1989

Simple regression methods: chapter 8

Makridakis, Spyros

Wiley, John & Sons

Forecasting Methods for Management

SPYROS MAKRIDAKIS
The European Institute of Business Administration (INSEAD)

STEVEN C. WHEELWRIGHT
Graduate School of Business Administration, Harvard University

FIFTH EDITION

WILEY

JOHN WILEY & SONS
New York · Chichester · Brisbane · Toronto · Singapore

To our children

Aris and Petros

and Marianne, Michael, Melinda, Kristen,
Matthew, and Spencer

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional service. If legal advice or other expert assistance is required, the services of a competent professional person should be sought. *From a Declaration of Principles jointly adopted by a Committee of the American Bar Association and a Committee of Publishers.*

Library of Congress Cataloging in Publication Data:

Makridakis, Spyros G.
Forecasting methods for management.

Includes index.
1. Economic forecasting. 2. Economic forecasting — methodology 3. Business forecasting. I. Wheelwright, Steven C. II. Title.
ISBN 0-471-60063-6

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Reproduced from "Forecasting Methods for Management"
By S.Makridakis and S.C.Wheelwright
Fifth Edition

CHAPTER 8

SIMPLE REGRESSION METHODS

In the preceding three chapters, several major classes of time-series forecasting methods were examined—exponential smoothing, decomposition, autoregressive/moving average, and filters. Various models within each class were presented, models appropriate for different patterns of data and different conditions.

In this chapter and the next two, we will examine another approach to forecasting—explanatory methods. It is one thing to fit a model (such as an exponential smoothing model) to a time series. It is quite another to come up with other variables that relate to the data series of interest and to develop a model that expresses the way the various variables are related.

Thus these chapters introduce a new concept in the attempt to forecast. A forecast will be expressed as a function of a certain number of factors or variables that influence its outcome. Such forecasts do not necessarily have to be time dependent. Developing an explanatory model facilitates a better understanding of the situation and allows experimentation with different combinations of inputs to study their effects on the forecasts. In this way, explanatory models can, by their basic formulation, be geared toward intervention—influencing the future through decisions made today.

In the application of simple regression, the focus of this chapter, we assume that a relationship exists between the variable we want to forecast (the "dependent" variable) and another variable (the "independent" variable).

Furthermore, we assume that the basic relationship is linear. Thus when we discuss simple linear regression, we mean relationships where Y, the item to be forecast, is a linear function of X, the independent variable. Obviously there are many situations in which this is not a valid assumption; for example, if we were forecasting monthly sales and it was believed that those sales varied according to the seasons of the year, such an approach would be inappropriate (unless the nonlinear seasonality were first transformed into a linear form). It may be, however, that if we were forecasting the same sales items, but on an annual basis, these sales could be modeled using a linear

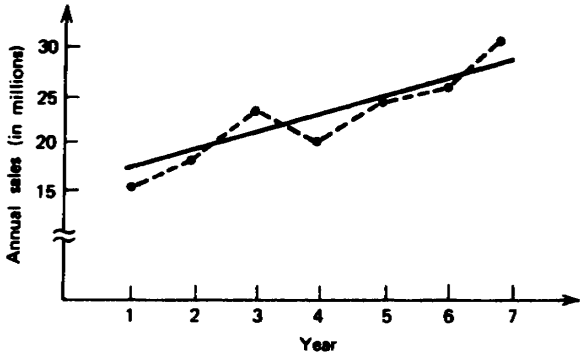

Figure 8-1 Projecting Annual Sales.

relationship. Figure 8-1 represents the pattern of data points that might exist when we look at annual sales.

It can be seen in Figure 8-1 that a straight line sloping upward (that is, a trend going upward) as we go from left to right would give a fairly good approximation of future sales. This was the kind of pattern that we saw in Chapter 5 that could be handled with Holt's linear exponential smoothing. In this chapter, however, we consider a method of handling the same type of pattern, which has different characteristics than the linear smoothing technique.

Figure 8-1 and those that we have considered in earlier chapters all involve forecasting some variable in terms of the time period. Thus, if we were to graph each of these situations, we would have the time variable on the horizontal axis and the variable that we wish to forecast on the vertical axis. Simple regression analysis is a technique that can deal with this type of relationship. We therefore assume that the independent variable X is the time (for example, X can take the values of 1, 2, 3, 4, 5, 6, 7, or 1982, 1983, 1984, 1985, 1986, 1987, 1988, in Figure 8-1).

There are a number of situations where we may want to forecast one variable on the basis of its relationship with two or more independent variables, one of which may be time. Multiple regression analysis, the topic of the next chapter, deals with this type of situation.

Simple regression also can be used when the single independent variable is not time. Consider, for example, a forecasting need faced by a large mail-order house. Each day a tremendous amount of mail is received, much of it containing orders that have to be filled. The mailing department has learned through experience that the number of orders to be filled seems to be related to the weight of the mail. They feel that it would be extremely useful to them if they could weigh the mail when it arrives in the morning and use that weight to predict the number of orders that will have to be filled that day, so they can schedule the time of the people who will fill those orders and decide whether overtime will be necessary.

As a first step in determining whether a relationship exists between mail weight and orders, they have kept a record over several days of the weight of the mail each day and the corresponding number of orders. These pairs of values can be plotted on a graph to identify the relationship between the weight of the mail and the number of orders, if such a relationship exists. It is evident from Figure 8-2 that there is a relationship between the weight of mail and the number of orders. This means that as the weight of mail increases, so does the number of orders. Furthermore, such a relationship is called linear because increases (or decreases) in the amount of mail bring proportional increases (or decreases) in the number of orders received.

The linearity of the relationship can be seen by looking at Figure 8-2 and realizing that the best way of describing the relationship between the weight of mail and the number of orders is the straight line that passes through the middle of all the points that denote weight of mail and number of orders. Figure 8-2 is called a scatter plot or diagram. It helps us visualize, graphically, relationships (or lack thereof) between pairs of variables. The straight line in Figure 8-2 is called the regression line, which can be computed statistically (see below), thus allowing us to measure the extent of the relationship between weight of mail and number of orders, that is, how much the number

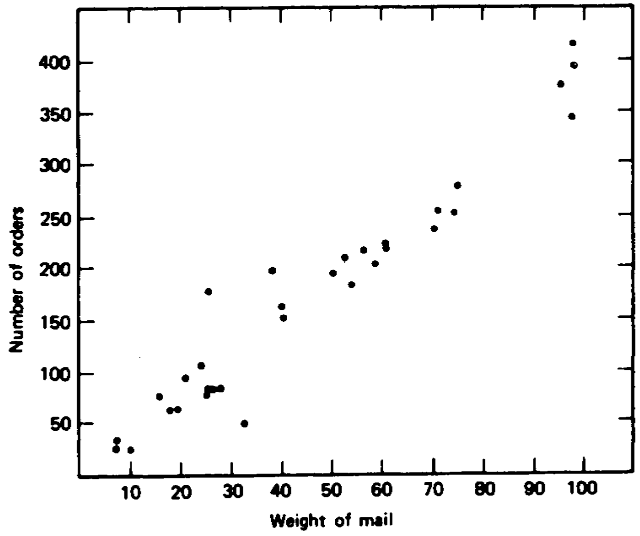

Figure 8-2 Plot of the Number of Orders and Mail Weight.

of orders will increase (or decrease) when the weight of mail increases (or decreases). If such a relationship can be measured, it is obvious that it can be used by the department to forecast the number of orders to be filled each day by the weight of mail received. Such knowledge can improve planning by assigning the right number of people to process orders each day.

This requires an explanatory model between weight and the number of orders. The method we shall examine in this chapter will allow us to obtain such a model through the statistical method of simple regression.

In the use of simple regression, the starting point is the assumption that a basic relationship exists between two variables and can be represented by some functional form. Mathematically it can be written as

$$Y = f(X).$$

This states that the value of Y is a function of (or depends on) the value of X. If it is assumed that the relationship is linear, it can be written as

$$\hat{Y} = a + bX.^{*} \tag{8-1}$$

Since this is the general form of any linear relationship, it is important that the reader understand just what this means. Suppose that the value of X is 0. In such a case $\hat{Y}$ would have the value a. Thus a is the point at which the straight line intersects the Y axis. If we refer again to Figure 8-2, this would mean that when the weight is 0 pounds, the number of orders would have the value of a, which we would reason to be 0, because if no mail is received, no orders are received (by mail). The value of b in Equation (8-1) is called the regression coefficient and indicates how much the value of $\hat{Y}$ changes when the value of X changes one unit. Thus, if we are comparing the number of orders from 40 pounds of mail with the number from 41 pounds of mail, we would expect an increase of b orders from that additional pound of mail. In the next section of this chapter we discuss exactly how the values of a and b can be computed for the mail-order example. Before doing so, however, it is useful to discuss briefly the concept of a linear relationship between two variables.

In many instances the relationship between two variables with which the manager is concerned is linear. In others, although the relationship does not appear to be linear when plotted, it might be possible to make it linear through some appropriate transformation of one of the variables, which would result in a new variable that does have a linear relationship with the other variable.

*We use $\hat{Y}$ to indicate an estimated or forecast value for Y. We use Y to indicate an actual or observed value.

A simple example will help to illustrate this point. Suppose we have two variables Y and W, whose relationship can be written as $\hat{Y} = a + b/W$. There is not a linear relationship between Y and W. However, if we let $X = 1/W$, this equation can be rewritten as $\hat{Y} = a + bX$, which is a linear relationship.

It is also possible to transform exponential relationships into linear ones through the use of logarithms. Many other nonlinear relationships can therefore be made (transformed to become linear). Although we realize that the topic of transforming nonlinear relationships is not an easy one, we want to point out that such transformations are possible and enhance the applicability of regression analysis to many more types of relationships than just linear ones. The interested reader can find more details in Makridakis, Wheelwright, and McGee (1989).

There are many situations in which regression analysis can be applied successfully and appropriately. The two main strengths to its application in forecasting, are (1) regression analysis can be used to explain what happens to the dependent variable through changes in the independent variable(s), and (2) it uses a statistical model to discover and measure the relationship if one exists. Later on in this chapter we shall see just how this can be done.

DETERMINING THE PARAMETERS OF A STRAIGHT LINE

In the preceding section the notion of a linear relationship that could be represented mathematically as $\hat{Y} = a + bX$ was presented. However, unless the values of a and b can be estimated, the mangaer cannot use the relationship between X and Y to forecast. What is needed, therefore, is a means of estimating the values of a and b. These values are referred to as parameter estimates in the equation for a straight line. Several methods can be used to estimate these parameters. Perhaps the most straightforward technique is to plot the historical observations, as in Figure 8-2 for the mail-order example, and to draw visually a line that passes through the middle of these points. In the mail-order case the line would begin at 0.0 and pass approximately midway among the historical points. Once this is done, the values of the parameters a and b could be read off the graph. Since a is the point at which the line intersects the Y axis, its value would be 0 in this example, and the value of b would be the increase in Y (the number of orders) for a unit (one-pound) increase in X.

Although the graphical method can work fairly well in this example, we often have several hundred observations that are widely scattered. Thus it can be difficult to draw a straight line that in some sense will give the best approximation of the relationship. What is needed is a technique for determining the values of a and b that can be used consistently and gives the "best" result. Regression analysis uses such a method, referred to as the *method of least squares*. To see how it fits a straight line to historical observations, we consider a simple example that includes only four observations,* the values of which are plotted in Figure 8-3. The dependent variable (the item we want to forecast) is the cost of production per unit, and the independent variable (the item that determines the cost of production) is the number of units produced. Thus we would like to determine the relationship between the cost and the number of units in such a form that when we specify the number of units to be produced we can forecast (estimate) their cost.

The dashed line in Figure 8-3 approximates the straight line whose equation is $\hat{Y} = a + bX$ and for which we shall determine the values of a and b. We shall use the method of least squares to determine these values in such a way that the line represents the "best" linear relationship for these four points.

The rationale of the method of least squares is that the distance between the actual observations Y and the corresponding points on the line $\hat{Y}$ should be minimized. More precisely, the criterion is that the sum of the squared errors between the actual cost Y and the estimated $\hat{Y}$ found through the regression line should be kept as small as possible by appropriately choosing a and b. To see what this involves, we consider Figure 8-4, in which the observed values (costs) are labeled Y_1, Y_2, Y_3, and Y_4, the deviations (errors) from the regression line (estimated costs) are labeled e_1, e_2, e_3, and e_4, and the points (costs) estimated by the regression line are labeled $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$, and $\hat{Y}_4$. The latter points are what we would forecast by using the regression line with values of X_1, X_2, X_3, and X_4, respectively.

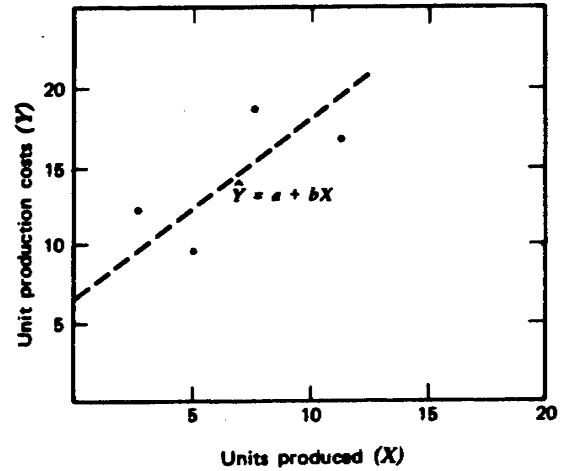

Figure 8-3 Forecasting Production Costs Based on the Number of Units Produced.

In this figure each of the deviations (errors) can be computed as $e_i = Y_i - \hat{Y}_i$, and each of the values of the regression line can be computed as $\hat{Y}_i = a + bX_i$. The method of least squares determines the values of a and b in such a way that the sum of the squared deviations $\Sigma e_i^2 = \Sigma(Y_i - \hat{Y}_i)^2$ is minimized (hence the name *least squares*).

Through calculus we can determine the values for a and b in such a way as to minimize the sum of squared errors, that is, Σe_i^2. Such values for a and b are found by applying the following two formulas:

$$b = \frac{\Sigma XY/n - \bar{X}\bar{Y}}{\Sigma X^2/n - \bar{X}^2} \qquad (8\text{-}2)$$

$$a = \bar{Y} - b\bar{X} \qquad (8\text{-}3)$$

where

$$\bar{Y} = \frac{\Sigma Y}{n} \quad \text{and} \quad \bar{X} = \frac{\Sigma X}{n}$$

and n is the number of observations (data points) with which the regression is estimated.

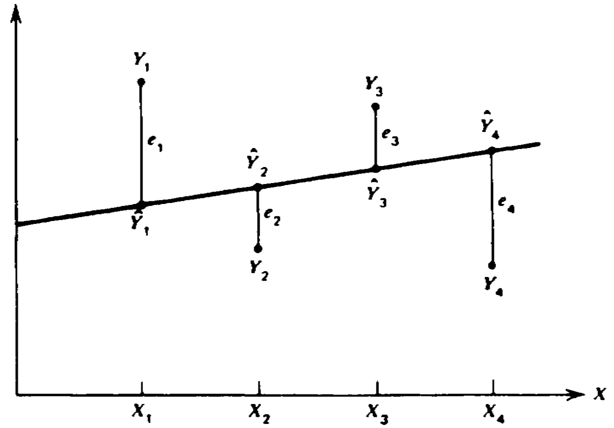

Figure 8-4 Quantities Used in the Method of Least Squares.

*The reader should note that for practical purposes, regression analysis can be applied successfully only when many data points are available. Here we have chosen an example with only four data points to keep the arithmetic at a minimum and to focus attention on the concepts.

To see how these computations can be made in practice, we return to the example given in Figure 8-3. The data, the relevant computations, and the resulting equation for the regression line are given below. The data and computations to find the values of a and b are as follows:

Cost Y	Units X	Y^2	X^2	XY
8	3	64	9	24
11	2	121	4	22
16	5	256	25	80
15	7	225	49	105
$\Sigma Y_i = 50$	$\Sigma X_i = 17$	$\Sigma Y_i^2 = 666$	$\Sigma X_i^2 = 87$	$\Sigma X_i Y_i = 231$

$$\bar{Y} = 50/4 = 12.5$$

$$\bar{X} = 17/4 = 4.25.$$

Then

$$b = \frac{\Sigma XY/n - \bar{X}\bar{Y}}{\Sigma X^2/n - \bar{X}^2} = \frac{231/4 - 12.5(4.25)}{87/4 - (4.25)^2} = 1.254$$

$$a = Y - bX = 12.5 - 1.254(4.25) = 12.5 - 5.33 = 7.17$$

Thus

$$\hat{Y} = 7.17 + 1.254X.$$

The values found using the regression equation $\hat{Y} = 7.17 + 1.254X$ are as follows:

$\hat{Y}$	e	$\hat{Y} - \bar{Y}$	$Y - \bar{Y}$
10.93	−2.93	−1.57	−4.5
9.68	1.32	−2.82	−1.5
13.44	2.56	0.94	+3.5
15.05	−0.95	3.45	+2.5
	$\Sigma e_i = \Sigma(\hat{Y}_i - \bar{Y}) = 0$		$\Sigma(Y_i - \bar{Y}) = 0$

To illustrate further how the method of least squares can be applied, we return to the mail-order example discussed earlier in this chapter. Table 8-1 lists some of the relevant data needed in calculating a and b.

Using Equations (8-2) and (8-3), we obtain

$$b = \frac{\Sigma XY/n - \bar{X}\bar{Y}}{\Sigma X^2/n - \bar{X}^2} = \frac{337{,}987/30 - 47.3(177.4)}{89{,}592.9/30 - (47.3)^2} = 3.84$$

$$a = 177.4 - 3.84(47.3) = -4.36.$$

Thus

$$\hat{Y} = -4.4 + 3.8X.$$

Now if a manager wants to forecast what the number of orders will be for a given weight of mail, he or she can do so using this equation. For example, if the company received 45 pounds of mail, the manager would estimate that the number of orders $\hat{Y}$ would be

$$\hat{Y} = -4.4 + 3.8(45) = 166.6.$$

The reader will recall that earlier we reasoned that if no mail were received, no orders would be received (by mail). If we applied the foregoing equation to the extreme, however, it would say that if no mail were received, the company would receive −4.4 orders, which is clearly absurd. This points up the need for applying managerial judgment when using regression analysis results and for identifying a range within which the linear relationship holds but outside of which it may have little meaning. In this example the company probably receives several pounds of mail each day unrelated to actual orders, which would explain why on an average more than one pound of mail must be received before any orders are received, as stated by the regression equation.

Furthermore, it is possible to determine the regression line in such a way that the value of a is equal to 0. The reasoning for doing so can be seen in the next section when the statistical significance of the values of a and b is discussed. At this point we can say that if the manager wants a to equal 0, he or she can recalculate the regression equation (most computer programs allow the user to force the regression equation to pass through 0, that is to make a equal to 0), which becomes:

$$\hat{Y} = \frac{b}{A}X.$$

Table 8-1 Data for Applying the Method of Least Squares in the Mail-Order Example

(1) Observation	(2) Weight of Mail X (Pounds)	(3) Number of Orders Y	(4) X^2	(5) Y^2	(6) XY
1	70.6591	254.272	4991.30	64,654.10	17,964.10
2	48.3784	199.311	2340.47	39,725.00	9,642.35
3	19.8185	64.693	392.77	4,185.22	1,282.12
4	39.8517	162.644	1588.16	26,453.20	6,481.65
5	50.1270	195.664	2512.71	38,284.20	9,808.02
6	68.4879	238.021	4690.59	56,653.80	16,301.50
7	55.7871	219.089	3112.21	48,000.00	12,222.40
8	96.7574	392.389	9361.99	153,969.00	37,966.60
9	40.5927	151.125	1647.76	22,838.70	6,134.55
10	52.7543	210.728	2783.01	44,406.30	11,116.80
11	7.2491	33.963	52.55	1,153.49	246.20
12	73.2380	256.993	5363.80	66,045.50	18,821.70
13	96.3148	347.976	9276.54	121,087.00	33,515.20
14	18.1598	65.733	329.77	4,320.86	1,193.70
15	93.5000	377.330	8742.26	142,378.00	35,280.40
16	25.4237	83.491	646.36	6,804.89	2,097.24
17	24.0767	82.211	579.68	6,758.64	1,979.37
18	74.6311	281.124	5569.80	79,030.70	20,980.60
19	53.1468	180.763	2824.58	32,675.20	9,606.97
20	28.5609	86.153	815.72	7,422.41	2,460.62
21	10.5201	23.890	110.67	570.74	251.32
22	21.1329	98.291	446.59	9,661.13	2,077.17
23	58.9518	206.030	3475.31	42,448.50	12,145.90
24	60.6061	221.245	3673.10	48,949.40	13,408.80
25	24.4888	105.029	599.70	11,031.00	2,572.03
26	16.7616	77.943	280.95	6,075.26	1,306.47
27	7.6762	24.240	58.92	587.62	186.08
28	60.8193	224.298	3698.99	50,309.60	13,641.70
29	94.3726	369.987	8906.18	136,891.00	34,916.60
30	26.8395	88.626	720.35	7,854.70	2,378.70
	$\Sigma X = 1419.6$ $\bar{X} = 1419.6/30 = 47.3$ $\bar{Y} = 5322.0/30 = 177.4$	$\Sigma Y = 5322.0$	$\Sigma X^2 = 89{,}592.9$	$\Sigma Y^2 = 1{,}281{,}220$	$\Sigma XY = 337{,}987$

The versatility and real power of simple regression have been explained only partly so far. Although we have developed the equations necessary to specify the most appropriate linear relationship, two other questions are of concern. First, what is the reliability of the forecasts based on a given regression line? For example, if this forecasting method indicates in the mail-order example that the number of orders will be 166.6, how certain can the manager be that the actual number of orders will not fluctuate between, say, 146.6 and 186.6? Second, when is it not appropriate to say that one variable is influenced by another because no real relationship between the two can be established? (These two questions are discussed in the next section).

THE PRECISION AND SIGNIFICANCE OF A REGRESSION EQUATION

It is possible to make statistical statements about the significance of the regression equations. The use of the statistical properties will also allow us to make statements about the likelihood that future values will vary from the forecast by certain amounts, the confidence that we can place in having determined the most appropriate straight line, and the accuracy of the coefficients *a* and *b*.

Several questions concerning the significance of an application of regression analysis can be dealt with in statistical terms. We consider three of them at this point:

1. Is the regression coefficient *b* significantly different from 0, or did it just occur by chance? The same question can be asked about the value of *a*.
2. What level of confidence can be placed in the regression coefficients *a* and *b*, that is, how precise is the estimate of *a* or *b*? For what range of values around *a* or *b* can the manager be confident that the true values of *a* or *b* are within those ranges?
3. How confident can the manager be when making a forecast $\hat{Y}$ that the actual value of *Y* will lie within a range around that forecast value, that is, what is the precision of $\hat{Y}$?

Turning first to the question of the significance of the regression coefficient *b*, we would like to know whether the true value of *b* is really different from 0. (If *b* is not different from 0, the best model will be $Y = \bar{Y}$.) Since we have estimated the value of *b* on the basis of a limited number of observations, we might have found a value different from 0 merely by chance. Thus what we would like to do, using statistics, is to say: if we suppose that the true value of *b* is 0, what is the likelihood (or chance) that we could have had our specific value of *b*?

The statistic needed to determine the significance of a regression coefficient is the standard error of that coefficient. For *b* this can be computed using Equation (8-4)

$$\text{SE}_b = \frac{\sqrt{\Sigma(Y_i - \hat{Y}_i)^2/(n - 2)}}{\sqrt{\Sigma(X_i - \bar{X})^2}} \tag{8-4}$$

The numerator in this equation is called the standard deviation of regression. This is the square root of the sum of the squared errors ($e_i = Y_i - \hat{Y}_i$) adjusted for degrees of freedom. The denominator is the square root of the sum of the squared deviations of *X* from the mean $\bar{X}$. Because there is a lot of arithmetic involved in computing SE_b, the standard error of *b*, this computation is usually done as an integral part of a computer program. To see how this standard error can be applied in practice, we can compute it for the mail-order example:

$$\text{SE}_b = \frac{\sqrt{\Sigma(Y_i - \hat{Y}_i)^2/(n - 2)}}{\sqrt{\Sigma(X_i - \bar{X})^2}} = \frac{14.63}{149.3} = 0.098.$$

The value of the standard error tells us something about the precision of the estimated regression coefficient *b* (whose value was found to be 3.8). By making the assumption that the various values of *b* (which can be found from many different samples, of size 30, of weight of mail and number of orders) are normally distributed, we can establish the maximum possible error in the estimated value of *b* (it is related to the value of the standard error of 0.098). Thus, in the worst case scenario, which would include practically all cases, the real value of *b* cannot be outside the range of $b \pm 3\text{SE}_b$, or 3.8 ± 0.294. That is, *b* can vary from about 3.5 to 4.1. This means that the real value of *b* cannot be 0 for any practical purposes.

Alternatively, we can compute the number of standard errors our value of *b* is from 0. Since $b = 3.8$ we divide 3.8 by 0.098 and obtain about 38. Thus the *b* value we computed is approximately 38 standard errors from 0. Using a table of *t* values, we find that the likelihood of computing a regression coefficient in error by 40 standard errors or more is essentially 0. Therefore, we can conclude with an almost 100% certainty that the regression coefficient is significantly different from 0 in this case.

Although the statistical theory behind standard errors may seem complicated, its application is straightforward, particularly when a computer program is used. Table 8-2, for instance, shows the computer output for the mail-order example using ISP (Interactive Statistical Program). The Y (denoting Yes) next to the regression coefficient b (POUNDSM) means that the real value of b is significantly different from 0. That is, increases or decreases in the weight of mail influence the number of orders not by chance but in a consistent, statistical manner. Furthermore, under the heading P Value, the value of 0 signifies the probability that the value of the computed coefficient is equal to 0.

The same computations can be found concerning the value of a. In this case the standard error of a is equal to 5.35, which means that the computed t-test for $a = -4.4$ is -0.82. Furthermore, the P Value is equal to 0.42, which indicates that the likelihood that the real value of a will be equal to 0 is high.

Table 8-3 shows the same computer output as Table 8-2, except that the value of a has been forced to become 0. The regression equation can be seen to be

$$\hat{Y} = bX$$

$$\hat{Y} = 3.77X.$$

The final point that we deal with here is the significance of an individual forecast. We would like to know, once we have found the values of a and b in our regression equation and have substituted a value of X into that equation, how confident we can be that the true value of Y will be around the value computed with $\hat{Y} = a + bX$. That is, we would like to have a confidence interval around this computed value of $\hat{Y}$.

Table 8-2 Regression Results of the Mail-Order Example Using ISP

Variable	Coefficient	Standard Error	t Test	Signif	P Value
a	−4.374343	5.348527	−0.818	N	0.420
POUNDSM	3.84204	0.09787115	39.256	Y	0.000

Critical t value from table ($\alpha = 0.05$) = 2.048.
$R^2 = 0.982$; R^2(adjusted) = 0.982; $R = 0.991$;*
F^2 test = 1541.04; standard deviation of regression = 14.65156;
observations = 30; degrees of freedom for numerator = 1, for denominator = 28.
F value from table ($\alpha = 0.05$) = 4.17.

*The symbol for the square of the correlation coefficient above is R^2 while in equations (8-6) and (8-7) we use r^2. This is so because the computer does not distinguish between simple regression where r^2 is used and multiple regression (see next chapter) where R^2 is used.

Table 8-3 Regression Results of the Mail-Order Example Using ISP

Variable	Coefficient	Standard Error	t Test	Signif	P Value
a	0.0				
POUNDSM	3.772725	0.04953009	76.170	Y	0.000

Critical t value from the table ($\alpha = 0.05$) = 2.048.
$R^2 = 0.982$; R^2(adjusted) = 0.981; $R = 0.991$;
F Test = 1504.44; standard deviation of regression = 14.82550;
observations = 30; degrees of freedom for numerator = 1, for denominator = 28.
F value from table ($\alpha = 0.05$) = 4.17.

The basis for establishing the confidence interval of a specific forecast value is the standard error of forecast SE_f, which can be computed using

$$\text{SE}_f = \left(\sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{n - 2}}\right)\left(\sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}}\right). \tag{8-5}$$

As with the standard error of the regression coefficient SE_b, it can be seen that the first factor in Equation (8-5) is simply the standard deviation for the errors e_i (adjusted for degrees of freedom), or the standard deviation of regression. The second factor in expression (8-5) is an adjustment for how far we are from $\bar{X}$ in making our forecast (that is, the difference between X_f and $\bar{X}$), and the number of observations n used in determining the regression equation. Note that if the values of X_i are not close to the mean $\bar{X}$ and n is large, the second factor in Equation (8-5) is approximately equal to 1 and the standard error of forecast, SE_f is simply the standard deviation of the regression. If, however, n is small, or if the values of X_i are close to the mean $\bar{X}$, or if the value being forecast X_f is far from $\bar{X}$, then the second factor in Equation (8-5) will be greater than 1, and the standard error of forecast will be larger than the standard deviation.

In the forecast we made earlier, the value of X was 45 pounds. Thus we can substitute this value for X_f in Equation (8-5) and obtain

$$\text{SE}_f = 14.6\sqrt{\left[1 + 0.03 + \frac{(45 - 47.3)^2}{2446}\right]} = 14.6(1.03) = 14.9.$$

In this example there is little adjustment in the standard deviation value of 14.6 because X is close to the mean value $\bar{X}$. If we want to establish a confidence interval for our estimate of Y (where we had $\hat{Y} = -4.4 + 3.8(45) = 166.6$), this interval will be $166.6 \pm 2(14.9)$. Thus, we are 95% certain that the true value of Y (that is, orders to be filled) will be between 136.8 and 196.4 when we receive 45 pounds of mail. Note that we use a value of 2 to multiply the standard error because we assume a 95% confidence interval, in which case the t value required to multiply the standard error is close to 2 (the exact value is 2.045). If we had wanted a 99.8% confidence interval, we should have multiplied the value of the standard error by about 3. Alternatively, if we had wanted a 68% confidence interval, we should have multiplied the standard error by the approximate value of 1. (See Figure 8-5 for confidence intervals of 68%, 95%, and 99.8%.)

The fact that the confidence interval surrounding a forecast varies according to the distance we are from X is shown in Figure 8-5. As a practical matter, the usable range of X values—where the regression can be applied confidently—depends on the amount of data used, the spread of X values around their mean, and how well the regression model captures the variation in these data.

From the mail-order example the amount of information that can be obtained from making these tests of significance should be evident. In this instance we found that the regression coefficient b was significantly different from 0 and that its true value would lie between 3.5 and 4.1. In addition, we found that in estimating the number of orders included in 45 pounds of mail we could be 95% sure that the orders would be between 136.8 and 196.4. Since this is a wide range, it would require management to contemplate the uncertainty surrounding the most likely forecast of 166.6 and take steps to deal with situations where the orders to be filled are around the two extreme ranges.

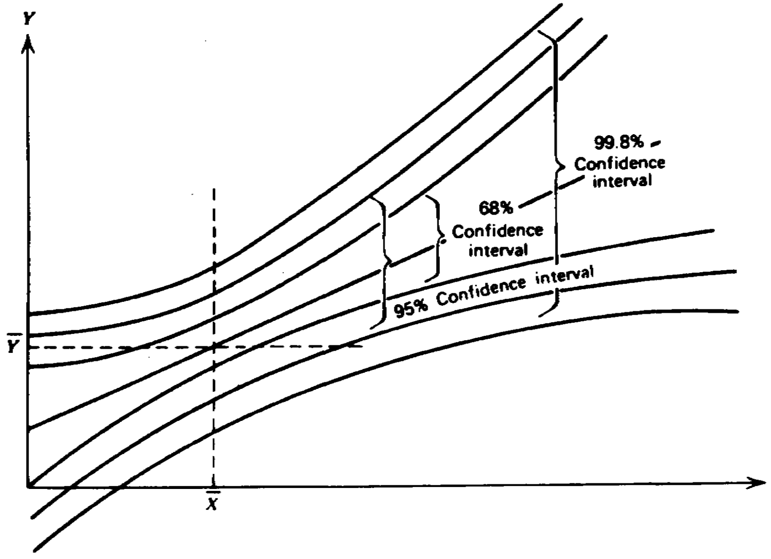

Figure 8-5 Ranges of Confidence of Individual Forecasts.

In concluding this section on signifiance, we want to emphasize one point: the role that the sample size plays in this area. (The sample size is simply the number of observations used in determining the regression line.) As a sample size becomes larger, the width of the confidence interval becomes smaller, as can be seen in Equation (8-5). Thus if we had 100 observations in the mail-order example, the confidence interval on a specific forecast (for example, for 45 pounds of mail) would usually be narrower than it was when we had only 30 observations. Any time the manager can obtain additional observations before computing a regression line, he or she should do so.

One of the advantages of simple regression analysis is that once the relationship $\hat{Y} = a + bX$ has been determined, it can be used to make any number of forecasts simply by inserting the value of X for which a forecast of Y is desired. One caution: the basic relationship should be assessed periodically. If the manager has some reason to believe that a change may have taken place, it will be necessary to collect a new set of data and recompute the values of a and b.

SIMPLE CORRELATION

The assumption made in the three preceding sections that dealt with simple regression is that one variable is dependent on another. Often two variables may be related, although it is not appropriate to say that the value of one of the variables depends on or is influenced by the value of the other. In such a situation, the correlation between the two variables can be found. The coefficient of correlation r is a relative measure of the degree of relationship that may exist between two variables. This coefficient can vary from 0 (which indicates no correlation) to ± 1 (which indicates perfect correlation). When the correlation coefficient is greater than 0, the two variables are said to be positively correlated; when it is less than 0, they are said to be negatively correlated. For simple regression, the sign of the correlation coefficient is always the same as the sign of the regression coefficient b.

The correlation coefficient can be found using the formula

$$r = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{\sqrt{n\Sigma X_i^2 - (\Sigma X_i)^2}\sqrt{n\Sigma Y_i^2 - (\Sigma Y_i)^2}} \qquad (8\text{-}6)$$

To compute the correlation coefficient we need to substitute the various sums in Equation (8-6) by their values as shown in Table 8-1. The value of r, therefore, is

$$r = \frac{30(337{,}987) - 1419.6(5322)}{\sqrt{30(89{,}592.9)(1419.6)^2}\sqrt{30(1{,}281{,}220) - (5322)^2}}$$

$$= \frac{10{,}139{,}610 - 7{,}555{,}111.2}{\sqrt{672{,}522.84}\sqrt{10{,}112{,}916}} = \frac{2{,}584{,}498.8}{2{,}607{,}904.7} = 0.991.$$

Since the value of r equals 0.991 (very close to 1), there is a very strong positive correlation (association, relationship) between the weight of mail received and the number of orders to be filled. As one goes up, so does the other.

The advantage of a correlation coefficient is that it is a relative measure and, as noted, it varies between 0 and ± 1. Thus it is easy to recognize the strength of a relationship between any two variables. A correlation coefficient close to 1 means a very strong relationship. One that is close to 0 indicates a weak relationship or none at all. On the other hand, if we want to measure how much X influences Y, then the value of the regression coefficient b is needed. In this sense the correlation coefficient r and the regression coefficient b provide us with complementary information. This is why all simple regression computer programs compute both b and r.

Another useful statistical measure in regression analysis is r^2. The formula for finding r^2 is given by

$$r^2 = \frac{\text{explained variation*}}{\text{total variation}} = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}. \qquad (8\text{-}7)$$

r^2 varies between 0 and 1 and is a measure of the goodness of fit. It tells us [see Equation (8-6)] the percentage of the total variation explained by the regression line. In a perfect regression, where the errors between actual Y_i and estimated $\hat{Y}_i$ are 0, the value of r^2 will be equal to 1. If r^2 is close to 0, the regression equation $\hat{Y}_i = a + bX_i$ does not explain much more than if the mean ($\hat{Y} = \bar{Y}$) is used as an alternative to the regression line.

Table 8-4 shows the total variations between the production costs Y_i used previously and their mean $\bar{Y} = 12.5$. If each of these four distances $Y_i - \bar{Y}$ is squared and the values are summed, the sum will provide us with the denominator of Equation (8-7). The calculations can be seen in Table 8-4. the explained variations (deviations) can also be seen in Table 8-4 as the difference between $\hat{Y}_i$ and $\bar{Y}$. If these variations are squared and the values summed, the sum will provide us with the numerator of equation (8-7). The ratio of the two squared sums is the value of r^2. Thus,

$$r^2 = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2} = \frac{23.204}{41} = 0.566.$$

This means that the regression line $\hat{Y} = 1.254 + 7.17X$ explains 56.6% of the total variation in costs by the variation in the amount of units produced. Obviously, with only four data points this statement may not mean much. However, if a larger amount of data ia available, r^2 will tell us what percentage of the total variation in Y is explained by variations in X.

Let us now compute r^2, or the coefficient of determination, as r^2 is also known, for the mail-order example of the preceding sections. Again we will use Equation (8-7),

$$r^2 = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2} = \frac{808}{823} = 0.982.$$

The appropriate interpretation here is that for the 30 sample observations that were used in fitting the regression line, 98.2% of the variation from the mean $\bar{Y}$ was explained by that regression line $\hat{Y} = -4.4 + 3.8X$. Thus, almost all changes in the number of orders received every day can be explained by the weight of the mail received that day. The remaining 1.8%, which cannot be explained, is caused by random factors (for examples, some people might use heavy envelopes and paper) or by other variables in addition to weight (such as requests for information as a result of an advertising campaign).

In simple regression the square root of r^2 is equal to the simple correlation coefficient. This can be verified in the case of the mail-order example. The square root of 0.982 is equal to 0.991, the correlation coefficient computed using Equation (8-6).

Another useful statistical measure used in regression analysis is the F ratio. This is given by

*It is called *explained* because it improves (that is, reduces) the variation or error over the alternative to the regression line, which is to use the mean value $\bar{Y}$, as a way of forecasting. The amount of improvement, that is, $\hat{Y}_i - \bar{Y}$, is then the *explained variation* of regression, over and above that of the mean.

Table 8-4 Total and Explained Variations for the Production Cost Example

Observation	Actual Cost Y	Mean Cost $\bar{Y}$	Value Estimated by Regression Line $\hat{Y}$	Total Variation $(Y_i - \bar{Y})^2$	Explained Variation $(\hat{Y}_i - \bar{Y})^2$
1	8	12.5	10.93	20.25	1.465
2	11	12.5	9.68	2.25	7.952
2	16	12.5	13.44	12.25	0.884
4	15	12.5	15.95	6.25	11.903
				$\Sigma(Y_i - \bar{Y})^2 = 41$	$\Sigma(\bar{Y}_i - \bar{Y})^2 = 23.204$

$$F = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2/(k-1)}{\Sigma(Y_i - \hat{Y}_i)^2/(n-k)} \tag{8-8}$$

where n = sample size (number of observations)
k = number of variables ($k = 2$ for simple regression).

It should be readily apparent from Equation (8-8) that the value of the F statistic is similar to the r^2 calculation—except that it uses the "unexplained" variance in the denominator and it is dependent on the value of the sample size n. As n gets larger, the number in the denominator in the F statistic gets smaller and the F statistic increases in value.

The F test for the mail-order example is

$$F = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2/(k-1)}{\Sigma(Y_i - \hat{Y}_i)^2/(n-k)}$$

$$= \frac{808/1}{14.63/28}$$

$$= 1540.$$

This value of F must be compared with the appropriate entry in a statistical table of F-test values to determine its significance for a given confidence level. However, because most computer programs provide this value, it is not necessary to know how to use such a table. What is necessary is that the F value computed using Equation (8-8) be greater than the corresponding value from the table, which in our case (using a 95% confidence interval) is equal to 4.18. In the mail-order example, where the computed F is 1540, which is much larger than 4.18, we conclude that the relationship between the number of orders and the weight of mail is statistically significant (that is, it is not the result of chance).

Let us now determine the significance of the relationship we computed for the simple example of four data points. The reader will recall that we found that $r^2 = 0.566$, or 56.6% of the variation was explained. Using equation (8-8), we can compute the F statistic for this example as

$$F = \frac{23.2/1}{17.8/2} = 2.61.$$

To determine whether the relationship $\hat{Y} = -4.4 + 3.8X$ is significantly different from 0, we have to compare this F value of 2.61 with the corresponding value from the F table. At the 95% level and for four observations such a value is 6.61. Thus in this example, even though $r^2 = 0.566$, we *cannot* say with any confidence that the regression line is significantly different from 0. That is, the relationship between units produced and cost may well be the result of chance, because the computed F of 2.61 is smaller than the corresponding value of 6.61 from the F table.

The importance of computing the overall significance of a regression equation should be emphasized. Only when the manager can say that it is significant does it make sense to use the regression equation to forecast. That is, if the sample size is so small or the relationship so weak that it is not significant, even though the coefficient of determination may be close to 1, the manager should not base a forecast on those data and the corresponding regression line.

THE REGRESSION EQUATION AS A MODEL

The regression equation $\hat{Y} = a + bX$, like any other form of equation, can be thought of as an abstract model that represents some aspect of reality. When, for example, we say that Y represents sales and X represents time, what we are actually doing is making an abstract model. We try to simplify reality and represent it in terms of the interaction of two factors only. As managers are well aware, however, this is a gross simplification; reality is much more complex. Sales are not influenced by time alone, but by a myriad of other factors such as GNP, prices, competitors' actions, transportation costs, production costs, advertising, government policies, or even the illness of a salesperson. Then how can we ignore them?

In any modeling effort there is a choice: we can either construct a simple model that may not completely duplicate reality, or we can build a complex model that is more accurate, but also requires a large amount of effort and resources to be developed and manipulated. Even if the most sophisticated model could be developed, there would still be some part of reality that could not be explained by the model. The number of factors in real-life phenomena is infinite.

To capture the fact that a part of the real process cannot and will not be explained by a regression model, we can use the term u to denote the variations unexplained by the model. This term is often called the *disturbance term*, or *white noise*, and it plays an important role in most forecasting methods.

Thus to be precise, our simple regression equation is not $\hat{Y} = a + bX$, but $Y = a + bX + u$, even though the term u is seldom needed for calculation or any other practical purpose. Its theoretical meaning is that the forecast can vary from the estimated value $\hat{Y} = a + bX$ by an amount u, the error, which we can estimate in probabilistic terms, as we learned in the preceding sections of this chapter. It becomes obvious, however, that the magnitude of the error u will vary from model to model. Theoretically the more variables we introduce, the smaller the range of values taken on by u. There is a limit, however, to the number of variables we can employ, since they introduce more complexity and higher cost. Thus we want to introduce the smallest number of variables (the principle of parsimony) and at the same time achieve a range of values for u as small as possible.

For the regression equation $Y = a + bX + u$ to be statistically correct, u must have the following properties:

1. The mean value of u must be equal to 0, because many factors that influence Y are not included in the regression equation. However, their influence is of opposite directions, so they tend to offset one another on average.
2. The error term u must be a random variable. At any time period, some of the factors not included in the equation will influence Y more than others. This may result in a positive or negative u. However, as long as the individual values of u at each time period are random (not the result of any systematic pattern), their effect on the estimated value of Y can be determined probabilistically. Basically, we want the errors to be independent of one another.
3. The disturbance term u must be *normally distributed.* (A *normal distribution* is commonly referred to as a bell-shaped curve dispersed around the mean value of 0.) This is a consequence of the large number of factors that influence Y that are not included in the regression equation. In this case, as in many others, it is more probable that extreme variations will cancel themselves out and thus be observed only infrequently, whereas in the majority of cases the errors will be clustered around the mean value. Such a pattern results in a normal distribution.
4. The variance of u must be constant. This means that the error term must neither increase nor decrease within the entire range of observations.

Violations in these properties of u can result in serious trouble, since the complete regression model

$$Y = a + bX + u$$

will no longer be correct. As we shall see in Chapter 9, however, there are ways of minimizing the likelihood that any of these properties will be violated in a specific application.

From now on, as we have done before, we imply $Y = a + bX + u$ when we write a regression equation, even though u will not be included. Thus we use the term $\hat{Y}$. That is

$$Y = a + bX + u$$

but we will use

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

The same is true for multiple regression, where

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n + u$$

but we will use

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n.$$

Finally, when we use the sample data to estimate $\hat{Y}$, the difference between the estimate $\hat{Y}$ and the actual Y is the regression, or residual error. We have denoted such a value by e. The residual error e is different from u, because it is the result of a specific and limited sample, whereas u would exist if it were possible to have an extremely large sample or even to include all the data. Like u, e must fulfill the four properties mentioned above.

In the next chapter the concept of regression will be extended to include more than one independent variable, and methods of testing the assumptions concerning the four properties required of u or e will be presented.

SELECTED REFERENCES FOR FURTHER STUDY

Chatterjee, S., and B. Price, 1977. *Regression Analysis by Example*, Wiley, New York.

Draper, N., and H. Smith, 1981. *Applied Regression Analysis*, 2nd ed., Wiley, New York.

Intrilligator, M. D., 1978. *Econometric Methods, Techniques, and Applications*, Prentice-Hall, Englewood Cliffs, NJ.

Johnson, J., 1972. *Econometric Methods*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.

Makridakis, S., S. C. Wheelwright, and V. E. McGee, 1989. *Forecasting: Methods and Applications*, 3rd ed., Wiley, New York, chaps. 5 and 6.

Pindyck, R. S., and D. L. Rubenfeld, 1976. *Econometric Models and Economic Forecasts*, McGraw-Hill, New York.

Wetherill, G. B., 1986. *Regression Analysis with Applications*, Chapman and Hall, London.

Wonnacott, H., and R. J. Wonnacott, 1986. *Regression: A Second Course on Statistics*, Krieger, Melbourne.