From: Kachigan, S. K. (1991). *Multivariate Statistical Analysis: A Conceptual Introduction* (2nd ed.). New York: Radius Press.

Chapter 4

Regression Analysis

1. Introduction

Whereas correlation analysis provides us with a summary coefficient of the *extent* of relationship between two variables, *regression analysis* provides us with an equation describing *the nature of the relationship* between two variables. In addition, regression analysis supplies variance measures which allow us to assess the accuracy with which the regression equation can predict values on the criterion variable, making it more than just a curve-fitting technique.

While the basic model underlying regression analysis is designed for experimental data in which the levels of the predictor variable are selected or fixed by the investigator, with objects then assigned at random to these levels, the technique can be, and usually is, used to describe the relationship between correlated random variables, where the investigator has no control over the values assumed by the objects on the predictor variable; e.g., the relationship between student grade averages and intelligence test scores, or between the crime rates and unemployment rates of cities, or between crop yields and rainfall levels. There are no severe consequences to this type of application of the basic regression technique, provided the predictor variables are measured with high accuracy.

Examples of the types of *experimental* relationships that can be studied with regression analysis are many. For example, we might want to determine the nature of the relationship between crop yield and levels of fertilizer application, between student test performance and hours of instruction, between disease duration and drug dosage, between blood pressure and controlled dietary salt levels, between product sales and systematically varied advertising levels, between maladaptive behavior frequencies and hours of therapy, between bacteria growth and culture medium concentrations, etc.

While regression analysis can be used with both correlational and experimental data, we will concentrate in this chapter on its application to the former

type, in order to capitalize on the discussions of the preceding chapter. The treatment of experimental data will be addressed in the following chapter on Analysis of Variance, an analytical approach related to regression analysis in its purpose.

2. Overview of Regression Analysis

The recurrent theme of *prediction* appeared throughout our discussion of correlation analysis. While it seemed intuitively clear that the greater the degree of correlation between two variables, the more likely we would be to accurately predict values on one from a knowledge of values on the other, we never learned a specific procedure for accomplishing the prediction. In regression analysis we have such a technique.

The concept of regression analysis—which could well be called prediction analysis—will be easy to understand since much of the spade work has already been done in our study of correlation analysis. Not only will correlation analysis help us in our understanding of regression analysis, but regression analysis will deepen our understanding of correlation analysis.

Just as there is the simple correlation coefficient to measure the degree of relationship between two variables, and the multiple correlation coefficient to measure the degree of relationship between a set of predictor variables and a criterion variable, there is both *simple* and *multiple* regression analysis. In simple regression we are interested in predicting an object's value on a criterion variable, given its value on *one* predictor variable. In the case of multiple regression we are interested in predicting an object's value on a criterion variable when given its value on each of *several* predictor variables. We will begin our study with simple regression, and then discover how we can generalize the concept to the multiple regression situation.

Objectives. The overall objectives of regression analysis can be summarized as follows: (1) to determine whether or not a relationship exists between two variables, (2) to describe the nature of the relationship, should one exist, in the form of a mathematical equation, (3) to assess the degree of accuracy of description or prediction achieved by the regression equation, and (4) in the case of multiple regression, to assess the relative importance of the various predictor variables in their contribution to variation in the criterion variable.

We will touch upon each of these issues, though not in the stated order, since for expositional reasons it is more logical to start with the second named objective—identifying the mathematical equation which describes the relationship between the variables in question.

What might a regression equation look like? Let us begin by looking at the values of several objects on two variables to see if we can gain some insight

into the nature of the desired equation:

	Value on variable x	*Value on variable y*
Object 1	8	31
Object 2	5	22
Object 3	11	40
Object 4	4	19
Object 5	14	49

The objects and variables x and y in this example are left unidentified but could well correspond to any of those named in the introductory section. Before reading on, it will be worthwhile to study the above pairs of values to see if you can identify a systematic relationship between an object's value on variable y and its value on variable x.

It could be noticed, for instance, that the value on variable y *is greater than* the corresponding value on variable x. But we can be more precise. We can say that the value on variable y is *more than triple* the value on variable x. A closer examination will show that the value on variable y is exactly *three times* the value on variable x *plus seven*. In short-hand notation we can specify the relationship as $y=3x+7$, which is nothing more than to say that an object's value on variable y is equal to three times its value on variable x plus seven. By convention, though, we often write the equation with the constant term first, $y=7+3x$.

Linear equations. For those who have not already recognized it, based on a study of elementary algebra, $y=7+3x$ is the equation of a specific straight line. Figure 1 presents a refresher course on the characteristics of a straight line, or *linear function*. The equation $y=7+3x$ is but a specific instance of the general equation of a line

$$y=a+bx$$

The value of b is referred to as the *slope* of the line; i.e., its inclination, or the rate of change in the value of the variable y for a unit change in value of the variable x. The higher the value of b, the steeper the slope.

The value of a, or the constant term, represents the value of y when $x=0$, or in other words the value of y where the line intercepts the y axis. It is, in fact, referred to as the *y-intercept*. In the equation $y=7+3x$, the slope is 3 and the y-intercept is 7. Notice, however, in Figure 1*b* that the physical appearance of the slope of the line is very arbitrary, depending to a great extent on how stretched out or compressed we make the scale on the y or x axes.

Returning now to the five pairs of values that were found to be related

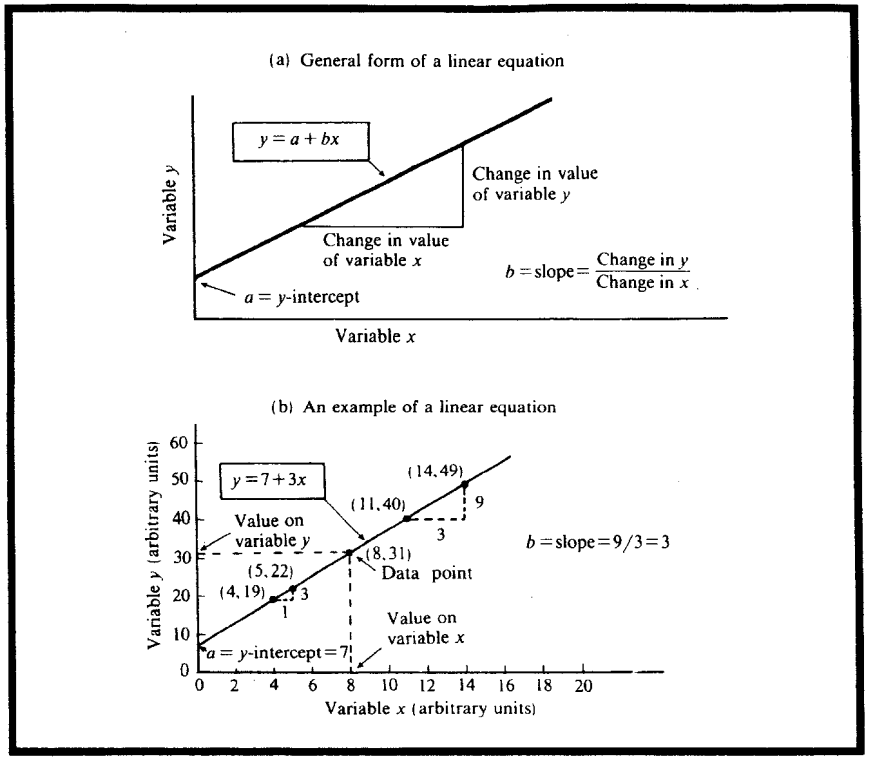

Figure 1 Characteristics of a linear equation.

exactly by the equation $y=7+3x$, note in Figure 1*b* how they lie exactly on the line representing the equation. In real life, however, rarely do we find data that are so perfectly related. More often, we will only be able to say that the value on variable y is *approximately equal* to $7+3x$, as illustrated by the pairs of values shown below:

	Value on variable x	*Value on variable y*
Object 6	7	26
Object 7	3	18
Object 8	6	30
Object 9	8	28
Object 10	7	30

As an exercise, plot these five data points onto Figure 1*b* to see how they deviate from the given line; and yet see how the line does represent a fair

description of the relationship between the pairs of values.

The importance of the linear equation is that we will be limiting our discussion to data that are linearly related, excluding such non-linear relationships as shown earlier in Figure 5*b* of Chapter 3. But this is not too severe a limitation since many relationships that we encounter are linear in form, and of those that are not, many can be made linear with appropriate data transformations. For example, the values of a variable y may not be linearly related to the values of variable x, but they may turn out to be linearly related to, say, the logarithm of x, or maybe the square root of x, or perhaps the reciprocal of x, or any of a number of other possible transformations.

3. The Regression Line

Given a scatter diagram as shown in Figure 2, how do we go about choosing the "best" of all the possible lines that could pass through the cluster of data points? In other words, compared to the line that is shown, why could we not choose one that was a little steeper or perhaps one not so steep. Or what about one that was positioned a little higher in the cluster of points, or a little lower? There are an infinite number of possible lines, $y = a + bx$, differing in slope b and/or y-intercept a, that could be drawn through the points in Figure 2.

We could say that the line shown passes right through the middle of the cluster of points. But what is the middle? Do we have an objective criterion for determining the middle? Perhaps we should choose the line that passes through the point $(\bar{x}, \bar{y})$ that represents the means of the two variables. This is a good start, but there are still an infinite number of lines that can pass through that point, each differing in slope. What would be the best inclination or slope of the line? The fact is, that on the basis of inspection alone, no two individuals are likely to agree on exactly the same line. For this reason we need a well-defined procedure for choosing a "best fitting" regression line

$$\boxed{y' = a + bx} \tag{1}$$

where y' (read "y prime") represents the predicted value of the criterion variable for a given value of the predictor variable x, and a and b are the y-intercept and slope, respectively, of the line and must be determined from the data. The prime sign (′) is used to distinguish a predicted value y' from an observed value y.

Least squares criterion. While there are a number of plausible criteria for choosing a best-fitting line, one of the most useful is the *least squares criterion*. In Figure 2 the data points deviate from the given line by varying amounts, the

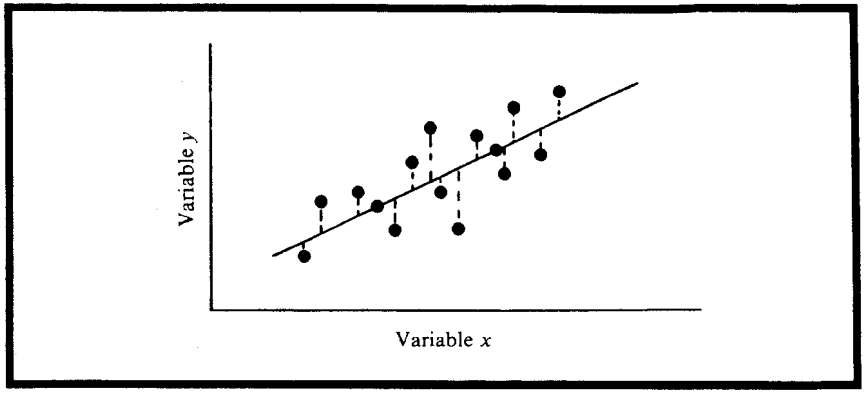

Figure 2 Deviations of data points from a linear function.

extent of deviation indicated by the dashed lines. With the least squares criterion we choose that particular line, among all possible, that results in the *smallest sum of squared deviations* of the data points from the line. Since the general form of the regression line is $y' = a + bx$, our task is to identify the values of a and b which will minimize $\Sigma(y_i - y_i')^2$, where y_i is an observed value and y_i' is the predicted value.

Slope and y-intercept. It would seem to be a very laborious procedure if we were to determine such a best-fitting line by trial and error alone—positioning and repositioning the line, each time tabulating the squared deviations of the sample data points from the line, until we discovered that line which resulted in the smallest sum of the squared deviations. But as we have seen so often before, the problem of identifying this best-fitting line has a purely mathematical solution. The slope b of the best-fitting line, based on the least squares criterion, can be shown to be

$$\boxed{b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}} \tag{2}$$

where the summation is over all n pairs of (x_i, y_i) values.

The value of a, the y-intercept, can in turn be shown to be a function of b, $\bar{x}$, and $\bar{y}$; i.e.,

$$\boxed{a = \bar{y} - b\bar{x}} \tag{3}$$

The derivations of (2) and (3) require calculus techniques and can be found in any advanced mathematical statistics text.

Table 1 Sample calculations for obtaining the slope b and intercept a of the best-fitting regression line using the definitional formulas.

x_i Shelf space	y_i Spice sales	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
340	71	40	1	40	1600
230	65	−70	−5	350	4900
405	83	105	13	1365	11025
325	74	25	4	100	625
280	67	−20	−3	60	400
195	56	−105	−14	1470	11025
265	57	−35	−13	455	1225
300	78	0	8	0	0
350	84	50	14	700	2500
310	65	10	−5	−50	100
Sums: 3000	700	0	0	4490	33400
$\bar{x} = 300.0$	$\bar{y} = 70.0$				

$$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{4490}{33400} = .1344$$

$$a = \bar{y} - b\bar{x} = 70.0 - (.1344)(300) = 29.68$$

$$y' = a + bx$$

$$y' = 29.68 + .1344x$$

To get a better idea of the numerical calculations involved in determining the values of b and a, we will reintroduce the spice sales vs. shelf space example of the preceding chapter. Table 1 shows the paired values of spice sales and shelf space occupied by the spice line in ten randomly selected stores, as well as the calculations for determining a and b. First we obtain the deviation of each shelf space measure from the mean of that variable; $x_i - \bar{x}$. Then we obtain the deviation of each spice sales measure from the mean of that variable; $y_i - \bar{y}$. Next we take the product of these pairs of deviations; $(x_i - \bar{x})(y_i - \bar{y})$. Finally, we square the deviations of the x_i measures from their mean; $(x_i - \bar{x})^2$. Now, summing the latter two sets of calculations, found in the last two columns of Table 1, across all ten stores, we have

$$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{4{,}490}{33{,}400} = .1344$$

Using this value of the slope b, along with $\bar{x}$ and $\bar{y}$, we can easily determine a, the y-intercept, as

$$a = \bar{y} - b\bar{x} = 70.0 - (.1344)(300.0) = 29.68$$

These calculations, then, demonstrate the applications of the definitional formulas (2) and (3) for the slope and y-intercept, respectively.

Substituting these values of a and b into the general form of the regression line, $y' = a + bx$, we have

$$y' = 29.68 + .1344x$$

as the best-fitting line through the set of ten data points according to the least squares criterion.

The line is shown graphically in Figure 3*a*, and it does appear to fit the data nicely. The reason the line does not intercept the y-axis at $a = 29.68$, as expected, is due only to the fact that we have curtailed the x and y axes.

If we now wanted to predict the sales of the spice line in a store in which it occupied, say, $x = 250$ in^2 of shelf space, we would simply plug that value into the regression equation $y' = 29.68 + .1344x$ and get

$$y' = 29.68 + .1344(250) = 63.28 \text{ dollars}$$

as our prediction of that store's spice sales. We must be cautioned, though, against applying the equation for values of x which are beyond those used to develop the equation, for the relationship may not be linear for those values of x.

Alternative formulas for *b*. It is interesting to note that the slope b can also be expressed

$$\boxed{b = r\left(\frac{s_y}{s_x}\right)} \tag{4}$$

where r is the correlation coefficient between the x and y variables, while s_x and s_y are their respective standard deviations.

Yet another equivalent formula for the slope b is given by

$$b = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \tag{5}$$

where the summation is across the $i = 1, 2, \ldots, n$ pairs of observations. The

Table 2 Sample calculations for obtaining the slope b of the best-fitting regression line using an alternative computational formula.

x_i Shelf space	y_i Spice sales	x_i^2	$x_i y_i$
340	71	115,600	24,140
230	65	52,900	14,950
405	83	164,025	33,615
325	74	105,625	24,050
280	67	78,400	18,760
195	56	38,025	10,920
265	57	70,225	15,105
300	78	90,000	23,400
350	84	122,500	29,400
310	65	96,100	20,150
Sums: 3,000	700	933,400	214,490
$\bar{x}=300$	$\bar{y}=70$		

$$b=\frac{n\Sigma x_i y_i-\Sigma x_i \Sigma y_i}{n\Sigma x_i^2-(\Sigma x_i)^2}=\frac{10(214{,}490)-(3{,}000)(700)}{10(933{,}400)-(3{,}000)^2}=.1344$$

$$a=\bar{y}-b\bar{x}=70.0-(.1344)(300)=29.68$$

$$y'=a+bx$$

$$y'=29.68+.1344x$$

application of this formula to the spice sales example is shown in Table 2, and again we find that $b=.1344$, although by a different route. Computationally, this is sometimes an easier method for calculating b.

Standardized regression equation. It is also interesting to note that the raw score regression equation $y'=a+bx$ simplifies to the standardized form

$$\boxed{z'_y=rz_x} \tag{6}$$

when the x and y variables are expressed as standardized z scores; i.e., with means of zero and standard deviations of one. The value of r is again the correlation coefficient between x and y, and corresponds to the slope of the line relating their standardized scores. There is no intercept term since the equation passes through the origin (0,0) corresponding to the means of the respective z variables.

The standard score form of the regression equation for the spice sales

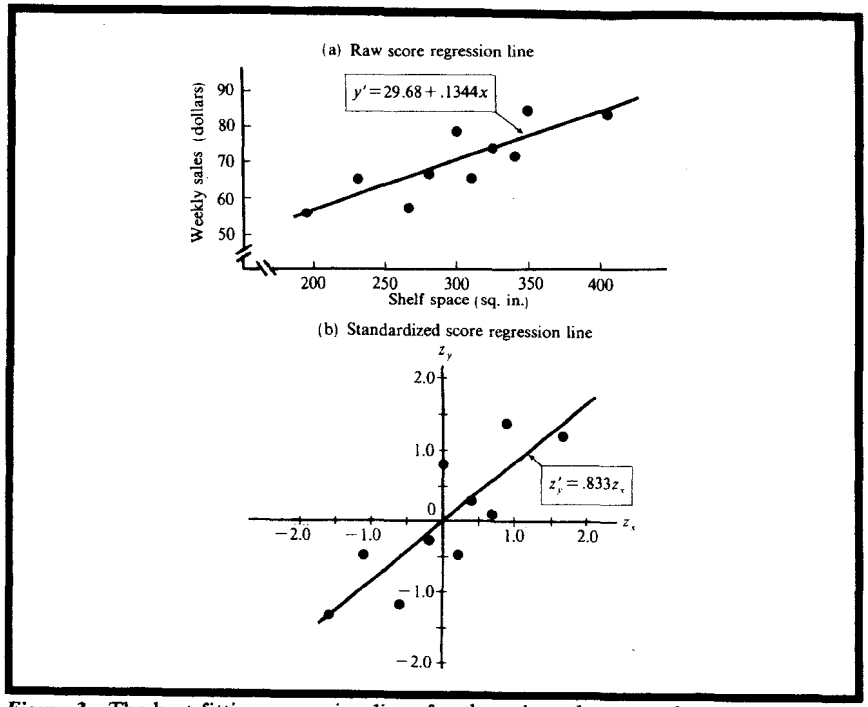

Figure 3 The best-fitting regression lines for the spice sales example.

example is portrayed in Figure 3*b*. The data points and the slope $r=.833$ were taken from Table 5 of Chapter 3. The difference in appearance of the slope and dispersion of data points in parts *a* and *b* of Figure 3 is due to the difference in scales for the two graphs; i.e., raw vs. standardized scores.

4. The Regression Model

In the preceding example we mechanically applied the formulas for the slope b and y-intercept a to obtain an equation for the best-fitting line through the given set of data points. However, for the resulting regression equation to be properly interpreted, a number of assumptions must be met concerning the populations of data we are studying.

The essence of the linear regression model is shown graphically in Figure 4. Specifically, the assumptions of the model are as follows:

1. For each value of the predictor variable x, there is a probability distribution of independent values of the criterion variable y. From each of these y distributions, one or more values is sampled at random.

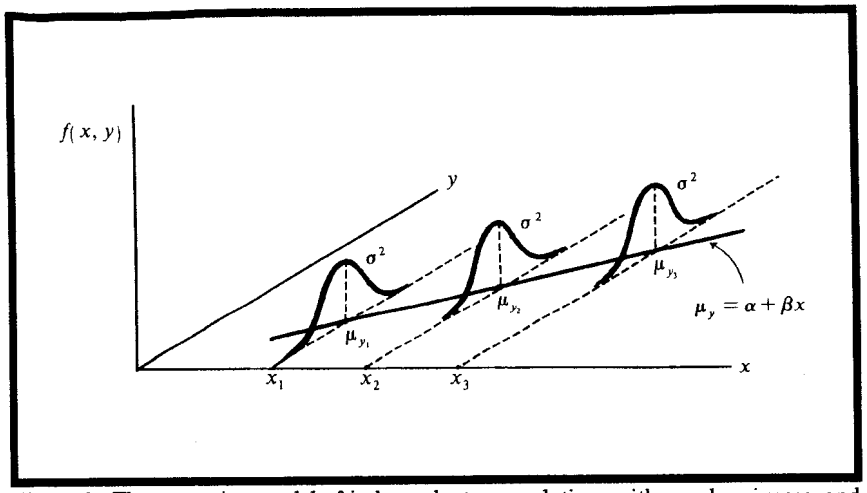

Figure 4 The regression model of independent y populations with equal variances, and with means falling on a straight line.

2. The variances of the y distributions are all equal to one another, a condition referred to as *homoscedasticity*.
3. The means of the y distributions fall on the regression line $\mu_y = \alpha + \beta x$; where μ_y is the mean of a y distribution for a given value of the predictor variable x, β (beta) is the slope of the line, and α (alpha) is the y-axis intercept of the line.

What we see, then, is that for any given value of the predictor variable x, the values of the criterion variable y vary randomly about the regression line. Consequently, any individual observation of the criterion variable, y_i, will deviate from the population regression line by a certain amount, call it e_i, where the value of e_i can be either positive or negative, depending upon whether the observation falls above or below the true regression line. Since these e_i's represent deviations from the mean of a y distribution, their average value will be zero.

Based on the above assumptions, we can characterize an individual observation of the criterion variable, y_i, as being equal to

$$y_i = \alpha + \beta x + e_i$$

That is, the observed value, y_i, is the sum of a fixed part dictated by the true regression line, $\alpha + \beta x$, plus a random part, e_i, due to the natural variation of the y values about the regression line. It follows, then, that for any given value of the predictor variable x, the variation of the y_i values is identical to the

variation of the e_i's, and it is assumed that this variation is the same regardless of the value of x.

The importance of the e_i's lies in the fact that they represent the primary source of error in trying to predict values of the criterion variable y. In the following section we will learn how to estimate the variance of these deviations from the true regression line.

5. Accuracy of Prediction

If the assumptions of the above regression model are met, we can be assured that the least squares method will yield a sample regression line, $y' = a + bx$, which is an unbiased estimate of the true, but unknown, population regression line, $\mu_y = \alpha + \beta x$. However, the a and b estimates of α and β are subject to sampling error just like any other sample statistics, so they will be sources of error in trying to predict the criterion variable value from a given value of the predictor variable. But, by far the one greatest source of error in attempting to predict individual values of the criterion variable y does not lie in the errors of estimating the slope and y-axis intercept of the regression line, but in the random variation of the y_i's about the regression line—the e_i's of the preceding section.

Standard error of estimate. The variation of the y_i values about the population regression line can be estimated by assessing their variation about the sample regression line. The standard deviation of the observed y_i values about the predicted values y_i' is referred to as the *standard error of estimate*, designated $s_{y \cdot x}$, and is given by the formula

$$s_{y \cdot x} = \sqrt{\frac{\Sigma (y_i - y_i')^2}{n-2}} \tag{7}$$

where the summation is across the $i = 1, 2, \ldots, n$ sample observations. The reason we divide by $n-2$, instead of $n-1$ as was customary with other sample standard deviations, is due to the fact that there are two constraints on the data—the slope and y-axis intercept which were used to obtain the predicted values y_i'.

Although $s_{y \cdot x}$ is referred to as the standard error of estimate, it is not a standard error in our conventional use of the term as a measure of the standard deviation of the sampling distribution of a statistic. Rather, it is an estimate (when squared) of the variance of the y populations about the true regression line, as shown in Figure 4. It might, more appropriately, be called the "standard deviation about regression." In this terminology the subscript notation of $s_{y \cdot x}$ is also more meaningful, signifying the standard deviation of y

Table 3 Calculations for obtaining the standard error of estimate.

x_i Shelf space	y_i Spice sales	$y_i' = 29.68 + .1344x_i$	$y_i - y_i'$	$(y_i - y_i')^2$
340	71	75.38	−4.38	19.184
230	65	60.59	4.41	19.448
405	83	84.11	−1.11	1.232
325	74	73.36	.64	.410
280	67	67.31	−.31	.096
195	56	55.89	.11	.012
265	57	65.30	−8.30	68.890
300	78	70.00	8.00	64.000
350	84	76.72	7.28	52.998
310	65	71.34	−6.34	40.196
$\bar{x} = 300$	$\bar{y} = 70$	$\bar{y}' = 70$	0.00	266.466
$s_x = 60.92$	$s_y = 9.83$	$s_{y'} = 8.19$		

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(y_i - y')^2}{n-2}} = \sqrt{\frac{266.466}{8}} = 5.77$$

for a given x.

Table 3 shows the calculations involved in determining the value of $s_{y \cdot x}$ for our spice sales vs. shelf space example. First we substitute each value of x into the sample regression equation $y' = 29.68 + .1344x$ to arrive at the y' estimates. The differences between the observed and predicted values, $y_i - y_i'$, are then squared and summed across the $n = 10$ stores, and finally divided by $n - 2$ before taking the square root; specifically,

$$s_{y \cdot x} = \sqrt{\frac{266.456}{8}} = 5.77$$

This, then, is our *estimate* of the variation of the y populations about the true regression line.

Confidence bands. At first glance we might think that $s_{y \cdot x}$ could be used to create a confidence band about the sample regression line, reflecting the maximum expected error in predicting y from x, with a given probability, say .95 or .99, similar to other confidence intervals based on sample statistics. But this is not so, since the errors in predicting y are not only due to $s_{y \cdot x}$, which estimates the random variation of y about the true regression line, but there are also two other sources of error: (1) the error in estimating the overall elevation or y-axis intercept of the true regression line, α, and (2) the error in estimating the slope β of the true regression line.

Furthermore, the error due the second of the above two factors—the error in estimating the slope—becomes more pronounced the more the predictor value x_i deviates from the average x value under study. The consequence of this ever-increasing error the further we move from the average value of the predictor value, is a *bowed* confidence band about the sample regression line. This can be seen most easily from a study of Figure 5. In part *a* of the figure, a sample regression line is superimposed upon the true but unknown regression line. Imagine, instead, that an infinite number of such sample regression lines were in the figure. Each would vary in slope due to sampling error, but the net effect would be the same: The further we move from the mean of the x variable, the larger the discrepancy between the sample and true regression line.

Figure 5*b* shows the resulting confidence band with its bowed feature. Its width, measured vertically at any given value of the predictor variable x, is a function of the three sources of error outlined above: the natural variation of y about the true regression line, $s_{y \cdot x}$; the error in estimating the y-axis intercept, α; and the error in estimating the slope β of the line. We have studied the first component, $s_{y \cdot x}$, in some detail, but to adequately probe the formulas and interpretations of the other two error sources would require extensive discussion, and is best left for advanced study. However, for reference purposes, the relevant expression appears in Figure 5*b*.

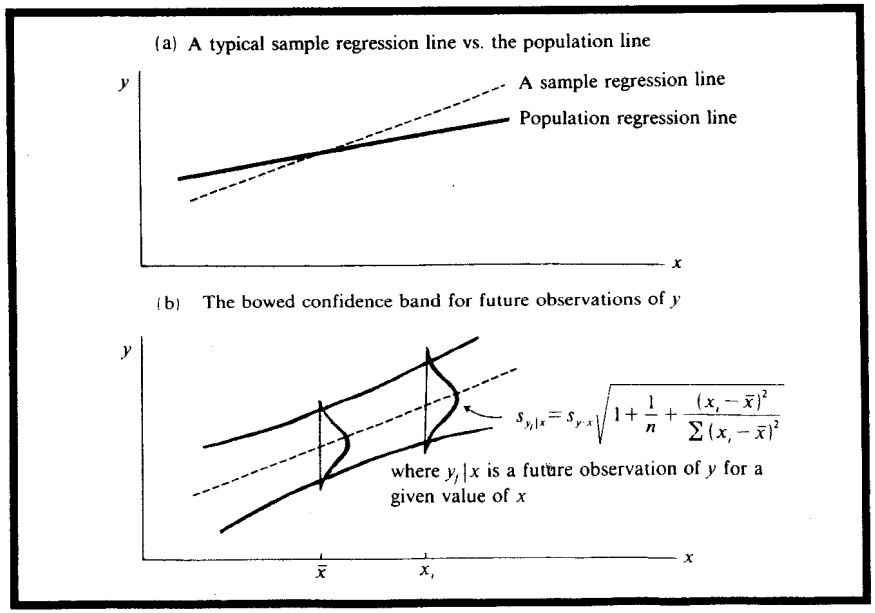

Figure 5 Errors of prediction resulting from a sample regression equation (See text).

Although $s_{y \cdot x}$ alone is not sufficient to precisely estimate the expected magnitudes of our prediction errors, it is fortunate that as n becomes large, say greater than 100, it can provide approximate confidence intervals, since the errors in the estimation of α and β become small relative to $s_{y \cdot x}$. This will be apparent from a study of the formula in Figure 5*b*: As n becomes large, the latter two terms become negligible. Under these large sample conditions, we can then expect that approximately 95% of our prediction errors are within $\pm 1.96 s_{y \cdot x}$ of the sample regression line, and that approximately 99% are within $\pm 2.58 s_{y \cdot x}$ of it, provided, of course, we make the further assumption that the y populations for each predictor value x are normally distributed in addition to having equal variance.

Reduction of prediction errors. An understanding of how well the regression equation predicts the criterion variable y—compared to simply predicting its overall mean value $\bar{y}$, regardless of the value of x—can be had by studying the graph of the equation among the data points of a correlational scatter diagram.

Parts *a* through *d* of Figure 6 show progressively higher degrees of correlation between two hypothetical variables. Beginning with a zero correlation, we see that the variation in prediction errors (indicated by the bold arrows) is exactly equal to the variation of the criterion variable itself (indicated by the double arrow).

This state of affairs can be contrasted with the situation portrayed in parts *b*, *c*, and *d* of Figure 6, where we see that with increasing degrees of correlation the deviations of the observed scores from the regression line get smaller and smaller. That is, the errors of prediction are reduced as the degree of correlation between the variables increases. The limiting situation, of course, would be the case of a perfect correlation in which all the observed points would lie right on the regression line and consequently there would be no errors of prediction.

If it is recognized that the bold arrows in Figure 6 reflect the variance of the y values about the regression line, $s^2_{y \cdot x}$, and that the double arrows reflect the overall variance of the y values, s^2_y, then the preceding relationship can be summarized concisely as

$$\frac{s^2_{y \cdot x}}{s^2_y} \doteq 1 - r^2$$

where $\doteq$ is the symbol for "is approximately equal to" and r^2 is the square of the correlation coefficient between variables x and y. The relationship would be exact were it not for the fact that $n-1$ is used in the definition of s_y, whereas $n-2$ is used in the definition of $s_{y \cdot x}$.

The spice sales vs. shelf space example will illustrate the above relationship. Recalling that $r = .833$, and obtaining the needed variance estimates from Table 3, we have

$$\frac{(5.77)^2}{(9.83)^2} \doteq 1 - (.833)^2$$

or

$$.345 \doteq .306$$

In situations where n is large the approximation will be much closer. In this example the various statistics were based on a sample of only $n = 10$. Notice that if we multiply the left side of the equation by $(n-2)/(n-1)$—i.e., $8/9$—we have an exact relationship $(8/9)(.345) = .306$. Thus, we conclude that the knowledge of the regression equation between x and y has reduced the variance of our errors of prediction to just over 30% of what it would be if we simply predicted the average value of y all the time, regardless of the value of x, or if we did not know the value of x. This is also one context in which

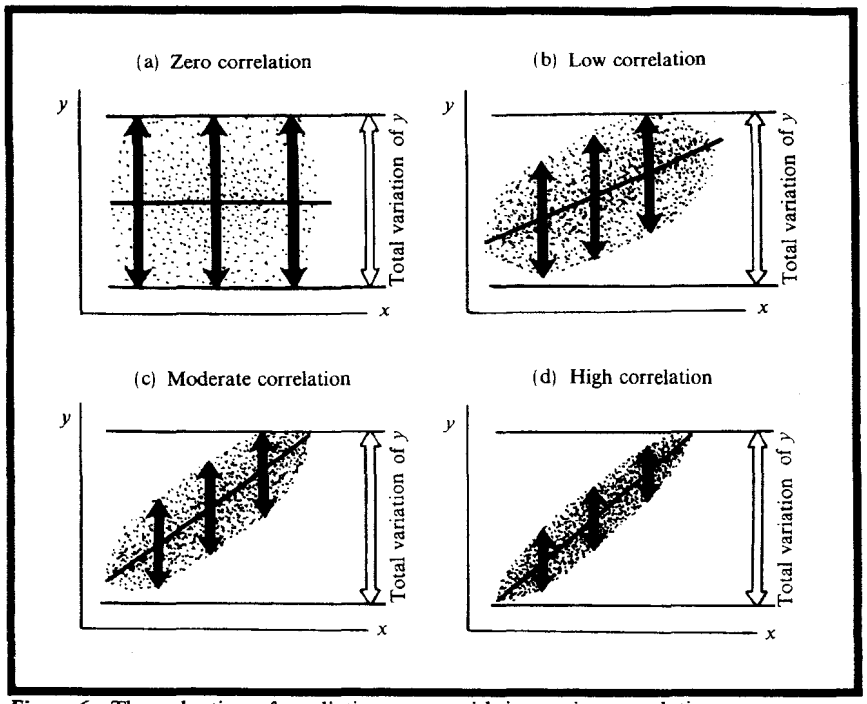

Figure 6 The reduction of prediction errors with increasing correlation.

meaning is given to the interpretation of r^2 as a measure of the proportion of variance in one variable accounted for by variation in the other.

Proportion of variance explained. Another interpretation of the efficacy of a regression equation, and its relation to r^2 as a measure of the amount of variance in one variable accounted for by the variance in another variable, can be seen graphically in Figure 7. For clarity of illustration only a few data points are shown, rather than the swarm of data used in Figure 6.

For each observed data point in the figure there is a corresponding predicted point. The observed data points are shown as open circles, while the predicted points as solid circles. Notice that if we project the observed data points and the corresponding predicted points against the y axis, we can compare their respective variations. In part *a* of Figure 7, portraying a high degree of correlation, the variation of the predicted y values is almost the same as the variation of the observed y values. In other words, nearly all the variation in the y variable is accounted for, or predictable by, the variation in the x variable which gave rise to the predicted scores through the regression equation. We see further in parts *b*, *c*, and *d* of the figure that the variance of the predicted scores compared to the variance of the observed scores gets smaller and smaller as the degree of correlation decreases, until we reach the limiting case of zero correlation, in which case none of the variance in the y variable is predictable from the variance in the x variable, because there is absolutely no variation in the predicted y values; for when there is a zero correlation between two variables, the best we can do in terms of prediction is to predict the mean value of the criterion variable regardless of the value of the predictor variable. This is the situation of a horizontal regression line, one with a slope of zero.

The relationship described above can be described in mathematical terms. If we calculated the variance of the observed and predicted y values, the ratio of the latter to the former would be nothing other than the value of r^2. That is,

$$\frac{s_{y'}^2}{s_y^2} = r^2$$

where the numerator of the ratio is the variance of the *predicted* y values, while the denominator is the variance of *observed* y values.

Again, we can illustrate the above relationship using the spice sales vs. shelf space data from Table 3. Substituting the appropriate values, we can confirm that

$$\frac{(8.19)^2}{(9.83)^2} = (.833)^2$$

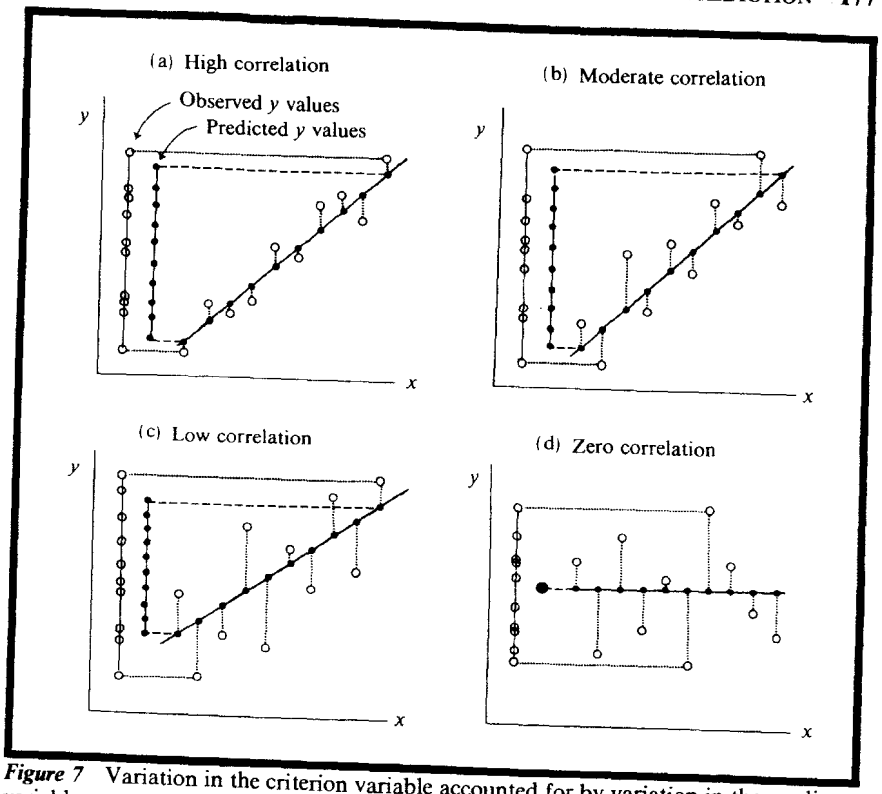

Figure 7 Variation in the criterion variable accounted for by variation in the predictor variable.

does in fact satisfy the equality.

We see, then, that r^2 represents the proportion of variance in the criterion variable accounted for by variance in the predictor variable which gives rise to the predicted y' values via the regression equation. In the above example $r^2 = .69$, signifying that 69% of the variation in spice sales is accounted for by variation in shelf space occupied by the product.

It should be noted that if the data were experimental rather than correlational in nature, then the value of r^2 is no longer the square of the "correlation coefficient"—i.e., an estimate of the population ρ^2—but only resembles it in its mathematical definition. In such instances, r^2 is often referred to as the *coefficient of determination*, and its value will be influenced by the particular values of the predictor variable chosen by us for study. Remember that in the case of correlational data, we have no influence over the values of the predictor variable, while in the case of experimental data, we choose the values of x. An example of the latter situation would be if we purposefully and randomly

varied the amount of shelf space occupied by the spices in a random sample of ten stores, rather than simply observing how the two variables covaried in a natural setting.

Figures 6 and 7 should be studied to see the complementary relationship between these two interpretations of r^2 vis-a-vis the regression line and the predicted scores. In one case, as the correlation increases, the errors of prediction decrease. Alternatively, as the correlation increases, we are able to account for more of the variation in the criterion variable with values predicted from the regression equation. It should be clear, then, that the task of regression analysis does not end with the development of the regression equation, but further involves an assessment of the accuracy with which the relationship is described.

6. Significance Test of the Slope

Although its discussion has been reserved until now, one of the first things we want to do upon obtaining the sample regression equation is to test its slope b.

If there is no relationship between the variables x and y, then the slope of the regression equation would be expected to be zero. To test the hypothesis H_0: $\beta = 0$, we need to know the standard error of the sample slope b, which is the estimate of β, and it is given by

$$s_b = \frac{s_{y \cdot x}}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{8}$$

where $s_{y \cdot x}$, the standard error of estimate, is given by formula (7).

Choosing a significance level α beforehand, we can then test the null hypothesis with the t variable

$$t = \frac{b - \beta}{s_b}$$

which is distributed with degrees of freedom $df = n - 2$.

For the spice sales vs. shelf space example the standard error of the slope b is given by

$$s_b = \frac{s_{y \cdot x}}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{5.77}{\sqrt{33{,}400}} = .032$$

where the denominator and numerator values were obtained from Tables 1 and 3, respectively.

The significance test of the sample slope $b = .1344$ is then

$$t = \frac{b - \beta}{s_b} = \frac{.1344 - 0}{.032} = 4.20$$

which with degrees of freedom $df = 8$ is well beyond the critical value of $t = 1.86$ needed to reject the null hypothesis at the $\alpha = .05$ significance level using a one-tailed test. Consequently, we can be confident that the observed linear equation was not simply a chance departure from a horizontal line, the situation when there is no relationship between the two variables. To use the t test, however, we must make the more stringent assumption that the y populations not only have equal variances, but are also normal in form.

7. Analysis of Residual Errors

If violations of the above assumptions of the regression model are not evident from a knowledge of the data source or from an inspection of the plot of the y values against the x values, then a graph of the prediction or *residual errors*, $y - y'$, will help to point out possible deviations from the assumptions.

Figure 8 presents four examples of such residual plots. In part *a* of the figure the residual errors are evenly distributed and not related to the value of x, suggesting that the relationship between y and x is indeed linear, as required, and that the variance of y for each value of x is the same, as required by the homoscedasticity assumption.

In part *b* of Figure 8 the residual errors increase in variance as x increases, suggesting that the homoscedasticity assumption has been violated by the data.

Figure 8*c* shows a curvilinear pattern for the residual errors, reflecting a curvilinear relationship between the x and y variables themselves, invalidating the linear regression model.

Part *d* of Figure 8 shows the residual prediction errors increasing as x increases and also becoming more dispersed. Such a pattern indicates a violation of either the linearity or homoscedasticity assumptions, or quite possibly both.

This type of residual analysis, along with an inspection of the graph of the original data, will prevent to a large extent the misapplication of the linear regression model and help us to avoid incorrect conclusions based on a purely mechanical application of the technique to a body of data.

As for the assumption of independent y distributions for the various values of x, it is best verified from a logical analysis of the data source. If the same or related objects contribute to more than one data point—either within or between the y distributions—then the observations, and consequently the prediction errors, are not likely to be statistically independent.

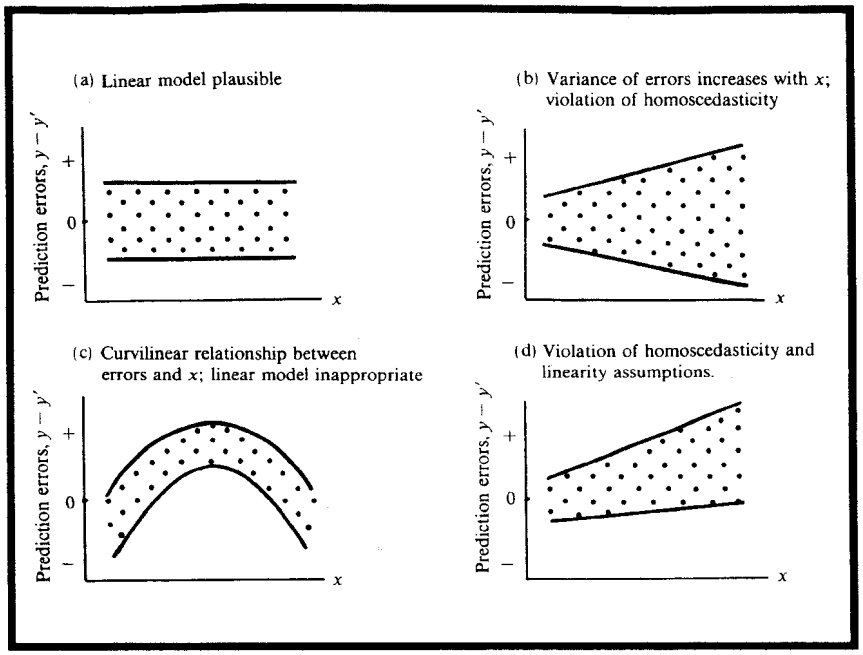

Figure 8 Examples of residual prediction error plots and likely interpretations.

8. Multiple Regression

Multiple regression is an extension of the concept of simple regression. Rather than using values on one predictor variable to estimate values on a criterion variable, we use values on *several* predictor variables. In using many predictor variables instead of just one, our aim is to reduce even further our errors of prediction; or, equivalently, to account for more of the variance of the criterion variable.

The input data for a multiple regression analysis is similar to that for a multiple correlation analysis; namely, a random sample of objects measured on some criterion variable of interest, as well as on k predictor variables. While the multiple correlation analysis requires that the predictor variables are random variables—as opposed to being determined by the researcher—there are multiple regression models to cover both types of situations.

An example of the type of problem to which multiple regression analysis lends itself would be the prediction of college grade point average based on predictor variables such as high school grade point average, aptitude test scores, household income, scores on various entrance exams, etc. A number of other examples of the application of multiple regression analysis will be introduced at the end of the chapter, after the basics of the technique have been studied.

Multiple regression equation. The multiple regression equation will be recognized as similar to the simple regression equation, but instead of a single predictor variable x we have several predictor variables $x_1, x_2, \ldots, x_k$. The general form of the equation is

$$\boxed{y' = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k} \tag{9}$$

where y' is the predicted value of the criterion variable and the values of a and the b coefficients must be determined from the sample data. Since it is based on sample observations, equation (9) must be thought of as an estimate of the true but unknown population equation

$$\mu_y = \alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k \tag{10}$$

Equations (9) and (10) do not represent straight lines as in the case of simple regression where we have only one predictor variable, but rather represent *planes* in multi-dimensional space, a concept admittedly difficult to conceive and virtually impossible to portray graphically. However, its application is easy enough.

As in simple regression, the least squares solution is used to determine the best multiple regression equation; i.e., the values of $a, b_1, b_2, \ldots, b_k$ that will yield values of y' such that the sum of the squared deviations of the predicted y' values from the actual observed y values—$\Sigma(y - y')^2$—is at a minimum. Alternatively, we can think of the least squares solution as that *weighted sum of values on the various predictor variables* that correlates most highly with the values on the criterion variable. For example, if the least squares criterion yielded the following equation for a three-predictor variable problem

$$y' = 24.3 + 7.1x_1 + 6.2x_2 + 91.5x_3$$

we would know that no other equation would yield predictions y' which would correlate more highly with the observed values of y; or, equivalently, no other equation would result in a smaller value of $\Sigma(y - y')^2$, the sum of the squared differences between the actual and predicted values of y.

Regression coefficients. The values of $b_1, b_2, \ldots, b_k$ in the regression equation $y' = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$ are alternatively referred to as b *coefficients* or as *regression coefficients*. In the following section, we will get a better idea of the interpretation, and limits on the interpretation, of the b coefficients; where they will be contrasted with *beta* coefficients based on the regression equation in standardized z score form.

9. Importance of the Predictor Variables

The multiple correlation coefficient R tells us the correlation between the weighted sum of the predictor variables and the criterion variable. Consequently, the squared multiple correlation coefficient R^2 tells us what proportion of the variance of the criterion variable is accounted for by *all the predictor variables combined*.

Still, it would also be worthwhile to know how much each of the *individual* predictor variables contributes to the total explained variance, or, alternatively, to the total reduction in prediction errors. For example, a multiple R^2 of .70 signifies that 70% of the variance of the criterion variable is accounted for, or predictable by, a given set of predictor variables. If, say, there were five predictor variables in this particular situation, would we be able to determine how much of that 70% could be attributed to each of the five predictor variables?

The fact is, there is no satisfactory method for determining the *absolute* contributions of individual predictor variables to their combined effect in accounting for the variance of a criterion variable, when we are dealing with correlational rather than experimental data. The problem lies in the fact that the predictor variables are usually correlated among themselves. We could sooner unscramble an omelette.

If the predictor variables were *uncorrelated* with each other, the problem would be simple. We would merely take the square of the correlation coefficient of a predictor variable with the criterion variable, r^2, as the measure of that predictor variable's contribution to the multiple R^2. In this situation of *independent* predictor variables, the r^2 of the individual predictor variables with the criterion variable will sum to R^2, and then it would be simple arithmetic to determine their percentage contribution to the sum.

When the predictor variables are correlated among themselves, however, the sum of the individual r^2 will be greater than R^2, since most of the predictor variables are *duplicating* the predictive power contained in another predictor variable. In other words, much of the explained variance of the criterion variable would be counted more than once.

What about the possibility of *eliminating* a particular predictor variable from our regression analysis and observe the extent to which R^2 drops in value. While this procedure seems appealing on the surface, it will not accomplish our purpose. If we added the decrements in the value of R^2 resulting from the withdrawal of each predictor variable, the sum would again exceed R^2. And for the same reason as before. Each time we remove a predictor variable we are removing some predictive ability that is in common with other predictor variables, and consequently we end up tabulating it more than once. This procedure, furthermore, can result in the gross misinterpretation of the predictive capacity of a variable. Imagine, for example, that the removal of a particular predictor variable results in a negligible decrease in the value of R^2. Are we to conclude that this variable is unrelated to the criterion variable? Not necessarily. It may be highly correlated with another predictor variable, and in that sense was superfluous for the analysis, but in the absence of that other variable may well have resulted in a substantial drop in the value of R^2.

Beta coefficients. About the best we can do in assessing the relative importance of the various predictor variables is to look at their coefficients in the multiple regression equation when all variables are in their *standardized z score* form; i.e., each with a mean of zero and a standard deviation of one. The coefficients of the standardized predictor variables are referred to as *beta coefficients* or *beta weights*, and the general form of such a prediction equation can be written

$$\boxed{z'_y = beta_1 z_1 + beta_2 z_2 + \cdots + beta_k z_k} \qquad (11)$$

where z'_y is the predicted standardized score on the criterion variable. The beta's are spelled out in (11) so as not to confuse them with the β's in equation (10) where they refer to the theoretical parameters of the population equation in raw score form. The beta's in equation (11) on the other hand, are actually empirical "beta estimates" of the corresponding coefficients of the population equation in standardized z score form. Nonetheless, through common usage they have come to be called simply beta coefficients or beta weights, and are the same as those discussed in conjunction with multiple correlation in the preceding chapter.

The beta coefficients are also sometimes referred to as *partial regression coefficients*. The term "partial" derives from the fact that these regression coefficients are related to the partial correlation coefficients (see Chapter 3) between the respective predictor variables and the criterion variable. That is, the value of the coefficient of each predictor variable x is a function of the correlation between that predictor variable and the criterion variable *as well as the correlations that exist among the predictor variables themselves*. As we have seen in our study of the partial correlation coefficient, it expresses the correlation between two variables under the condition that all other concommitantly measured variables are held constant; that is, it statistically extracts the effects of other variables which correlate with the two variables with which we are concerned—the criterion variable and a given predictor variable.

Since each variable in the standardized form of the multiple regression equation (11) has exactly the same standard deviation and mean, the absolute values of the beta coefficients will tell us the *rank order* of importance of the predictor variables. For example, in the equation

$$z'_y = .44z_1 + .09z_2 + .27z_3$$

All regressions. An alternative method of identifying a concise regression equation is to consider *every possible* regression equation that could be constructed from our set of predictor variables. For this *all-regressions* approach it can be shown that if we have k predictor variables there are 2^k possible regression equations. For example, assume we had five predictor variables: The first one *could* or *could not* appear in the equation, the second one *could* or *could not* appear in the equation, etc. The two possibilities for each of the five variables results in $2\times2\times2\times2\times2$ or $2^5=32$ possibilities.

Modern computers are so fast that every one of the possible regression equations can be computed within seconds, provided the number of variables is not too large. Rather than looking at every single equation though, which would number over 1,000 with as few as ten variables, we could ask the computer center statistician to provide us with the "best" equation when using *one* variable, when using *two* variables, when using *three* variables, and so on. By "best" we mean the ones associated with the largest R^2's.

The stepwise and all-regressions procedures allow us to identify a regression equation based on relatively few predictor variables, yet which accounts for virtually all the variance that could be explained if we used the entire set of predictor variables. This is desirable from the standpoint of parsimony of explanation and economy of data collection.

On the other hand, there is the danger that we might select variables for inclusion in the regression equation based purely on chance relationships. Therefore, as stressed in our discussion of multiple correlation, we should apply our chosen regression equation to a fresh sample of objects to see how well it does in fact predict values on the criterion variable. This validation procedure is absolutely essential if we are to have any faith at all in future applications of the regression equation.

11. Applications of Regression Analysis

As stressed throughout the chapter, two key benefits to be derived from the application of regression analysis include (1) the prediction of values on a criterion variable based on a knowledge of values on predictor variables, and (2) the assessment of the relative degree to which each predictor variable accounts for variance in the criterion variable.

In terms of specific applications of the technique, the possibilities are near limitless. In business and economics there is interest in identifying predictors of sales, productivity, unemployment rates, inflation rates, strike activity, etc. Researchers in education and psychology are interested in predictors of academic achievement, career success, aptitudes, personality traits, mental health, etc. Sociologists, psychologists, and anthropologists are interested in predictors of crime, marriage, divorce, and birth rates. Predictors of crop yield,

animal behaviors, disease durations, and bodily functions such as blood pressure or skin temperature are of interest to researchers in biology and medicine. These are just a few of the criterion variables which might be studied in various fields, and many more could be identified, each with a long list of potential predictor variables.

What is important to realize is that alternative regression analyses can be applied to the same basic analytical problem, depending upon the objects we choose to study. Consider, for example, in the world of business, which perhaps has the widest range of objects for study, an automobile manufacturer interested in the criterion variable "sales of Model *A*." Now this is a very broadly stated problem and needs sharper definition. To be more specific, the manufacturer is interested in knowing which variables are related to the sales of Model *A*, information which can then be used to possibly increase its sales. While we recognize that on the surface this problem lends itself to a regression analysis, we must formulate the problem more specifically in order to apply the analytical technique. It is at this point that the ingenuity of the investigators comes into play.

A number of options are available to the planners of the study. The criterion variable has been identified as the sales of Model *A*, so the next task is to identify a set of *objects* on which to measure this variable. We might consider the individual *dealerships*, since they surely vary with respect to the criterion variable. Next, we need to identify a set of potential predictor variables, characteristics of the dealerships that might be related to our criterion variable of Model *A* sales.

A management, sales, and research team will brainstorm to come up with a comprehensive set of potential predictor variables: e.g., population density in a fifteen-mile radius, distance to nearest dealer of competitor *B*, local advertising expenditure, number of feet of street frontage, average gasoline price in a fifty-mile radius, number of sales personnel, number of service personnel, number of years at existing site, average trade-in allowance, number of autos on hand, etc.

As can be easily seen, the list could go on and on. Notice, also, that the nature of the objects chosen for the analysis, dealerships, more or less dictates the nature of the predictor variables. Note further that some of the predictor variables are under the manufacturer's control (e.g., number of sales personnel, trade-in allowance, number of autos on hand, etc.) while others are not (e.g., average gasoline price, years at existing site, population density, etc.). However, even in those instances in which the manufacturer has no direct control over the predictor variable, the regression analysis could still be beneficial in selecting future dealership sites. In the case of the predictor variables which are under the manufacturer's control, they can later be manipulated on a test basis to determine if they do indeed *cause* changes in sales.

The auto marketer can attack the very same problem from a completely different angle. With sales of Model *A* still the criterion variable, we may choose as objects not individual dealerships, but individual *sales persons*. The predictor variables could then include dimensions such as years of experience, age, scores on a personnel test, grade in a training course, height to weight ratio, etc. The regression analysis will then tell us how these characteristics are related to Model *A* sales for individual sales persons.

Or, alternatively, the manufacturer could use marketing *territories* as objects, which would dictate predictor variables such as population, number of Model *A* dealers, number of dealers of competitor Model *B*, advertising expenditure, and many of the same variables that were applicable in the dealership analysis.

Yet another possibility is to use *time periods* as the objects of our analysis. Sales during the various time periods would be related to such variables as number of used cars sold, advertising expenditure, rainfall, weeks since new model introduction, unemployment rate, sales of Model *B*, size of payroll, total hours open to public, etc.

We can even use *prospective car buyers* as objects. The criterion variable could be a numerical rating of interest in purchasing Model *A*, and the predictor variables could be ratings on a series of image characteristics such as "attractive styling," "roomy interior," comfortable ride," "good resale value," "economical to drive," etc. The extent to which the ratings on the various image dimensions predict ratings of purchase interest would shed further light on the factors that influence sales of Model *A*.

What we have seen in the above examples is that a single criterion variable can be studied with a number of alternative regression analyses. By understanding the extent to which characteristics of dealerships, sales personnel, sales territories, time periods, and consumers are related to Model *A* sales, the marketing efforts of the automobile manufacturer can be adapted to improve the sales of Model *A*, the criterion variable.

This type of analysis—identifying a criterion variable of interest, selecting an appropriate set of objects on which to measure it, and identifying a set of potential predictor variables—is applicable to the widest possible range of analytical problems, whether the criterion variable is sales, crop yields, attitudes, academic achievement, job success, strike activity, disease levels, crime rates, or life span. It should also be clear from the preceding examples that regression analysis is more than the mechanical application of a statistical technique to a matrix of data. The formulation of the problem, the identification of criterion and predictor variables and the objects upon which they are measured, and the interpretations of the resulting regression equation and the accompanying R^2 and beta weights, will determine how useful the analysis will be. And, of course, we must be satisfied that the raw data conforms to the statistical assumptions of the given regression model.

Also, we have seen that there is no one regression analysis that is most appropriate for understanding a criterion variable, but rather the greatest understanding is most likely to result from a number of alternative analyses, each viewing the problem from a different angle.

To round out our discussion of regression analysis the balance of the chapter will touch briefly on several special topics related to the application and interpretation of the technique.

12. Collinearity Problem

A particularly vexing problem in the application of multiple regression analysis arises from the situation in which two or more predictor variables are very highly correlated with each other. This is referred to as the *multicollinearity* problem, or simply as *collinearity*. Under such conditions the computer attempting to analyze the data according to its stated instructions is likely to go awry. Exactly when this will happen is not always identifiable, otherwise it could be prevented. We should be forewarned, though, to use some common sense in our selection of predictor variables so as not to include groups of variables that we know on logical grounds must be highly correlated with each other. For example, we would not include the variables of sales, costs, and profits into a regression analysis since any two will automatically determine the third.

Related to the collinearity problem is the situation in which we include a predictor variable that is really not a predictor variable as such but rather a slight variation of the criterion variable. For example, if our criterion variable was defined as the sales performance of a set of sales reps, and among our predictor variables we included the commissions earned by the reps, it is unlikely that we would gain any information on the other predictor variables we studied, since commissions would account for virtually all the variance in the criterion measure. If there is no variance left to account for, how can we assess the importance of the remaining predictor variables? Either we should have left sales commissions out of the problem or let it stand as a criterion variable. In situations such as this, we cannot expect the computer to think for us.

13. Dummy Variables

In order to use qualitative predictor variables (such as sex) in a regression analysis, we can transform the variables into quantitative *dummy variables*. Essentially what we do is convert each level of a qualitative variable into a binary variable. For example, the qualitative variable of sex (male vs. female)

could be made into a dummy variable representing *maleness*—i.e., male vs. not male—with the respective numerical values 1 and 0. Or we could construct the dummy variable representing *femaleness*—i.e., female vs. not female—again with the respective numerical values 1 and 0. The three levels of the qualitative political affiliation variable—Democrat, Republican, and Independent—could be made into three dummy variables: Democrat vs. not Democrat, Republican vs. not Republican, and Independent vs. not Independent. In each case, one level of the dummy variable could take on the value of 1 and the other level a value of 0.

The benefit of such a transformation is that the quantitative dummy variables can now be introduced as predictors into a regression analysis. For example we could determine if the sales reps' sex was related to their sales performance. Or we could determine if the dominant political affiliation of a voter district was related to the district's crime rate.

However, when we use dummy variables we must keep the collinearity problem in mind. We cannot introduce dummy variables for every level of the qualitative variable, since they are not independent of one another. If we know that an individual has a value of 0 on the dummy variable of "maleness" we can predict perfectly the individual's value on the dummy variable "femaleness"—namely, it must be 1. Therefore we need to include only one of these two dummy variables. In the case of dominant political affiliation of a voter district, if we know that a district scores 0 on "Democratness", and 1 on "Republicanness" we know for sure that it must score 0 on "Independentness." So, we need to include only two of these three dummy variables. Knowing a district's value on any two of the three dummy variables will automatically inform us of its value on the remaining dummy.

In general, when we construct dummy variables from a qualitative variable, we will always want to use *one less* than the number we can create. For example, if we have classified voters into ten occupation categories, we can create nine dummy variables for use in a regression analysis for predicting frequency of voting based on occupation.

14. Autoregression

An interesting application of regression analysis is to predict values on the criterion variable based on values of the same criterion variable obtained earlier in time. Imagine the price of Stock *A* on each of 100 trading days. Now let us pair each of these prices with the price on the immediately preceding day, as explained in our discussion of serial correlation. We can now try to predict the price of Stock *A* on a given day based on its price the previous day. In fact, we could turn it into a multiple regression equation by introducing as additional predictor variables the stock's price two days back, three days back, etc.

While we should be so lucky as to be able to predict the future in the stock market, the autoregression technique is useful in identifying dependencies among data collected sequentially which we may wish to extract before submitting the data to further analysis. The technique is also useful for projecting time series data such as crime rates, fertility rates, strike activity, etc.

15. Regression to the Mean

The expression "regression" originated from the observation that exceptionally tall fathers tended to sire sons who, when matured, tended to be shorter than their father's height. Similarly, exceptionally short fathers had sons who tended to be taller than their fathers. While a full interpretation of such a finding would require theories of genetics and the dynamics of mate selection, we can attribute it partly to the phenomenon of *regression to the mean*.

To understand the concept consider that any empirical measurement of a characteristic is composed of two parts—the *true value* of the characteristic plus or minus some *error*. On repeated measurements the true value remains the same but the error component fluctuates. We know that when we measure a large number of objects with respect to a characteristic, some of the objects will score high, some low, and some in between. Now wherever the object scores, part of the score is due to an error component. Thinking in terms of conditional probabilities, we can ask ourselves whether those objects that scored exceptionally high were not benefiting from a large error component; and, similarly, those that scored exceptionally low, were in the receipt of a large negative error component. Cast in a different light, suppose we knew only the size of an object's error component: What would we predict as the object's total score if we knew it had an exceptionally large positive error component—would it tend to be above or below the mean. The dynamics of this phenomenon become apparent when we *remeasure* our set of objects on the same characteristic and compare their respective values on the two measurements. What we find is that those that scored exceptionally high (or low) on the first measurement score closer to the mean on the second measure; that is, there is a *regression to the mean*. The greater the error component, the greater will be the regression or "turning back" to the mean.

This phenomenon is worth bearing in mind whenever exceptional scores on a single measurement are singled out for attention; especially when the objects possessing these scores are to receive special due, as in academic, medical, or business settings. For example, a year after introduction of a new product, two cities are singled out as having exceptionally high sales. During the next year these cities receive all manner of special attention and marketing

expenditure. After the second year it is found that their performance has dropped compared to their first year performance. The cities which had the best second-year performance gain were those with lackluster first-year performance. In other words, much of the initial variation in sales from city to city was due purely to chance variation, and the cities that performed best during the first year just happened to be recipients of a larger positive error than the other cities. On the other hand, this need not be so: It could well be that the variation was not due to error at all, but to fundamental causal factors operative in each city. This is the variance that regression analysis tries to tap.

Consider as another example a mutual money market fund that boasts having the best performance of all the leading investment funds during the most recent year. We should not be too impressed with this performance. After all, of the many funds, and of the many starting each year, one of them *had* to do better than all the others. This is a truism. Again, we want to know the reliability of this performance. Will it duplicate its performance next year, or will another fund claim the leadership role, while the other regresses to the mean.

16. Self-Fulfilling Prophesy

The true validity of the predictions arising from a regression analysis cannot always be ascertained. Since the analysis is not purely an intellectual exercise but the basis for action, the outcome of the analysis may often provoke activities that make the predictions come true—the phenomenon of *self-fulfilling prophesy*. If sales for certain stores in a chain are predicted to be above average, these stores may enjoy special promotions and other attention they might not normally receive, and consequently live up to expectations but for the wrong reasons. The same is likely to happen when students are placed into special classes based on achievement or aptitude test scores.

On the other hand, we might experience a *self-negating prophesy* in which dire predictions are forestalled through corrective actions. For example, predictions of falling sales, poor academic achievement, or disease onset may result in special remedial efforts to avoid such possibilities. In instances such as these, it becomes a very philosophical question as to whether our regression analysis has validity, for while our dire predictions did not come true, we surely benefited from the analysis.

17. Concluding Comments

In this chapter we have touched upon the basic concepts of regression analysis, a technique for describing the mathematical relationship between a criterion variable and one or more predictor variables. We also discovered how r^2, beta coefficients, and the measures of prediction error help us to interpret

the practical value of a regression equation. While regression analysis can be applied to problems in which the predictor variables are either random variables or fixed experimental variables, we concentrated our examples on the former type since they are so commonplace. Experimental variables, as they are related to a criterion variable, will be discussed at greater length in the following chapter on Analysis of Variance, although it should be recognized that they can also be described by the regression approach.