SAGE **UNIVERSITY** PAPERS

Series: Quantitative Applications in the Social Sciences

RS
ries, please write

Series / Number 07-057

UNDERSTANDING REGRESSION ANALYSIS

An Introductory Guide

LARRY D. SCHROEDER
Syracuse University

DAVID L. SJOQUIST
Georgia State University

PAULA E. STEPHAN
Georgia State University

SAGE PUBLICATIONS
The International Professional Publishers
Newbury Park London New Delhi

For information address:

SAGE Publications, Inc.
2455 Teller Road
Newbury Park, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

International Standard Book Number 0-8039-2758-4

Library of Congress Catalog Card No. 85-063790

95 96 97 98 99 20 19 18 17 16 15

When citing a university paper, please use the proper form. Remember to cite the current Sage University Paper series title and include the paper number. One of the following formats can be adapted (depending on the style manual used):

(1) SCHROEDER, LARRY D , SJOQUIST, DAVID L., and STEPHAN, PAULA E. (1986) Understanding Regression Analysis. An Introductory Guide. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-057. Newbury Park, CA: Sage.

OR

(2) Schroeder, Larry D., Sjoquist, David L., & Stephan, Paula E. (1986). *Understanding regression analysis: An introductory guide* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-057). Newbury Park, CA: Sage.

CONTENTS

To our children,

Leanne
Nathan
Jennifer
David

Series Editor's Introduction

Researchers in the social sciences, business, policy studies and other areas rely heavily on the use of linear regression analysis. The frequency with which the technique is employed is demonstrated by a review of articles in professional journals such as the *American Economic Review, Journal of Finance, American Political Science Review, Journal of Policy Analysis and Management, Journal of Marketing, Journal of Educational Research,* and *American Sociological Review.* The use of linear regression is so common because this research tool adds considerably to the understanding of economic, political, and social phenomena.

Frequently, instructors would like to supplement their courses with materials, such as articles from professional journals, that use regression analysis. To students unfamiliar with regression, however, research based on the technique can be incomprehensible. For those who have yet to take a statistics course, this book is intended to provide the background needed to understand much of the empirical work relying on linear regression analysis. The book provides a heuristic explanation of the basic procedures and terms used in regression analysis. Written at the most elementary level and assuming only a minimal mathematics background, the book focuses on the intuitive and verbal interpretation of regression coefficients, associated statistics, and hypothesis tests. Other terminology often encountered in today's literature is also explained, including standardized regression coefficients, dummy variables, interaction terms, and transformations. Brief discussions of some of the major problems encountered in regression analysis are also presented.

The book can be used as a supplementary text in a variety of courses in numerous fields. Examples given in the text encompass the fields of demography, economics, education, finance, marketing, policy analysis, political science, public administration, and sociology. Instructors in any of these areas are likely to find the text useful.

The authors do not intend for this book to serve as a substitute for a course or textbook in statistics. It is not designed to teach the use of

regression analysis, but rather to fill the void that exists when the student encounters empirical papers before taking a statistics course. On the other hand, the level of exposition makes the volume suitable as an introductory supplement in applied statistics courses where students are encountering linear regression for the first time.

This book is an outgrowth of material previously prepared by the authors for students in intermediate economics courses who did not have a background in statistics. An earlier, more limited version of the book was published by General Learning Press under the title, *Interpreting Linear Regression Analysis: A Heuristic Approach.* This version has been expanded to encompass the many other disciplines that use regression analysis.

—*Richard G. Niemi*
Series Co-Editor

Acknowledgments

We are especially grateful to Theodore C. Boyden for providing the encouragement to undertake this project. Special thanks go to the following individuals who provided suggestions for examples and clarified various arguments: Kenneth Bernhardt, Michael Binford, Libby Dalton, Benoit Deschamps, Louis Ederington, Kirk Elifson, Charles Jeret, Ralph LaRossa, Taylor Little, Jr., Dileep Mehta, Donald Reitzes, and Frank Whittington.

We also want to thank Esther Gray, Bee Hutchins, Marian Mealing, Billie Shook, and Carla Thomas for their expert typing, David Amis for help with the illustrations, and Richard G. Niemi for his support.

UNDERSTANDING REGRESSION ANALYSIS

LARRY D. SCHROEDER
Syracuse University
DAVID L. SJOQUIST
Georgia State University
PAULA E. STEPHAN
Georgia State University

1. LINEAR REGRESSION

Hypothesized Relationships

The two statements, "The more a political candidate spends on advertising, the larger the percentage of the vote he will receive" and "Mary is taller than Jane," express different types of relationships. The first statement implies that the percentage of the vote that a candidate receives is a function of, or is caused by, the amount of advertising, while in the second statement no causality is implied. More precisely, the former expresses a *causal* or *functional* relationship while the latter does not. A functional relationship is thus a statement (often in the form of an equation) of how one variable, called the *dependent* variable, depends on one or more other variables, called *independent* variables. In the example, the share of the vote a candidate receives is dependent on (is a function of) the amount of advertising, which is independent of the percentage of the vote received. Another independent variable that might be included is the number of prior years in office, in which case the functional relationship would be stated as, "The candidate's share of the vote depends on the amount of advertising as well as the candidate's prior years in office."

Other examples of functional relationships are: (1) "If he allows his hair to grow longer, he will become stronger," (2) "If she studies more, her grades will improve," and (3) "If the price of oranges increases, individuals will purchase fewer oranges."

One of the activities of researchers is testing the validity or falsity of hypothesized functional relationships, called *hypotheses*[1] or *theories.* This volume discusses one tool used in testing hypotheses—linear regression.

Linear regression analysis is applicable to a vast array of subject matter. Consider the following situations in which regression analysis has been employed: a study of the effect of shelf space devoted to a particular product on the sales of that product (Curhun, 1972); a study of the effect of the size of the dividend paid by a corporation on the value of the corporation's stock (Durand, 1959); a study of the effect of school quality on academic achievement (Coleman et al., 1966); a study of the effect of age on the probability that an individual or family will move (Polachek and Horvath, 1977).

All of these examples are cases in which the application of regression analysis was useful, although the application was not always as straightforward as the example to which we now turn.

A Numerical Example

To facilitate the discussion of *linear regression analysis*, the following food consumption example will be referred to throughout the book. Suppose one were asked to investigate by how much a typical family's food expenditure increases as a result of an increase in its income. While most would agree that there is a relationship between the amount spent on food and income, the example is in fact an investigation of an economic theory. The theory suggests that the consumption of food is a function of family income;[2] that is, $C = f(I)$, read "C is a function of I", where C (the dependent variable) refers to the consumption of food and I (the independent variable) refers to income. Throughout the book we will refer to the theory that C increases as I increases as the hypothesis.

The investigation of the relationship between C and I allows for both testing the theory that C increases as a result of increases in I and obtaining an estimate of how much food consumption changes as income changes. One can therefore consider the investigation as an analysis of two related questions: (1) Does spending on food increase when a family's income increases? (2) By how much does spending on

food change when income increases or decreases? As will be seen in Chapter 3, these questions cannot be answered with certainty. However, since the material in this section can be more easily understood by assuming that answers to these questions can be provided with certainty, we shall proceed initially under this assumption.

At least two strategies for analyzing these questions are feasible. One can observe various families over time and note how their consumption of food changes as their income changes, or one can observe income and food consumption differences among several families and note how differences in food consumption are related to differences in income. We have adopted the latter approach, employing the hypothetical data given in columns 1 and 2 of Table 1, which represent annual income and food consumption information from a sample of 50 families in the United States for one year. Assume that this sample was chosen randomly from the population of all families in the United States.[3] The associated levels of these two variables have been plotted as the 50 points in Figure 1.

Casual observation of the points in Figure 1 suggests that C increases as I increases. However, the magnitude by which C changes as I changes for the 50 families is not obvious. For this reason the presentation of data in tabular or graphical form is not by itself a particularly useful format from which to draw inferences. These formats are even less desirable as the number of observations and variables increases. Thus we seek a means of summarizing or organizing the data in a more useful manner.

Any functional relationship can be most conveniently expressed as a mathematical equation. If one can determine the equation for the relationship between C and I, one can use this equation as a means of summarizing the data. Since an equation is defined by its form and the values of its parameters,[4] the investigation of the relationship between C and I entails learning something from the data about the form and parameters of the equation.

The economic theory that suggests that C is a function of I does not indicate the form of the relationship between C and I. That is, it is not known whether the equation is of a linear or some other, more complex form. In some problems the general form of the equation is suggested by the theory, but since this is not so in the food expenditure problem, it is necessary to specify a particular form. We shall assume that the form of the equation for our problem is that of a straight line, which is the simplest and most commonly used functional form.[5]

TABLE 1
Food Consumption, Family Income, and Family Size Data

(1) *Food Consumption*	*(2)* *Income*	*(3)* *Family Size*	*(4)* *Live on Farm*
$ 723.52	$ 8,246	1	No
780.70	8,742	4	No
990.74	9,048	6	No
1,634.98	10,584	7	No
1,189 40	10,626	2	No
1,295.64	10,984	2	No
1,025.52	11,822	1	No
1,792.18	12,532	2	No
1,328.00	12,952	5	No
780.06	13,220	2	Yes
1,366.14	13,386	6	No
2,950.72	13,746	8	No
1,273.34	13,946	2	No
1,953 58	14,206	2	No
866.62	14,388	1	No
2,125.30	14,622	4	No
2,372.00	15,032	2	No
2,477 34	15,172	5	No
1,148.24	16,284	1	No
2,108 14	16,664	3	No
1,810.96	17,124	2	No
1,776.58	17,302	2	No
2,295.04	18,254	3	No
877.52	18,908	1	Yes
1,284.00	18,922	2	No
1,502.94	19,330	2	Yes
1,939.00	20,108	3	No
2,443.06	20,600	3	No
2,003.44	21,238	4	No
1,682.36	22,120	2	No
2,308.16	22,452	7	No
1,472.44	23,288	2	No
2,534.66	23,316	4	No
2,194.76	23,588	2	No
1,638.26	23,708	3	No
2,612.00	23,830	6	No
2,328.96	23,908	2	No
1,666.90	24,216	3	No
2,560.22	25,422	1	No
3,103.54	25,504	9	No
2,819.06	26,286	5	No
975.10	26,590	2	No

(continued)

TABLE 1 (Continued)

(1) *Food Consumption*	*(2)* *Income*	*(3)* *Family Size*	*(4)* *Live on Farm*
2,122 52	26,852	1	No
1,068.38	27,146	3	Yes
2,253.46	27,936	6	No
2,763.40	28,556	5	No
1,904.66	28,874	3	No
2,111.50	29,450	4	No
3,211.64	29,624	1	No
2,665.78	29,690	4	No

SOURCE: Hypothetical data.

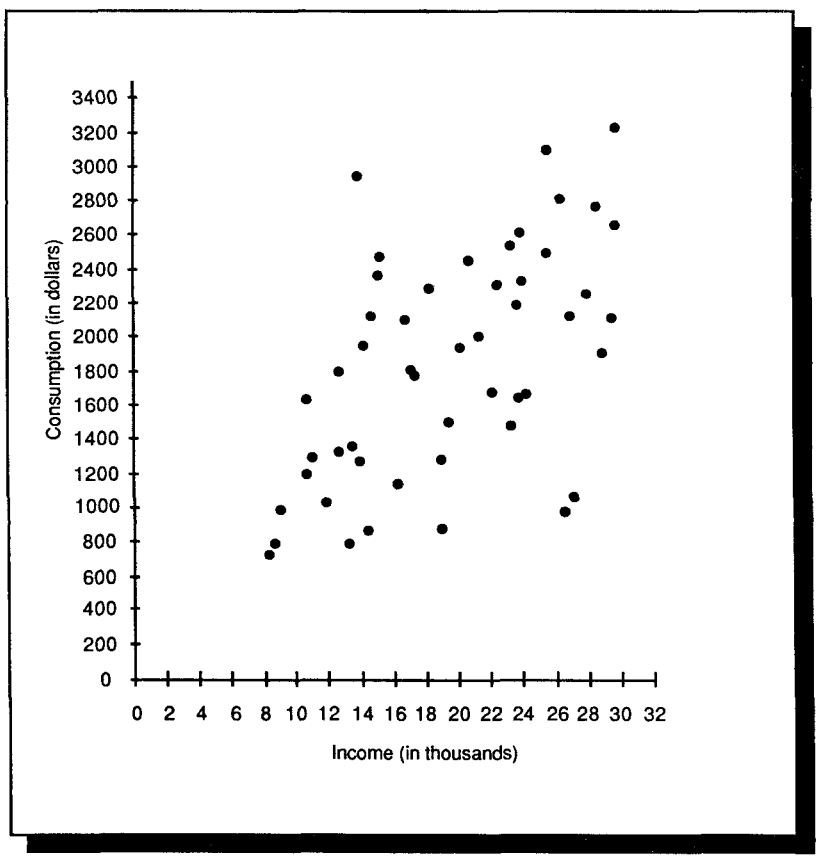

Figure 1: Scatter Diagram of Family Income and Food Consumption

Given this assumption, one can express the functional relationship that exists between C and I for all U.S. families as

$$C = \alpha + \beta I \qquad [1]$$

where α (the Greek letter alpha) and β (the Greek letter beta) are the unknown parameters assumed to hold for the population of U.S. families and are referred to as the *population parameters.*[6] (See also Figure 2.)

Given the assumption that the form of the equation of the possible relationship between C and I can be represented by a straight line, what

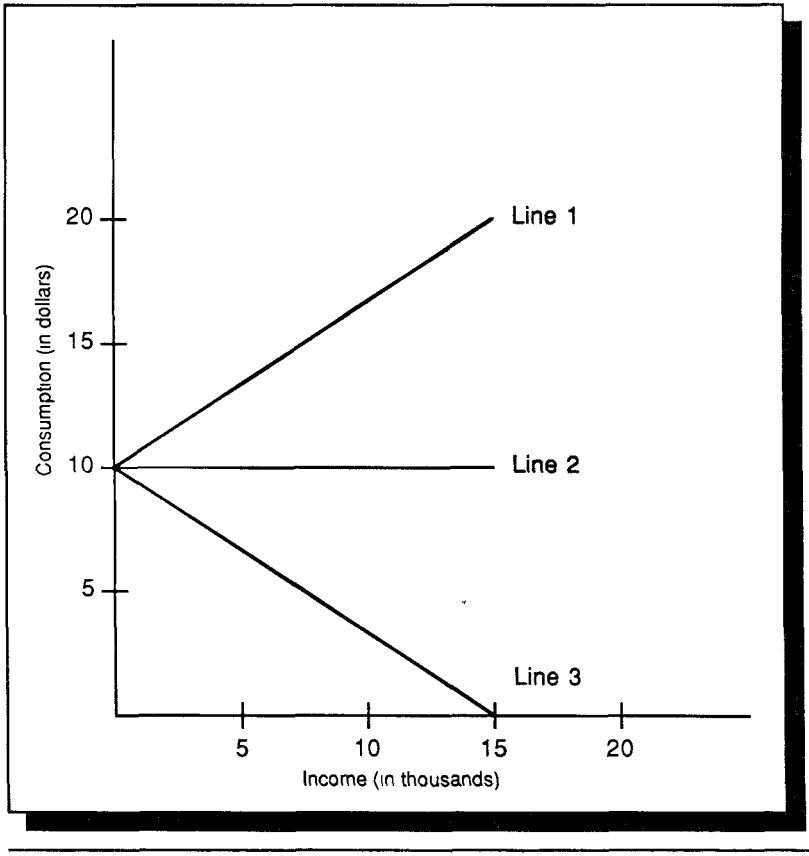

Figure 2: Illustration of Different Slopes

remains is to estimate the values of the population parameters of the equation using our sample of 50 families. The two questions posed earlier refer to the value of the slope—that is, the value of β. The first question asks whether β is greater than zero, while the second asks the value of β. By obtaining an estimate of the value of β, a statement can be made as to the effect of changes in income on the level of food consumption for the 50 families in our sample. Further, from this estimate of β inferences can be drawn about the behavior of all families in the population.

Before proceeding, it is important to note the following. The actual or "true" form of the relationship between I and C is not known. We have simply assumed a particular form for the relationship in order to summarize the data in Figure 1. Further, we do not know the values of the population parameters of the assumed linear relationship between C and I. The task is to obtain estimates of the values of α and β. We will denote these estimates as a and b.

Estimating a Linear Relationship

The question that may come to mind at this point is, how can it be stated that income and food consumption are related by a precise linear equation when the data points in Figure 1 clearly do not lie on a straight line? The answer comprises three parts. First, the equation is only a summary of the data points and does not imply that C and I are related in precisely this manner. Second, the hypothesis is based on the implicit assumption that only income and consumption differ between these families. However, other things, such as family size and tastes, are not likely to be the same and no doubt affect the amount of food consumed. Third, there is randomness in people's behavior; that is, an individual or family, for no apparent reason, may buy more or less food than some other family that appears to be in exactly the same situation with regard to income, taste, and the like. Thus one would not expect the data points to lie consistently on a straight line, even if the line did represent the average response to changes in income.

As noted previously, from the data points in Figure 1 it is not obvious how much C increases as I increases; that is, it is uncertain what the position of the line summarizing the data points should be. To see this, consider the two solid lines that have been arbitrarily drawn through the points in Figure 3. Line 1 has the equation $C = 1{,}000 + 0.01I$, and line 2 has the equation $C = 200 + 0.10I$. Which of these two lines is the better

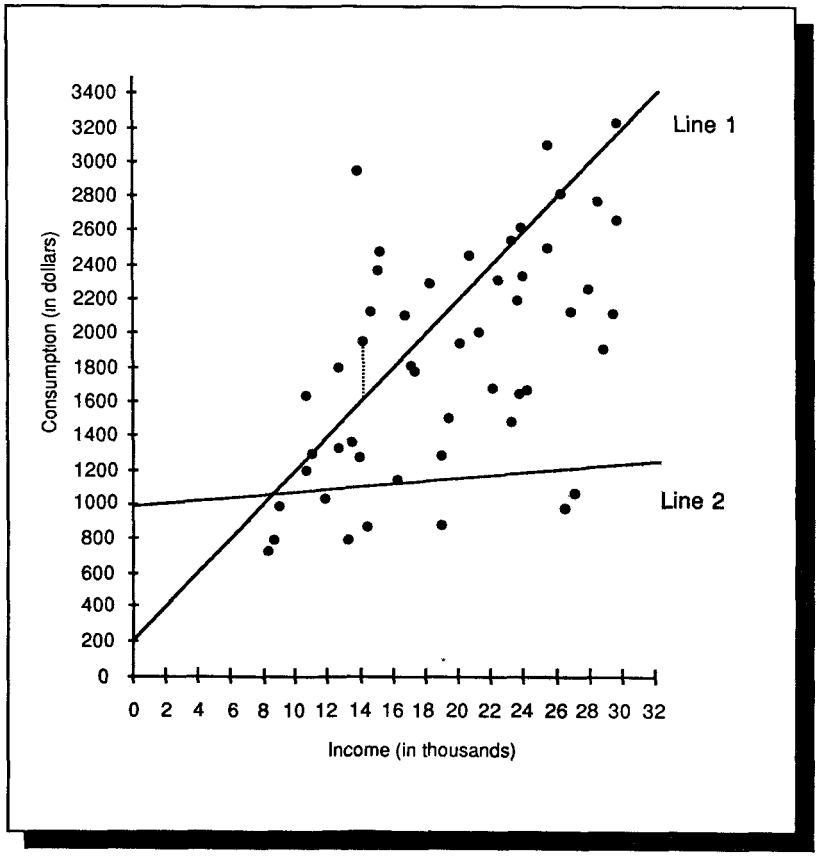

Figure 3: Two Possible Summaries of the Income-Consumption Relationship

estimate of how food consumption changes as income changes? This is the same as asking which of the two equations is better at summarizing the relationship between C and I found in Table 1. More generally, which line among all the straight lines that it is possible to draw in Figure 3 is the "best" in terms of summarizing the relationship between C and I? Regression analysis, in essence, provides a procedure for determining the *regression line,* which is the best straight line (or linear) approximation of the relationship between C and I. This procedure is equivalent to finding particular values for the slope and intercept.

An intuitive idea of what is meant by the process of finding a linear approximation of the relationship between the independent and depen-

dent variables can be obtained by taking a string or pencil and trying to "fit" the points in Figure 1. Move the string up or down, or rotate it until it takes on the general tendency of the points in the graph.

What property should this line possess? If asked to select which of the two solid lines in Figure 3 is better at summarizing (estimating) the relationship between income and food consumption, one would undoubtedly choose line 1 because it is "closer" to the points than line 2. (This is not to imply that line 1 is the regression line.)

Closeness or distance can be measured in different ways. Two possible measures are the vertical or horizontal distance between the observed points and a line. In the normal case, where the dependent variable is plotted along the vertical axis, distance is measured vertically as the differences between the observed points and the line. This is shown in Figure 3, where the vertical dotted line drawn from the data point to line 1 measures the distance between the observed data point and the line. In this case distance is measured in dollars of consumption, not in feet or inches. The choice of the vertical distance stems from the theory stating that the value of C depends on the value of I. Thus, for a particular value of income, it is desired that the regression line be chosen so as to predict a value of food consumption that is as close as possible to the value of food consumption observed at that income level.

The regression line cannot minimize the distance for all points simultaneously. In Figure 3 it can be seen that some points are closer to line 1 while others are closer to line 2. Thus a means of averaging or summing up all these distances is needed to obtain the best fitting line.

Although several methods exist for summing these distances, the most common method in regression analysis is to find the sum of the squared values of the vertical distances. This is expressed as

$$\sum_{i=1}^{N} (C_i - \hat{C}_i)^2,$$

where C_i is the value of C that would be estimated by the regression line and is read "C hat sub i."[7]

Least Squares Regression

In the most common form of regression analysis, the line that is chosen is the one that minimizes

$$\sum_{i=1}^{N} (C_i - \hat{C}_i)^2,$$

which is called the *sum of the squared errors*, frequently denoted SSE. For each observation, the distance between the observed and the predicted level of consumption can be thought of as an error, since the observed level of consumption is not likely to be predicted exactly but is missed by some amount $(C_i - \hat{C}_i)$. This error may be due, for example, to randomness in behavior or other factors such as differences in family size. Because the squares of the errors are minimized, the term *least squares regression analysis* is used.

The reason for selecting the sum of the squared errors lies in statistical theory that is beyond the scope of this book. However, an intuitive rationale for its selection can be presented. If the errors were not squared, distances above the line would be canceled by distances below the line. Thus it would be possible to have several lines, all of which minimized the sum of the nonsquared errors.[8] It is implicit that closeness is good, while remoteness is bad. It can also be argued that the undesirability or remoteness increases more than in proportion to the error. Thus, for example, an error of four dollars is considered more than twice as bad as an error of two dollars. One way of taking this into account is to weight larger errors more than smaller errors, so that in the process of minimizing it is more important to reduce larger errors. Squaring errors is one means of weighting them.

Let a and b represent the estimated values of α and β for the still unknown regression line. Thus $\hat{C}_i$ can be expressed as $\hat{C}_i = a + bI_i$. Substituting $a + bI_i$ for $\hat{C}_i$, the expression for SSE can be rewritten as

$$\sum_{i=1}^{N} (C_i - a - bI_i)^2 \qquad [1]$$

Using the calculus, expressions for a and b can be found that minimize the value of expression 2 and hence give the least squares estimates of α and β, which in turn define the regression line (see Appendix A for the derivation of the formulas).

For the given set of data, the a and b that minimize

$$\sum_{i=1}^{50} (C_i - a - bI_i)^2$$

are a = 714.58 and b = +0.058 (see Appendix A for the calculation of these values). Therefore, the least squares line, which is drawn in Figure 4, has the equation

$$C = 714.58 + 0.0581 \qquad [3]$$

These results mean, for example, that the estimate of consumption for a family whose annual income is $10,000 is $1294.94—that is, $1294.24 = $714.58 + 0.058($10,000). Remember, this is an estimate of C and not

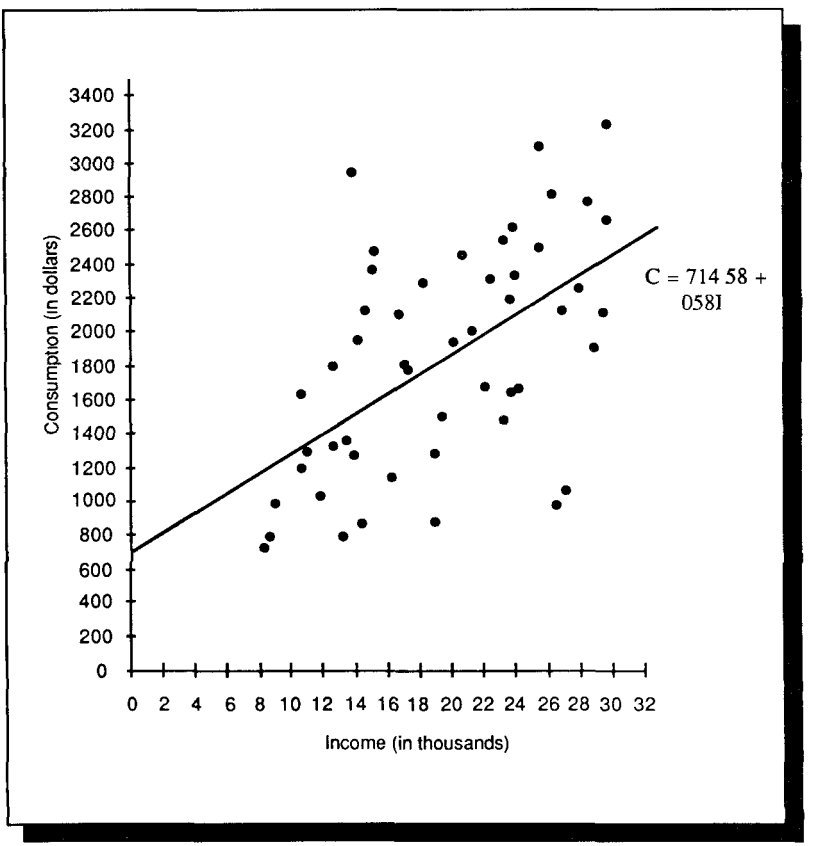

Figure 4: "Best Fitting" Regression Line

necessarily the amount one would observe for a specific family with an income of $10,000. The value of a, $714.58, is the estimated food consumption for a family with zero income. The value of b, 0.058, implies that for this sample, each dollar change in family income results in a change of $0.058 in food consumption in the same direction (note the positive sign for b).

These conclusions, of course, hold only for this particular sample. When the least squared technique is applied to additional samples of consumers, one would obtain additional (generally different) estimates of α and β.

It is important to point out that regression analysis does not prove causation. Our estimate of β is consistent with the theory that an increase in income causes an increase in food consumption. However, it does not prove causation. Note that we could have reversed the equation, making I depend on C, and argued that higher food consumption makes for healthier and more productive workers who thus have higher incomes. Since I and C increase together, this relationship would also be supported. It would take some alternative experiment or test to determine the direction of the causation. Our estimate of β, however, is not consistent with the theory that food consumption decreases with increases in income.[9]

Examples

Before proceeding, three examples are presented to illustrate how regression analysis is used.

EXAMPLE 1—INFLATION AND STOCK PRICES

Are stocks of major corporations a hedge against inflation—that is, does the return on a portfolio of stocks increase with the rate of inflation? Jaffe and Mandellzen (1976) address this question, as part of a broader study, by estimating the following regression equation

$$R_t = .0168 - 3.014I_t$$

where R_t is the rate of return on a market portfolio of stocks in month t and I_t is the rate of inflation in month t.[10] The estimate of the regression coefficient on I_t is –3.014, which implies that an increase in the inflation rate of one percentage point is associated with a reduction in the rate of return of 3.014 percentage points. Thus, for this portfolio, stocks do not appear to be a hedge against inflation.

EXAMPLE 2—HOME STATE ADVANTAGE

Has the advantage held by a U.S. presidential candidate in his home state diminished over time as elections have become more nationalized? This question was addressed by Lewis-Beck and Rice (1984). The regression equation they obtained is

$$H = 2.03 + .18T$$

where H is the home state advantage, measured in percentage points of the state popular vote, and T is an election year counter (e.g., for 1904 T = 1, for 1908 T = 2, and so on). Notice that the coefficient on T is positive, which suggests that the home state advantage has not declined over time.

EXAMPLE 3—PAY PREMIUM FOR VETERANS

In a recent article, De Tray (1982) argues that veterans receive a pay premium because employers, in evaluating the potential of employees, realize that veterans have had to pass mental and physical exams and survive a period of military service before being honorably discharged. He further argues that the quality of information provided by veteran status depends on the percentage of an age group that served in the military. Men who did not serve during war years, when virtually all able-minded and able-bodied men were drafted, may be less productive on the average than men who did not serve during peacetime, when few were called up. Therefore, De Tray hypothesizes that the veteran premium is positively related to the percentage in an age group that served in the military. To test this hypothesis, De Tray computed the veteran premium, w, for each of several age groups and regressed it on the percentage of each age group that served in the military, V. He found that the regression equation is equal to

$$w = -.078 + .165V$$

indicating that the premium increases as the percentage of the age group that served in the military increases. It should be noted that this is only part of a larger study.

The Linear Correlation Coefficient

In the first part of this chapter, we demonstrated how regression analysis can be used to summarize the relationship between a dependent

and independent variable. We turn now to an explanation of descriptive statistics designed to evaluate (1) the degree of association between variables and (2) how well the independent variable has explained the dependent variable.

The correlation coefficient measures the degree of linear association between two variables.[11] To understand what statisticians mean by linear association, consider Figure 5, which has the same 50 points as Figure 1. The average (or mean) level of food consumption is represented by the dotted line, while the solid line represents the mean level of income. The two lines divide the figure into the four quadrants denoted

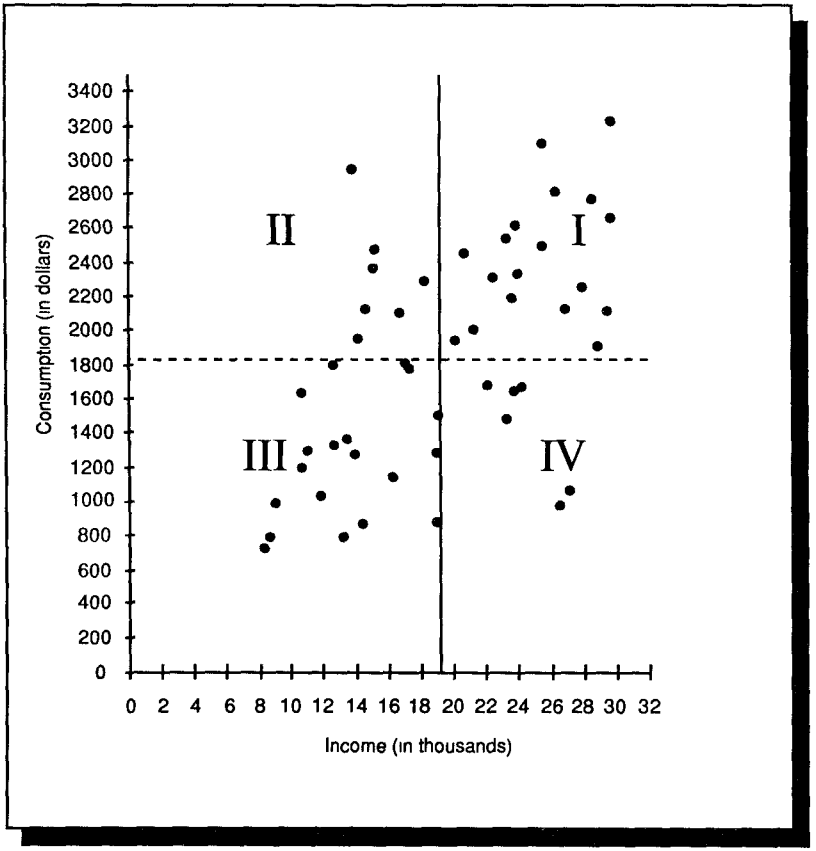

Figure 5: Linear Correlation Analysis: The Food Expenditure Problem

by Roman numerals. Levels of C that are greater than the average of 1842.45 lie above the dashed line in quadrants I and II, while less than average levels lie below, in quadrants III and IV. Similarly, income levels greater than the average lie to the right of 19,399 in quadrants I and IV, while those less than average lie to the left in quadrants II and III.

Figure 5 demonstrates that a majority of the points in the sample lie in quadrants I and III. Because of this pattern, the variables C and I are said to be *positively correlated.* Put differently, C and I are said to be positively correlated when Cs above (below) the mean value of food consumption, denoted $\bar{C}$, are associated with Is above (below) the mean value of income, denoted $\bar{I}$. On the other hand, if the Cs below $\bar{C}$ had been associated with the I's above I (and vice versa), one would have said that the variables were *negatively correlated.* The reader should be able to demonstrate that in this case the data points would have been clustered in quadrants II and IV. Another possibility exists: If the data points had been spread fairly evenly throughout the four quadrants, one would have said that C and I were *uncorrelated.*

The particular descriptive statistic that measures the degree of linear association between two variables is called the *correlation coefficient* and is denoted r. Although we offer no proof, r always lies between the values of -1 and $+1$ ($-1.0 \leq r \leq +1.0$). When there is little association between two variables (when two variables are relatively uncorrelated), r is close to zero. In the presence of strong correlation, r is close to 1 ($+1$ for positive correlation, -1 for negative correlation).

Although a positive correlation coefficient of .554 was found in the food example, where it was hypothesized that changes in income caused changes in food expenditures, the presence of either a positive or negative correlation does not always indicate causality. In particular, because the correlation coefficient only measures the degree of association between two variables, a cause-and-effect relationship is but one of four reasons why the presence of correlation may be observed. In addition, variables may appear correlated if both variables affect each other, if the two variables are both related to a third variable, or if the variables are systematically associated by coincidence.

An example of the first condition is that IQ scores and student achievement scores are likely to be positively correlated. Although it seems reasonable that IQ influences achievement, many educators believe that this is only part of the story. Indeed, it seems likely that the IQ measure also reflects the level of achievement. An example of the second

condition is the positive correlation that exists across cities between the number of churches and the number of bars. Although churches may spring up in response to bars (or bars in response to churches), the positive association most likely results because both variables are related to some other variable, such as population. A good example of the last condition is the positive correlation of .609 found between the number of letters in the names of the teams in the American Football Conference and the number of wins during the 1984 regular season.[12]

The Coefficient of Determination

Recall that for any problem, the regression line is defined to be the line lying closest to the data points (closest in the sense that the line minimizes the sum of the squared error term). Often, for comparative purposes, it is useful to know just how close is "close"; in other words, it is helpful to be able to evaluate what is referred to as the *goodness of fit* of the regression line.

An intuitive feeling for what is meant by goodness of fit is given in Figure 6, in which two distinct sets of data points have been plotted along with the two lines that minimize the sum of the squared errors. The regression line in panel A of Figure 6 clearly fits the data points more closely than the line in panel B.

The measure of relative closeness used by statisticians for evaluating goodness of fit is called the *coefficient of determination.* Because of its relationship to the correlation coefficient, this measure is generally referred to as the r^2. (The coefficient of determination is the square of the correlation coefficient.) The r^2 statistic measures closeness as the percentage of total variation in the dependent variable explained by the regression line. Formally, the measure is defined as

$$r^2 = \sum_{i=1}^{N} (\hat{C}_i - \bar{C})^2 / \sum_{i=1}^{N} (C_i - \bar{C})^2 \qquad [4]$$

To measure variation in a family's food consumption, we want some common base from which to measure differences in C. To the extent that families consume more or less than the mean food consumption, $\bar{C}$, there is variation in food consumption. Thus we use $\bar{C}$ as the base for measuring variations in C between families.

The denominator of equation 4 is a measure of the total variation in the dependent variable about its mean value $\bar{C}$. For example, consider a household with an income of $20,108 and observed consumption of

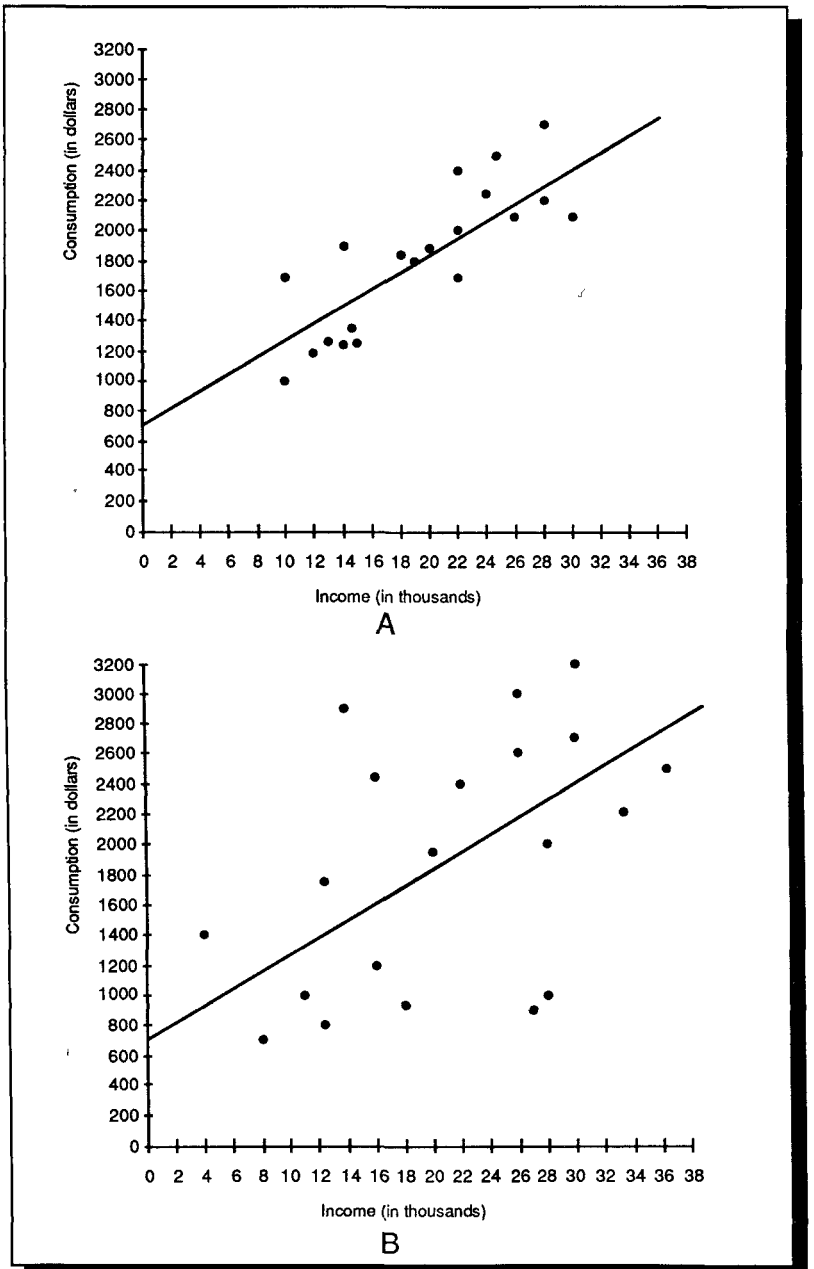

Figure 6: Comparison of Goodness of Fit for Two Regression Lines

$1939.00 (shown in Table 1). Since the mean value of consumption is $1842.45, the observed variation of C from the mean is $96.55 for this observation ($96.55 = 1939.00 – 1842.45). So that negative variations do not cancel positive variations, the individual variations are squared before they are summed.

The numerator of equation 4 is a measure of the total variation explained by the regression line. For example, from regression equation 3, it follows that the best estimate of food consumption for the family with an income of $20,108 is $1880.84 (1880.84 = 714.58 + .058($20,108). Since this is $38.39 from the mean ($38.39 = $1880.84 – $1842.45), it is said that $38.39 is the variation explained by the regression line for this observation. The total explained variation is found by summing the square of these variations for the entire sample.

For the food expenditure problem, the value of the r^2 is .307, and one can say that the regression line explains 30.7 percent of the total variation in food expenditures. Stated somewhat differently, it can be said that 30.7 percent of the variation (about the mean) in the dependent variable has been explained by variation (about the mean) in the independent variable.

Notice that if the data points were all to lie directly on the regression line, the observed values of the dependent variable would be equal to the predicted values, and the r^2 would be equal to 1. As the independent variable explains less and less of the variation in the dependent variable, the value of r^2 falls toward zero. Hence, as would be expected, the r^2 for the data in panel A of Figure 6, .783, is greater than that for the data in panel B of Figure 6, .198.

For the three examples presented earlier, the coefficients of determination, r^2, are .0269 for the relationship between stock prices and inflation, .025 for the presidential home state advantage, and .45 for the veteran's premium equation. Note the differences in their values.

Regression and Correlation

It is important to note that linear regression, the correlation coefficient, and the coefficient of determination are all related but that they provide different amounts of information and are based on different assumptions. First, as indicated previously, the coefficient of determination is simply the square of the correlation coefficient. An examination of Figure 5 should also convince the reader that if two variables are positively (negatively) correlated, the regression coefficient will have a positive (negative) sign.[13]

While this general relationship between r and b will always hold, one might ask if one of these two measures provides more information than the other. The answer is that the regression coefficient is more informative since it indicates by how much the dependent variable changes as the independent variable changes, whereas the correlation coefficient indicates only whether or not the two variables move in the same or opposite directions and the degree of linear association. This additional information from regression is obtained, however, only at the cost of a more restrictive assumption—namely, that the dependent variable is a function of the independent variable. It is not necessary to designate which is the dependent and which the independent variable when a correlation coefficient is obtained.

2. MULTIPLE LINEAR REGRESSION

In Chapter 1, variations in the dependent variable were attributed to changes in only a single independent variable. This is known as *simple linear regression.* Yet theories frequently suggest that several factors simultaneously affect a dependent variable. *Multiple linear regression analysis* is a method for measuring the effects of several factors concurrently.

There are numerous occasions where the use of multiple regression analysis is appropriate. In economics it is argued that the quantity of a good that will be purchased by an individual depends on both income and the price of the product (Manning and Phelps, 1979). The likelihood that a family will move depends on both the age of the head of the household as well as the family's income (Fields, 1979). In determining the effect of advertising on the sales of some product, it is important to include not only the amount of advertising during the current period but also the amount in earlier periods (Simon, 1969). The proportion of the vote a congressional incumbent gets in an election is influenced by several factors, including the health of the local economy, the incumbent's performance in obtaining federal funds for the district, and how long the incumbent has been in office (Felman and Jondrow, 1984).

Estimating Regression Coefficients

In the food consumption example only a single variable, income, was hypothesized as a determinant of family food expenditures. One recognizes, however, that even though two families have identical

incomes, their food expenditures may differ greatly. For example, the families may differ in size, in the availability of homegrown items which can decrease out-of-pocket food costs, or in taste. Therefore, it is reasonable to hypothesize that variables, in addition to income, affect the amount spent on food. One likely hypothesis is that the amount of food consumed is positively related to the family's size, S. Multiple linear regression analysis is used to estimate the effect of S on food consumption while at the same time taking into account the effect of income.

The concept of multiple regression analysis is identical to that of simple regression analysis except that two or more independent variables are used simultaneously to explain variations in the dependent variable. When family size is added to income to explain food consumption, the newly hypothesized relation can be written as

$$C = \alpha + \beta_1 I + \beta_2 S \qquad [5]$$

where α, β_1, and β_2 must be estimated from observed values of consumption, income, and family size. For any observed combination of values for I and S, it is still desired to find values for the coefficients that minimize the distance between the corresponding observed and estimated values of C.

A graphical presentation of these concepts is now more difficult, since with two independent variables, three-dimensional drawings are required. Minimizing distance in this context means minimizing the length of line segments drawn between the observed values of the dependent variable and its estimated value lying on the plane corresponding to $C = \alpha + \beta_1 I + \beta_2 S$. Algebraically, this means finding the values of a, b_1, and b_2 that minimize the value of

$$\sum_{i=1}^{N} (C_i - a - b_2 I_i - b_2 S_i)^2.$$

As in the case of simple regression analysis, a technique exists which ensures that the resulting estimates of α, β_1, and β_2 are those that minimize the sum of squared errors and thus give the best estimates of the coefficients. When this technique is applied to the data in Table 1, the estimated regression equation obtained is

$$C = 330.77 + 0.056I + 129.62S \qquad [6]$$

Interpretation of these results is similar to simple regression analysis. For example, the coefficients derived from the data indicate that the estimate of food consumption for a family of four with an income of \$10,000 is \$1409.25, since \$1409.25 = \$330.77 + 0.056(\$10,000) + \$129.62(4).

More generally, the estimated coefficient on any independent variable estimates the effect of that variable *while holding the other independent variable(s) constant.* Thus the results shown in equation 6 indicate that holding income constant, an increase of one in family size is associated with a \$129.62 increase in food consumption.[14] Similarly, the results suggest that a dollar increase in income will increase food expenditures by 5.6 cents, holding family size constant. One can also consider the effect of a simultaneous change in S and I. For example, the estimated effect of a decrease in income of \$1000 at the same time family size increases by one would be +\$73.62 = 0.056(−1000) + 129.62(1).

The coefficient on income in equation 6 is slightly different from that reported in the simple linear regression case, where a one-dollar change in income resulted in a 5.8-cent change in food consumption. In some cases when another independent variable is introduced, this change in the value of the estimated coefficient may be large. This issue is discussed in more detail in Chapter 5.

Multiple regression results come closer to showing the pure effect of income on food consumption since they explicitly recognize the influence of family size on food expenditures. It is for this reason that in formal studies it is not proper to exclude a variable such as family size when the theory indicates that the variable should be included. To simplify the presentation, we have not followed this proper practice.

Finally, note that multiple linear regression is not limited to only two independent variables. Rather, it applies to any case when two or more independent variables are used simultaneously to explain variations in a single dependent variable.

Standardized Coefficients

In the multiple regression example, we noted by how much food consumption would change for a given change in income holding family size constant, and by how much food consumption would change for a given change in family size, holding income constant. A question that may arise is whether income or family size has the greater impact on food consumption. If we simply compared the size of the estimated parameters, it is obvious that b_2 is much greater than b_1, suggesting that

family size has a greater effect on C or is more important than income. But that is not an appropriate comparison, since income is measured in dollars and family size is measured in persons. Comparing b_1 with b_2 is comparing the effect of a one-dollar change in income to the effect of a one-person change in family size. Relative to the range of income levels, a one-dollar change in income is very small, while for family size a one-person change is quite large.

Instead of determining the effect of a one-dollar change in income or a one-person change in family size, suppose we use a standardized unit to measure changes in income and family size. One such measure, the *standard deviation,* measures the dispersion of the values of a particular variable about its mean.[15] Look at the values of income and family size in Table 1 and notice that income is spread out over a wider range of values (from $8,246 to $29,690) than is family size (from 1 to 9). This dispersion is reflected in the standard deviations, which for income is $6,382 and for family size 2.00. Thus using the standard deviation as the unit of measure takes into account that a one-person change in family size is very important relative to the spread of values for family size, while a one-dollar change in income is rather unimportant relative to the dispersion in income levels.

Frequently researchers report *standardized coefficients,* also referred to as *beta coefficients* (do not confuse the beta coefficient with β, the population parameter). These standardized coefficients measure the change in the dependent variable (measured in standard deviations) that results from a one-standard-deviation change in the independent variables.

For the regression reported in equation 6, the standardized coefficients are .535 for income and .386 for family size. Thus changing income by one standard deviation ($6,382), while holding family size constant, would change food consumption by .535 standard deviations. Changing family size by one standard deviation, holding income constant, would change food consumption by .386 standard deviations. When viewed in this way, a change in income has a greater relative effect on food purchases than does a change in family size, a finding just opposite to that suggested by the regression coefficient.

Associated Statistics

Just as there is a great deal of similarity between the interpretation of simple and multiple regression coefficients, so are many of the associated statistics for the two regression methods also similar.

The *coefficient of multiple correlation,* often denoted as R, is similar to r in that both measure the degree of associated variations in variables. Rather than measuring the association between two variables, the value of R indicates the degree to which variation in the dependent variable is associated with variations in the several independent variables taken simultaneously. Similarly, R^2, the *coefficient of multiple determination,* measures the percentage of the variation in the dependent variable which is explained by variations in the independent variables taken together.

For regression equation 6, R^2 is .456, indicating that 45.6 percent of the variation in C about its mean is explained by variations in I and S about their respective means. Note that the addition of the second independent variable has increased the explanatory value of the regression over that of the simple linear regression case. It is also evident, however, that even this regression equation does not explain all the variation in food expenditures.

It cannot be overemphasized that although the coefficient of determination is of interest, it should never be the sole determinant of the "goodness" or "badness" of a regression result. The maximization of R^2 is not the purpose of regression analysis.

The value of the coefficient of determination will never decrease when another variable is added to the regression. Although the additional variable may be of no use whatsoever in explaining variations in the dependent variable, it cannot reduce the explanatory value of the previously included variables. Thus, by carefully choosing additional independent variables, an investigator can increase the value of R^2 greatly without improving his or her knowledge of what affects the value of the dependent variable. For instance, the amount spent on food is partly reflective of the amount spent on meat. If a researcher were to include the dollar value of meat purchases as another independent variable, the R^2 statistic would probably increase greatly. However, such an equation would not increase our understanding of why food consumption expenditures differ across families. The moral is: If a variable has no place in the theory, it should not be included in the regression analysis.

Since including additional variables can never decrease the value of R^2 and normally increases it, analysts commonly report the *adjusted* R^2, denoted $\bar{R}^2$. This term is R^2 adjusted for the number of independent variables used in the regression.[16] Thus it is possible that by adding another independent variable to the regression, the adjusted R^2 will decrease although R^2 actually increases. For this reason, $\bar{R}^2$ is some-

times used to determine whether including another independent variable increases the explanatory power of the regression.

Examples

To illustrate the use of multiple regression, consider the following three examples:

EXAMPLE 1—PREMARITAL COHABITATION

What is the effect of premarital cohabitation with one's future spouse on marital satisfaction? This question was addressed by DeMaris and Leslie (1984) through the use of multiple regression analysis. Using data from 309 recently married couples, a multiple regression equation, summarized in Table 2, was estimated for wives.

The dependent variable is a measure of marital satisfaction. The independent variable of greatest interest is "having cohabited," which takes on only two values—zero if the couple did not cohabit, and one if they did. The coefficient on cohabitation is negative, suggesting for this sample that premarital cohabitation reduces marital satisfaction. To see this, note that cohabitation can be interepreted as meaning that the

TABLE 2
Regression Equation for Cohabited Equation

Variables	*b*	*Beta*
Father's occupation is white-collar	−.18	−.01
Education	.16	.02
No religious preference	−2.55	−.07
Church attendance	.33	.04
Differences in education	.10	.01
Small difference in church attendance	−5.68**	−.17
Large difference in church attendance	−.42	−.01
Husband is 5-8 years older than wife	5.66*	.14
Husband is 9 or more years older than wife	1.37	.02
Sex-role traditionalism	.11	.10
Having been previously married	3.76	.12
Presence of minor children at home	−4.55*	−.15
Having cohabited	−4.61**	−.14
$R^2 = .13$		
Number of observations = 262		

SOURCE: DeMaris and Leslie (1984). Reprinted by permission.
*$p < .05$.
**$p < .01$.

value of "having cohabited" increases from zero to one. Changing "having cohabited" from zero to one changes the value of the dependent value by –4.61, the value of the coefficient on "having cohabited."

Since people who do and do not cohabit may differ in other ways that might also affect marital satisfaction, it was necessary to control for these factors by including other variables in the regression equation. Notice that many of these variables, including cohabitation, are yes/no variables, usually called *dummy variables* (these are discussed in Chapter 4). While the authors report the standardized coefficients (beta), they do not report the intercept. The value of R^2 is .13. The asterisks are explained in Chapter 3.

EXAMPLE 2—HOUSEWORK TIME

A question that Gronau (1977) has studied is what determines how people spend their limited time. As part of a larger study, Gronau estimated the regression equation presented in Table 3 for a sample of 621 married white women who were not employed outside the home. The dependent variable is the amount of time in a year that was spent doing housework, such as cooking and cleaning.

Notice that older and more educated women spend less time at housework. As the husband's wage and the family's other income increases, less time is spent at housework. This could result from eating out more often or by using cleaning services, both of which could increase as the family's income increases. The coefficient on the husband's wage suggests that an increase in his wage of one dollar an hour

TABLE 3
Regression Equation for Allocation of Time

Variable	*b*	*t-Ratio*
Constant	1,669.40	–
Wife's age	–1.165	.37
Wife's education	–53.469	3.28
Husband's education	22.668	1.82
Husband's wage ($/hour)	–16.129	2.21
Income from sources other than work (year)	–.044	2.23
Children aged 0-17	327.654	6.94
Children at school	–125.196	2.86
Rooms in house	83.251	3.17
R^2 = .26		
Number of observations = 621		

SOURCE: Gronau (1977). Reprinted by permission.

TABLE 4
Regression Equation for Job Satisfaction

Variable	*b*	*Standard Error*
Satisfaction with pay	−.003	.005
Satisfaction with promotion	−.010	.004
Satisfaction with co-workers	.003	.007
Satisfaction with work	−.034	.007
Satisfaction with supervision	−.021	.007
R^2 = .270		
Number of observations = 263		

SOURCE: Futrell and Parasuraman (1984). Reprinted by permission.

reduces the time spent on housework by 16.129 hours per year. On the other hand, the greater the number of children and the larger the house, the more time spent doing housework. The meaning of the t ratio is explained in Chapter 3.

EXAMPLE 3—JOB SATISFACTION

The relationship of job satisfaction to the propensity to leave a job was investigated by Futrell and Parasuraman (1984). Using a questionnaire administered to salespersons, the authors determined the individual's level of satisfaction with various aspects of his or her job and the extent to which the individual was seeking to change jobs, with the latter being used to measure the propensity to leave. The regression equation presented in Table 4 was estimated for a sample of 263 salespersons. With the exception of co-worker satisfaction, the coefficients have the expected signs; a higher level of satisfaction is associated with a lower propensity to leave. The standard error is discussed in Chapter 3.

3. HYPOTHESIS TESTING

Introduction

In the food expenditure problem, the hypothesis was advanced that family food consumption increases as income increases. Since the estimated coefficient was found to be a positive number, one might immediately conclude that we have proven our case. Unfortunately, drawing such inferences is not so easy, since our hypothesis concerns the *population* of all food consumers, not just the 50 persons in our *sample*.

However, the hypothesis-testing procedure allows us to make statements about the entire population from our sample, not just statements about the particular sample we happened to draw. In order to make such inferential statements—that is, to infer from the sample something about the population—we must develop some statistical theory. Therefore, before turning to testing hypotheses about population regression coefficients, we consider a slightly less complex example.

Suppose you were browsing through the library and came across a document indicating that the average height of *all* students who attended your university or college in 1920 was 5 feet 4 inches (64 inches). Suppose further that you became interested in learning whether the students enrolled in your school today are taller than those of three generations ago. One way to attack this problem would be to measure the height of all students currently enrolled. While that procedure might work well in a small liberal arts college with only a few hundred students, the task would be enormous if you were a student at a large state university. Fortunately, statistical theory allows one to make inferences about the mean height of the entire population using only information on the average height of students computed from a single random sample of the student population. After this inference has been made, comparisons can be made with the height for the population of students in 1920.

To continue with the example, suppose you measure the height of a *random sample* of 200 students and find that their mean height is 67 inches. Your sample of 200 is only one of many such samples that could be drawn from students on a large university campus. Therefore, even though the mean of 67 inches is greater than 64, you should not immediately conclude that today's student body is taller than the 1920 group. Instead, the hypothesis-testing procedure must account for the fact that, since your particular sample is only one of a large number of possible samples, the 67-inch mean is only one of a number of possible sample means. Some samples may yield sample means less than 64 inches.

The theory of hypothesis testing provides a method for making inferences about the entire population from sample data. The method recognizes that, since the inferential statement is based on sample information, we can never be totally certain of the validity of the inference about the population.[17] Instead, one must allow for some probability that an incorrect conclusion has been drawn. Statistical theory allows us to define the likelihood of making such an incorrect inference. For example, based on the sample mean of 67 inches, you might conclude

that today's student body is taller than the 1920 student body but that there is a 1 percent chance that you have drawn an incorrect conclusion. Inferential statements based on sample data never yield conclusions about the population values that are 100 percent certain.

In the food expenditure regression problem, the hypothesis was advanced that family food consumption increases as income increases. Since hypotheses are stated in terms of the values of the population parameter, this hypothesis is equivalent to the hypothesis that β is greater than zero.[18] The discussion now turns to the hypothesis-testing procedure, a technique that allows one to draw inferences about the population parameter from a sample estimate of that parameter.

In order to understand hypothesis testing, it is important to reiterate that we have been working with only one sample from the population. Just as one could have multiple samples of students, it is possible to draw multiple samples of families. If we did this, the regression procedure outlined in Chapter 1 could be used to generate additional estimates of β which would probably not be identical to our earlier estimate, since the samples are different. Some of these b's will be very good in the sense that they lie close to the true, but unobservable, β. Others will be bad in the sense that they lie some distance from β. Our problem is that we have no way of knowing if ours is a good or bad estimate of β.

Suppose that a method existed to compute what we will call a test value, tv, such that there was only a 5 out of 100 chance of getting an estimate that overstates β by more than this test value. In other words, out of every 100 samples drawn, only 5 would generate b's that overstate β by more than tv. If β were zero, this implies that only 5 out of every 100 estimates would be so bad that they would yield a value of b greater than this test value. Thus we could argue that if β were zero, the probability of getting an estimate of β the size of tv or greater would be very low—explicitly, 5 percent. Suppose that for our data set the value of tv is .022 (we show later how this number is derived).

For the food consumption problem, we wish to investigate the possibility that there is no relationship between consumption and income—that is, that β is zero—versus the possibility that food consumption increases as income increases—that is, that β is greater than zero. In our simple regression equation, we obtained a b of .058, which is clearly greater than zero.[19] The test value tells us that if the population value of β is zero, there is only a 5 percent chance of obtaining estimates of β greater than .022. Therefore, if β is zero, it is quite unlikely that the estimated regression coefficient would be greater than .022. Our b is

greater than .022. Based on the low probability of this occurring *if β is zero,* we say that we are willing to reject the statement that β is zero in favor of the statement that β is greater than zero. There is at most a 5 percent chance that we have rejected the statement that β is zero when indeed it is zero. In the language of hypothesis testing, we have rejected the *null hypothesis* that food consumption is invariant to income level ($\beta = 0$) in favor of the *alternate hypothesis* that food consumption increases as income increases ($\beta > 0$).

Hypothesis testing is analogous to decisions reached in courts of law. Under the court system, a defendant is brought to trial and he or she is *assumed* to be not guilty. For the judge or jury to reject the assumption of not guilty in favor of the alternate finding of guilty, sufficient evidence must be produced. In the court system, errors can be made; innocent defendants can be found guilty and guilty individuals can be found not guilty. Under a legal system where the evidence must show "beyond a shadow of doubt" that the assumption of nonguilt is to be rejected, there is a primary concern for the inferential error of the first type—that is, of convicting an innocent person.[20]

Just as the defendant is assumed not guilty until proven guilty, in hypothesis testing the null hypothesis is assumed true until there is sufficient evidence that it is not true. Likewise, just as inferential errors can occur in courts of law, inferential errors can also occur in hypothesis testing. Again, we are particularly concerned with an inferential error of the type that occurs if one rejects the null hypothesis in favor of the alternate when the null hypothesis is actually true. Instead of simply stating that the analyst should reject the assumption that the null is true in favor of the alternate if the evidence suggests it "beyond a shadow of a doubt," the hypothesis-testing procedure allows the investigator to specify an exact probability of making an inferential error—that is, allows the investigator to define how big the "shadow of a doubt" is. Most commonly, 1, 5, and 10 percent probabilities are chosen; however, there is nothing that prevents the analyst from using other probabilities of this type of inferential error.[21] When the researcher can reject the null hypothesis that $\beta = 0$ in favor of the alternate, the regression coefficient is said to be *significant,* which is short for significantly different from zero at a stated probability. The *level of significance* depends on the probability the investigator has assigned to rejecting the null when it is indeed true.

In Table 2, the double asterisks next to the coefficient on the cohabitation variable imply that this coefficient is significant at the 1 percent

level of significance (this is how "$p < .01$" in that table is to be read). This means that, in rejecting the null hypothesis that cohabitation has no effect on marital satisfaction ($\beta = 0$) in favor of the alternate that there is an effect, there is at most a 1 percent chance that we have rejected the null hypothesis that $\beta = 0$ when indeed β is zero. Likewise, as will be seen, the t ratios reported beside the regression coefficients in the housework example of Table 3 can be used to determine whether or not a coefficient is significant.

The Testing Procedure

The formal procedure used to test hypotheses concerning the value of the population parameter is comparable to the procedure discussed earlier. First, a hypothesis concerning the value of the population parameter is formulated. This hypothesis is referred to as the *null hypothesis,* denoted H_0, and is assumed to hold unless sufficient evidence is found to reject it. The null hypothesis in the food consumption problem is that β is equal to zero (this is written as $H_0{:}\beta = 0$). Second, the test value method (to be discussed later) is used to compute a number, tv, such that if H_0 is true, there is a low prespecified probability of obtaining an estimate that overstates β by more than tv. The chosen probability is referred to as the level of significance; we will use 5 percent for the time being. Thus, on average no more than 5 percent of all samples will produce b's that are greater than the population parameter by more than this test value when the null hypothesized value of β is the actual value of β. Third, the difference between b and the hypothesized value of β is computed. Finally, the following criterion is used to test the null hypothesis:

(1) Reject the null hypothesis if this computed difference is greater than the test value.
(2) Do not reject the null hypothesis if this difference is less than or equal to the test value.

Statement 1 in the criterion says that if the difference between the estimate and the hypothesized value is greater than the test value, the null hypothesis is to be rejected, since there is only a 5 percent chance that, if the null is true, an incorrect inference about the population parameter will be made. If, on the other hand, the difference is less than or equal to the test value (statement 2 of the criterion above), one cannot feel confident in rejecting the null hypothesis, since 95 percent of the

samples will produce b's that vary by no more than this amount from β when the null hypothesized value of β is the actual value of β.

Note from the above criterion that only rejection or nonrejection of the null hypothesis is possible. Nonrejection does not imply that one accepts the null hypothesis. This is because the procedure outlined previously only tells us the probability of rejecting the null hypothesis when it is true. This is analogous to the court example where the finding is "not guilty" instead of "innocent." The level of significance does not tell us anything about the probability of accepting the null when it is false. On the other hand, if the null hypothesis is rejected, it is usually stated that the alternate hypothesis, often denoted H_a, is accepted. It is for this reason that the relationship that the researcher predicts between the independent and dependent variable is stated as the alternate hypothesis.

We have now formulated the concept of the null hypothesis and the criterion used to test that hypothesis. The hypothesis-testing procedure will be complete once the method for constructing the test value (tv) has been presented. As will be shown, the test value depends on (1) the estimated variability of the estimates of β from sample to sample and (2) a probability distribution.

The Standard Error of the Estimated Coefficient

The *standard error* of the regression coefficient is a measure of the amount of variability that would be present among different b's estimated from samples drawn from the same population. While it is true that equation 3 in Chapter 1 provides a unique estimate of β, it is also the case that if a different set of data were drawn from the population, a different estimate of β would probably result. Statistical theory allows us to estimate how much variability there would be among all these estimates (that is, allows us to estimate the standard error) just by taking information from one sample.

In essence, the standard error measures how sensitive the estimate of the parameter is to changes in a few observations in the sample. To understand what is meant by sensitive, consider Figure 7. Panel A presents two samples from population A, panel B presents two samples from population B, and panel C presents two samples from population C. In each case the ordinary least squares regression lines are also presented. The figure is constructed so that, with the exception of the circled observations, the data points are the same for any given panel

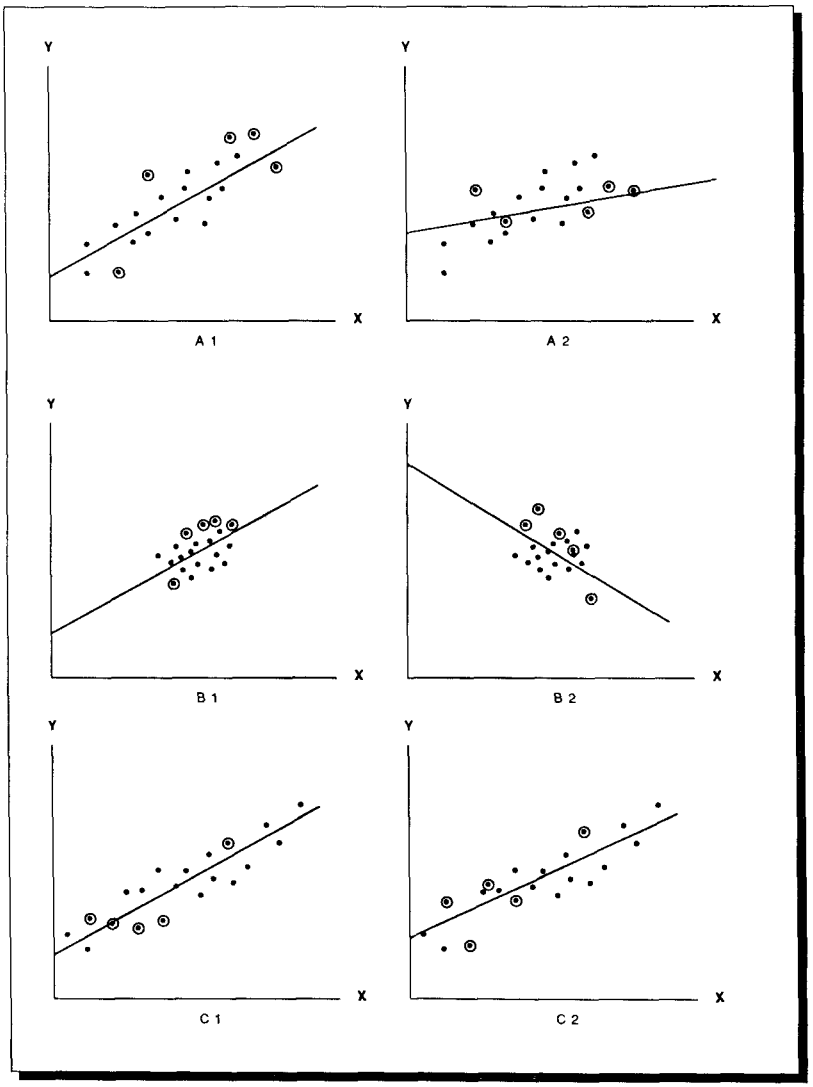

Figure 7: Sensitivity of Regression Line to Changes in Observations

(i.e., within lettered pairs). In the case of the circled observations, within a given panel the values of the X's have remained unchanged while the associated Y values have changed. It is apparent that regression coefficients estimated from either population A or B are extremely sample-de-

pendent. In both situations a change in a few of the observations results in a large change in the slope of the regression line and hence a large change in b. The data drawn from population C, however, are neither scattered nor clustered. In this instance, a change in a few of the observations will not alter b substantially.

What characteristics do the data in panels A and B have which do not appear in panel C? In A the amount of variability of the dependent variable Y (measured on the vertical axis) which cannot be attributable to variations in X is great relative to that in data set C. In panel B the variations in X are considerably less than the comparable variations in the independent variables shown in Panel C. Each of these characteristics is positively related to the standard error of a regression coefficient and creates additional uncertainty regarding the true parameter β.

The measure of the standard error[22] allows one to make inferences about how sensitive the estimate of β is to changes in sample composition without taking another sample. Because a large standard error casts doubt on the estimate, the magnitude of the test value depends positively on the size of the standard error. The standard error, generally represented as s_b, is often reported along with the regression coefficients, as in Table 4.

The Student's t Distribution

A probability distribution[23] is also used in the hypothesis-testing procedure. To better understand the role that probability plays in the testing procedure, reconsider what has been said thus far about regression parameters. First, it has been stressed that the population parameter can never be observed. Second, it has been noted that the estimate of the parameter from any sample is but one possible estimate; additional samples from the population yield additional, probably different estimates. Not all estimates are equally "close" to the population parameter. Finally, it is desired to draw inferences about the population parameter from one estimate of the parameter. In the food consumption problem, the b of .058 is to be used to make inferences about the population β. Thus one would like to know if .058 is one of the estimates that is close to β.

A question of this nature can never be answered, since the value of the population parameter is unobservable and hence unknown. A statement can, however, be made regarding the probability of obtaining an estimate with a given degree of closeness to the assumed, null hypothesized, value of β. Analogously, probabilistic statements can be made concerning the degree of closeness associated with a given probability.

These statements can be made because statisticians have determined the probability distribution of the fraction $(b - \beta)/s_b$. In general, this fraction is distributed according to what is known as the *Student's t distribution.* (A discussion of how statisticians are able to determine the probability distribution of $(b - \beta)/s_b$ is beyond the scope of this book.) The Student's t distribution allows one to make probabilistic statements concerning the size of the fraction $(b - \beta)/s_b$. The distribution relates the probability that the fraction will be no larger than what is known as the t statistic, denoted t_s.

For a stated probability, the t statistic depends on the *degrees of freedom*, defined as the number of observations in the problem (the size of the sample) minus the number of coefficients estimated. Values for the Student's t distribution are given in Appendix B. In the consumption problem, there are 48 degrees of freedom, since two coefficients (a and b) were estimated and there are 50 observations.[24] (See also Figure 8.)

For any given problem with 48 degrees of freedom, the t distribution states that for 5 percent of the samples, the fraction $(b - \beta)/s_b$ will be larger than 1.677. This implies that the probability is 5 percent that the following inequality holds:[25]

$$(b - \beta)/s_b > 1.677 \qquad [7]$$

Multiplying this inequality by s_b yields

$$(b - \beta) > 1.677s_b \qquad [8]$$

Inequality 8 means that if the null hypothesis is true, only 5 percent of the estimates will exceed the null hypothesized value by more than $1.677s_b$. Thus 95 percent will overstate the null hypothesis by less than this value.

Forming Test Values

The expression $1.677s_b$ is an example of a test value. More generally, a test value is formed by multiplying the appropriate t statistic by the standard error of the estimator. In the food expenditure problem, s_b = .013. Since $t_s s_b = (1.677)(.013) = .022$, the test value is .022. The null hypothesis can be rejected *if the difference between the estimated coefficient and the hypothesized value is greater than this test value.* In the case where the hypothesized value is zero, this difference is always equal

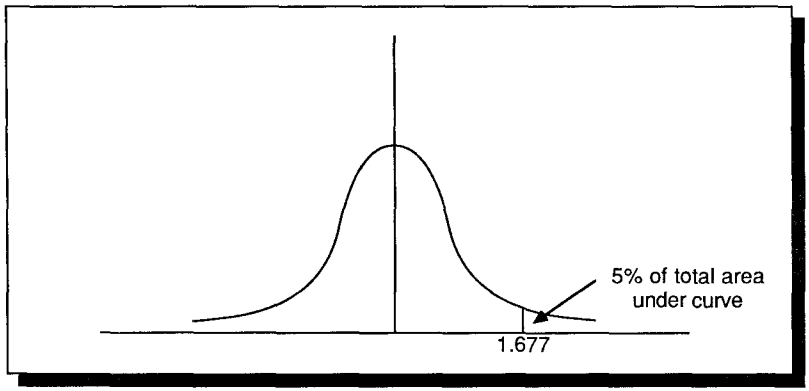

Figure 8: t Distribution

to the estimated coefficient, b, in this case .058. Thus, for the food expenditure problem, the null hypothesis can be rejected in favor of the alternate hypothesis that a positive relationship exists between income and food expenditure, since $.058 > .022$. More generally, it follows that the null hypothesis that $\beta = 0$ can be rejected in favor of the alternate hypothesis that it is greater than zero if

$$b > s_b t_s \qquad [9]$$

The testing procedure can also be used to test hypotheses concerning hypothesized values of β other than zero.[26] Suppose, for example, that one wished to test the hypothesis that a one-dollar increase in income is associated with a 4-cent increase in family food expenditure against the hypothesis that it is associated with a larger increase. In this case, the null hypothesis is $H_0: \beta = .04$, and the alternate hypothesis is $H_a: \beta > .04$. The difference between .04 and our estimate of .058 is .018. Given that this is less than the test value of .022, one cannot reject the null hypothesis. On the other hand, the reader should be able to verify that the null hypothesis, that $\beta = .03$, could be rejected at the 5 percent level of significance in favor of the alternate hypothesis that $\beta > .03$. In this instance we say that the coefficient is significantly greater than .03.

The Role of Standard Error and Sample Size

The statistical inference made about the population parameter from its estimate clearly depends on the size of the test value, which in turn

depends on the size of the standard error of the estimated coefficient and on the size of the appropriate t statistic. A larger test value means, other things being equal, that it is harder to reject the null hypothesis in favor of the alternate. If the standard error in the food expenditure problem had been larger, the test value would also have been larger and different inferences might have been drawn about the population parameter.

As noted in the discussion of the t distribution, for a given level of significance, the size of the t statistic, and hence the size of the test value, is influenced by the size of the sample.[27] That the number of observations in the sample will influence the size of the interval is reasonable, since a small sample is less likely to be representative of the population than a larger sample. The t statistics given in Appendix B illustrate that as the degrees of freedom decrease, the t statistic increases. Thus, for example, if the food expenditure sample size had been smaller, the appropriate t statistic would have been larger. As a result, the test value would also have been larger and different inferences might have been drawn about the population parameter.

Changing the Level of Significance

Although the 5 percent level of significance is suitable for much empirical research, in some instances it is desirable to have a smaller probability of rejecting the null hypothesis when it is true. As can be seen from Appendix B, for a given number of degrees of freedom the t statistic (and hence the size of the test value) increases as the level of significance decreases.[28] Applying the method discussed earlier, one finds that for the food expenditure problem, at the 2.5 percent level of significance the test value is $.026 = t_s s_b = (2.011)(.013)$. In a similar fashion, at the 1 percent level of significance the test value is .031. Notice that it might be possible to reject a hypothesis at the 5 percent level of significance but not at a lower level of significance. Often researchers will indicate at what level a variable is significant. In the cohabitation example of Table 2 the single asterisk indicates that a coefficient is significant at the 5 percent level; the double asterisk indicates significance at the 1 percent level. The lowest level at which a null hypothesis can be rejected is called by some authors the *prob value* or *p value* of a test (for an example of this, see Table 2).

t Ratio

Simple algebraic manipulation allows us to rewrite equation 9 as

$$(b/s_b) > t_s \qquad [10]$$

The expression b/s_b is referred to as the *t ratio*. The reader can check that for the food consumption problem it is 4.462. Researchers often report this number in lieu of the standard error. Thus, for example, the numbers beside the regression coefficients in the housework time example (Table 3) are t ratios and not standard errors.

The null hypothesis that $\beta = 0$ can easily be tested by computing the t ratio and comparing it to the appropriate t statistic. If the t ratio is greater than the appropriate t statistic, the null hypothesis can be rejected at the specified level of significance. In addition, the t ratio provides a way of determining the level of significance at which the null hypothesis can be rejected. For example, Appendix B demonstrates that for the food expenditure problem, the hypothesis that $\beta = 0$ can be rejected at the 0.5 percent level of significance. (For 48 degrees of freedom, the t statistic at the 0.5 percent level of significance is 2.682, substantially less than the t ratio of 4.462.) For a similar reason, the t ratio of 3.17 reported beside the number-of-rooms variable in the housework time example (Table 3) implies that the null hypothesis that $\beta = 0$ can be rejected at the 0.5 percent level.

Just as the examples of Chapter 2 do not provide a uniform format for tests of significance, neither do computerized regression programs. For example, as can be seen from Appendix C, SPSS output provides information on standard errors, while SAS output provides information on t ratios as well as standard errors.

Left-Tail Tests

The reader will note that all of the alternate hypotheses presented thus far have taken the form, "β is greater than some number." In order to test the corresponding null hypothesis and make inferences about the alternate hypothesis, we have computed by how much our estimate overstates the null hypothesized value and then compared this difference to the test value. This type of test is called a *right-tail test*. It gets its name from the fact that in this instance the alternate hypothesis is positive and lies to the right of the null hypothesized value. There are, of course, instances in which one is interested in alternate hypotheses that concern negative values. In this case a left-tail test is in order. *Left-tail tests* are appropriate when the alternate hypothesis is of the form that the population parameter is less than some specified number, such as zero. In such a case, we would have: $H_0: \beta = 0$, $H_a: \beta < 0$.

A test value for a left-tail test can be computed in the same manner as a test value for a right-tail test. For example, for a left-tail test with 48 degrees of freedom, only 5 percent of the sample will yield b's that understate the population parameter by more than $-1.677s_b$. Note that once again we are comparing the difference between the estimate and the null hypothesized value to some test value. Here, however, if we use $(b - \beta)$ as a measure of "understatement," the difference is negative since the alternate hypothesis lies to the left of the null hypothesis, not to the right. Thus we are saying that in only 5 percent of the cases is this difference *more negative* than $-1.677s_b$; that is, in only 5 percent of the cases is $(b - \beta) < -1.677s_b$.[29]

Just as we computed a t ratio for a right-tail test, we can also compute a t ratio for a left-tail test. In this case, however, we reject the null hypothesis that the population parameter is zero if $b/s_b < t_s$.[30]

Two-Tail Tests

Occasionally theory does not suggest the direction of the relationship between the dependent and independent variables. In this case a *two-tail test* is appropriate. A good example of where this arises is found in the relationship between cohabitation and marital satisfaction. It could be argued that because cohabitation before marriage allows couples to work through various problems, a positive relationship exists between cohabitation and marital satisfaction. On the other hand, cohabitation prior to marriage may decrease marital satisfaction because couples tire of each other or because the "newness" of the relationship has worn off. Thus we are not sure whether to argue for a positive or a negative relationship between marital satisfaction and cohabitation. This is an example of an instance where a two-tail test is appropriate. In such a test, the null hypothesis is $H_0: \beta = 0$, and the alternate hypothesis is $H_a: \beta \neq 0$.

A two-tail test must consider the possibility that the estimate *over-* or *under*states β. From the previous discussion, we know that with 48 degrees of freedom, there is a 5 percent chance that an estimate overstates the population parameter by more than $1.677s_b$. Likewise, there is a 5 percent chance that it understates the parameter by more than $-1.677s_b$. Combining these statements, we can say that there is a 10 percent chance that the estimate differs either positively or negatively from the population parameter by more than $1.677s_b$. In *absolute value terms,* this means that there is a 10 percent chance that $|b - \beta| > 1.677s_b$.[31]

If β is zero, this implies that 10 percent of all possible samples will generate estimates of β that in absolute value terms are greater than $1.677s_b$. Similarly, the reader can verify from Appendix B that with 48 degrees of freedom, if β is zero, 5 percent of all samples will generate estimates of β that in absolute value terms are greater than $2.011s_b$. (Look at 48 degrees of freedom and the 2.5 percent level of significance.) More generally, when we use a two-tail test rather than a one-tail test, for a given t statistic we must double the probability of rejecting the null hypothesis when it is in fact true. In the cohabitation example (Table 2), the authors report that the coefficient on the cohabitation variable is statistically different from zero at the 1 percent level of significance. Since a two-tail test is appropriate here, the authors used a t statistic associated with the 2 percent level of significance for a one-tail test.

Table 5 presents a summary of right-tail, left-tail, and two-tail tests. The procedure for computing t ratios is also summarized. The reader is cautioned to remember that for any t statistic, the level of significance is always twice as large in a two-tail test than in a one-tail test.

Confidence Intervals

Two-tail tests are sometimes made by creating what are called *confidence intervals* rather than by using the test value method outlined here. Just as we can discuss the probability that the estimate differs from the population parameter by more than a certain amount, we can also discuss the probability of the difference being less than or equal to this value. For example, with 48 degrees of freedom, we know that 10 percent of all estimates will, in absolute value terms, differ from β by more than $1.677s_b$; 90 percent will differ by $1.677s_b$ or less. This implies that in 90 percent of the cases

$$|\beta - b| \leq 1.677s_b \qquad [11]$$

By rewriting inequality 11 to remove the absolute value signs, and by adding b to each term, it follows that there is a 90 percent probability attached to the statement

$$b - 1.677s_b \leq \beta \leq b + 1.677s_b \qquad [12]$$

Statement 12 is in the form of a confidence interval. It says that 90 percent of the intervals defined by the end points $b - 1.677s_b$ and

TABLE 5
Testing Procedure

Test Value Procedure			
Where test appropriate	Use when theory suggests population parameter is less than some specified number, β^*. β^* is often zero.	Use when theory suggests population parameter is greater than some specified number, β^*. β^* is often zero.	Use when theory does not suggest a direction between the dependent and independent variable. β^* is often zero.
Test procedure			
Step 1	Form null hypothesis: $H_o: \beta = \beta^*$ Alternate hypothesis: $H_a: \beta < \beta^*$	Form null hypothesis: $H_o: \beta = \beta^*$ Alternate hypothesis: $H_a: \beta > \beta^*$	Form null hypothesis: $H_o: \beta = \beta^*$ Alternate hypothesis: $H_a: \beta \neq \beta^*$
Step 2	Determine desired level of significance and given the degrees of freedom find the appropriate t statistic for a left-tail test.	Determine desired level of significance and given the degrees of freedom find the appropriate t statistic for a right-tail test.	Determine desired level of significance and given the degrees of freedom find the appropriate t statistic for a two-tail test.[b]
Step 3	Compute the difference $(b - \beta^*)$.	Compute the difference $(b - \beta^*)$.	Compute the difference $\lvert b - \beta^* \rvert$.
Step 4	Compute the test value $t_s s_b$.	Compute the test value $t_s s_b$.	Compute the test value $\lvert t_s s_b \rvert$.
Step 5	*Compare $(b - \beta^*)$ to test value.*	*Compare $(b - \beta^*)$ to test value.*	*Compare $\lvert b - \beta^* \rvert$ to test value*
Step 6	If $(b - \beta^*) < t_s s_b$, reject the null hypothesis in favor of the alternate. *Remember that both h sides of this expression are negative.*	If $(b - \beta^*) > t_s s_b$, reject the null hypothesis in favor of the alternate.	If $\lvert b - \beta^* \rvert > \lvert t_s s_b \rvert$, reject the null hypothesis in favor of the alternate.
t Ratio Procedure			
Where appropriate	Use when null hypothesis is that $\beta = \beta^* = 0$, altnerate is that $\beta < 0$.	Use when null hypothesis is that $\beta = \beta^* = 0$, alternate is that $\beta > 0$.	Use when null hypothesis is that $\beta = \beta^* = 0$, alternate is that $\beta \neq 0$.
Test procedure	If $b/s_b < t_s$, the null hypothesis is rejected in favor of the alternate at the appropriate level of significance. *Remember that both sides of this expression are negative.*[a]	If $b/s_b > t_s$, the null hypothesis can be rejected in favor of the alternate at the appropriate level of significance.	If $\lvert b/s_b \rvert > \lvert t_s \rvert$, the null hypothesis can be rejected at the appropriate level of significance.[b]

a. It should be noted that some researchers present t ratios as absolute values. If this is done, the ratio is always positive and should be compared to an appropriate positive-valued t statistic. In this instance, the null hypothesis is rejected if $\lvert b/s_b \rvert > t_s$.
b. Remember that for any given t statistic, the level of significance is always twice as large in a two-tail test than in a one-tail test.

$b + 1.677s_b$ will bracket the value of the population β. If instead we had wished to construct a 95 percent confidence interval, the end points of the interval would have been defined as $b - 2.011s_b$ and $b + 2.011s_b$. More generally, the form of the confidence interval is $(b - t_s s_b, b + t_s s_b)$. The test criterion is: (1) reject the null hypothesis if the null hypothesized value does not lie in the confidence interval, (2) do not reject the null hypothesis if the null hypothesized value lies in the confidence interval.

F Statistic

In the case of multiple regression analysis, there are instances when one might wish to test hypotheses about all or some subset of the regression coefficients considered simultaneously. This is especially true if the investigator finds that it is not possible to reject the null hypothesis that the individual coefficients differ from zero yet feels that, taken simultaneously, the independent variables significantly affect the dependent variable.

In multiple regression analysis an investigator anticipates that each of the independent variables included in the equation will influence the dependent variable. It is of course possible that none of the independent variables are found to be significantly related to the dependent variable. More explicitly, if there were two independent variables in the equation but, using the above techniques, neither was found to be significantly different from zero at acceptable levels of significance, we could not reject either $H_0{:}\beta_1 = 0$ or $H_0{:}\beta_2 = 0$.

Independently testing the two null hypotheses $H_0{:}\beta_1 = 0$ and $H_0{:}\beta_2 = 0$ is not the same thing as testing the null hypothesis that $H_0{:}\beta_1 = \beta_2 = 0$. The latter is a test of whether all of the coefficients taken together are simultaneously equal to zero, while the former tests whether each individually is equal to zero. In regression analysis it is possible not to reject the hypothesis that the coefficients individually are zero while at the same time rejecting the notion that simultaneously the coefficients are all zero. To fail to reject the null hypothesis that simultaneously the coefficients are zero means that there is reason to believe that the entire model is not statistically significant. The test for the simultaneous equality of all regression coefficients (or some subset thereof) equaling zero is done through the use of the *F statistic.*

One might wonder how it is possible to reject the null hypothesis $H_0{:}\beta_1 = \beta_2 = 0$ when it is not possible to reject either the null hypothesis $H_0{:}\beta_1 = 0$ or the null hypothesis $H_0{:}\beta_2 = 0$. As one explanation, consider

the following example: Suppose that in our food consumption example we had used family size and the number of children (under the age of 21) as the only two independent variables. These two variables are highly correlated.[32] As will be seen in Chapter 5 in the discussion of multicollinearity, when two independent variables are correlated, the estimated standard errors of the regression coefficients are larger than they would be in the absence of one or the other correlated independent variable. Thus we may be unable to reject the two null hypotheses $H_0:\beta_1 = 0$ and $H_0:\beta_2 = 0$. Instead of testing each coefficient separately, we could test whether, taken together, the two independent variables affect food consumption. Here the null hypothesis $H_0:\beta_1 = \beta_2 = 0$ is expected to be rejected in favor of the alternate hypothesis that one or the other of the independent variables is different from zero.

Just as hypothesis testing regarding a single regression coefficient depends on the sample data and the Student's t distribution, so the F statistic relies on the sample and a probability distribution called the F distribution. The use and interpretation of the F statistic are similar to those of the t statistic. Just as a t ratio can be computed to aid in hypothesis testing, an F ratio can also be constructed and compared to an F statistic obtainable from a table published in most statistics books (see Appendix D for a list of such books). The F ratio is related to the degree of explanatory power of the entire regression equation, since it is equal to

$$\left(\frac{R^2}{(1-R^2)}\right)\left(\frac{(N-k-1)}{k}\right) = F \qquad [13]$$

where N is the number of observations and k is the number of independent variables in the regression (excluding the intercept term).

If the F ratio is greater than the value of the F statistic, found in the table, one can reject the null hypothesis that the regression coefficients taken in combination are equal to zero. In the consumption example, the value of the F ratio is 19.66, while the F statistic is 3.19 for the 5 percent level of significance with degrees of freedom of (2, 47). (The degrees of freedom are expressed as two numbers separated by a comma. The first represents the number of coefficients being tested simultaneously, while the second is the number of observations used in the regression minus the number of regression coefficients estimated in

the multiple regression). One can thus reject the null hypothesis that $\beta_1 = \beta_2 = 0$ at the 5 percent level of significance, since $19.66 > 3.19$.

What Tests of Significance Can and Cannot Do

Before turning to Chapter 4, it is important to emphasize the strengths and limitations of the hypothesis-testing procedure. Its strength is that in the presence of randomness, the procedure allows us to draw inferences about the population parameter. Since any estimate of a population parameter is likely to have some random component, this is a substantial benefit. In this analysis we have stressed randomness due to sampling error, but other sources of randomness also exist. For example, measurement error could lead to some randomness even if one had information on the entire population (this is discussed in Chapter 5).

The weakness of the method is that researchers may forget what exactly it is they have tested. Finding that a coefficient is significantly different from zero does not imply that the corresponding variable is necessarily important. Statistical significance does not necessarily imply political, social, or economic significance. The relationship found may be so small—even though statistically significant—that the variable is of little consequence. For example, most researchers have found that persons with more education earn higher incomes. The more relevant question is how large a relationship exists. If an additional dollar spent on schooling succeeds in increasing annual income by only 2.5 cents, education may not be a valuable economic investment. To answer the question of importance, one needs some a priori idea of how big the relationship need be to justify the conclusion that education is an important determinant of income. In the education example, one might conclude that a 2.5 percent return on the investment is of little consequence, since the individual can get a substantially higher return on a dollar invested in virtually any other type of investment.[33]

4. EXTENSIONS TO THE MULTIPLE REGRESSION MODEL

In the food consumption example, observations were (hypothetically) made on a set of families at one point in time with the measured

values of family income and family size used to derive the results and test hypotheses. As the several examples drawn from various disciplines suggest, linear regression is not restricted to one form of data, nor is it limited to hypothesis testing. This chapter addresses these extensions.

Types of Data

The data used in the food consumption example are known as *cross-sectional data,* since they have been generated by a slice or cross-section of the population. A second important data form is *time series data,* in which variables are measured at different points in time. Annual or quarterly gross national product (GNP) data and national divorce rates for the past 30 years each constitute time series data sets. Several of the examples presented earlier were based on time series data.

Regression estimation techniques and interpretation of the results are exactly the same for time series data as for cross-sectional data. Consider, for example, a study of the relationship across time between imports into the United States and the level of GNP. One might hypothesize that imports into a country during a year are positively related to the country's GNP in that year. If the relationship is assumed to be linear, it can be written as

$$M_t = \alpha + \beta GNP_t$$

where M_t denotes the dollar value of imports observed in year t and GNP_t represents the level of GNP during that same year. Using the techniques discussed in previous chapters, historical values of M and GNP can be used to estimate α and β.

When studying behavior over time, it is sometimes hypothesized that the value of a variable in one time period is dependent on its value in the previous period. This is reasonable if behavior is conditioned by habits that persist over time. In such cases the previous period's value of the dependent variable can be used as an independent variable and is called a *lagged dependent variable.* For example, in the previous problem one might specify that imports in year t depend on both the level of GNP in year t and on the level of M in year t–1. That is,

$$M_t = \alpha + \beta_1 GNP_t + \beta_2 M_{t-1}$$

A more complex form of data can be created when cross-sectional information is combined over time to form *longitudinal data* sets.

Observations on a set of families over time or financial data collected from a sample of counties in the United States observed for several years would each constitute longitudinal data sets.

Longitudinal data can be analyzed in a variety of ways. If observations from only one time period are used, the data constitute a simple cross-section. Alternatively, one observational unit (e.g., a family) may be analyzed across time, thereby creating a time series analysis. Finally, researchers sometimes analyze a longitudinal database by combining all of the cross-sections into a *pooled cross-sectional analysis.* While statistical procedures needed to carry our such an analysis are more complex than the procedure outlined earlier, the underlying principles still hold.

In addition to their time dimension, data can also be classified according to the degree of aggregation across behavioral units. *Micro data* measure variables within the behavioral unit itself (e.g., the family); *aggregate data* measure behavior for a group of such behavioral units.[34] A sample of 1980 GNP data for a set of countries forms a cross-section of aggregate data, while the GNP for Mexico during the period 1950-1980 would constitute a time series of aggregate information. The food consumption example consists of micro cross-sectional data; if one observed wheat sales of an individual farm for the period 1930-1980, the resulting data set would constitute micro time series data. The form of the data does not, in general, alter the procedures nor the interpretation of results. Certain of the statistical problems discussed in Chapter 5 are, however, more frequently associated with the particular form of the data.

The R^2 statistics obtained from different types of data are likely to differ. First, since behavior is often conditioned by past actions, there is generally less randomness when a unit is observed across time than when a cross-section of units is studied. For example, the amount of driving you do this year is probably not too different from the amount you did last year. On the other hand, if one were to observe miles driven for a cross-section of individuals, the data set might contain salesmen who travel for a living and retired persons who drive only to church on Sundays. Because of this phenomenon, one will generally find higher R^2 values with time series data than with cross-sectional information.

Second, aggregate data from many firms or households hide certain differences in behavior among these units, since "high" and "low" values cancel each other. This "averaging" means that there is less variability in the dependent variable to be explained by the independent variable(s) and often results in higher R^2 values for the aggregate information than for comparable micro data.

These possible differences in the variability of the data constitute a major reason that one should not simply look at the R^2 results of studies and praise those in which the ratio is "high" while scoffing at those with low R^2 values. It is quite possible for all regression coefficients to be significantly different from zero, and yet the coefficient of determination may be very small. If testing hypotheses about the regression coefficients is the aim of the study, the coefficient of determination should be considered only as additional information, not as the summary indicator of the quality of results.

Dummy Variables

Most of the independent variables discussed thus far are *continuous variables,* since they can generally assume an infinite number of values. Often, however, *dummy independent variables* are employed in regression analysis. Such variables, sometimes called *categorical, dichotomous,* or *binary variables,* take on only the values of zero or one. The use of such a variable is appropriate whenever the theory implies that behavior differs between two different time periods (e.g., during Republican and Democratic administrations), or between two groups within a cross-section (e.g., married and unmarried individuals). In the cohabitation study (DeMaris and Leslie, 1984), a dummy variable (having cohabited) was the focus of the analysis.

In the food consumption problem, theory may lead one to hypothesize that the purchase of food differs between farm families and nonfarm families. The independent variable K can then be added to the regression equation where K takes the value of 1 if the spending unit resides on a farm and 0 if it is a nonfarm family (see Table 1, column 4). Assume that one is interested only in the effects of income, I, and farm status, K, on C. Estimates of the parameters can be derived using the techniques of multiple linear regression analysis. The results from such an analysis are $C = 742.84 + .060I - 599.16K$. The coefficient on K indicates that, based on the sample, food expenditures for farm families are estimated to be \$599.16 less than for nonfarm families *with the same income.* This can be seen by substituting the two possible values of K (0 and 1) into the estimated equation. For farm families (K = 1), the resulting equation is simply $742.84 + 0.060I - 599.16$ (or $143.68 + 0.060I$). The estimated relationships in Figure 9 illustrate that farm and nonfarm groups are assumed to respond in the same way to changes in income. That is, the regression lines have identical slopes, but the intercept term for farm families lies \$599.16 below that for nonfarm inhabitants.[35]

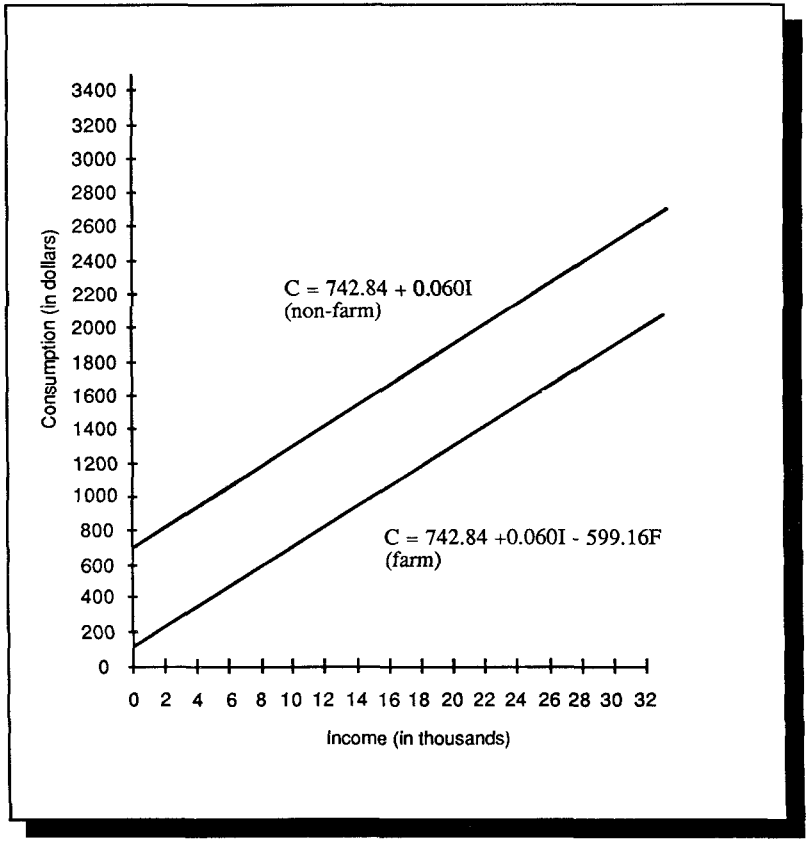

Figure 9: Farm and Nonfarm Regression Lines

At times there may be more than two mutually exclusive categories which a variable can assume. For example, the race/ethnicity of a survey respondent may be classified as white, black, Hispanic, or other. Again, dummy variables may be used to capture possible differences in the dependent variable across these groups or time periods.[36]

In such situations, *all but one* of the possible groupings of the classification variable are used as dummy variables. Thus, in the four-way grouping on race/ethnicity, three different dummy variables would be formed; one group is "excluded" and serves as a reference group against which comparisons can be made. It does not matter which group is chosen as the reference group; the implications of the results will remain the same. For example, if whites are chosen as the reference

group, then three different dummy variables—Black, Hispanic, and other—would be formed. The variable "Black" would be equal to 1 only if the respondent was Black; otherwise, it would be 0. The variable "Hispanic" would be equal to 1 if the respondent was Hispanic, 0 otherwise, and similarly for the "other" race group. The resulting equation of a dependent variable Y regressed against one continuous independent variable X and these dummy variables representing the race groups would appear as

$$Y = \alpha + \beta_1 X + \beta_2 \text{Black} + \beta_3 \text{Hispanic} + \beta_4 \text{Other}$$

Multiple linear regression analysis would yield coefficient estimates on each of the included dummy variables. The intercept term reflects the value of the dependent variable for the reference group, since for this group all the dummy variables would be equal to zero. The coefficient on each of the dummy variables is the estimate of the difference in the value of the dependent variable between the group in question and the reference group. Thus the coefficient on "Black" would estimate the difference in the dependent variable between blacks and whites (the reference group). The t ratio associated with the coefficient on a particular dummy variable can be used to test whether or not that group differs statistically from the reference group.

Interaction Variables

Another extension of the linear regression model occurs when *interaction effects* are included in an analysis. Two common types of interaction effects are interactions between a continuous variable and a dummy variable, and interaction between two continuous variables.

DUMMY INTERACTION EFFECTS

The food consumption equation used earlier assumed that, as I increases by one dollar, food consumption spending for both farm and nonfarm families will increase in an identical fashion (i.e., by about 6 cents). However, this may not always be a reasonable assumption. Dummy interaction variables allow an investigator to posit that the response to a change in a continuous independent variable differs between classified groups.

Consider again the food consumption example with income and farm/nonfarm status as independent variables. A dummy interaction term yields the model

$$C = \alpha + \beta_1 I + \beta_2 K + \beta_3 (I)(K)$$

The coefficient β_1 estimates the effect of a one-dollar change in income on food consumption for nonfarm dwellers, while for farm dwellers the estimated effect of income is $\beta_1 + \beta_3$, since K = 1 for this group. The estimate of β_3 would therefore be the differential effect of a one-dollar change in income on food expenditures between farm and nonfarm families. Using the same data but including an interaction term between the dummy variable K (farm/nonfarm residence) and the continuous variable I yields the following regression results:

$$C = 746.44 + 0.059I - 666.81K + 0.003(K)(I)$$

This implied graphical relationship between C and I is shown in Figure 10. Note that unlike Figure 9, the regression lines are not parallel when an interaction effect is included.

INTERACTION EFFECTS BETWEEN TWO CONTINUOUS VARIABLES

There are also instances in which analysts expect that two continuous variables interact in their influence on a dependent variable. One example of an interaction between two continuous variables which is certainly felt in winter is that produced by wind speed and temperatures on the "wind chill." At any temperature, increased wind speed will lower the wind chill measure; likewise, at a given wind speed, lower temperatures result in lower wind chills. The additional interaction effect means that the effect of lower temperatures on wind chill is greater at higher wind speeds.

Transformations

The previous case of interaction terms is one instance in which an independent variable has been *transformed.* Since linear regression worked well in that instance, it should not be surprising to find that other types of transformation can also be used. Probably the most common form of transformation is one that converts a nonlinear relationship between variables into a linear one.

Students of economics are probably familiar with "U-shaped" average cost curves which imply that the cost of producing a unit of output declines at low levels of output and subsequently begins to rise at higher levels. The resulting plot of average costs on the vertical axis and output on the horizontal axis result in a graph that takes on the general shape of

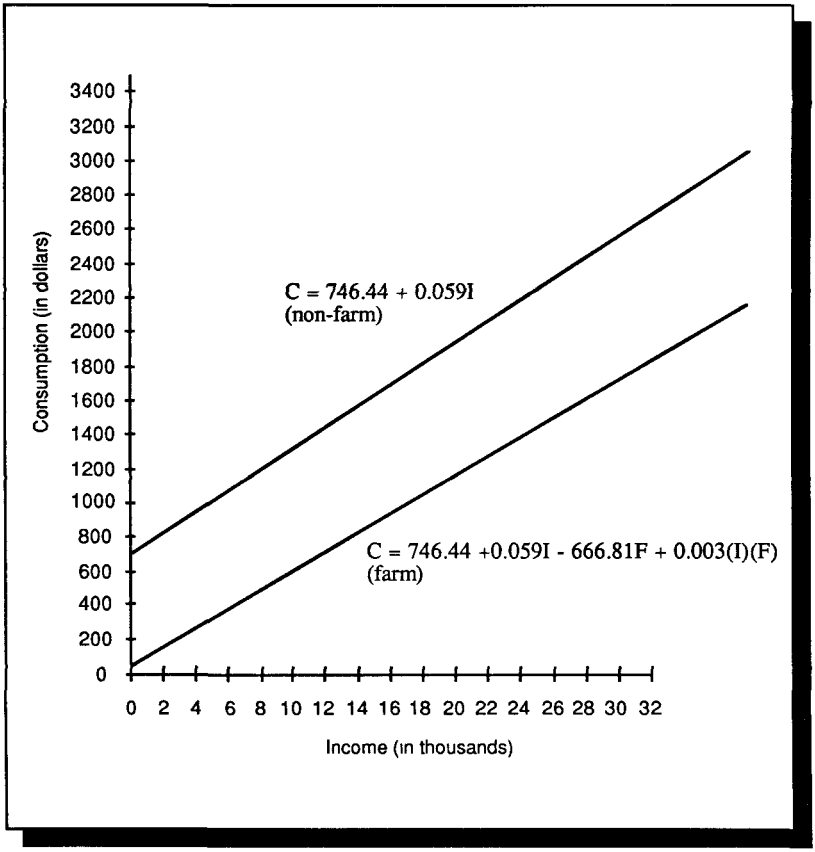

Figure 10: Farm and Nonfarm Regression Lines Allowing for Interaction

a U. Likewise, if a country's population growth rate is 2 percent a year, a plot of population against time will result in a curve that rises nonlinearly. Since the 2 percent increase in population is being applied to a larger and larger base, as time passes larger absolute annual increases in population will result. Fortunately, in many such instances linear regression analysis can be used by transforming the nonlinear relationship into an equivalent, but linear, form.

Suppose that two variables, L and M, are theorized to be related in the following nonlinear manner:

$$L = \alpha' M^{\beta} \qquad [14]$$

where α' and β are two unknown parameters.[37] It is possible to rewrite equation 14 in a linear form by taking the natural logarithm (abbreviated ln) of both sides of the equality. This yields[38]

$$\ln L = \ln \alpha' + \beta \ln M \qquad [15]$$

By redefining the terms in 15 as $Y = \ln L$, $\alpha = \ln \alpha'$, and $X = \ln M$, equation 15 can be expressed as

$$Y = \alpha + \beta X \qquad [16]$$

Since this equation is identical to the simple linear regression equation 1 in Chapter 1, the techniques discussed there will yield estimates of α and β as well as their associated statistics. If the estimated value of β in 16 is found, for example, to be -1.2, the implication is that a one-unit increase in the *natural logarithm* of M is associated with a 1.2 unit decrease in the *natural logarithm* of L. Another interpretation of the coefficient 1.2 is that for each 1 percent increase in M there is an associated 1.2 percent decrease in L.[39]

Another method of handling nonlinear relationships with linear regression is by squaring an independent variable. The resulting relationship is termed a *polynomial model,* since it results in the following polynomial equation

$$Y = \alpha + \beta_1 X + \beta_2 X^2 \qquad [17]$$

This is a particularly interesting form of a nonlinear relationship, since it suggests that the change in Y for each unit change in X depends on the value of X.[40] Such a model can be used if an analyst believes, for example, that the effect of age on a dependent variable declines as the respondent ages. Likewise, equation 17 can trace out U- or inverted U-shaped relationships between an independent and dependent variable. Hence, this function would be used if an analyst expected housing rents to increase as one moved away from the congestion of the central business district (CBD), but after some distance away from the CBD rents might begin to decline due to the costs of the long commute to work. Higher-order polynomial functions can be estimated in a similar manner.

Prediction

Besides testing hypotheses, linear regression results can also be used for purposes of predicting the value of a dependent variable for particular values of the independent variable(s). For the food consumption example, the result of a multiple linear regression using the three independent variables—income, family size, and farm or nonfarm residence—is

$$C = \$375.25 + 0.058I + 123.10S - 533.74K \qquad [18]$$

Equation 18 can be used for prediction. For example, for a farm family of five with an income of $13,000, the prediction would be $1211.01 = [$375.25 + (.058)($13,000) + ($123.10)(5) + (–$533.74)(1)].

Although regression results can be used for purposes of prediction, several aspects of this usage deserve elaboration. Regression findings may not be particularly useful for predicting values of the dependent variable, even though the results indicate that the variables are significantly related to a dependent variable. A small R^2 indicates that only a small proportion of the total variability in the dependent variable can be accounted for by the independent variables used in the equation. This suggests that numerous other unmeasured or random factors also influence the size of the dependent variable. In such instances it is heroic to predict particular values of the dependent variable on the basis of such results. Likewise, if the t ratios for the regression coefficients are quite low, one cannot have much confidence in the predicted results, since a low t ratio implies considerable uncertainty about the true population regression coefficient.

Since a set of regression coefficients is estimated from a single group of data, one should be suspicious of predictions based on extreme extrapolations from those data. For example, while one might use the food consumption results to predict Canadian food consumption, one would be ill advised to predict behavior in Cuba with these results. Likewise, predictions for the year 2010 based on data collected over the period 1960-1980 may prove to be extremely inaccurate.

An additional aspect of using regression results for forecasting is that it may require predicting values for the independent variables. Errors in estimating the values of these variables for the future will result in forecasting errors for the dependent variable, even if the model itself is perfect.

Examples

A wide variety of applications of multiple linear regression using different types of data and alternative forms of variables are available in the literature. Here we consider only three to demonstrate this range of applicability.

EXAMPLE 1—COMPUTER LITERACY

What factors influence "computer literacy"? This is the question addressed by Lockheed, Nielson, and Stone (1985) in a study which evaluated 413 New Jersey high school students enrolled in a computer course. A "pretest" was given at the first meeting of the course in order for increases in knowledge to be measured. Pretests also allow investigators to standardize for differentials in knowledge at the outset of an educational experience.

Multiple regression techniques were used to analyze the difference between the final test score and the pretest, also termed the "gain score." Table 6 reports on two different specifications of the gain-score determinations of ninth- and tenth-grade students. The results suggest that those with higher pretest scores had higher gains in competency, that females had smaller increments to their scores than did similar male students, and that being in an accelerated math class had a positive influence on test scores. In the second specification, access to an outside computer was also found to affect test scores significantly, while the other variables were not different from 0 at a 5 percent level of significance.

EXAMPLE 2—SEASONALITY IN FERTILITY

Birth rates in the United States have consistently been higher in September (conception in December) than in May (conception in August). This has led some demographers to hypothesize that the weather influences the frequency of conception and hence the monthly birth rate.

Seiver (1985) examined monthly birth rates in the United States for the period 1947-1976 and discovered that there was a significant reduction over time in the magnitude of the April-May "trough" (in birth rates). He then analyzed this change in the May seasonal effect using cross-section regression techniques based on the change in birth rates in the various states between 1960 and 1970. He hypothesized that changes

TABLE 6
Determinants of Gain Scores in Computer Literacy Among New Jersey 9th- and 10th-Grade Students

	Model[a]	
Variable	*1*	*2*
Intercept	7.248	5.826
Pretest score	0.345	0.295
	(4.00)	(3.38)
Female[b]	−0.805	−1.020
	(2.38)	(2.41)
In accelerated math class[b]	2.128	2.120
	(5.61)	(5.62)
Access to computer outside class[b]		0.978
		(2.53)
Use school computer only in class[b]		0.499
		(1.09)
Play computer games[c]		0.126
		(0.75)
Ask teacher for help[d]		0.229
		(1.66)
R^2	.246	.280
$\bar{R}^2$	.236	.258
Number of observations = 231		

SOURCE: Lockheed, Nielson, and Stone (1985). Reprinted by permission.
a. Numbers in parentheses are absolute values of t ratios.
b. A dummy variable set equal to one if respondent had this attribute, zero otherwise.
c. A 5-point scale of frequency of playing computer games (1 = never, 5 = several times a week).
d. A 5-point scale of frequency of asking teacher for help (1 = never, 5 = several times a day).

in the state labor force participation rates of women between 1960 and 1970 (LFP), changes in median family income during the decade (INC), changes in the proportion of high school graduates in a state between 1960 and 1970 (HS), and the increase in air conditioning use in a state (AC) would all affect the change in the May birth rate. His resulting regression equation was

$$\text{MAY BIRTH} = 0.19 - 1.74\text{LFP} + 0.014\text{INC} - 0.18\text{HS} + 1.76\text{AC}$$
$$(0.38)\ (1.48) \qquad (0.046) \qquad (1.63) \qquad (0.25)$$

where the numbers in parentheses are the estimated standard errors of the regression coefficients. An examination of these results suggests that the only variable found to be statistically significant was the use of air conditioning. Apparently, as air conditioning has provided a more pleasant environment during the summer months, April-May births have tended to be more in line with other months.

EXAMPLE 3—EFFECTS OF AUTOMOBILE SAFETY STANDARDS

The effectiveness of public policies can sometimes be evaluated through regression analysis. The National Traffic and Motor Vehicle Safety Act of 1966 required certain safety features such as padded dashboards and head restraints to be installed in all new vehicles. Graham and Garber (1984) analyzed the effects on highway death rates of this legislation. Since such legislation did not affect older vehicles, they used the proportion of miles driven by regulated cars as a primary independent variable to explain annual auto, truck, and bus death rates in the United States during the period 1947-1980. Numerous other variables were used as well in their regression equation, including time to account for the long-term historical decrease in highway deaths per million miles of vehicle travel.

Table 7 reproduces one set of Graham and Garber's results which indicate that, indeed, the Vehicle Safety Act did significantly reduce death rates on the highway. To show this even more clearly, the authors also "predicted" what traffic death rates would have been during the period after 1966 if all other variables had remained unaltered but there had been no safety regulation. They conclude that the act reduced the death rate by 19-29 percent. Such counterfactual analysis provides a convenient method for summarizing the effects of particular independent variables on a dependent variable.

5. PROBLEMS AND ISSUES OF LINEAR REGRESSION

The advent of the computer and numerous computer packages has made linear regression analysis accessible to nearly everyone. The use of such computer packages is normally very easy; however, their purely mechanical application is not appropriate. Although the preceding

TABLE 7
Regression Estimates for Death Rate Equation

Independent Variable	*Regression Coefficient (t ratio in parentheses)*
Intercept	−2.84
	(−0.85)
Proportion of miles driven by regulated cars	−1.31
	(−2.36)
Average speed on main rural highways	0.02
	(0.43)
Per capita alcohol consumption by adults	1.38
	(1.51)
Proportion of licensed drivers under age 25	12.20
	(0.99)
Proportion of miles driven by trucks	12.93
	(3.50)
Proportion of miles by compacts and subcompacts	0.74
	(0.36)
Proportion vehicles in no-fault states	−0.93
	(−0.73)
Cost of accident index	0.14
	(0.86)
Real earned income per working age adult	0.59
	(2.14)
Percentage urbanized	−2.78
	(−1.09)
Time	−0.16
	(−6.14)
$R^2 = 0.981$	
Number of observations = 34	

SOURCE: Graham and Garber (1984). Reprinted by permission.

discussion may seem to suggest that regression analysis is a straightforward exercise without pitfalls, unfortunately this is not the case.

Regression analysis, especially hypothesis testing, is based on several important assumptions. Among them are (1) that the correct equation is being used—that is, the proper variables were included as independent variables and the proper functional form was used; (2) that the variables are measured accurately; (3) that the independent variables are independent of each other; (4) that the data constitute a random sample; and (5) that the residual error term is "well-behaved." (Recall that the residual error term refers to the difference between the observed value of the dependent variable and its value as predicted from the estimated regression equation.) Difficulties arise in regression analysis when any of these

assumptions are violated. The computer packages do not automatically solve these difficulties; it is up to the researcher to handle them.

Many analysts recognize the shortcomings of linear regression and often attempt to overcome the resulting problems. This final chapter addresses some of the more common problems associated with linear regression, the implications each problem has on the outcome, and some of the methods that analysts use to circumvent the difficulties. We begin with the issue that faces any analyst—specification of the model. After examining the issues associated with data, we discuss various problems related to the form of the error term in the regression equation. Appendix D contains a list of books that discuss regression analysis. Some of the more advanced books discuss the issues presented in this chapter.

Specification

Although all of the issues discussed in this chapter are, loosely speaking, associated with specification, we limit ourselves to the specification problems that analysts face when deciding which variables to include and exclude in a regression equation and the functional form of that equation. Unfortunately, omitting a relevant variable, even an irrelevant variable, or using an improper functional form can produce undesirable effects on the results.

OMITTING A RELEVANT VARIABLE

When a variable is omitted from a regression equation, the regression coefficients on the included variables will, in general, be unreliable or invalid, since they will be "biased" estimates of the true population regression coefficients.[41] While this conclusion stems from statistical theory and is not proven here, the idea underlying this result is intuitively plausible. Suppose that two variables, income and family size, are the sole determinants of food consumption and that all other variability in food purchases across families is purely a random occurrence. If the analyst uses only income to explain variability in food consumption and if income and family size are correlated, the estimated coefficient on income will reflect the effects of both income and family size on food purchases. This is why the addition of family size to the food consumption equation altered the estimated effect of a change in income from 5.8¢ to 5.6¢.

Since the task of regression analysis is to estimate the response in the dependent variable to changes in an independent variable, an incorrect estimate of that response may be serious. Unfortunately, there is little an

analyst can do to detect whether an important variable has been left out of the equation. Because of the uncertainty regarding omitted variables, researchers often include results from two or more different specifications of the same phenomenon. If, under alternative specifications, there is little change in the size of the estimated coefficients, the estimates are said to be *robust*. Such experimentation strengthens the analyst's belief in the model used; even then one can never be absolutely certain that a relevant variable has not been omitted.

INCLUSION OF AN IRRELEVANT VARIABLE

One might think that, since omitting a relevant variable is "bad," the solution to the problem is to throw every available variable into the equation. Of course, this solution also has pitfalls. If a variable is included in the equation but is not in fact relevant, the estimates of the coefficients will be unbiased. However, if the irrelevant variable is correlated with the included relevant variables, the size of the estimated standard errors of the coefficients of the relevant variables will increase. This in turn means that the ratios will be smaller than if the correct specification were used. Hence, the analyst is more likely to conclude that the coefficient on a relevant variable is not significantly different from zero, (i.e., the researcher will not be able to reject the null hypothesis that there is no association with the dependent variable). Thus, adding unnecessary variables causes a loss in precision of the estimated coefficients on the relevant variables.

INCORRECT FUNCTIONAL FORM

In the previous section it was shown that least squares linear regression is not restricted to simple linear relationships among variables. There are, in fact, myriad possible functional forms that are amenable to estimation using least squares techniques. The issue is which form to use.

If the underlying relationship between variables is actually nonlinear but a linear function is estimated, the resulting coefficient will be biased. Consider the previously mentioned case of the value of houses located at different distances from the central business district. It is reasonable to expect that the relationship would be nonlinear. If an analyst simply estimated the linear function, Value = $\alpha + \beta$ Distance, the estimated coefficient on distance might be very close to zero and would suggest a flat value-distance relationship. Such an estimate of β would be a biased

or misleading indicator of the functional relationship between value and distance.

One way in which nonlinearities may be detected is to plot the residual error (the difference between the actual value of the dependent variable and its value as estimated from the equation). If there are large negative (positive) residuals at low and high values of an independent variable and large positive (negative) residuals at intermediate levels of the independent variable, a nonlinear relationship is suggested.

STEPWISE REGRESSION

Since decisions regarding which of numerous possible variables to include in a regression equation are difficult, *stepwise regression* techniques are sometimes used. These techniques allow the computer to experiment with different combinations of independent variables.

In one method of stepwise regression, the compuer first estimates simple linear regressions using each of all the possible independent variables specified by the analyst. For example, if there were 20 possible independent variables, the computer program would estimate 20 different simple linear regressions. From the set of 20 results the program would choose which one is "best." This selection, which is a part of the computer program, usually relies on the coefficient of determination, R^2.

In step 2 the program would try each of the 19 remaining independent variables together with the variable chosen in step 1 and produce 19 different regression results, each with two independent variables. Again, the rule regarding which of these 19 is "best" would be invoked and results from this second step would be printed. This process continues until either all 20 variables are included in the equation or no remaining variable increases the R^2 statistic sufficiently to permit the inclusion of additional variables.

Although R^2 statistics can be tested using an F distribution (see equation 13 in Chapter 3), it should be recognized that changes in R^2 attributable to any particular variable usually depend on what variables are already in the equation. For example, when income alone is used in the food consumption example, the R^2 is 0.307. If family size is the sole regressor used to explain food spending, the R^2 is 0.170; adding income as a second regressor increases the R^2 to 0.456. This second approach would suggest that income explains only an additional 29 percent (= 0.456 – 0.170) of the total variability in food consumption, rather than 31 percent as indicated above. Incremental changes in R^2 values should

therefore be interpreted in terms of which other variables have already been included in the model. Without careful thought, stepwise regression analysis can turn into a fishing expedition that is void of theory.

In summary, specification is one of the most perplexing problems faced by most analysts. Misspecification can produce misleading or imprecise results. Furthermore, computational techniques relying heavily on computers and devoid of theory do not provide the solution. It is still innovative thought and theory that must be relied on most to surmount problems.

Proxy Variables and Measurement Error

While theorizing about appropriate variables is not always easy, actually observing some variables and measuring them accurately can be equally difficult. Appropriate data are often not available. In such cases analysts often turn to alternative, second-best measures of the phenomenon at hand. The variables chosen are termed *proxy variables* since they are being used to approximate the real thing. The degree of approximation will influence the estimated impact of the variable of actual interest.

There are many examples of uses of proxy variables in the literature. Whenever dummy variables are substituted for what is really a continuous variable, a proxy is being used. For example, some analysts of political behavior may theorize that the "liberalism" of the president affects particular types of behavior, but, in the absence of a direct measure of liberal tendencies, they use a dummy variable set equal to 1 if the president is a Democrat and 0 if a Republican.

Attitudes are seldom easy to measure directly. For that reason, numerous *scaling variables* have been developed which are constructed from responses to attitudinal surveys. Examples of such scales are found in DeMaris and Leslie's (1984) study of cohabitation where the dependent variable, level of marital satisfaction, was constructed from questions asked of the respondents.

Variables that are available are often substituted for unobserved variables. For example, even though theory may suggest that work experience influences wages, experience may not be available in a data set. In such instances researchers often substitute age under the assumption that the older the worker, the greater his or her work history. This measure, or a derivative thereof (such as age less the years of education less five), may be reasonably accurate for males with continuous labor market experiences. It is, however, less accurate in cases where individ-

uals, especially women, have had discontinuous formal labor market work histories.

Use of imperfect proxy variables can introduce errors of measurement into the analysis. Another form of measurement error is simply mismeasurement of the variables that are available. For example, respondents to a survey may deliberately understate their age or not report accurately the candidate for whom they voted. Measurement error can also occur if survey questions are asked in an ambiguous way.

Measurement errors can result in biased estimates of regression coefficients. Sometimes these errors can be avoided through more accurate data collection procedures; however, when analysts use data collected by others, it is unlikely that much can be done to improve the quality of the numbers. Instead, cognizance should be taken of the probable measurement errors and how systematic over- or underreporting of either the independent or dependent variables might influence the estimated coefficients.

Selection Bias

There are instances in which, even though every variable is measured accurately, the nature of the sample is such that the observations are for a nonrepresentative sample of the population. All results based on questionnaires that can be completed by anyone who is willing to put forth the effort are potentially nonrepresentative, since the participants have been self-selected. Similarly, when studying women's wages, women not in the labor force are systematically excluded from the analysis. In such a case the results of the regression analysis cannot readily be used to predict the wage that a woman currently not working could get is she were to get a job. This is because there is likely some systematic difference between women who are working in the labor market and those who are not working for wages. Any regression based only on the former group will not capture this influence. If the regression results from the *censored sample* (working women) are to be used to make inferences about all women, it is necessary to adjust for the selection bias that exists.

Multicollinearity

A final problem associated with data used in a regression is multicollinearity. It arises whenever two or more independent variables used in a regression are not independent but are correlated. Unfortunately, in the social sciences this problem arises often, since many socioeconomic

variables such as education, social status, political preference, income, and wealth are likely to be interrelated. Time series data are also likely to exhibit multicollinearity. Many economic series tend to move in the same direction (e.g., production, income, and employment data).

When two or more independent variables are correlated, the statistical estimation techniques discussed earlier are incapable of sorting out the independent effects of each on the dependent variable. For example, New York State imposed a mandatory seat belt law at about the same time that law enforcement agencies in the state cracked down on drunken drivers. For this reason, any subsequent decline in auto fatalities cannot be attributed exclusively to either one or the other of these policy decisions.

While regression coefficients estimated using correlated independent variables are unbiased, they tend to have larger standard errors than they would have in the absence of multicollinearity. This in turn means that the t ratios will be smaller. Thus it is more likely that one will find the regression coefficients not to be significant than in the case where no multicollinearity plagues the data. In essence, there is less precision associated with estimated coefficients.

Multicollinearity is probably present in all regression analysis, since the independent variables are unlikely to be totally uncorrelated. Thus whether or not multicollinearity is a problem depends on the degree of collinearity. The difficulty is that there is no statistical test that can determine whether or not it really is a problem. One method to search for the problem is to look for "high" correlation coefficients between the variables included in a regression equation. Even then, however, this approach is not foolproof, since multicollinearity also exists if linear combinations of variables are used in a regression equation.[42] There is no single preferable technique for overcoming multicollinearity, since the problem is due to the form of the data. If two variables are measuring the same thing, however, one of the variables is often dropped, since little information is lost by doing so.

Autocorrelation

Measurement errors, selection bias, and multicollinearity are all attributable to the data available to a researcher. The next set of issues pertains to assumptions regarding the residual error term.

Recall that the residual error term is the difference in the observed value of the dependent variable for the ith observation, Y_i, and the value

of the dependent variable predicted from the estimated regression for the ith observation, $\hat{Y}_i$. The discussion of regression analysis in Chapters 1 and 2 is based on the *ordinary least squares* (OLS) *regression model.* This model assumes that (1) even though some errors are small and others are large, some are positive and others are negative, they have a mean of zero; (2) the error term associated with one observation is uncorrelated with the error term associated with all other observations; (3) while some of the error terms may be small and others large, the variability of the error terms is in no way related to the independent variables used; and (4) the error term is not correlated with the independent variables. Violations of any of these assumptions produce undesirable properties in the results obtained when regression coefficients are estimated without regard for these assumptions. While a full discussion of all these topics is beyond our purpose here, it is useful to review the most common problems that arise in the course of regression analysis and to indicate the steps that analysts take in response to these problems.

The first of these issues is termed *autocorrelation* or *serial correlation.* Autocorrelation refers to the case in which the residual errors terms from different observations are correlated. If the terms are positively correlated, *positive autocorrelation* is said to exist, while if they are negatively correlated, *negative autocorrelation* is present.

Autocorrelation and the problems it presents are more likely to appear with time series data, and most commonly the problem is restricted to error terms associated with successive time periods. To illustrate a stylized example of positive autocorrelation, consider a hypothetical time series regression analysis of total spending by a school district, Ed, as a function of the personal income of residents in the district. The data in Table 8 are constructed for 12 consecutive years, with column 2 indicating the actual level of Ed in each year, Ed_t, and column 3 showing the level of predicted Ed, $\hat{Ed}_t$, found from a regression equation. Column 4 is the simple difference between columns 2 and 3, $Ed_t - \hat{Ed}_t$. The entries in column 5 are the values from 4 but lagged one year. That is, the value in column 5 for year two is the same as in column 4 in year one, and so on through the entire set of data.

The residuals and their associated lagged values are plotted in Figure 11. One observes that the pattern of residuals is the same as would be found for any two positively correlated variables. Positive autocorrelation thus means that there is positive correlation between successive error terms. For negative autocorrelation, just the opposite holds; thus,

TABLE 8

Error Terms in the Case of Positive Autocorrelation

(1)	*(2)* *Observed Value*	*(3)* *Predicted Value*	*(4)* *Residual*	*(5)* *Lagged Residual*
Period	*Ed*	$\hat{Ed}_t$	$Ed_t - \hat{Ed}_t$	$Ed_{t-1} - \hat{Ed}_{t-1}$
1	110	105	5	–
2	115	108	7	5
3	126	120	6	7
4	129	126	3	6
5	129	127	2	3
6	130	131	−1	2
7	133	136	−3	−1
8	137	142	−5	−3
9	149	150	−1	−5
10	155	155	0	−1
11	166	162	4	0
12	176	171	5	4

SOURCE: Hypothetical data.

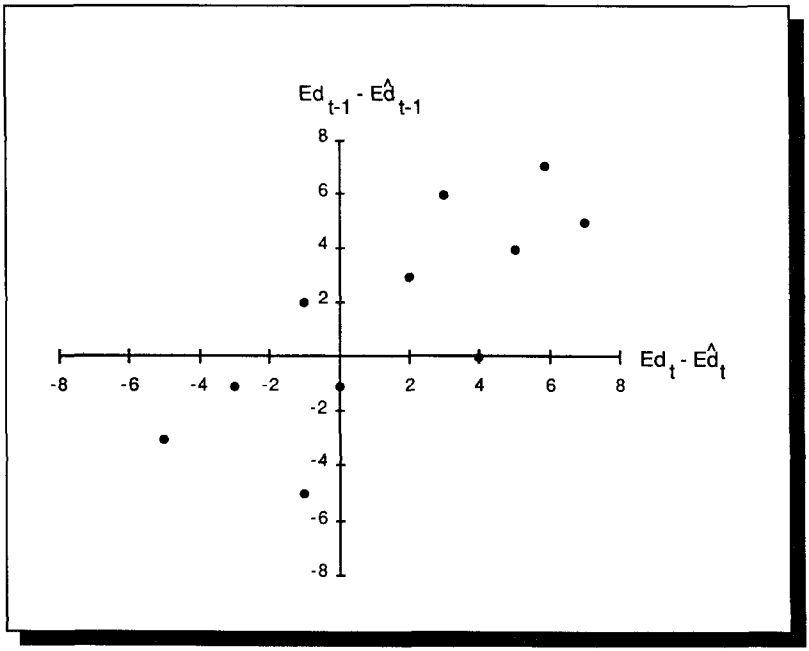

Figure 11: Plot of Residuals and Lagged Residuals with Positive Autocorrelation

if the error is positive in one observation, it is highly likely that it will be negative in the adjacent observation.[43]

Autocorrelation can be caused by several factors, including omission of an important explanatory variable or the use of an incorrect functional form. It may also simply be due to the tendency of effects to persist over time or for dependent variables to behave cyclically. Whatever the cause, autocorrelation influences the outcome of the hypothesis-testing procedure. The effect of positive autocorrelation is underestimation of the standard error of the estimated coefficient, s_b. This in turn yields an inflated t ratio, which means that it is possible that coefficients will be found to be significantly different from zero when in fact they are not.

While simply looking at the residual terms may provide some clue to the existence of autocorrelation, many authors report a test statistic called the *Durbin-Watson coefficient,* especially when time series data are being analyzed. This coefficient can be used to test the null hypothesis that successive error terms are not autocorrelated.

When serially correlated error terms are detected, there are special techniques available to circumvent the problem. Many analysts use a technique called *generalized least squares* (GLS) regression to overcome the problem. This method is based on ordinary least squares regression techniques but uses variables that have been transformed.

Heteroskedasticity

Heteroskedasticity refers to another nonrandom pattern in the residual error term. Assumption (3) in the discussion of the OLS regression model is that the variability in the error term does not depend on any factor included in the analysis. This assumption is known as the assumption of *homoskedastic errors;* when it is violated, heteroskedasticity is said to exist. The problem arises most frequently in the analysis of cross-sectional data.

Consider the relationship between the number of employees in an organization and the number of supervisors. One might specify that the number of supervisors is a function of the number of employees. While a general positive relationship will probably be found (i.e., organizations with larger labor forces have greater numbers of supervisors), it may also be the case that some large organizations have numerous supervisors whereas other large organizations have relatively few. Such a situation is sketched in Figure 12, where the variability in the residual

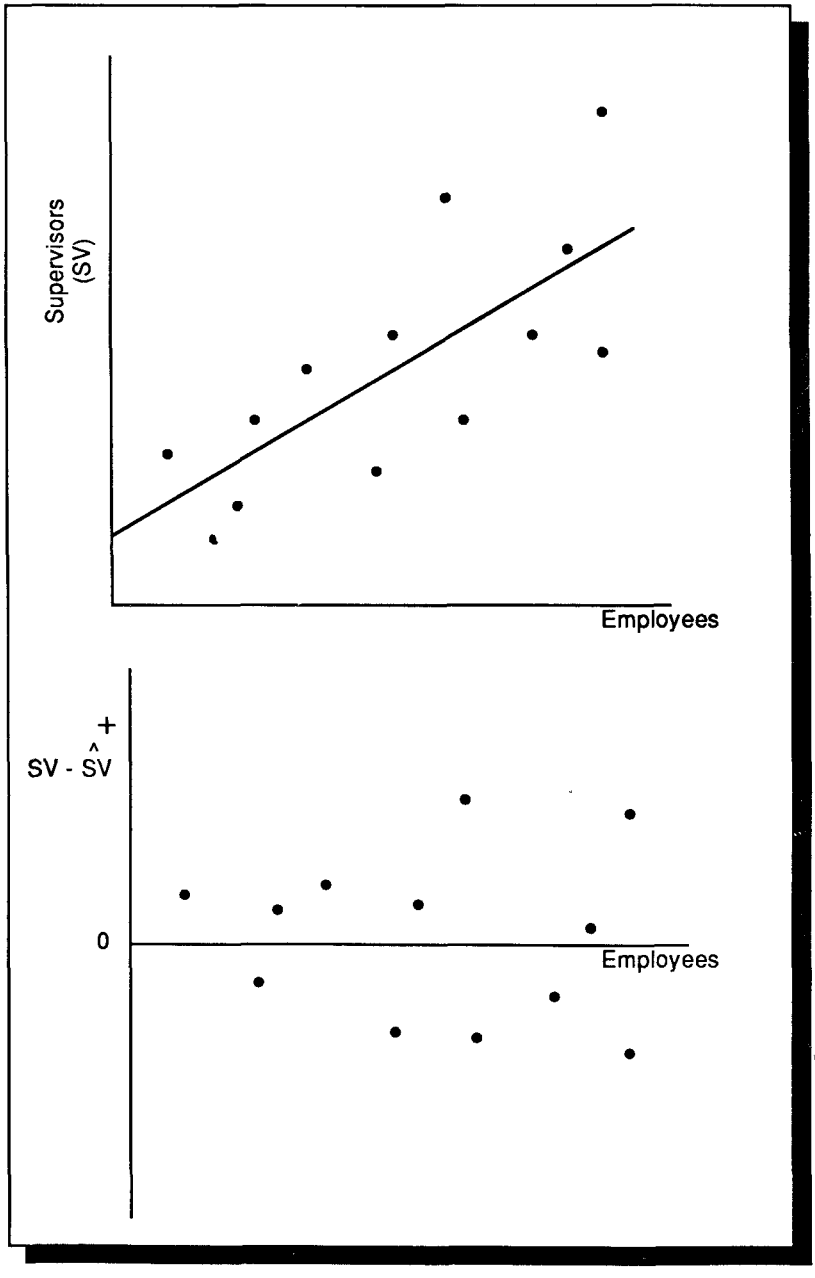

Figure 12: Case of Heteroskedasticity

error terms is not constant for all values of the independent variables. The residuals are said to be *heteroskedastic.*

As with autocorrelation, heteroskedasticity affects the size of the standard error of the regression coefficient, thereby biasing hypothesis-test results. The effect on s_b will depend on the exact manner in which the heteroskedasticity was formed. Several different tests are available for detecting the problem of heteroskedasticity. All depend on an examination of the residuals. Again, when the problem is detected, generalized least squares can be used to give differential weights to the observations and thereby circumvent its effects on tests of hypothesis.

Simultaneous Equations

As noted earlier, the linear regression model assumes that the residuals are purely random variables. Contemporaneous correlation arises when the residual and the independent variable(s) are correlated.[44] This problem can arise for a variety of reasons, but it most commonly occurs when simultaneous phenomena are under investigation.

Even though we warned that causality is never proven by regression analysis, when a researcher specifies that Y = f(X), an implicit causal linkage is assumed. In general, this functional relationship runs from X to Y (i.e., the value of Y is dependent on the value of X). But in many situations the dependency may run both ways (i.e., X is also a function of Y).

One common example of a simultaneous process occurs in the area of criminology. Those cities with higher crime rates are likely to put more resources into crime fighting: Police = f(Crime). At the same time, if police protection is effective, the crime rate should be decreased: Crime = f(Police). In these instances, simple linear regression will yield biased estimates of the phenomenon under investigation.

The solution to simultaneity is rather complex and well beyond the scope of this book. It is worthwhile, however, to consider briefly two primary issues that are commonly mentioned by analysts investigating simultaneous phenomena: identification and estimation.

Although a variety of methods for estimating simultaneous relationships are available, these techniques require that the coefficients of the model be mathematically obtainable or *identifiable.* Consider the case of the price and quantity of wheat sold each year. Economic theory holds that market prices and quantities are determined by the simultaneous actions of suppliers and demanders. One could easily obtain data on the total wheat marketed each year in the United States and the

annual price of wheat for the last several decades. One might then estimate the equation, Quantity = $\alpha + \beta$ Price. While one would obtain estimates of the parameters, there is no way of knowing or identifying whether the estimated relationship is the demand or the supply relationship, since quantity and price are involved in both relationships.

In order for identification to be possible in such a case, the model must be expanded in some manner. For example, one might argue that quantity demanded (Q_d) is a function of price (P), income (I), and the U.S. population (Pop), while quantity supplied (Q_s) is a function of price and the cost of producing wheat. This would yield the following simultaneous equation model, in which the coefficients are identifiable:

$$Q_D = \alpha_1 + \beta_1 P + \beta_2 I + \beta_3 POP$$

$$Q_S = \alpha_2 + \beta_4 P + \beta_5 Cost$$

A variety of techniques are available to investigators when models are identifiable. Most of the methods also yield associated statistics similar to those discussed earlier so that hypothesis tests can also be performed. One commonly used method is called *two-stage least squares,* a technique highly regarded because of its simplicity, ease of computation, and fairly desirable statistical properties. Other estimation techniques include *three-stage least squares* and *maximum likelihood* methods. Advanced study of statistics is required, however, for an understanding of these techniques. Nevertheless, the methods are applied in a variety of circumstances, including estimation of multiple equation macroeconomic models which are used to forecast the course of the economy.

EXAMPLE 1—EXPENDITURES ON POLICE

It was noted earlier that criminal behavior and policies concerning police protection can be considered as another example of a simultaneous model. This approach was taken in a recent study of determinants of police expenditures in a cross-section of 79 cities with populations greater than 100,000 (Bahl, Gustely, and Wasylenko, 1978). A three-equation model was built. One equation specified that total police spending depends on, among other factors, the level of employment in the local police department. A second equation specified that police employment depends on several variables, including the crime rate in the city. Finally, a "crime" equation was specified which hypothesized

that criminal activity depends on such socioeconomic factors as the local unemployment rate, the average prison sentence served in the state in which the city is located, as well as the level of police employment.

Two-stage least squares regression was used to estimate the model, with all continuous variables expressed as logarithms. Because of this functional form, the resulting coefficients are elasticity estimates (percentage change in the dependent variable relative to a percentage change in the independent variable). The resulting elasticity of police employment with respect to the crime rate was 0.378, implying that cities with higher crime rates employed greater numbers of police; the elasticity of the crime rate with respect to police employment was found to be –0.231, suggesting that cities with greater police employment had lower crime rates.

Limited Dependent Variables

Analysts often wish to study behavior which is observed only as a binary indicator. Examples include whether or not a person is in the labor force, whether or not an applicant was admitted to a university, or whether or not an otherwise qualified voter is registered to vote.

In Chapter 4 we reviewed the use of 0-1 dummy variables as independent variables in regression analysis. While such dichotomous indicators are appropriate as explanatory variables, ordinary least squares regression analysis is not appropriate when a 0-1 or other limited choice variable is the dependent variable.

Several problems arise in the case of a 0-1 dependent variable which make the ordinary least squares regression inappropriate. Consider the simple case of a model which specifies that higher income persons have a greater probability to be registered as Republican voters. The dependent variable, R, in the model would then be equal to 1 if the person were a Republican and 0 if not. One might estimate the equation, $R = \alpha + \beta I$, using least squares regression analysis. (You might sketch a two-dimensional graph showing I on the horizontal axis and the 0-1 variable R on the vertical axis.)

While the techniques of Chapter 1 would yield estimates of α and β, several problems can arise. First, it is possible that for certain values of I (together with the estimated a and b) the predicted value of R would be either less than zero or greater than one. But since $\hat{R}$ can be interpreted as the probability of being a Republican, such values do not make sense. Second, the variability of residuals obtained from such an estimation

will depend on the size of the independent variable, suggesting that heteroskedasticity is a problem here. Finally, while we have not stressed it, the theory that underlies the hypothesis-testing procedure is based on the assumption of normally distributed residuals, which is certainly not the case in this instance.

While ordinary least squares regression is inappropriate in such instances, nonlinear estimation techniques have been developed to overcome the major statistical difficulties outlined earlier. Two techniques are most commonly used in such instances. One is called *probit analysis,* while the second is termed *logit analysis.* The primary theoretical difference between the two concerns the probability distributions that underlie the process being analyzed. Nevertheless, each is capable (after some manipulation of the results) of providing estimates of the effect of unit changes in the independent variable(s) on the probability of an event.[45]

For completeness, we should also mention that recent developments in econometrics have extended the special models to allow for analysis of situations in which there are a small number of mutually exclusive outcomes in a choice process (e.g., choice of college). In addition, some analysts have used another special technique, termed *Tobit analysis,* when faced with a situation in which many participants in a choice process choose a zero outcome while others choose some positive number that is unlimited in size. An example of such a case is the amount of money a family spends on new car purchases in one year. Many families buy no new car at all (Y = 0), while others make purchases anywhere in the range from, say, $5,000 to $35,000.

Conclusions

Linear regression provides a powerful method for analyzing a wide variety of behavioral situations. At the same time, this technique relies on a set of assumptions that may or may not hold in different applications. As this technique becomes more widely known, and as computational facilities become more accessible through the use of computers, we anticipate that the use of linear regression analysis will become even more widespread. A few hours spent perusing journals within a particular social science discipline will reveal that the use of this statistical technique has increased greatly over time. While the problems of forecasting the future based on the past have already been discussed, our own a priori expectations are that regression's importance as an analytical technique will increase in the foreseeable future.

APPENDIX A:
DERIVATION OF a AND b

The purpose of this appendix is to show how to obtain the values of a and b that minimize the sum of squared error term SSE. From Chapter 1, the sum of the squared errors is given by

$$SSE = \Sigma(C_i - a - bI_i)^2 \qquad [A1]$$

where Σ implies summation from i = 1 to N. The values of a and b that minimize equation A1 are found by taking the partial derivative of SSE with respect to a and b and setting the resulting derivatives equal to zero. This yields

$$\partial SSE/\partial a = (-2)\Sigma(C_i - a - bI_i) = 0 \qquad [A2]$$

$$\partial SSE/\partial b = (-2)\Sigma[(I_i)(C_i - a - bI_i)] = 0 \qquad [A3]$$

Dividing through both A2 and A3 by –2 and rearranging term yields

$$\Sigma C_i = aN + b\Sigma I_i \qquad [A4]$$

$$\Sigma(C_iI_i) = a\Sigma I_i + b\Sigma(I_i^2) \qquad [A5]$$

Equations A4 and A5 are in the standard form of the normal equations for a straight line. The terms ΣC_i, ΣI_i, $\Sigma(I_iC_i)$, and $\Sigma(I_i^2)$ can be computed from the data set. Equations A4 and A5 can then be solved simultaneously for a and b. The resulting values of a and b minimize SSE.

Equations A4 and A5 can also be solved to obtain formulas for the values of a and b. The formula for b is

$$b = \Sigma[(I_i - \bar{I})(C_i - \bar{C})] \;/\; \Sigma(I_i - \bar{I})^2 \qquad [A6]$$

where C and I represent the means of C and I, respectively.

Once b is known, a can be obtained by using the expression

$$\bar{C} = a + b\bar{I} \qquad [A7]$$

which is obtained by dividing equation A4 by N. The derivation of equation A6 is tedious but not difficult and is presented in most statistics books. Notice that equation A7 says that the regression line passes through the point defined by the mean values of C and I.

Equations A4 and A5 can be used to obtain estimates of a and b for the food consumption problem. From the data in Table 1, the following values can be obtained:

$\Sigma I_t = 969{,}984$

$\Sigma C_t = 92{,}122.45$

$\Sigma(I_t^2) = 20{,}813{,}307{,}472$

$\Sigma(C_t I_t) = 1{,}903{,}186{,}495.00$

$N = 50$

Substituting these values into equations A4 and A5 yields

$$92{,}122.45 = a(50) + b(969{,}984) \qquad [A4']$$

$$1{,}903{,}186{,}495.00 = 969{,}984 + b(20{,}813{,}307{,}472) \qquad [A5']$$

These two equations are then solved simultaneously to yield a = 714.58 and b = 0.058.

APPENDIX B:
CRITICAL VALUES FOR STUDENT'S t DISTRIBUTION
Level of Significance (percentage) Values for Right-Tail Test[a]

Degrees of Freedom	*10%*	*5%*	*2.5%*	*1%*	*.5%*
1	3.0777	6.3138	12.7062	31.8207	63.6574
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0322
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1315	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5177	2.8314

APPENDIX B (Continued)

Degrees of Freedom	*10%*	*5%*	*2.5%*	*1%*	*.5%*
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.3163	1.7081	2.0595	2.4851	2.7874
26	1.3150	1.7056	2.0555	2.4786	2.7787
27	1.3137	1.7033	2.0518	2.4727	2.7707
28	1.3125	1.7011	2.0484	2.4671	2.7633
29	1.3114	1.6991	2.0452	2.4620	2.7564
30	1.3104	1.6973	2.0423	2.4573	2.7500
31	1.3095	1.6955	2.0395	2.4528	2.7440
32	1.3086	1.6939	2.0369	2.4487	2.7385
33	1.3077	1.6924	2.0345	2.4448	2.7333
34	1.3070	1.6909	2.0322	2.4411	2.7284
35	1.3062	1.6896	2.0301	2.4377	2.7238
36	1.3055	1.6883	2.0281	2.4345	2.7195
37	1.3049	1.6871	2.0262	2.4314	2.7154
38	1.3042	1.6860	2.0244	2.4286	2.7116
39	1.3036	1.6849	2.0227	2.4258	2.7079
40	1.3031	1.6839	2.0211	2.4233	2.7045
41	1.3025	1.6829	2.0195	2.4208	2.7012
42	1.3020	1.6820	2.0181	2.4185	2.6981
43	1.3016	1.6811	2.0167	2.4163	2.6951
44	1.3011	1.6802	2.0154	2.4141	2.6923
45	1.3006	1.6794	2.0141	2.4121	2.6896
46	1.3002	1.6787	2.0129	2.4102	2.6870
47	1.2998	1.6779	2.0117	2.4083	2.6846
48	1.2994	1.6772	2.0106	2.4066	2.6822
49	1.2991	1.6766	2.0096	2.4049	2.6800
50	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316
∞	1.2816	1.6449	1.9600	2.3263	2.5758

SOURCE: Owen (1962, courtesy Atomic Energy Commission, Washington, DC).
a. For a left-tail test the appropriate t statistic will be negative. Thus, for 10 degrees of freedom and at the 1 percent level of significance, the t statistic is −2.7638. For a two-tail test the level of significance must be doubled. This implies, for example, that with 50 degrees of freedom the t statistic of 1.6759 is associated with a significance level of 10 percent, not 5 percent.

APPENDIX C: REGRESSION OUTPUT FROM SAS AND SPSS

Two regression programs currently in use in most universities and research centers are part of the SAS (Statistical Analysis System) and SPSS (Statistical

Package for the Social Sciences) statistical packages. These packages consist of different computerized routines. The purpose of this appendix is to indicate where on SPSS or SAS computer output one finds the sorts of statistics that are discussed in the text. It should be noted that other regression programs, including those preprogrammed for microcomputers, provide many of the same kinds of information as that shown here.

To facilitate the discussion, the information of most interest to us here has been circled on the output and coded with an uppercase letter. In what follows we note the meaning of each. To facilitate the discussion, the food consumption data from Table 1 were used to generate these results based on the specification of equation 5. Slightly different types of information are provided by the two programs and, due to rounding, the results may differ slightly.

DEP VARIABLE: C

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	SAS PROB>F
MODEL	2	10025264	5012632	19.662-G	0.0001-H
ERROR	47	11981936	254935		
C TOTAL	49	22007200			

ROOT MSE	504.911	R-SQUARE	0.4555-I
DEP MEAN	1842.449	ADJ R-SQ	0.4324-J
C.V.	27.40432		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR HO: PARAMETER=0	PROB > |T|
INTERCEP	1	330.767-A	254.209	1.301	0.1995
I	1	0.056141 > B	0.011315 > C	4.961 > D1	0.0001 > E
S	1	129.622	36.143888	3.586	0.0009

DEPENDENT VARIABLE. C

VARIABLE(S) ENTERED ON STEP NUMBER 1 . I
S

MULTIPLE R	0 67494	ANALYSIS OF VARIANCE	DF	SUM OF SQUARES	MEAN SQUARE	F
R SQUARE	0 45554-I	REGRESSION	2	10025265 44517	5012632 72258	19 6624-G
ADJUSTED R SQUARE	0 43238-J	RESIDUAL	47	11981935 13276	254934 79006	
STANDARD ERROR	504 91068					

VARIABLES IN THE EQUATION					VARIABLES NOT IN THE EQUATION				
VARIABLE	B	BETA	STD ERROR B	F	VARIABLE	BETA IN	PARTIAL	TOLERANCE	F
I	0 5614086D-01	0 53465	0 01132	24 616					
S	129 6220	0 38646	36 14389	12 861					
(CONSTANT)	330 7666-A								

B > F > C > D2

ALL VARIABLES ARE IN THE EQUATION

STATISTICS WHICH CANNOT BE COMPUTED ARE PRINTED AS ALL NINES.

A

FILE NONAME (CREATION DATE = 08/27/85)

Code Meaning

A The estimate of the intercept coefficient, a.

B The estimate of the regression coefficients on each of the independent variables used in the regression, such as b_1 and b_2. (Note that SPSS uses a "scientific notation." The "D-01" at the end of the coefficient on the I variable means that one should move the decimal point one place to the left. That is, the estimate is 0.05614086.)

C The estimates of the standard errors of the regression coefficients, s_b. (SPSS does not provide this information for the intercept, or constant, coefficient.)

D Due to programming differences, SPSS provides different, but equivalent, information regarding the statistical significance of the regression coefficients.

D1 [SAS]
SAS provides the user with the value of the t ratio computed under the null hypothesis that the population regression coefficient is equal to zero (H_o:$\beta = 0$).

D2 [SPSS]
The entries under the heading of the column marked "F", for F ratio, in the SPSS output are exactly the same as those under the heading of "T FOR HO" in SAS except for the fact that those in SPSS are the squares of those in SAS. (To see this, multiply the SAS entries times themselves—e.g., $4.961 \times 4.961 = 24.616$.) The SPSS program is based on the fact that an F distribution with 1 numerator degree of freedom and d degrees of freedom in the denominator is equal to the square of a t distribution with d degrees of freedom.

E Available in SAS (but not SPSS) is the level of significance at which one can reject the null hypothesis that the regression coefficient is equal to zero. It is important to note that a two-tailed test is assumed here.

F The standardized regression coefficient, or beta coefficient, discussed in Chapter 2. SPSS always provides this information; it is also available in SAS, but only as a special option.

G The F ratio used to test the null hypothesis H_o:$\beta_1 = \beta_2 = 0$. The appropriate degrees of freedom to be used to test this null hypothesis can be found to the left of this ratio on each output under the heading DF.

H Again, SAS provides the user with the lowest level of significance at which the null hypothesis can be rejected. As shown here, one can reject the null at the 0.0001 level.

I The estimated coefficient of determination, R^2.

J $\bar{R}^2$, the coefficient of determination adjusted for degrees of freedom.

APPENDIX D:
SUGGESTED TEXTBOOKS

There is a long list of textbooks available which focus to some extent on linear regression analysis. Most introductory statistics texts contain at least one chapter devoted to the subject, while econometrics textbooks tend to focus nearly exclusively on linear regression. Among the potential books in these areas are the following:

Introductory Statistics

Hamburg, M. (1985) Basic Statistics: A Modern Approach (3rd ed.). San Diego, CA: Harcourt Brace Jovanovich.

Koopmans, L. H. (1981) An Introduction to Contemporary Statistics. Boston: Duxbury Press.

Wonnacott, T. and R. J. Wonnacott (1984) Introductory Statistics of Business and Economics (3rd ed.). New York: John Wiley.

There are also introductory textbooks for statistics designed for specific disciplines such as political science, sociology, or education:

Regression-Oriented Texts

Draper, N. R. and H. Smith (1981) Applied Regression Analysis (2nd ed.). New York: John Wiley.

Kleinbaum, D. G. and L. L. Kupper (1978) Applied Regression Analysis: Another Multivariable Method. North Scituate, MA: Duxbury Press.

Younger, M. S. (1979) A Handbook for Linear Regression. North Scituate, MA: Duxbury Press.

In addition, many of the books in this Sage series, Quantitative Applications in the Social Sciences, focus exclusively on linear regression.

Econometrics

Johnston, J. (1984) Econometric Methods (3rd ed.). New York: McGraw-Hill.

Maddala, G. S. (1977) Econometrics. New York: McGraw-Hill.

Pindyck, R. S. and D. L. Rubinfeld (1981) Econometric Models and Economic Forecasts (2nd ed.). New York: McGraw-Hill.

NOTES

1. Hypotheses need not be functional relationships, since it can be hypothesized that Mary is taller than Jane without implying causation. However, the hypotheses that we discuss are statements of functional relationships.

2. Economic theory states that the consumption of a product is a function of income, the price of the product, the prices of related products, and the tastes of the consumer. When everything except income is held constant, changes in the consumption of the product become a function of changes in income alone.

3. In its use in statistics, population refers to a collection of data, not necessarily to people. If we were interested in the advertising expenditures of firms, we would draw a sample from the population consisting of all firms.

4. There are many different functional forms for an equation relating two variables. For example, the two equations $Y = a + bX$ and $Y = a(X)^b$ represent, respectively, the linear and hyperbolic forms. The Xs and Ys represent variables—that is, symbols that take on any value within some specified set of values. The a's and b's represent parameters—that is, symbols that take on only one value in each equation.

5. Note that problems may arise when one specifies a functional form when a different form should have been specified, meaning that care must be taken in selecting the particular form used. Further, linearity may imply more than just a straight line. These topics are discussed in Chapters 4 and 5.

6. When I equals zero, the value of C equals α. In a graph, α is the intercept on the ordinate axis—that is, the axis on which the dependent variable is measured, usually the vertical axis. The slope of the line is β, and it describes how C changes with each unit change in I. The slope of a line or equation is defined as the change in the dependent variable divided by the change in the independent variable. In Figure 2, the three lines illustrate three different general values for the slope. For line 1 the slope is positive, meaning that increases (decreases) in I are associated with increases (decreases) in C; for line 2 the slope is zero, meaning that C does not change when I changes; and for line 3 the slope is negative, meaning that the increases (decreases) in I are associated with decreases (increases) in C. In each case the value of α is 10.

7. If we used the regression equation and calculated the value of consumption for income equal to I_i, the estimated level of consumption would be denoted by $\hat{C}_i$. The symbol Σ (the Greek letter sigma) is the standard symbol for summation. For example,

$$\sum_{i=1}^{3} C_i$$

means to sum the first three values of C (from Table 1); that is,

$$\sum_{i=1}^{3} C_i = \$723.52 + \$780.70 + \$990.74 = \$2,494.96.$$

8. To see this, consider a sample with three observations:

C	I
1	1
2	2
3	3

Plot the three observations and draw a line through all three points. Now draw a different straight line which passes through the second observation but not the first and third. For both lines, the sum of the nonsquared distances is zero.

9. This statement anticipates the discussion of hypothesis testing in Chapter 3. As will be seen there, the last statement depends on more than the sign of the estimated parameter.

10. The observations are monthly values for the rate of return and the rate of inflation for the period January 1953 through December 1971.

11. There are several different measures of correlation, depending on the nature of the variables. An explanation of the necessary conditions for calculating the correlation coefficient discussed here is beyond the scope of the book.

12. The correlation coefficient in this case is, of course, affected greatly by the fact that the Dolphins and the Seahawks had a good year while the Jets, Colts, and Bills did not. For the National Football Conference, the equivalent correlation coefficient is .328.

13. The relationship between the regression coefficient and the correlation coefficient can be shown to be $b = r(s_y/s_x)$, where s_y is the standard deviation of the dependent variable and s_x is the standard deviation of the independent variable. Standard deviation is a measure of the dispersion, about the mean, of the distribution of some variable. The further the values of a variable are spread out from the mean, the greater the value of the standard deviation. The formula for the standard deviation is

$$\sqrt{\Sigma (X_i - \bar{X})^2 / (n-1)}$$

where n is the number of observations.

14. The reader can verify this proposition in the following way: The estimated value of C for a family of four with a $10,000 income is $1,409.25; for a family of five with a $10,000 income, C is estimated to be $1,538.87. This is a change in C of $129.62 = $1,538.87 – $1,409.25 associated with a one-unit change in S while holding income constant.

15. See note 13 for the formula for the standard deviation.

16. Specifically,

$$\bar{R}^2 = \frac{(N-1)R^2 - k}{N-k-1}$$

where N is the number of observations and k is the number of independent variables in the regression equation.

17. Hypothesis testing is relevant whenever there is a random component to the estimate. Although our discussion focuses on sampling as the source of this randomness, other sources for randomness exist.

18. All hypotheses investigated in this study are stated with respect to the slope parameter (β for the food consumption problem), since most research focuses on changes in the dependent variable associated with changes in the independent variable. The analysis, however, applies equally to the intercept parameter (α in the food expenditure problem.)

19. The hypothesis-testing procedure can be used for testing coefficients obtained from a multiple regression equation. We are using the simple regression coefficient only to expedite the discussion.

20. It can be shown that by minimizing this type of error, the court system makes it more likely that defendants who actually did commit a crime will be set free.

21. In the hypothesis-testing procedure, the analyst sets the probability of making the inferential error of rejecting the null hypothesis when it is true. The procedure, however, does not consider the possibility that the null hypothesis will not be rejected even though it is *not* true.

22. The standard error is equal to

$$\frac{\sqrt{\frac{1}{N-2}\,\Sigma\,(C_i-\hat{C}_i)^2}}{\sqrt{\Sigma\,(I_i-\bar{I})^2}}.$$

23. A probability distribution relates the probability associated with a given event. Following is an example of the probability distribution of T, the event defined as the number of tails thrown out of four tosses of a fair coin:

T	Probability of T
T = 0	1/16
T = 1	4/16
T = 2	6/16
T = 3	4/16
T = 4	1/16

24. While the underlying mathematics of the t distribution are far more complex than we wish to develop here, a picture of what the t distribution looks like may be useful. The distribution sketched in Figure 8 is bell-shaped and centered at zero. The construction of the distribution is such that the total area under the curve is equal to 1. Hence, any portion of the area under the curve can be thought of as some proportion of 1. The exact shape of the t distribution depends on the degrees of freedom. For each possible t distribution, statisticians have computed the proportion of the area under the curve lying on either side of any value of t, denoted t_s. For example, as shown later, if t_s is equal to 1.677, then 5 percent of the area under the curve will lie to the right of t_s and 95 percent of the area will lie to the left of 1.677 (when there are 48 degrees of freedom). It is particularly convenient that probabilities are always stated as positive fractions in the range from zero to one. Hence,

one can say that the probability is 0.05 that a variable distributed as a t distribution will be greater than 1.677 and 0.95 that the variable will be less than 1.677.

25. The expression $>$ is read "greater than."

26. Strictly speaking, from any one sample and its estimate, these methods can be used legitimately only to test one hypothesis about the value of a coefficient. The examples presented here are for expository purposes only.

27. The role of degrees of freedom is encompassed in the t distribution. Without mathematical proof, as the degrees of freedom decline, the shape of the t distribution becomes less fat. But this in turn means that in order to leave only 5 percent of the area under the curve but to the right of the t statistic, a higher value of the t statistic must be utilized than when there are more degrees of freedom. For example, with only 20 degrees of freedom, a t statistic of 1.725 must be used instead of 1.677 to allow 5 percent of the area under the curve to lie to the right of t_s.

28. Note that if one wants to have a smaller probability associated with the shaded area under the curve, a larger value for the t statistic would have to be used. For example, if one wanted only a 2.5 percent probability of making an inferential error, the t distribution (as shown in Appendix B) would imply that a t statistic of 2.011 would have to be used. This in turn would yield a larger test value and would make it more difficult to reject the null hypothesis.

29. Figure 8 shows that the t distribution is symmetric. This means, for example, that 5 percent of the area under the curve will be to the right of 1.677 and 5 percent of the area under the curve will be to the left of -1.677.

30. Sometimes t ratios for left-tailed tests are presented as absolute values. If this is done, the ratio is always positive and should be compared to an appropriate positive-valued t statistic. In this instance the null hypothesis is rejected if $|b/s_b| > t_s$.

31. The expression $|b - \beta|$ is to be read as the absolute value of the difference between b and β. In absolute value terms, $|\beta - b|$ is equivalent to $|b - \beta|$.

32. They are highly correlated because family size is normally equal to the number of children plus one or two adults.

33. For a discussion of the limitations of tests of significance, see McCloskey (1985).

34. The distinction between aggregate and micro data is somewhat artificial. For example, families consisting of more than one member can be considered aggregate units, and a firm's sales are probably due to the combined efforts of several persons. Nevertheless, it is important when observing regression results to recognize the degree of aggregation implied by the data.

35. Note that there would be no change in the implications of the results had the dummy variable K been defined to equal 1 for nonfarm families and 0 for farmers. The estimated equation would have been $C = 143.68 + 0.060I + 599.16K$. That is, the estimated response to the one-dollar increase in income would still be 0.060, and nonfarm families (where $K = +1$) would still be shown to consume $599.16 more than farm families with equal income.

36. The reader should note that any continuous variable can be transformed into a classification variable. For example, while age is a continuous variable, surveys often report ages according to groups (e.g., less than 20 years of age, 20-65, or older than 65). The present discussion applies in those instances as well.

37. If, for example, $\alpha' = 1$ and $\beta = 2$, the resulting equation would be $L = M^2$. The reader should experiment with different values of M to verify that L and M are related nonlinearly.

38. Equation 15 relies on the following characteristics of mathematical operations on logarithms: $\log(XY) = \log(X) + \log(Y)$, and $\log(X^c) = c \log(X)$ where X and Y are any two positive real numbers and C is a real number. The symbol ln denotes the special case of a logarithm to the base e.

39. Students of economics will recognize that the ratio of the percentage change in L relative to a percentage change in M is the definition of the elasticity of L with respect to M. Thus this transformation provides a convenient way to estimate elasticity coefficients. Note too that in this case the assumption is made that the elasticity is the same at all points along the relationship.

40. From the calculus we know that $dY/dX = \beta_1 + 2\beta_2 X$; that is, the change in Y associated with a change in X depends on β_1, β_2, *and* X.

41. Bias is a statistical property and refers to whether or not, in numerous samples from a population, the estimates of the parameters will, on average, be equal to the population parameter. An unbiased estimator will, on average, yield estimates equal to that parameter.

42. Without proof, this is the reason that, when dummy variables representing three or more classes of outcomes are being analyzed, one group is omitted from the analysis. Without such omission, the results would be plagued with perfect multicollinearity.

43. Other patterns are also possible (e.g., correlations between residuals lagged two periods).

44. Note that in the case of heteroskedasticity, it is the variability of the residuals that is related to the independent variables; here it is the residuals themselves.

45. In fact, some studies have shown that there is not a great deal of difference in the results between these techniques and the results obtained from the ordinary least squares models. On the other hand, the latter models cannot directly yield predictions that will necessarily conform to the laws of probability and still face the issues discussed earlier.

REFERENCES

BAHL, R., R. D. GUSTELY, and M. J. WASYLENKO (1979) "The determinants of local government police expenditures: A public employment approach." National Tax Journal 31: 67-69.

COLEMAN, J. et al. (1966) Equality of Educational Opportunity. Washington, DC: Government Printing Office.

CURHAN, R. C. (1972) "The relationship between shelf space and unit sales in supermarkets." Journal of Marketing Research 9: 406-412.

DEMARIS, A. and G. R. LESLIE (1984) "Cohabitation with the future spouse: Its influence upon marital satisfaction and communication." Journal of Marriage and the Family 46: 77-84.

De TRAY, D. (1982) "Veteran status as a screening device." American Economic Review 72: 133-142.

DURAND, D. (1959) "The cost of capital and the theory of investment: Comment." American Economic Review 49: 49.

FELDMAN, P. and J. JONDROW (1984) "Congressional elections and local federal spending." American Journal of Political Science 28: 147-163.

FIELDS, G. (1979) "Place-to-place migration: Some new evidence." Review of Economic and Statistics 41: 21-32.

FUTRELL, C. M. and A. PARASURAMAN (1984) "The relationship of satisfaction and performance to salesforce turnover." Journal of Marketing 48: 33-40.

GRAHAM, J. D. and S. GARBER (1984) "Evaluation effects of automobile safety regulation." Journal of Policy Analysis and Management 3: 206-224.

GRONAU, R. (1977) "Leisure, home production, and work—The theory of allocation of time revisited." Journal of Political Economy 85: 1099-1123.

JAFFE, J. and G. MANDELKER (1976) "The 'Fisher Effect' for risky assets: An empirical investigation." Journal of Finance 31: 447-456.

LEWIS-BECK, M. S. and T. W. RICE (1983) "Localism in presidential elections: The home state advantage." American Journal of Political Science 27: 548-556.

LOCKHEED, M. E., A. NIELSON, and M. STONE (1985) "Determinants of microcomputer literacy in high school students." Journal of Educational Computing Research 1: 81-96.

MANNING, W. G., Jr., and C. E. PHELPS (1979) "The demand for dental care." Bell Journal of Economics 10: 503-525.

McCLOSKEY, D. M. (1985) "The loss function has been mislaid: The rhetoric of significance tests." American Economic Review 75: 201-205.

MENDENHALL, W., L. OTT, and R. L. SCHAAFFER (1971) Elementary Survey Sampling. Belmont, CA: Duxbury Press.

OWEN, D. B. (1962) Handbook of Statistical Tables. Reading, MA: Addison-Wesley.

POLACHEK, S. and F. HORVATH (1977) "A life cycle approach to migration," in R. G. Ehrenberg (ed.), Research in Labor Economics, Vol. 1. Greenwich, CT: JAI Press.

SEIVER, D. A. (1985) "Trend and variation in the seasonality of U.S. fertility, 1947-1976." Demography 22: 89-100.

SIMON, J. L. (1969) "The effect of advertising on liquor brand sales." Journal of Marketing Research 6: 301-313.

SPILLER, P. T. (1983) "The differential impact of airline regulations on individual firms and markets: An empirical analysis." Journal of Law and Economics 26(3): 755-689.

LARRY D. SCHROEDER *is Professor of Public Administration and Economics in the Maxwell School, Syracuse University, and Director of the Metropolitan Studies Program. He holds a Ph.D. from the University of Wisconsin and has taught quantitative methods courses at the Maxwell School and at Georgia State University. His substantive areas of interest are in public finance and policy, particularly state and local government finance and administration. His recent work has focused primarily on forecasting government revenues and expenditures, and on financial administration in developing countries.*

DAVID L. SJOQUIST *is Professor of Economics at Georgia State University and holds a Ph.D. from the University of Minnesota. He has taught microeconomic theory, public finance, and statistics. His primary research interests are in public finance and urban economics. He has published numerous articles in scholarly journals, including the* American Economic Review, Public Choice, *and the* National Tax Journal.

PAULA E. STEPHAN *is Professor of Economics and Industrial Relations at Georgia State University, where she teaches courses in labor economics. She holds a Ph.D. from the University of Michigan. As a member of the Small Grants Panel of the U.S. Department of Labor and of the editorial board of the* Southern Economic Journal, *she has reviewed many empirical proposals and manuscripts. Much of her research focuses on issues of labor supply. Recently she has been examining the productivity of U.S.-trained scientists.*