudes to increase the use of seat belts. But just because attitude and behavior are correlated does not mean that attitude causes behavior. Using regression coefficients causally is even more dangerous than using them for prediction. More will be said about the causal interpretation of measures of relationship in the next chapter.

Regression Toward the Mean

The variance in the predicted scores is never larger than the variance in the criterion. This is one of the first statistical facts ever discovered. Galton in 1890 was surprised to learn that tall fathers tended to have sons shorter than themselves. If father's height is used to predict child's height, the predicted child's height is closer to the mean height than is the father's. This also worked when Galton used the son's height to predict the father's. That is, if child's height is used to predict father's height, the predicted height for the father is closer to the mean height than the son's. Galton labeled this phenomenon as *regression toward the mean.*

When the slope is zero, the predicted scores take on one value: the mean of the criterion. In this case the predicted scores have no variance. When there are no errors in prediction, the predicted score equals the criterion score and hence the two have equal variance.

Summary

Two variables are said to covary if differences in one variable are related in a systematic fashion to differences in a second variable. The most common form of a relationship is a linear one. In a *linear relationship*, the strength of the relationship does not depend on the values of the variables. Linear relationships can be positive or negative. In a *positive relationship* as one variable increases, the other variable also increases. In a *negative relationship* as one variable increases, the other decreases.

The *slope* is a measure of linear association. It measures the effect of a change in one variable as a function of a one-unit change in another variable. When measuring the slope, one variable is denoted as the predictor and the other as the criterion. The slope is given by

$$b = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum X^2 - (\sum X)^2/n}$$

where X is the predictor and Y the criterion. The *intercept* measures the predicted value for the criterion when the predictor is zero. In other words, it is the point at which the regression line intersects the Y axis and is given as follows:

$$a = \bar{Y} - b\bar{X}$$

The *variance in errors* is given by the following:

$$s_{Y \cdot X}^2 = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2)$$

The regression line can be used for either prediction or explanation. In prediction one variable is used to predict the other. In explanation, one variable is assumed to produce changes in the other.

Problems

1. There is a country saying that one can determine the temperature (Fahrenheit) by counting the number of cricket chirps in 14 seconds and adding 40. Consider cricket chirps in 14 seconds as the predictor variable and temperature as the criterion. What, according to the saying, is the slope and intercept of this regression equation?

2. An industrial psychologist uses number of cigarettes smoked per day (S) to predict the number days absent during the year (A). Her regression equation is:

$$A = 2.23 + .081S + e$$

   a. How many days absent does the equation predict for someone who does not smoke? Someone who smokes 20 cigarettes a day? Someone who smokes 40 cigarettes a day?
   b. If the company were able to lower the number of cigarettes that each of its employees smoked by 15, and the company employs 94 people, how many fewer lost days per year would it be predicted to have?

3. For the following pairs of variables, which should be treated as the predictor and which as the criterion?

   a. marital satisfaction and similarity
   b. effort and performance
   c. sleep and efficiency
   d. health and mood

4. For the following pairs of variables what is the likely direction of the relationship, positive or negative?

   a. religious belief and church attendance
   b. vocabulary and intelligence
   c. rainfall and outdoor activity
   d. criminal behavior and alcoholism

e. hours studied and midterm grade
f. odometer mileage and repair costs
g. grade point average and number of parties attended
h. price of beer and enjoyable taste

5. The following scores are the height and weight of five persons.

Height (inches)	*Weight (pounds)*
60	140
64	170
72	210
68	180
70	150

Treat height as the predictor variable and weight as the criterion.

a. What is the slope, intercept, and variance of the errors? Interpret each statistic.
b. Compute the errors for the five scores and verify that their mean is zero.
c. Draw a scatterplot and plot the regression line.

6. For the following pairs of variables what would be the object across which the measure of association would be computed?

a. literacy rate and gross national product
b. population and crime rate
c. sense of control and happiness
d. leadership and productivity

7. Describe each of the following relationships as either predictive or causal.

a. predictor: cigarette smoking; criterion: lung cancer
b. predictor: beer consumption; criterion: wine consumption
c. predictor: child's height; criterion: child's reading skill
d. predictor: physical attractiveness; criterion: popularity
e. predictor: presence of smoke; criterion: fire

8. Harrison (1984) conducted a questionnaire study of crowding, privacy, and loneliness among female dormitory residents. Satisfaction with privacy in the dormitory (X) was measured on a scale from one to seven, with higher scores indicating greater satisfaction. Subjects were also asked how often they avoided people other than friends in the dormitory (Y). Again, responses were given on a seven-point scale, with higher scores indicating more frequent avoidance behavior. The data of 20 of the subjects, randomly chosen, are given below, along with some calculations.

Subject	*Satisfaction (X)*	*Avoidance (Y)*	X^2	Y^2	XY
1	4	2	16	4	8
2	6	1	36	1	6
3	6	1	36	1	6
4	5	2	25	4	10
5	2	6	4	36	12
6	6	2	36	4	12
7	6	1	36	1	6
8	5	2	25	4	10
9	6	5	36	25	30
10	4	1	16	1	4
11	6	1	36	1	6
12	4	4	16	16	16
13	4	1	16	1	4
14	5	4	25	16	20
15	7	2	49	4	14
16	3	2	9	4	6
17	4	3	16	9	12
18	2	7	4	49	14
18	6	1	36	1	6
20	6	1	36	1	6
Total	97	49	509	183	208

a. Construct a scatterplot. Does the plot tend to show a positive or negative relationship?
b. Compute the slope of the regression equation, with satisfaction as the predictor and avoidance as the criterion. Interpret the slope. Compute the intercept. What does it mean?
c. Compute the variance of X and of Y. Compute the variance of the errors.

9. From the results of problem 8, state the regression equation. Compute the predicted avoidance scores for each of the observed satisfaction scores. Compute the variance of the predicted scores. How does it compare with the variance of the observed avoidance scores? What is this change in variance called?

10. For the regression equation

$$Y = 10.3 + .6X + e$$

find predicted scores for the following values of X: 10, 12, 15, and 31.

11. If $s_Y^2 = 15$, $s_X^2 = 10$, $n = 25$, and $b_{YX} = .5$ find the following:

a. b_{XY} b. $s_{Y \cdot X}^2$ c. $s_{X \cdot Y}^2$

12. Below is the temperature in Hartford, Connecticut, and the expected number of cars that will have starting difficulties.

Temperature (Fahrenheit)	*Number of Disabled Cars*
40	1159
32	1288
25	1519
20	2276
15	2941
10	3296
5	4481
0	5665
–5	7210
–10	8858

Treat temperature as the predictor and number of disabled cars as the criterion and estimate the slope and intercept. Interpret each. Compute the errors for each observation. Using the equation, how many cars will have starting difficulties when the temperature is –8 degrees? When it is 70 degrees?