MULTIPLE REGRESSION IN BEHAVIORAL RESEARCH

EXPLANATION AND PREDICTION

THIRD EDITION

ELAZAR J. PEDHAZUR

CHAPTER 1

Overview

Remarkable advances in the analysis of educational, psychological, and sociological data have been made in recent decades. Much of this increased understanding and mastery of data analysis has come about through the wide propagation and study of statistics and statistical inference, and especially from the analysis of variance. The expression "analysis of variance" is well chosen. It epitomizes the basic nature of most data analysis: the partitioning, isolation, and identification of variation in a dependent variable due to different independent variables.

Other analytic statistical techniques, such as multiple regression analysis and multivariate analysis, have been applied less frequently until recently, not only because they are less well understood by behavioral researchers but also because they generally involve numerous and complex computations that in most instances require the aid of a computer for their execution. The recent widespread availability of computer facilities and package programs has not only liberated researchers from the drudgery of computations, but it has also put the most sophisticated and complex analytic techniques within the easy reach of anyone who has the rudimentary skills required to process data by computer. (In a later section, I comment on the use, and potential abuse, of the computer for data analysis.)

It is a truism that methods per se mean little unless they are integrated within a theoretical context and are applied to data obtained in an appropriately designed study. "It is sad that many investigations are carried out with no clear idea of the objective. This is a recipe for disaster or at least for an error of the third kind, namely 'giving the right answer to the wrong question'" (Chatfield, 1991, p. 241). Indeed, "The important question about methods is not 'how' but 'why'" (Tukey, 1954, p. 36).

Nevertheless, much of this book is about the "how" of methods, which is indispensable for appreciating their potentials, for keeping aware of their limitations, and for understanding their role in the overall research endeavor. Widespread misconceptions notwithstanding, data do *not* speak for themselves but through the medium of the analytic techniques applied to them. It is important to realize that analytic techniques not only set limits to the scope and nature of the answers one may obtain from data, but they also affect the type of questions a researcher asks and the manner in which the questions are formulated. "It comes as no particular surprise to discover that a scientist formulates problems in a way which requires for their solution just those techniques in which he himself is especially skilled" (Kaplan, 1964, p. 28).

Analytic techniques may be viewed from a variety of perspectives, among which are an analytic perspective and a research perspective. I use "analytic perspective" here to refer to such aspects as the mechanics of the calculations of a given technique, the meaning of its elements and the interrelations among them, and the statistical assumptions that underlie its valid application. Knowledge of these aspects is, needless to say, essential for the valid use of any analytic technique. Yet, the analytic perspective is narrow, and sole preoccupation with it poses the threat of losing sight of the role of analysis in scientific inquiry. It is one thing to know how to calculate a correlation coefficient or a t ratio, say, and quite another to know whether such techniques are applicable to the question(s) addressed in the study. Regrettably, while students can recite chapter and verse of a method, say a t ratio for the difference between means, they cannot frequently tell when it is validly applied and how to interpret the results it yields.

To fully appreciate the role and meaning of an analytic technique it is necessary to view it from the broader research perspective, which includes such aspects as the purpose of the study, its theoretical framework, and the type of research. In a book such as this one I cannot deal with the research perspective in the detail that it deserves, as this would require, among other things, detailed discussions of the philosophy of scientific inquiry, of theories in specific disciplines (e.g., psychology, sociology, and political science), and of research design. I do, however, attempt throughout the book to discuss the analytic techniques from a research perspective; to return to the question of why a given method is used and to comment on its role in the overall research setting. Thus I show, for instance, how certain elements of an analytic technique are applicable in one research setting but not in another, or that the interpretation of elements of a method depends on the research setting in which it is applied.[1]

I use the aforementioned perspectives in this chapter to organize the overview of the contents and major themes of this book. Obviously, however, no appreciable depth of understanding can be accomplished at this stage; nor is it intended. My purpose is rather to set the stage, to provide an orientation, for things to come. Therefore, do not be concerned if you do not understand some of the concepts and techniques I mention or comment on briefly. A certain degree of ambiguity is inevitable at this stage. I hope that it will be diminished when, in subsequent chapters, I discuss in detail topics I outline or allude to in the present chapter.

I conclude the chapter with some comments about my use of research examples in this book.

THE ANALYTIC PERSPECTIVE

The fundamental task of science is to explain phenomena. Its basic aim is to discover or invent general explanations of natural events (for a detailed explication of this point of view, see Braithwaite, 1953). Natural phenomena are complex. The phenomena and constructs of the behavioral sciences—learning, achievement, anxiety, conservatism, social class, aggression, reinforcement, authoritarianism, and so on—are especially complex. "Complex" in this context means that the phenomenon has many facets and many causes. In a research-analytic context, "complex" means that a phenomenon has several sources of variation. To study a construct or a variable scientifically we must be able to identify the sources of its variation. Broadly, a variable is any attribute on which objects or individuals vary. This means that when we apply an instrument that measures the variable to a sample of individuals, we obtain more or less different scores for each. We talk about the variance of college grade-point averages (as a measure of achievement) or the

[1] I recommend wholeheartedly Abelson's (1995) well-reasoned and engagingly written book on themes such as those briefly outlined here.

variability among individuals on a scale designed to measure locus of control, ego strength, learned helplessness, and so on.

Broadly speaking, the scientist is interested in explaining variance. In the behavioral sciences, variability is itself a phenomenon of great scientific curiosity and interest. The large differences in the intelligence and achievement of children, for instance, and the considerable differences among schools and socioeconomic groups in critical educational variables are phenomena of deep interest and concern to behavioral scientists.

In their attempts to explain the variability of a phenomenon of interest (often called the *dependent variable*), scientists study its relations or covariations with other variables (called the *independent variables*). In essence, information from the independent variables is brought to bear on the dependent variables. Educational researchers seek to explain the variance of school achievement by studying its relations with intelligence, aptitude, social class, race, home background, school atmosphere, teacher characteristics, and so on. Political scientists seek to explain voting behavior by studying variables presumed to influence it: sex, age, income, education, party affiliation, motivation, place of residence, and the like. Psychologists seek to explain aggressive behavior by searching for variables that may elicit it: frustration, noise, heat, crowding, exposure to acts of violence on television.

Various analytic techniques have been developed for studying relations between independent variables and dependent variables, or the effects of the former on the latter. In what follows I give a synopsis of techniques I present in this book. I conclude this section with some observations on the use of the computer for data analysis.

Simple Regression Analysis

Simple regression analysis, which I introduce in Chapter 2, is a method of analyzing the variability of a dependent variable by resorting to information available on an independent variable. Among other things, an answer is sought to the question: What are the expected changes in the dependent variable because of changes (observed or induced) in the independent variable?

In Chapter 3, I present current approaches for diagnosing, among other things, deviant or influential observations and their effects on results of regression analysis. In Chapter 4, I introduce computer packages that I will be using throughout most of the book, explain the manner in which I will be apply them, and use their regression programs to analyze a numerical example I analyzed by hand in earlier chapters.

Multiple Regression Analysis

When more than one independent variable is used, it is of course possible to apply simple regression analysis to each independent variable and the dependent variable. But doing this overlooks the possibility that the independent variables may be intercorrelated or that they may interact in their effects on the dependent variable. *Multiple regression* analysis (MR) is eminently suited for analyzing collective and separate effects of two or more independent variables on a dependent variable.

The bulk of this book deals with various aspects of applications and interpretations of MR in scientific research. In Chapter 5, I introduce the foundations of MR for the case of two independent variables. I then use matrix algebra to present generalization of MR to any number of independent variables (Chapter 6). Though most of the subject matter of this book can be mastered without resorting to matrix algebra, especially when the calculations are carried out by computer, I strongly recommend that you develop a working knowledge of matrix algebra, as it is extremely useful and general for conceptualization and analysis of diverse designs. To this end, I present an introduction to matrix algebra in Appendix A. In addition, to facilitate your acquisition of logic and skills in this very important subject, I present some topics twice: first in ordinary algebra (e.g., Chapter 5) and then in matrix algebra (e.g., Chapter 6).

Methods of statistical control useful in their own right (e.g., partial correlation) or that are important elements of MR (e.g., semipartial correlation) constitute the subject matter of Chapter 7. In Chapter 8, I address different aspects of using MR for prediction. In "The Research Perspective" section presented later in this chapter, I comment on analyses aimed solely at prediction and those aimed at explanation.

Multiple Regression Analysis in Explanatory Research

Part 2 of the book deals primarily with the use of MR in explanatory research. Chapters 9, 10, and 13 address the analyses of designs in which the independent variables are *continuous* or quantitative—that is, variables on which individuals or objects differ in degree. Examples of such variables are height, weight, age, drug dosage, intelligence, motivation, study time. In Chapter 9, I discuss various approaches aimed at partitioning the variance of the dependent variable and attributing specific portions of it to the independent variables. In Chapter 10, on the other hand, I show how MR is used to study the effects of the independent variables on the dependent variable. Whereas Chapters 9 and 10 are limited to linear regression analysis, Chapter 13 is devoted to curvilinear regression analysis.

There is another class of variables—*categorical* or qualitative—on which individuals differ in kind. Broadly, on such variables individuals are identified according to the category or group to which they belong. Race, sex, political party affiliation, and different experimental treatments are but some examples of categorical variables.

Conventionally, designs with categorical independent variables have been analyzed through the analysis of variance (ANOVA). Until recent years, ANOVA and MR have been treated by many as distinct analytic approaches. It is not uncommon to encounter students or researchers who have been trained exclusively in the use of ANOVA and who therefore cast their research questions in this mold even when it is inappropriate or undesirable to do so. In Part 2, I show that ANOVA can be treated as a special case of MR, and I elaborate on advantages of doing this. For now, I will make two points. (1) Conceptually, continuous and categorical variables are treated alike in MR—that is, both types of variables are viewed as providing information about the status of individuals, be it their measured aptitude, their income, the group to which they belong, or the type of treatment they have been administered. (2) MR is applicable to designs in which the independent variables are continuous, categorical, or combinations of both, thereby eschewing the inappropriate or undesirable practice of categorizing continuous variables (e.g., designating individuals above the mean as high and those below the mean as low) in order to fit them into what is considered, often erroneously, an ANOVA design.

Analytically, it is necessary to code categorical variables so that they may be used in MR. In Chapter 11, I describe different methods of coding categorical variables and show how to use them in the analysis of designs with a single categorical independent variable, what is often called simple ANOVA. Designs consisting of more than one categorical independent variable (factorial designs) are the subject of Chapter 12.

Combinations of continuous and categorical variables are used in various designs for different purposes. For instance, in an experiment with several treatments (a categorical variable), aptitudes of subjects (a continuous variable) may be used to study the interaction between these variables in their effect on a dependent variable. This is an example of an aptitude-treatments-interaction (ATI) design. Instead of using aptitudes to study their possible interactions with treatments, they may be used to control for individual differences, as in the analysis of covariance (ANCOVA). In Chapters 14 and 15, I show how to use MR to analyze ATI, ANCOVA, and related designs (e.g., comparing regression equations obtained from two or more groups).

In Chapter 16, I show, among other things, that when studying multiple groups, total, between-, and within-groups parameters may be obtained. In addition, I introduce some recent developments in multilevel analysis.

In all the designs I mentioned thus far, the dependent variable is continuous. In Chapter 17, I introduce logistic regression analysis—a method for the analysis of designs in which the dependent variable is categorical.

In sum, MR is versatile and useful for the analysis of diverse designs. To repeat: the overriding conception is that information from independent variables (continuous, categorical, or combinations of both types of variables) is brought to bear in attempts to explain the variability of a dependent variable.

Structural Equation Models

In recent years, social and behavioral scientists have shown a steadily growing interest in studying patterns of causation among variables. Various approaches to the analysis of causation, also called structural equation models (SEM), have been proposed. Part 3 serves as an introduction to this topic. In Chapter 18, I show how the analysis of causal models with observed variables, also called path analysis, can be accomplished by repeated applications of multiple regression analysis. In Chapter 19, I introduce the analysis of causal models with latent variables. In both chapters, I use two programs—EQS and LISREL—designed specifically for the analysis of SEM.

Multivariate Analysis

Because multiple regression analysis is applicable in designs consisting of a single dependent variable, it is considered a univariate analysis. I will note in passing that some authors view multiple regression analysis as a multivariate analytic technique whereas others reserve the term "multivariate analysis" for approaches in which multiple dependent variables are analyzed simultaneously. The specific nomenclature is not that important. One may view multivariate analytic techniques as extensions of multiple regression analysis or, alternatively, the latter may be viewed as a special case subsumed under the former.

Often, it is of interest to study effects of independent variables on more than one dependent variable simultaneously, or to study relations between sets of independent and dependent variables. Under such circumstances, multivariate analysis has to be applied. Part 4 is designed to serve as a introduction to different methods of multivariate analysis. In Chapter 20, I introduce discriminant analysis and multivariate analysis of variance for any number of groups. In addition, I show that for designs consisting of two groups with any number of dependent variables, the analysis may be carried out through multiple regression analysis. In Chapter 21, I present canonical analysis—an approach aimed at studying relations between sets of variables. I show, among other things, that discriminant analysis and multivariate analysis of variance can be viewed as special cases of this most general analytic approach.

Computer Programs

Earlier, I noted the widespread availability of computer programs for statistical analysis. It may be of interest to point out that when I worked on the second edition of this book the programs I used were available only for mainframe computers. To incorporate excerpts of output in the manuscript (1) I marked or copied them, depending on how much editing I did; (2) my wife then typed the excerpts; (3) we then proofread to minimize errors in copying and typing. For the current edition, I used only PC versions of the programs. Working in Windows, I ran programs as the need arose, without quitting my word processor, and cut and pasted relevant segments of the output. I believe the preceding would suffice for you to appreciate the great value of the recent developments. My wife surely does!

While the availability of user-friendly computer programs for statistical analysis has proved invaluable, it has not been free of drawbacks, as it has increased the frequency of blind or mindless application of methods. I urge you to select a computer program only after you have formulated your problems and hypotheses. Clearly, you have to be thoroughly familiar with a program so that you can tell whether it provides for an analysis that bears on your hypotheses.

In Chapter 4, I introduce four packages of computer programs, which I use repeatedly in various subsequent chapters. In addition, I introduce and use programs for SEM (EQS and LISREL) in Chapters 18 and 19. In all instances, I give the control statements and comment on them. I then present output, along with commentaries. My emphasis is on interpretation, the meaning of specific terms reported in the output, and on the overall meaning of the results. Consequently, I do not reproduce computer output in its entirety. Instead, I reproduce excerpts of output most pertinent for the topic under consideration.

I present more than one computer package so that you may become familiar with unique features of each, with its strengths and weaknesses, and with the specific format of its output. I hope that you will thereby develop flexibility in using any program that may be available to you, or one that you deem most suitable when seeking specific information in the results.

I suggest that you use computer programs from the early stages of learning the subject matter of this book. The savings in time and effort in calculations will enable you to pay greater attention to the meaning of the methods I present and to develop a better understanding and appreciation of them. Yet, there is no substitute for hand calculations to gain understanding of a method and a "feel" for what is going on when the data are analyzed by computer. I therefore strongly recommend that at the initial stages of learning a new topic you solve the numerical examples both by hand and by computer. Comparisons between the two solutions and the identification of specific aspects of the computer output can be a valuable part of the learning process. With this in mind, I present small, albeit unrealistic, numerical examples that can be solved by hand with little effort.

THE RESEARCH PERSPECTIVE

I said earlier that the role and meaning of an analytic technique can be fully understood and appreciated only when viewed from the broad research perspective. In this section I elaborate on some aspects of this topic. Although neither exhaustive nor detailed, I hope that the discussion will serve to underscore from the beginning the paramount role of the research perspective in determining how a specific method is applied and how the results it yields are interpreted. My presentation is limited to the following aspects: (1) the purpose of the study, (2) the type of research, and (3) the theoretical framework of the study. You will find detailed discussions of these and other topics in texts on research design and measurement (e.g., Cook & Campbell, 1979; Kerlinger, 1986; Nunnally, 1978; Pedhazur & Schmelkin, 1991).

Purpose of Study

In the broadest sense, a study may be designed for predicting or explaining phenomena. Although these purposes are not mutually exclusive, identifying studies, even broad research areas, in which the main concern is with either prediction or explanation is easy. For example, a college admissions officer may be interested in determining whether, and to what extent, a set of variables (mental ability, aptitudes, achievement in high school, socioeconomic status, interests, motivation) is useful in *predicting* academic achievement in college. Being interested solely in prediction, the admissions officer has a great deal of latitude in the selection of predictors. He or she may examine potentially useful predictors individually or in sets to ascertain the most useful ones. Various approaches aimed at selecting variables so that little, or nothing, of the predictive power of the entire set of variables under consideration is sacrificed are available. These I describe in Chapter 8, where I show, among other things, that different variable-selection procedures applied to the same data result in the retention of different variables. Nevertheless, this poses no problems in a predictive study. Any procedure that meets the specific needs and inclinations of the researcher (economy, ready availability of some variables, ease of obtaining specific measurements) will do.

The great liberty in the selection of variables in predictive research is countervailed by the constraint that no statement may be made about their meaningfulness and effectiveness from a theoretical frame of reference. Thus, for instance, I argue in Chapter 8 that when variable-selection procedures are used to optimize prediction of a criterion, regression coefficients should *not* be interpreted as indices of the effects of the predictors on the criterion. Furthermore, I show (see, in particular Chapters 8, 9, and 10) that a major source of confusion and misinterpretation of results obtained in some landmark studies in education is their reliance on variable-selection procedures although they were aimed at explaining phenomena. In sum, when variables are selected to optimize prediction, all one can say is, given a specific procedure and specific constraints placed by the researcher, which combination of variables best predicts the criterion.

Contrast the preceding example with a study aimed at *explaining* academic achievement in college. Under such circumstances, the choice of variables and the analytic approach are largely determined by the theoretical framework (discussed later in this chapter). Chapters 9 and 10 are devoted to detailed discussions of different approaches in the use of multiple regression analysis in explanatory research. For instance, in Chapter 9, I argue that popular approaches of incremental partitioning of variance and commonality analysis cannot yield answers to questions about the relative importance of independent variables or their relative effects on the dependent variable. As I point out in Chapter 9, I discuss these approaches in detail because they are often misapplied in various areas of social and behavioral research. In Chapter 10, I address the interpretation of regression coefficients as indices of effects of independent variables on the dependent variable. In this context, I discuss differences between standardized and unstandardized regression coefficients, and advantages and disadvantages of each. Other major issues I address in Chapter 10 are adverse effects of high correlations among independent variables, measurement errors, and errors in specifying the model that presumably reflects the process by which the independent variables affect the dependent variables.

Types of Research

Of various classifications of types of research, one of the most useful is that of experimental, quasi-experimental, and nonexperimental. Much has been written about these types of research, with special emphasis on issues concerning their internal and external validity (see, for example, Campbell & Stanley, 1963; Cook & Campbell, 1979; Kerlinger, 1986; Pedhazur & Schmelkin, 1991). As I pointed out earlier, I cannot discuss these issues in this book. I do, however, in various chapters, draw attention to the fact that the interpretation of results yielded by a given analytic technique depends, in part, on the type of research in which it is applied.

Contrasts between the different types of research recur in different contexts, among which are (1) the interpretation of regression coefficients (Chapter 10), (2) the potential for specification errors (Chapter 10), (3) designs with unequal sample sizes or unequal cell frequencies (Chapters 11 and 12), (4) the meaning of interactions among independent variables (Chapters 12 through 15), and (5) applications and interpretations of the analysis of covariance (Chapter 15).

Theoretical Framework

Explanation implies, first and foremost, a theoretical formulation about the nature of the relations among the variables under study. The theoretical framework determines, largely, the choice of the analytic technique, the manner in which it is to be applied, and the interpretation of the results. I demonstrate this in various parts of the book. In Chapter 7, for instance, I show that the calculation of a partial correlation coefficient is predicated on a specific theoretical statement regarding the patterns of relations among the variables. Similarly, I show (Chapter 9) that within certain theoretical frameworks it may be meaningful to calculate semipartial correlations, whereas in others such statistics are not meaningful. In Chapters 9, 10, and 18, I analyze the same data several times according to specific theoretical elaborations and show how elements obtained in each analysis are interpreted.

In sum, in explanatory research, data analysis is designed to shed light on theory. The potential of accomplishing this goal is predicated, among other things, on the use of analytic techniques that are commensurate with the theoretical framework.

RESEARCH EXAMPLES

In most chapters, I include research examples. *My aim is not to summarize studies I cite, nor to discuss all aspects of their design and analysis.* Instead, I focus on specific facets of a study

insofar as they may shed light on a topic I present in the given chapter. I allude to other facets of the study only when they bear on the topic I am addressing. Therefore, *I urge you to read the original report of a study that arouses your interest before passing judgment on it.*

As you will soon discover, in most instances I focus on shortcomings, misapplications, and misinterpretations in the studies on which I comment. In what follows I detail some reasons for my stance, as it goes counter to strong norms of not criticizing works of other professionals, of tiptoeing when commenting on them. Following are but some manifestations of such norms.

In an editorial, Oberst (1995) deplored the reluctance of nursing professionals to express publicly their skepticism of unfounded claims for the effectiveness of a therapeutic approach, saying, "Like the citizens in the fairy tale, we seem curiously unwilling to go on record about the emperor's obvious nakedness" (p. 1).

Commenting on controversy surrounding the failure to replicate the results of an AIDS research project, Dr. David Ho, who heads an AIDS research center, was reported to have said, "The problem is that too many of us try to avoid the limelight for controversial issues and avoid pointing the finger at another colleague to say what you have published is wrong" (Altman, 1991, p. B6).

In a discussion of the "tone" to be used in papers submitted to journals published by the American Psychological Association, the *Publication Manual* (American Psychological Association, 1994) states, "Differences should be presented in a professional non-combative manner: For example, 'Fong and Nisbett did not consider . . . ' is acceptable, whereas 'Fong and Nisbett completely overlooked . . . ' is not" (pp. 6–7).

Beware of Learning Others' Errors

With other authors (e.g., Chatfield, 1991, pp. 248–251; Glenn, 1989, p. 137; King, 1986, p. 684; Swafford, 1980, p. 684), I believe that researchers are inclined to learn from, and emulate, articles published in refereed journals, not only because this appears less demanding than studying textbook presentations but also because it holds the promise of having one's work accepted for publication. This is particularly troubling, as wrong or seriously flawed research reports are prevalent even in ostensibly the most rigorously refereed and edited journals (see the "Peer Review" section presented later in this chapter).

Learn from Others' Errors

Although we may learn from our errors, we are more open, therefore more likely, to learn from errors committed by others. By exposing errors in research reports and commenting on them, I hope to contribute to the sharpening of your critical ability to scrutinize and evaluate your own research and that of others. In line with what I said earlier, I do not address overriding theoretical and research design issues. Instead, I focus on specific errors in analysis and/or interpretation of results of an analysis. I believe that this is bound to reduce the likelihood of you committing the same errors. Moreover, it is bound to heighten your general alertness to potential errors.

There Are Errors and There Are ERRORS

It is a truism that we all commit errors at one time or another. Also unassailable is the assertion that the quest for perfection is the even debilitate, research. Yet, clearly, errors vary in severity and the potentially deleterious consequences to which they may lead. I would like to stress that my concern is not with perfection, nor with minor, inconsequential, or esoteric errors, but with egregious errors that cast serious doubt about the validity of the findings of a study.

Recognizing full well that my critiques of specific studies are bound to hurt the feelings of their authors, I would like to apologize to them for singling out their work. If it is any consolation, I would point out that their errors are not unique, nor are they necessarily the worst that I have come across in research literature. I selected them because they seemed suited to illustrate common misconceptions or misapplications of a given approach I was presenting. True, I could have drawn attention to potential errors without citing studies. I use examples from actual studies for three reasons: (1) I believe this will have a greater impact in immunizing you against egregious errors in the research literature and in sensitizing you to avoid them in your research. (2) Some misapplications I discuss are so blatantly wrong that had I made them up, instead of taking them from the literature, I would have surely been accused of being concerned with the grotesque or of belaboring the obvious. (3) I felt it important to debunk claims about the effectiveness of the peer review process to weed out the poor studies—a topic to which I now turn.

PEER REVIEW

Budding researchers, policy makers, and the public at large seem to perceive publication in a refereed journal as a seal of approval as to its validity and scientific merit. This is reinforced by, among other things, the use of publication in refereed journals as a primary, if not the primary, criterion for (1) evaluating the work of professors and other professionals (for a recent "bizarre example," see Honan, 1995) and (2) admission as scientific evidence in litigation (for recent decisions by lower courts, rulings by the Supreme Court, and controversies surrounding them, see Angier, 1993a, 1993b; Greenhouse, 1993; Haberman, 1993; Marshall, 1993; *The New York Times*, National Edition, 1995, January 8, p. 12). It is noteworthy that in a brief to the Supreme Court, The American Association for the Advancement of Science and the National Academy of Sciences argued that the courts should regard scientific "claims 'skeptically' until they have been 'subject to some peer scrutiny.' Publication in a peer-reviewed journal is 'the best means' of identifying valid research" (Marshall, 1993, p. 590).

Clearly, I cannot review, even briefly, the peer review process here.[2] Nor will I attempt to present a balanced view of pro and con positions on this topic. Instead, I will draw attention to some major inadequacies of the review process, and to some unwarranted assumptions underlying it.

Failure to Detect Elementary Errors

Many errors to which I will draw attention are so elementary as to require little or no expertise to detect. Usually, a careful reading would suffice. Failure by editors and referees to detect such errors makes one wonder whether they even read the manuscripts. Lest I appear too harsh or unfair, I will give here a couple of examples of what I have in mind (see also the following discussion, "Editors and Referees").

[2] For some treatments of this topic, see *Behavioral and Brain Sciences* (1982, *5*, 187–255 and 1991, *14*, 119–186); Cum-

Reporting on an unpublished study by Stewart and Feder (scientists at the National Institutes of Health), Boffey (1986) wrote:

> Their study . . . concluded that the 18 full-length scientific papers reviewed had "an abundance of errors" and discrepancies—a dozen per paper on the average—that could have been detected by any competent scientist who read the papers carefully. Some errors were described as . . . "so glaring as to offend common sense." . . . [Data in one paper were] so "fantastic" that it ought to have been questioned by any scientist who read it carefully, the N.I.H. scientists said in an interview. The paper depicted a family with high incidence of an unusual heart disease; a family tree in the paper indicated that one male member supposedly had, by the age of 17, fathered four children, conceiving the first when he was 8 or 9. (p. C11)

Boffey's description of how Stewart and Feder's paper was "blocked from publication" (p. C11) is in itself a serious indictment of the review process.

Following is an example of an error that should have been detected by anyone with superficial knowledge of the analytic method used. Thomas (1978) candidly related what happened with a paper in archaeology he coauthored with White in which they used principal component analysis (PCA). For present purposes it is not necessary to go into the details of PCA (for an overview of PCA versus factor analysis, along with relevant references, see Pedhazur & Schmelkin, 1991, pp. 597–599). At the risk of oversimplifying, I will point out that PCA is aimed at extracting components underlying relations among variables (items and the like). Further, the results yielded by PCA variables (items and the like) have loadings on the components and the *loadings may be positive or negative*. Researchers use the high loadings to interpret the results of the analysis. Now, as Thomas pointed out, the paper he coauthored with White was very well received and praised by various authorities.

> One flaw, however, mars the entire performance: . . . the principal component analysis was incorrectly interpreted. We interpreted the major components based strictly on high positive values [loadings]. Principal components analysis is related to standard correlation analysis and, of course, both positive *and negative* values are significant. . . . The upshot of this statistical error is that our interpretation of the components must be reconsidered. (p. 234)

Referring to the paper by White and Thomas, Hodson (1973) stated, "These trivial but rather devastating slips could have been avoided by closer contact with relevant scientific colleagues" (350). Alas, as Thomas pointed out, "Some very prominent archaeologists—some of them known for their expertise in quantitative methods—examined the White-Thomas manuscript prior to publication, yet the error in interpreting the principal component analysis persisted into print" (p. 234).[3]

I am hardly alone in maintaining that many errors in published research are (should be) detectable through careful reading even by people with little knowledge of the methods being used. Following are but some instances.

In an insightful paper on "good statistical practice," Preece (1987) stated that "within British research journals, the quality ranges from the very good to the very bad, and this latter includes statistics so erroneous that *non*-statisticians should immediately be able to recognize it as rubbish" (p. 407).

Glantz (1980), who pointed out that "critical reviewers of the biomedical literature consistently found that about half the articles that used statistical methods did so incorrectly" (p. 1), noted also "errors [that] rarely involve sophisticated issues that provoke debate among professional statisticians, but are simple mistakes" (p. 1).

Tuckman (1990) related that in a research-methods course he teaches, he asks each student to pick a published article and critique it before the class. "Despite the motivation to select perfect work (without yet knowing the criteria to make that judgment), each article selected, with rare exception, is torn apart on the basis of a multitude of serious deficiencies ranging from substance to procedures" (p. 22).

Editors and Referees

In an "Editor's Comment" entitled "Let's Train Reviewers," the editor of the *American Sociological Review* (October 1992, *57*, iii–iv) drew attention to the need to improve the system, saying, "The bad news is that in my judgment one-fourth or more of the reviews received by *ASR* (and I suspect by other journals) are not helpful to the Editor, and many of them are even misleading" (p. iii). Turning to his suggestions for improvement, the editor stated, "A good place to start might be by reconsidering a widely held assumption about reviewing—the notion that 'anyone with a Ph.D. is able to review scholarly work in his or her specialty' " (p. iii).[4]

Commenting on the peer review process, Crandall (1991) stated:

> I had to laugh when I saw the recent American Psychological Association announcements recruiting members of under represented groups to be reviewers for journals. The only qualification mentioned was that they must have published articles in peer reviewed journals, because *"the experience of publishing provides a reviewer with the basis for preparing a thorough, objective evaluative review"* [italics added]. (p. 143)

Unfortunately, problems with the review process are exacerbated by the appointment of editors unsuited to the task because of disposition and/or lack of knowledge to understand, let alone evaluate, the reviews they receive. For instance, in an interview upon his appointment as editor of *Psychological Bulletin* (an American Psychological Association journal concerned largely with methodological issues), John Masters is reported to have said, "I am consistently embarrassed that my statistical and methodological acumen became frozen in time when I left graduate school except for what my students have taught me" (Bales, 1986, p. 14). He may deserve an A+ for candor—but being appointed the editor of *Psychological Bulletin*? Could it be that Blalock's (1989, p. 458) experience of encountering "instances where potential journal editors were passed over because it was argued that their standards would be too demanding!" is not unique?

Commenting on editors' abdicating "responsibility for editorial decisions," Crandall (1991) stated, "I believe that many editors do not read the papers for which they are supposed to have editorial responsibility. If they don't read them closely, how can they be the editors?" (p. 143; see also, Tuckman, 1990).

In support of Crandall's assertions, I will give an example from my own experience. Following a review of a paper I submitted to a refereed journal, the editor informed me that he would like to publish it, but asked for some revisions and extensions. I was surprised when, in acknowledging receipt of the revised paper, the editor informed me that he had sent it out for another review. Anyway, some time later I received a letter from the editor, who informed me that though he "had all but promised publication," he regretted that he had to reject the paper *"given the fact that the technique has already been published"* [italics added]. Following is the *entire* review (with the misspellings of authors' names, underlining, and mistyping) that led to the editor's decision.

> The techniques the author discusses are treated in detail in the book *Introduction to Linear Models and the Design and Analysis of Experiments* by William Mendenhill [*sic.* Should be Mendenhall], Wadsworth Publishing Co. *1968*, Ch. 13, p. 384 and Ch. 4, p. 66, in detail and I may add are no longer in use with more sophisticated software statistical packages (e.g. Multivariance by Boik [*sic*] and Finn [should be Finn & Bock], FRULM by Timm and Carlson etc. etc. Under *no*/condition should this paper be published—not original and out of date.

I wrote the editor pointing out that I proposed my method as an alternative to a cumbersome one (presented by Mendenhall and others) that was then in use. In support of my assertion, I enclosed photocopies of the pages from Mendenhall cited by the reviewer and invited the editor to examine them.

In response, the editor phoned me, apologized for his decision, and informed me that he would be happy to publish the paper. In the course of our conversation, I expressed concern about the review process in general and specifically about (1) using new reviewers for a revised paper and (2) reliance on the kind of reviewer he had used. As to the latter, I suggested that the editor reprimand the reviewer and send him a copy of my letter. Shortly afterward, I received a copy of a letter the editor sent the reviewer. Parenthetically, the reviewer's name and address were removed from my copy, bringing to mind the question: "Why should the wish to publish a scientific paper expose one to an assassin more completely protected than members of the infamous society, the Mafia?" (R. D. Wright, quoted by Cicchetti, 1991, p. 131). Anyway, after telling the reviewer that he was writing concerning my paper, the editor stated:

> I enclose a copy of the response of the author. I have read the passage in Mendenhall and find that the author is indeed correct.
>
> On the basis of your advice, I made a serious error and have since apologized to the author. I would ask you to be more careful with your reviews in the future.

Why didn't the editor check Mendenhall's statements before deciding to reject my paper, especially when all this would have entailed is the reading of *two* pages pinpointed by the reviewer? And why would he deem the reviewer in question competent to review papers in the future? Your guesses are as good as mine.

Earlier I stated that detection of many egregious errors requires nothing more than careful reading. At the risk of sounding trite and superfluous, however, I would like to stress that to detect errors in the application of an analytic method, the reviewer ought to be familiar with it. As I amply show in my commentaries on research studies, their very publication leads to the inescapable conclusion that editors and referees have either not carefully read the manuscripts or have no knowledge of the analytic methods used. I will let you decide which is the worse offense.

As is well known, much scientific writing is suffused with jargon. This, however, should not serve as an excuse for not investing time and effort to learn the technical terminology required to understand scientific publications in specific disciplines. It is one thing to urge the authors of scientific papers to refrain from using jargon. It is quite something else to tell them, as does the

[4] A similar, almost universally held, assumption is that the granting of a Ph.D. magically transforms a person into an all-knowing expert, qualified to guide doctoral students on their dissertations and to serve on examining committees for

terminology in a paper should be understood by psychologists *throughout the discipline*" [italics added] (p. 27). I believe that this orientation fosters, unwittingly, the perception that when one does not understand a scientific paper, the fault is with its author. Incalculable deleterious consequences of the widespread reporting of questionable scientific "findings" in the mass media have made the need to foster greater understanding of scientific research methodology and healthy skepticism of the peer review process more urgent than ever.

CHAPTER 2

Simple Linear Regression and Correlation

In this chapter, I address fundamentals of regression analysis. Following a brief review of variance and covariance, I present a detailed discussion of linear regression analysis with one independent variable. Among topics I present are the regression equation; partitioning the sum of squares of the dependent variable into regression and residual components; tests of statistical significance; and assumptions underlying regression analysis. I conclude the chapter with a brief presentation of the correlation model.

VARIANCE AND COVARIANCE

Variability tends to arouse curiosity, leading some to search for its origin and meaning. The study of variability, be it among individuals, groups, cultures, or within individuals across time and settings, plays a prominent role in behavioral research. When attempting to explain variability of a variable, researchers resort to, among other things, the study of its covariations with other variables. Among indices used in the study of variation and covariation are the variance and the covariance.

Variance

Recall that the sample variance is defined as follows:

$$s_x^2 = \frac{\Sigma(X-\bar{X})^2}{N-1} = \frac{\Sigma x^2}{N-1} \qquad (2.1)$$

where s_x^2 = sample variance of X; Σx^2 = sum of the squared deviations of X from the mean of X; and N = sample size.

When the calculations are done by hand, or with the aid of a calculator, it is more convenient to obtain the deviation sum of squares by applying a formula in which only raw scores are used:

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \qquad (2.2)$$

terminology in a paper should be understood by psychologists *throughout the discipline*" [italics added] (p. 27). I believe that this orientation fosters, unwittingly, the perception that when one does not understand a scientific paper, the fault is with its author. Incalculable deleterious consequences of the widespread reporting of questionable scientific "findings" in the mass media have made the need to foster greater understanding of scientific research methodology and healthy skepticism of the peer review process more urgent than ever.

CHAPTER

2

Simple Linear Regression and Correlation

In this chapter, I address fundamentals of regression analysis. Following a brief review of variance and covariance, I present a detailed discussion of linear regression analysis with one independent variable. Among topics I present are the regression equation; partitioning the sum of squares of the dependent variable into regression and residual components; tests of statistical significance; and assumptions underlying regression analysis. I conclude the chapter with a brief presentation of the correlation model.

VARIANCE AND COVARIANCE

Variability tends to arouse curiosity, leading some to search for its origin and meaning. The study of variability, be it among individuals, groups, cultures, or within individuals across time and settings, plays a prominent role in behavioral research. When attempting to explain variability of a variable, researchers resort to, among other things, the study of its covariations with other variables. Among indices used in the study of variation and covariation are the variance and the covariance.

Variance

Recall that the sample variance is defined as follows:

$$s_x^2 = \frac{\Sigma(X-\bar{X})^2}{N-1} = \frac{\Sigma x^2}{N-1} \qquad (2.1)$$

where s_x^2 = sample variance of X; Σx^2 = sum of the squared deviations of X from the mean of X; and N = sample size.

When the calculations are done by hand, or with the aid of a calculator, it is more convenient to obtain the deviation sum of squares by applying a formula in which only raw scores are used:

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \qquad (2.2)$$

where ΣX^2 = sum of the squared raw scores; and $(\Sigma X)^2$ = square of the sum of raw scores. Henceforth, I will use "sum of squares" to refer to deviation sum of squares unless there is ambiguity, in which case I will use "deviation sum of squares."

I will now use the data of Table 2.1 to illustrate calculations of the sums of squares and variances of X and Y.

Table 2.1 Illustrative Data for *X* and *Y*

	X	X^2	Y	Y^2	XY
	1	1	3	9	3
	1	1	5	25	5
	1	1	6	36	6
	1	1	9	81	9
	2	4	4	16	8
	2	4	6	36	12
	2	4	7	49	14
	2	4	10	100	20
	3	9	4	16	12
	3	9	6	36	18
	3	9	8	64	24
	3	9	10	100	30
	4	16	5	25	20
	4	16	7	49	28
	4	16	9	81	36
	4	16	12	144	48
	5	25	7	49	35
	5	25	10	100	50
	5	25	12	144	60
	5	25	6	36	30
Σ:	60	220	146	1196	468
M:	3.00		7.30		

$$\Sigma x^2 = 220 - \frac{60^2}{20} = 40 \qquad \Sigma y^2 = 1196 - \frac{146^2}{20} = 130.2$$

$$s_x^2 = \frac{40}{19} = 2.11 \qquad s_y^2 = \frac{130.2}{19} = 6.85$$

The standard deviation (s) is, of course, the square root of the variance:

$$s_x = \sqrt{2.11} = 1.45 \qquad s_y = \sqrt{6.85} = 2.62$$

Covariance

The sample covariance is defined as follows:

$$s_{xy} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N-1} = \frac{\Sigma xy}{N-1} \tag{2.3}$$

where s_{xy} = covariance of X and Y; and Σxy = sum of the cross products deviations of pairs of X and Y scores from their respective means. Note the analogy between the variance and the covariance. The variance of a variable can be conceived of as its covariance with itself. For example,

$$s_x^2 = \frac{\Sigma(X-\bar{X})(X-\bar{X})}{N-1}$$

In short, the variance indicates the variation of a set of scores from their mean, whereas the covariance indicates the covariation of two sets of scores from their respective means.

As in the case of sums of squares, it is convenient to calculate the sum of the cross products deviations (henceforth referred to as "sum of cross products") by using the following algebraic identity:

$$\Sigma xy = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N \text{ = number of pairs}} \tag{2.4}$$

where ΣXY is the sum of the products of pairs of raw X and Y scores; and ΣX and ΣY are the sums of the raw scores of X and Y, respectively.

For the data of Table 2.1,

$$\Sigma xy = 468 - \frac{(60)(146)}{20} = 30$$

$$s_{xy} = \frac{30}{19} = 1.58$$

Sums of squares, sums of cross products, variances, and covariances are the staples of regression analysis; hence, it is essential that you understand them thoroughly and be able to calculate them routinely. If necessary, refer to statistics texts (e.g., Hays, 1988) for further study of these concepts.

(1 IV)

SIMPLE LINEAR REGRESSION

I said earlier that among approaches used to explain variability of a variable is the study of its covariations with other variables. The least ambiguous setting in which this can be accomplished is the experiment, whose simplest form is one in which the effect of an independent variable, X, on a dependent variable, Y, is studied. In such a setting, the researcher attempts to ascertain how induced variation in X leads to variation in Y. In other words, the goal is to determine how, and to what extent, variability of the dependent variable depends upon manipulations of the independent variable. For example, one may wish to determine the effects of hours of study, X, on achievement in vocabulary, Y; or the effects of different dosages of a drug, X, on anxiety, Y. Obviously, performance on Y is usually affected also by factors other than X and by random errors. Hence, it is highly unlikely that all individuals exposed to the same level of X would exhibit identical performance on Y. But if X does affect Y, the means of the Y's at different levels of X would be expected to differ from each other. When the Y means for the different levels of X differ from each other and lie on a straight line, it is said that there is a simple linear regression of Y on X. By "simple" is meant that only one independent variable, X, is used. The preceding ideas can be expressed succinctly by the following linear model:

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{2.5}$$

where Y_i = score of individual i on the dependent variable; α(alpha) = mean of the population when the value of X is zero, or the Y intercept; β(beta) = regression coefficient in the population, or the slope of the regression line; X_i = value of independent variable to which individual i was exposed; ϵ(epsilon)$_i$ = random disturbance, or error, for individual i.[1] The regression coefficient (β) indicates the effect of the independent variable on the dependent variable. Specifically, for each unit change of the independent variable, X, there is an expected change equal to the size of β in the dependent variable, Y.

The foregoing shows that each person's score, Y_i, is conceived as being composed of two parts: (1) a fixed part indicated by $\alpha + \beta X$, that is, part of the Y score for an individual exposed to a given level of X is equal to $\alpha + \beta X$ (thus, all individuals exposed to the same level of X are said to have the same part of the Y score), and (2) A random part, ϵ_i, unique to each individual, i.

Linear regression analysis is not limited to experimental research. As I amply show in subsequent chapters, it is often applied in quasi-experimental and nonexperimental research to explain or predict phenomena. Although calculations of regression statistics are the same regardless of the type of research in which they are applied, interpretation of the results depends on the specific research design. I discuss these issues in detail later in the text (see, for example, Chapters 8 through 10). For now, my emphasis is on the general analytic approach.

Equation (2.5) was expressed in parameters. For a sample, the equation is

$$Y = a + bX + e \tag{2.6}$$

where a is an estimator of α; b is an estimator of β; and e is an estimator of ϵ. For convenience, I did not use subscripts in (2.6). I follow this practice of omitting subscripts throughout the book, unless there is a danger of ambiguity. I will use subscripts for individuals when it is necessary to identify given individuals. In equations with more than one independent variable (see subsequent chapters), I will use subscripts to identify each variable.

I discuss the meaning of the statistics in (2.6) and illustrate the mechanics of their calculations in the context of a numerical example to which I now turn.

A Numerical Example

Assume that in an experiment on the effects of hours of study (X) on achievement in mathematics (Y), 20 subjects were randomly assigned to different levels of X. Specifically, there are five levels of X, ranging from one to five hours of study. Four subjects were *randomly* assigned to one hour of study, four other subjects were *randomly* assigned to two hours of study, and so on to five hours of study for the fifth group of subjects. A mathematics test serves as the measure of the dependent variable. Other examples may be the effect of the number of exposures to a list of words on the retention of the words or the effects of different dosages of a drug on reaction time or on blood pressure. Alternatively, X may be a nonmanipulated variable (e.g., age, grade in school), and Y may be height or verbal achievement. For illustrative purposes, I will treat the data of Table 2.1 as if they were obtained in a learning experiment, as described earlier.

Scientific inquiry is aimed at explaining or predicting phenomena of interest. The ideal is, of course, perfect explanation—that is, without error. Being unable to achieve this state, however, scientists attempt to minimize errors. In the example under consideration, the purpose is to explain achievement in mathematics (Y) from hours of study (X). It is very unlikely that students studying the same number of hours will manifest the same level of achievement in mathematics. Obviously, many other variables (e.g., mental ability, motivation) as well as measurement errors will introduce variability in students' performance. All sources of variability of Y, other than X, are subsumed under e in Equation (2.6). In other words, e represents the part of the Y score that is not explained by, or predicted from, X.

The purpose, then, is to find a solution for the constants, a and b of (2.6), so that explanation or prediction of Y will be maximized. Stated differently, a solution is sought for a and b so that e—errors committed in using X to explain Y—will be at a minimum. The intuitive solution of minimizing the sum of the errors turns out to be unsatisfactory because positive errors will cancel negative ones, thereby possibly leading to the false impression that small errors have been committed when their sum is small, or that no errors have been committed when their sum turns out to be zero. Instead, it is the sum of the squared errors (Σe^2) that is minimized, hence the name *least squares* given to this solution.

Given certain assumptions, which I discuss later in this chapter, the least-squares solution leads to estimators that have the desirable properties of being best linear unbiased estimators (BLUE). An estimator is said to be unbiased if its average obtained from repeated samples of size N (i.e., expected value) is equal to the parameter. Thus b, for example, is an unbiased estimator of β if the average of the former in repeated samples is equal to the latter.

Unbiasedness is only one desirable property of an estimator. In addition, it is desirable that the variance of the distribution of such an estimator (i.e., its sampling distribution) be as small as possible. The smaller the variance of the sampling distribution, the smaller the error in estimating the parameter. Least-squares estimators are said to be "best" in the sense that the variance of their sampling distributions is the smallest from among linear unbiased estimators (see Hanushek & Jackson, 1977, pp. 46–56, for a discussion of BLUE; and Hays, 1988, Chapter 5, for discussions of sampling distributions and unbiasedness). Later in the chapter, I show how the variance of the sampling distribution of b is used in statistical tests of significance and for establishing confidence intervals. I turn now to the calculation of least-squares estimators and to a discussion of their meaning.

The two constants are calculated as follows:

$$b = \frac{\Sigma xy}{\Sigma x^2} \tag{2.7}$$

$$a = \bar{Y} - b\bar{X} \tag{2.8}$$

Using these constants, the equation for predicting Y from X, or the *regression equation*, is

$$Y' = a + bX \tag{2.9}$$

where Y' = predicted score on the dependent variable, Y. Note that (2.9) does not include e $(Y - Y')$, which is the error that results from employing the prediction equation, and is referred to as the residual. It is the $\Sigma(Y - Y')^2$, referred to as the sum of squared residuals (see the following), that is minimized in the least-squares solution for a and b of (2.9).

For the data in Table 2.1, $\Sigma xy = 30$ and $\Sigma x^2 = 40$ (see the previous calculations). $\bar{Y} = 7.3$ and $\bar{X} = 3.0$ (see Table 2.1). Therefore,

$$Y' = 5.05 + .75X$$

[1] The term "linear" refers also to the fact that parameters such as those that appear in Equation (2.5) are expressed in linear form even though the regression of Y on X is nonlinear. For example, $Y = \alpha + \beta X + \beta X^2 + \beta X^3 + \epsilon$ describes the cubic regression of Y on X. Note, however, that it is X, not the β's, that is raised to second and third powers. I deal with such equations, which are subsumed under the general linear model, in Chapter 13.

In order, then, to predict Y, for a given X, multiply the X by b (.75) and add the constant a (5.05). From the previous calculations it can be seen that b indicates the expected change in Y associated with a unit change in X. In other words, for each increment of one unit in X, an increment of .75 in Y is predicted. In our example, this means that for every additional hour of study, X, there is an expected gain of .75 units in mathematics achievement, Y. Knowledge of a and b is necessary and sufficient to predict Y from X so that squared errors of prediction are minimized.

A Closer Look at the Regression Equation

Substituting (2.8) in (2.9),

$$\begin{aligned} Y' &= a + bX \\ &= (\bar{Y} - b\bar{X}) + bX \\ &= \bar{Y} + b(X - \bar{X}) \\ &= \bar{Y} + bx \qquad (2.10) \end{aligned}$$

Note that Y' can be expressed as composed of two components: the mean of Y and the product of the deviation of X from the mean of X (x) by the regression coefficient (b). Therefore, when the regression of Y on X is zero (i.e., $b = 0$), or when X does not affect Y, the regression equation would lead to a predicted Y being equal to the mean of Y for each value of X. This makes intuitive sense. When attempting to guess or predict scores of people on Y in the absence of information, except for the knowledge that they are members of the group being studied, the best prediction, in a statistical sense, for each individual is the mean of Y.

Such a prediction policy minimizes squared errors, inasmuch as the sum of the squared deviations from the mean is smaller than one taken from any other constant (see, for example, Edwards, 1964, pp. 5–6). Further, when more information about the people is available in the form of their status on another variable, X, but when variations in X are *not* associated with variations in Y, the best prediction for each individual is still the mean of Y, and the regression equation will lead to the same prediction. Note from (2.7) that when X and Y do not covary, Σxy is zero, resulting in $b = 0$. Applying (2.10) when $b = 0$ leads to $Y' = \bar{Y}$ regardless of the X values.

When, however, b is not zero (that is, when X and Y covary), application of the regression equation leads to a reduction in errors of prediction as compared with the errors resulting from predicting $\bar{Y}$ for each individual. The degree of reduction in errors of prediction is closely linked to the concept of partitioning the sum of squares of the dependent variable (Σy^2) to which I now turn.

Partitioning the Sum of Squares

Knowledge of the values of both X and Y for each individual makes it possible to ascertain how accurately each Y is predicted by using the regression equation. I will show this for the data of Table 2.1, which are repeated in Table 2.2. Applying the regression equation calculated earlier, $Y' = 5.05 + .75X$, to each person's X score yields the predicted Y's listed in Table 2.2 in the column labeled Y'. In addition, the following are reported for each person: $Y' - \bar{Y}$ (the deviation of the predicted Y from the mean of Y), referred to as deviation due to regression, and its square $(Y' - \bar{Y})^2$; $Y - Y'$ (the deviation of observed Y from the predicted Y), referred to as the residual, and its square $(Y - Y')^2$.

Table 2.2 Regression Analysis of a Learning Experiment

	X	Y	Y'	$Y' - \bar{Y}$	$(Y' - \bar{Y})^2$	$Y - Y'$	$(Y - Y')^2$
	1	3	5.80	−1.50	2.2500	−2.80	7.8400
	1	5	5.80	−1.50	2.2500	−.80	.6400
	1	6	5.80	−1.50	2.2500	.20	.0400
	1	9	5.80	−1.50	2.2500	3.20	10.2400
	2	4	6.55	−.75	.5625	−2.55	6.5025
	2	6	6.55	−.75	.5625	−.55	.3025
	2	7	6.55	−.75	.5625	.45	.2025
	2	10	6.55	−.75	.5625	3.45	11.9025
	3	4	7.30	.00	.0000	−3.30	10.8900
	3	6	7.30	.00	.0000	−1.30	1.6900
	3	8	7.30	.00	.0000	.70	.4900
	3	10	7.30	.00	.0000	2.70	7.2900
	4	5	8.05	.75	.5625	−3.05	9.3025
	4	7	8.05	.75	.5625	−1.05	1.1025
	4	9	8.05	.75	.5625	.95	.9025
	4	12	8.05	.75	.5625	3.95	15.6025
	5	7	8.80	1.50	2.2500	−1.80	3.2400
	5	10	8.80	1.50	2.2500	1.20	1.4400
	5	12	8.80	1.50	2.2500	3.20	10.2400
	5	6	8.80	1.50	2.2500	−2.80	7.8400
Σ:	60	146	146	.00	22.50	.00	107.7

Careful study of Table 2.2 will reveal important elements of regression analysis, two of which I will note here. The sum of predicted scores ($\Sigma Y'$) is equal to ΣY. Consequently, the mean of predicted scores is always equal to the mean of the dependent variable. The sum of the residuals $[\Sigma(Y - Y')]$ is always zero. These are consequences of the least-squares solution.

Consider the following identity:

$$Y = \bar{Y} + (Y' - \bar{Y}) + (Y - Y') \qquad (2.11)$$

Each Y is expressed as composed of the mean of Y, the deviation of the predicted Y from the mean of Y (deviation due to regression), and the deviation of the observed Y from the predicted Y (residual). For the data of Table 2.2, $\bar{Y} = 7.30$. The first subject's score on Y (3), for instance, can therefore be expressed thus:

$$\begin{aligned} 3 &= 7.30 + (5.80 - 7.30) + (3 - 5.80) \\ &= 7.30 + \quad (-1.50) \quad + (-2.80) \end{aligned}$$

Similar statements can be made for each subject in Table 2.2.

Earlier, I pointed out that when no information about an independent variable is available, or when the information available is irrelevant, the best prediction for each individual is the mean of the dependent variable ($\bar{Y}$), and the sum of squared errors of prediction is Σy^2. When, however, the independent variable (X) is related to Y, the degree of reduction in errors of prediction that ensues from the application of the regression equation can be ascertained. Stated differently, it is possible to discern how much of the Σy^2 can be explained based on knowledge of the regression of Y on X.

Approach the solution to this problem by using the above-noted identity—see (2.11):

$$Y = \bar{Y} + (Y' - \bar{Y}) + (Y - Y')$$

Subtracting $\bar{Y}$ from each side,

$$Y - \bar{Y} = (Y' - \bar{Y}) + (Y - Y')$$

Squaring and summing,

$$\begin{aligned}\Sigma(Y - \bar{Y})^2 &= \Sigma[(Y' - \bar{Y}) + (Y - Y')]^2 \\ &= \Sigma(Y' - \bar{Y})^2 + \Sigma(Y - Y')^2 + 2\Sigma(Y' - \bar{Y})(Y - Y')\end{aligned}$$

It can be shown that the last term on the right equals zero. Therefore,

$$\Sigma y^2 = \Sigma(Y' - \bar{Y})^2 + \Sigma(Y - Y')^2 \tag{2.12}$$

or

$$\Sigma y^2 = ss_{\text{reg}} + ss_{\text{res}}$$

where ss_{reg} = regression sum of squares and ss_{res} = residual sum of squares.

This central principle in regression analysis states that the deviation sum of squares of the dependent variable, Σy^2, is partitioned into two components: the sum of squares due to regression, or the regression sum of squares, and the sum of squares due to residuals, or the residual sum of squares. When the regression sum of squares is equal to zero, it means that the residual sum of squares is equal to Σy^2, indicating that nothing has been gained by resorting to information from X. When, on the other hand, the residual sum of squares is equal to zero, all the variability in Y is explained by regression, or by the information X provides.

Dividing each of the elements in the previous equation by the total sum of squares (Σy^2),

$$\frac{\Sigma y^2}{\Sigma y^2} = \frac{ss_{\text{reg}}}{\Sigma y^2} + \frac{ss_{\text{res}}}{\Sigma y^2}$$

$$1 = \frac{ss_{\text{reg}}}{\Sigma y^2} + \frac{ss_{\text{res}}}{\Sigma y^2} \tag{2.13}$$

The first term on the right-hand side of the equal sign indicates the proportion of the sum of squares of the dependent variable due to regression. The second term indicates the proportion of the sum of squares due to error, or residual. For the present example, $ss_{\text{reg}} = 22.5$ and $ss_{\text{res}} = 107.7$ (see the bottom of Table 2.2). The sum of these two terms, 130.2, is the Σy^2 I calculated earlier. Applying (2.13),

$$\frac{22.5}{130.2} + \frac{107.7}{130.2} = .1728 + .8272 = 1$$

About 17% of the total sum of squares (Σy^2) is due to regression, and about 83% is left unexplained (i.e., attributed to error).

The calculations in Table 2.2 are rather lengthy, even with a small number of cases. I presented them in this form to illustrate what each element of the regression analysis means. Following are three equivalent formulas for the calculation of the regression sum of squares. I do

not define the terms in the formulas, as they should be clear by now. I apply each formula to the data in Table 2.2.

$$\begin{aligned}ss_{\text{reg}} &= \frac{(\Sigma xy)^2}{\Sigma x^2} \\ &= \frac{(30)^2}{40} = 22.5\end{aligned} \tag{2.14}$$

$$\begin{aligned}ss_{\text{reg}} &= b\Sigma xy \\ &= (.75)(30) = 22.5\end{aligned} \tag{2.15}$$

$$\begin{aligned}ss_{\text{reg}} &= b^2\Sigma x^2 \\ &= (.75)^2(40) = 22.5\end{aligned} \tag{2.16}$$

I showed above that

$$\Sigma y^2 = ss_{\text{reg}} + ss_{\text{res}}$$

Therefore,

$$\begin{aligned}ss_{\text{res}} &= \Sigma y^2 - ss_{\text{reg}} \\ &= 130.2 - 22.5 = 107.7\end{aligned} \tag{2.17}$$

Previously, I divided the regression sum of squares by the total sum of squares, thus obtaining the proportion of the latter that is due to regression. Using the right-hand term of (2.14) as an expression of the regression sum of squares, and dividing by the total sum of squares,

$$r_{xy}^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} \tag{2.18}$$

where r_{xy}^2 is the squared Pearson product moment coefficient of the correlation between X and Y. This important formulation, which I use repeatedly in the book, states that the squared correlation between X and Y indicates the proportion of the sum of squares of Y (Σy^2) that is due to regression. It follows that the proportion of Σy^2 that is due to errors, or residuals, is $1 - r_{xy}^2$.

Using these formulations, it is possible to arrive at the following expressions of the regression and residual sum of squares:

$$ss_{\text{reg}} = r_{xy}^2 \Sigma y^2 \tag{2.19}$$

For the data in Table 2.2, $r_{xy}^2 = .1728$, and $\Sigma y^2 = 130.2$,

$$ss_{\text{reg}} = (.1728)(130.2) = 22.5$$

and

$$\begin{aligned}ss_{\text{res}} &= (1 - r_{xy}^2)\Sigma y^2 \\ ss_{\text{res}} &= (1 - .1728)(130.2) = 107.7\end{aligned} \tag{2.20}$$

Finally, instead of partitioning the sum of squares of the dependent variable, its *variance* may be partitioned:

$$s_y^2 = r_{xy}^2 s_y^2 + (1 - r_{xy}^2)s_y^2 \tag{2.21}$$

where $r^2_{xy}s^2_y$ = portion of the variance of Y due to its regression on X; and $(1-r^2_{xy})s^2_y$ = portion of the variance of Y due to residuals, or errors. r^2, then, is also interpreted as the proportion of the variance of the dependent variable that is accounted for by the independent variable, and $1-r^2$ is the proportion of variance of the dependent variable that is not accounted for. In subsequent presentations, I partition sums of squares or variances, depending on the topic under discussion. Frequently, I use both approaches to underscore their equivalence.

Graphic Depiction of Regression Analysis

The data of Table 2.2 are plotted in Figure 2.1. Although the points are fairly scattered, they do depict a linear trend in which increments in X are associated with increments in Y. The line that best fits the regression of Y on X, in the sense of minimizing the sum of the squared deviations of the observed Y's from it, is referred to as the regression line. This line depicts the regression equation pictorially, where a represents the point on the ordinate, Y, intercepted by the regression line, and b represents the slope of the line. Of various methods for graphing the regression line, the following is probably the easiest. Two points are necessary to draw a line. One of the points that may be used is the value of a (the intercept) calculated by using (2.8). I repeat (2.10) with a new number,

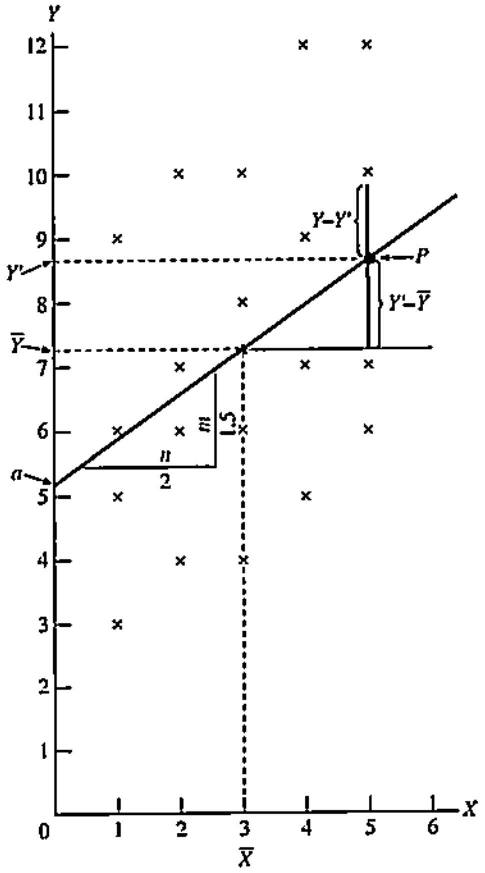


Figure 2.1

$$Y' = \bar{Y} + bx \qquad (2.22)$$

from which it can be seen that, regardless of what the regression coefficient (b) is, $Y' = \bar{Y}$ when $x = 0$—that is, when $X = \bar{X}$. In other words, the means of X and Y are always on the regression line. Consequently, the intersection of lines drawn from the horizontal (abscissa) and the vertical (ordinate) axes at the means of X and Y provides the second point for graphing the regression line. See the intersection of the broken lines in Figure 2.1.

In Figure 2.1, I drew two lines, m and n, paralleling the Y and X axes, respectively, thus constructing a right triangle whose hypotenuse is a segment of the regression line. The slope of the regression line, b, can now be expressed trigonometrically: it is the length of the vertical line, m, divided by the horizontal line, n. In Figure 2.1, $m = 1.5$ and $n = 2.0$. Thus, $1.5/2.0 = .75$, which is equal to the value of b I calculated earlier. From the preceding it can be seen that b indicates the rate of change of Y associated with the rate of change of X. This holds true no matter where along the regression line the triangle is constructed, inasmuch as the regression is described by a straight line.

Since $b = m/n$, $m = bn$. This provides another approach to the graphing of the regression line. Draw a horizontal line of length n originating from the intercept (a). At the end of n draw a line m perpendicular to n. The endpoint of line m serves as one point and the intercept as the other point for graphing the regression line.

Two other concepts are illustrated graphically in Figure 2.1: the deviation due to residual $(Y - Y')$ and the deviation due to regression $(Y' - \bar{Y})$. For illustrative purposes, I use the individual whose scores are 5 and 10 on X and Y, respectively. This individual's predicted score (8.8) is found by drawing a line perpendicular to the ordinate (Y) from the point P on the regression line (see Figure 2.1 and Table 2.2 where I obtained the same Y' by using the regression equation). Now, this individual's Y score deviates 2.7 points from the mean of Y $(10 - 7.3 = 2.7)$. It is the sum of the squares of all such deviations (Σy^2) that is partitioned into regression and residual sums of squares. For the individual under consideration, the residual: $Y - Y' = 10 - 8.8 = 1.2$. This is indicated by the vertical line drawn from the point depicting this individual's scores on X and Y to the regression line. The deviation due to regression, $Y' - \bar{Y} = 8.8 - 7.3 = 1.5$, is indicated by the extension of the same line until it meets the horizontal line originating from $\bar{Y}$(see Figure 2.1 and Table 2.2). Note that $Y' = 8.8$ for all the individuals whose $X = 5$. It is their residuals that differ. Some points are closer to the regression line and thus their residuals are small (e.g., the individual whose $Y = 10$), and some are farther from the regression line, indicating larger residuals (e.g., the individual whose $Y = 12$).

Finally, note that the residual sum of squares is relatively large when the scatter of the points about the regression line is relatively large. Conversely, the closer the points are to the regression line, the smaller the residual sum of squares. When all the points are on the regression line, the residual sum of squares is zero, and explanation, or prediction, of Y using X is perfect. If, on the other hand, the regression of Y on X is zero, the regression line has no slope and will be drawn horizontally originating from Y. Under such circumstances, $\Sigma y^2 = \Sigma(Y - Y')^2$, and all the deviations are due to error. Knowledge of X does not enhance prediction of Y.

TESTS OF SIGNIFICANCE

Sample statistics are most often used for making inferences about unknown parameters of a defined population. Recall that tests of statistical significance are used to decide whether the probability of obtaining a given estimate is small, say .05, so as to lead to the rejection of the null hypothesis that the population parameter is of a given value, say zero. Thus, for example, a small probability associated with an obtained b (the statistic) would lead to the rejection of the hypothesis that β (the parameter) is zero.

I assume that you are familiar with the logic and principles of statistical hypothesis testing (if necessary, review this topic in a statistics book, e.g., Hays, 1988, Chapter 7). As you are probably aware, statistical tests of significance are a major source of controversy among social scientists (for a compilation of articles on this topic, see Morrison & Henkel, 1970). The controversy is due, in part, to various misconceptions of the role and meaning of such tests in the context of scientific inquiry (for some good discussions of misconceptions and "fantasies" about, and misuse of, tests of significance, see Carver, 1978; Cohen, 1994; Dar, Serlin, & Omer, 1994; Guttman, 1985; Huberty, 1987; for recent exchanges on current practice in the use of statistical tests of significance, suggested alternatives, and responses from three journal editors, see Thompson, 1993).

It is very important to place statistical tests of significance, used repeatedly in this text, in a proper perspective of the overall research endeavor. Recall that all that is meant by a statistically significant finding is that the probability of its occurrence is small, assuming that the null hypothesis is true. But *it is the substantive meaning of the finding that is paramount*. Of what use is a statistically significant finding if it is deemed to be substantively not meaningful? Bemoaning the practice of exclusive reliance on tests of significance, Nunnally (1960) stated, "We should not feel proud when we see the psychologist smile and say 'the correlation is significant beyond the .01 level.' Perhaps that is the most he can say, but he has no reason to smile" (p. 649).

It is well known that given a sufficiently large sample, the likelihood of rejecting the null hypothesis is high. Thus, "if rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data" (Nunnally, 1960, p. 643; see also Rozeboom, 1960). Sound principles of research design dictate that the researcher first decide the effect size, or relation, deemed substantively meaningful in a given study. This is followed by decisions regarding the level of significance (Type I error) and the power of the statistical test (1 – Type II error). Based on the preceding decisions, the requisite sample size is calculated. Using this approach, the researcher can avoid arriving at findings that are substantively meaningful but statistically not significant or being beguiled by findings that are statistically significant but substantively not meaningful (for an overview of these and related issues, see Pedhazur & Schmelkin, 1991, Chapters 9 and 15; for a primer on statistical power analysis, see Cohen, 1992; for a thorough treatment of this topic, see Cohen, 1988).

In sum, the emphasis should be on the substantive meaning of findings (e.g., relations among variables, differences among means). Nevertheless, I do not discuss criteria for meaningfulness of findings, as what is deemed a meaningful finding depends on the characteristics of the study in question (e.g., domain, theoretical formulation, setting, duration, cost). For instance, a mean difference between two groups considered meaningful in one domain or in a relatively inexpensive study may be viewed as trivial in another domain or in a relatively costly study.

In short, criteria for substantive meaningfulness cannot be arrived at in a research vacuum. Admittedly, some authors (notably Cohen, 1988) provide guidelines for criteria of meaningfulness. But being guidelines in the abstract, they are, inevitably, bound to be viewed as unsatisfactory by

researchers when they examine their findings. Moreover, availability of such guidelines may have adverse effects in seeming to "absolve" researchers of the exceedingly important responsibility of assessing findings from the perspective of meaningfulness (for detailed discussions of these issues, along with relevant references, see Pedhazur & Schmelkin, 1991, Chapters 9 and 15). Although I will comment occasionally on the meaningfulness of findings, I will do so only as a reminder of the preceding remarks and as an admonition against exclusive reliance on tests of significance.

Testing the Regression of *Y* on *X*

Although formulas for tests of significance for simple regression analysis are available, I do not present them. Instead, I introduce general formulas that subsume simple regression analysis as a special case.

Earlier, I showed that the sum of squares of the dependent variable (Σy^2) can be partitioned into two components: regression sum of squares (ss_{reg}) and residual sum of squares (ss_{res}). Each of these sums of squares has associated with it a number of degrees of freedom (df). Dividing a sum of squares by its df yields a mean square. The ratio of the mean square regression to the mean square residual follows an F distribution with df_1 for the numerator and df_2 for the denominator (see the following). When the obtained F exceeds the tabled value of F at a preselected level of significance, the conclusion is to reject the null hypothesis (for a thorough discussion of the F distribution and the concept of df, see, for example, Hays, 1988; Keppel, 1991; Kirk, 1982; Walker, 1940; Winer, 1971). The formula for F, then, is

$$F = \frac{ss_{reg}/df_1}{ss_{res}/df_2} = \frac{ss_{reg}/k}{ss_{res}/(N-k-1)} \tag{2.23}$$

where df_1 associated with ss_{reg} are equal to the number of independent variables, k; and df_2 associated with ss_{res} are equal to N (sample size) minus k (number of independent variables) minus 1. In the case of simple linear regression, $k = 1$. Therefore, 1 df is associated with the numerator of the F ratio. The df for the denominator are $N - 1 - 1 = N - 2$.

For the numerical example in Table 2.2, $ss_{reg} = 22.5$; $ss_{res} = 107.7$; and $N = 20$.

$$F = \frac{22.5/1}{107.7/18} = 3.76$$

with 1 and 18 df.

Assuming that the researcher set α (significance level) = .05, it is found that the tabled F with 1 and 18 df is 4.41 (see Appendix B for a table of the F distribution). As the obtained F is smaller than the tabled value, it is concluded that the regression of Y on X is statistically not different from zero. Referring to the variables of the present example (recall that the data are illustrative), it would be concluded that the regression of achievement in mathematics on study time is statistically not significant at the .05 level or that study time does not significantly (at the .05 level) affect mathematics achievement. Recall, however, the important distinction between statistical significance and substantive meaningfulness, discussed previously.

Testing the Proportion of Variance Accounted for by Regression

Earlier, I said that r^2 indicates the proportion of variance of the dependent variable accounted for by the independent variable. Also, $1 - r^2$ is the proportion of variance of the dependent variable *not* accounted for by the independent variable or the proportion of error variance. The significance of r^2 is tested as follows:

$$F = \frac{r^2/k}{(1-r^2)/(N-k-1)} \tag{2.24}$$

where k is the number of independent variables.

For the data of Table 2.2, $r^2 = .1728$; hence,

$$F = \frac{.1728/1}{(1-.1728)/(20-1-1)} = 3.76$$

with 1 and 18 *df.* Note that the same F ratio is obtained whether one uses sums of squares or r^2. The identity of the two formulas for the F ratio may be noted by substituting (2.19) and (2.20) in (2.23):

$$F = \frac{r^2\Sigma y^2/k}{(1-r^2)\Sigma y^2/(N-k-1)} \tag{2.25}$$

where $r^2\Sigma y^2 = ss_{\text{reg}}$ and $(1-r^2)\Sigma y^2 = ss_{\text{res}}$. Canceling Σy^2 from the numerator and denominator of (2.25) yields (2.24). Clearly, it makes no difference whether sums of squares or proportions of variance are used for testing the significance of the regression of Y on X. In subsequent presentations I test one or both terms as a reminder that you may use whichever you prefer.

Testing the Regression Coefficient

Like other statistics, the regression coefficient, b, has a standard error associated with it. Before I present this standard error and show how to use it in testing the significance of b, I introduce the variance of estimate and the standard error of estimate.

Variance of Estimate. The variance of scores about the regression line is referred to as the variance of estimate. The parameter is written as $\sigma^2_{y.x}$, which denotes the variance of Y given X. The sample unbiased estimator of $\sigma^2_{y.x}$ is $s^2_{y.x}$, and is calculated as follows:

$$s^2_{y.x} = \frac{\Sigma(Y-Y')^2}{N-k-1} = \frac{ss_{\text{res}}}{N-k-1} \tag{2.26}$$

where Y = observed Y; Y' = predicted Y; N = sample size; and k = number of independent variables. The variance of estimate, then, is the variance of the residuals. It indicates the degree of variability of the points about the regression line. Note that the rightmost expression of $s^2_{y.x}$ is the same as the denominator of the F ratio presented earlier—see (2.23). The variance of estimate, then, is the mean square residual (*MSR*).

For the data in Table 2.2,

$$s^2_{y.x} = MSR = \frac{107.7}{18} = 5.983$$

The *standard error of estimate* is the square root of the variance of estimate, that is, the standard deviation of the residuals:

$$s_{y.x} = \sqrt{\frac{\Sigma(Y-Y')^2}{N-k-1}} = \sqrt{\frac{ss_{\text{res}}}{N-k-1}} \tag{2.27}$$

For our data, $s_{y.x} = \sqrt{5.983} = 2.446$.

The standard error of b, the regression coefficient, is

$$s_b = \sqrt{\frac{s^2_{y.x}}{\Sigma x^2}} = \frac{s_{y.x}}{\sqrt{\Sigma x^2}} \tag{2.28}$$

where s_b = standard error of b; $s^2_{y.x}$ = variance of estimate; $s_{y.x}$ = standard error of estimate; and Σx^2 = sum of squares of the independent variable, X. s_b is the standard deviation of the sampling distribution of b and can therefore be used for testing the significance of b:

$$t = \frac{b}{s_b} \tag{2.29}$$

where t is the t ratio with *df* associated with $s^2_{y.x}$: $N-k-1$ (N = sample size; k = number of independent variables).

For the data of Table 2.2, $b = .75$; $s^2_{y.x} = 5.983$; and $\Sigma x^2 = 40$ (see the previous calculations). Hence,

$$t = \frac{.75}{\sqrt{\frac{5.983}{40}}} = \frac{.75}{\sqrt{.1496}} = 1.94$$

with 18 *df* (20 − 1 − 1), $p > .05$. In simple linear regression, testing the significance of b is the same as testing the regression of Y on X by using sums of squares or proportions of variance. The conclusions are, of course, the same. Based on the previous test, you can conclude that the regression coefficient (b) is statistically not significantly different from zero (at the .05 level).

Recall that when, as in the present example, the numerator *df* for F is 1, $t^2 = F$. Thus, $1.94^2 = 3.76$, the F ratio I obtained earlier. There are, however, situations when the use of the t ratio is preferable to the use of the F ratio. First, although I used (2.29) to test whether b differs significantly from zero, it may be used to test whether b differs significantly from any hypothesized value. The formula takes the following form:

$$t = \frac{b-\beta}{s_b} \tag{2.30}$$

where β is the hypothesized regression coefficient.

Assume that in the numerical example under consideration I had reason to hypothesize that the regression coefficient in the population is .50. To test whether the obtained b differs significantly from the parameter, I would calculate

$$t = \frac{.75-.50}{.3868} = .65$$

with 18 *df.* This is obviously statistically not significant at the .05 level. In other words, the obtained b is statistically not significantly different from the hypothesized regression coefficient.

Second, using a t ratio, confidence intervals can be set around the regression coefficient. The use of confidence intervals in preference to tests of statistical significance has been strongly advocated by various authors (e.g., Hays, 1988; Nunnally, 1960; Rozeboom, 1960). Because of space considerations, I will only sketch some arguments advanced in favor of the use of confidence intervals. Probably the most important argument is that a confidence interval provides more information than does a statement about rejecting (or failing to reject) a null hypothesis, which is almost always false anyway. Moreover, a confidence interval enables one to test simultaneously all possible null hypotheses. The narrower the confidence interval, the smaller the range of possible null hypotheses, and hence the greater the confidence in one's findings. In view of the preceding, confidence intervals should become an integral part of research reports, or standard errors of statistics should be reported so that interested readers may use them in assessing the findings.

The confidence interval for b is

$$b \pm t_{(\alpha/2,\,df)}s_b$$

where t is the tabled t ratio at $\alpha/2$ with df associated with standard error of estimate, and s_b is the standard error of b. Assuming that I wish to obtain the 95% confidence interval in the present example, the tabled t at .05/2 (0.025) with 18 df is 2.101 (see the table of t distribution in statistics books, or take $\sqrt{F}$ with 1 and 18 df from Appendix B), $b = .75$, and $s_b = .3868$. The 95% confidence interval is

$$.75 \pm (2.101)(.3868) = -.0627 \text{ and } 1.5627$$

or

$$-.0627 \le \beta \le 1.5627$$

As is pointed out in various statistics books (e.g., Hays, 1988; Li, J. C. R., 1964; Snedecor & Cochran, 1967), it is inappropriate to conclude that the parameter lies within the given confidence interval. Rather, the confidence interval is meant to serve as "*an estimated range of values with a given high probability of covering the true population value*" (Hays, 1988, p. 206). Stated differently, what is implied by the construction of a confidence interval is that, if many such intervals were to be constructed in like fashion, $1 - \alpha$ (95% in the present example) of them would contain the parameter (β in the present example). It is hoped that the interval being constructed is one of them. Note that in the present example the interval includes zero, thereby indicating that β is statistically not significantly different from zero at the .05 level.

Third, using a t ratio, instead of F, one may apply one-tailed tests of significance. Assume that I had reason to test the b at .05 using a one-tailed test, then the t I obtained previously (1.94 with 18 df) would have been declared statistically significant (a t of 1.73 is required for a one-tailed test at .05 level with 18 df). For discussions of one-tailed versus two-tailed tests of significance, see Burke (1953), Cohen (1965), Guilford and Fruchter (1978), Kaiser (1960), Pillemer (1991).

FACTORS AFFECTING PRECISION OF THE REGRESSION EQUATION

Careful study of the formulas for tests of significance in regression analysis reveals that three factors affect them: (1) sample size (N); (2) the scatter of points about the regression line, indicated by $\Sigma(Y - Y')^2$; and (3) the range of values selected for the X variable, reflected by Σx^2.

To demonstrate these points, I repeat formulas I used earlier with new numbers:

$$F = \frac{ss_{reg}/k}{} \qquad (2.31)$$

Other things equal, the larger N the smaller the denominator, and the larger the F ratio. Holding N constant, the smaller the scatter about the regression line (i.e., the smaller the ss_{res}), the larger ss_{reg}, and consequently the larger the F ratio:

$$t = \frac{b}{\sqrt{\dfrac{s^2_{y.x}}{\Sigma x^2}}} \qquad (2.32)$$

Other things equal, the larger Σx^2, the smaller the s_b, and consequently, the larger the t ratio. Holding X constant, $s^2_{y.x}$ is a function of the scatter of points about the regression line. Therefore, the smaller $s^2_{y.x}$, the smaller the s_b, and the larger the t ratio. Similar reasoning applies also to formulas in which the proportion of variance accounted for is tested for significance.

I will illustrate the effects of the above-noted factors by selecting, in turn, different parts of the data of Table 2.2. These are reported in Table 2.3 and are plotted in Figure 2.2. Also given in Table 2.3, for easy reference, are some of the formulas I used in this chapter. I suggest that you repeat some of the calculations used in Table 2.3 as an exercise. Obtaining the same results by using one or more algebraic identities will help make the ideas of regression analysis part of your vocabulary.

Table 2.3 Four Sets of Illustrative Data

	(a)		(b)		(c)		(d)	
	X	Y	X	Y	X	Y	X	Y
	1	5	1	3	1	3	2	4
	1	6	1	9	1	5	2	6
	2	6	2	4	1	6	2	7
	2	7	2	10	1	9	2	10
	3	6	3	4	5	6	3	4
	3	8	3	10	5	7	3	6
	4	7	4	5	5	10	3	8
	4	9	4	12	5	12	3	10
	5	7	5	6			4	5
	5	10	5	12			4	7
							4	9
							4	12
N:	10		10		8		12	
ss:	20	20.9	20	108.5	32	59.5	8	70.67
a:	4.85		5.25		5.00		5.08	
b:	.75		.75		.75		.75	
r^2:	.54		.10		.30		.06	
ss_{reg}:	11.25		11.25		18.00		4.50	
ss_{res}:	9.65		97.25		41.50		66.17	
$s_{y.x}$:	1.10		3.49		2.63		2.57	
F:	9.33(1,8)		.93(1,8)		2.60(1,6)		.68(1,10)	
t:	3.05(8)		.96(8)		1.61(6)		.82(10)	

$$b = \frac{\Sigma xy}{\Sigma x^2} \qquad a = \bar{Y} - b\bar{X} \qquad ss_{reg} = b\Sigma xy = b^2\Sigma x^2 = r^2\Sigma y^2$$

$$ss_{res} = \Sigma y^2 - ss_{reg} = (1 - r^2)\Sigma y^2 \qquad F = \frac{ss_{reg}/k}{} \qquad t = \frac{b}{}$$

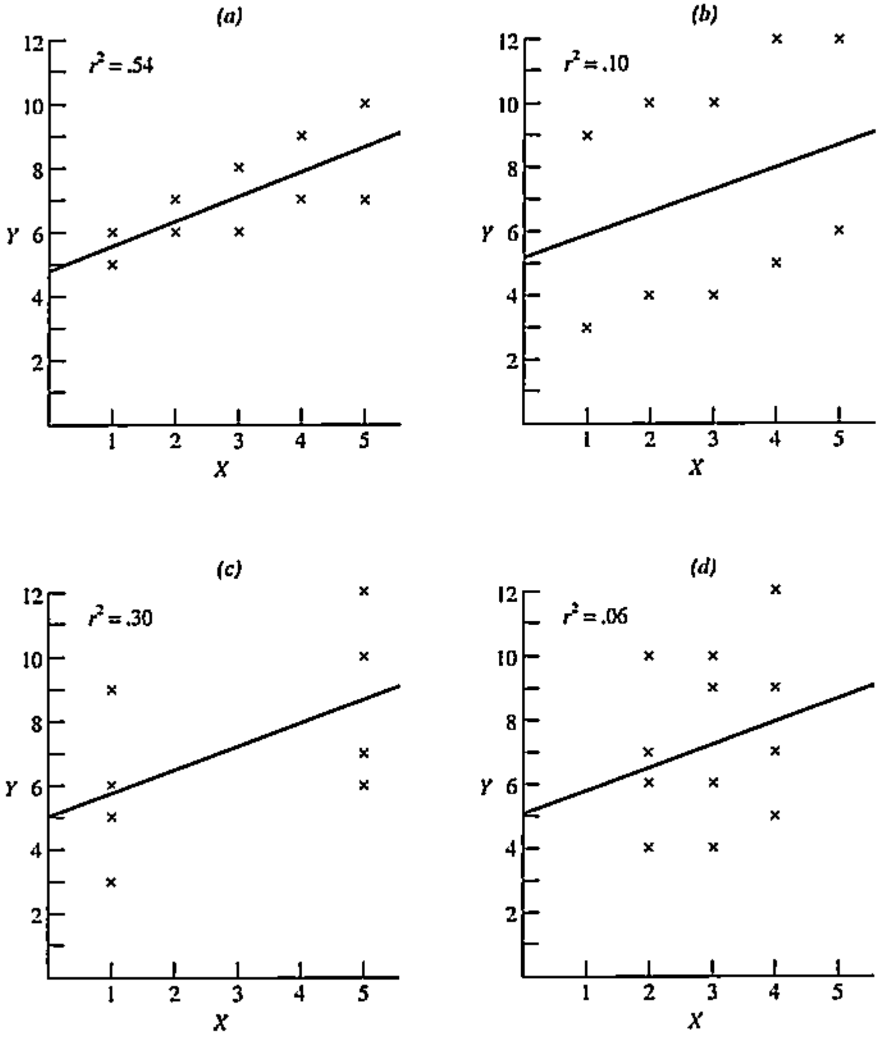


Figure 2.2

I will note several characteristics of Table 2.3 and Figure 2.2. The regression coefficient in the four sets of data is the same ($b = .75$). The F ratio associated with the b, however, is statistically significant only in (a). Compare and contrast (a) with (b): having an identical b and identical Σx^2 (20), the regression sum of squares ($b^2\Sigma x^2$) is the same in both (11.25). N is also the same in both. They differ in the residual sum of squares—9.65 for (a) and 97.25 for (b)—which reflects the scatter of points about the regression line. Consequently, the standard error of estimate, which is the standard deviation of the residuals, of (a) is about one-third of what it is for (b), and similarly for the standard errors of the two b's. Note also that the proportion of variance (r^2) accounted for in (a) is .54, whereas in (b) it is .10.

Compare and contrast (c) and (d). The former consists of the extreme values of X (1 and 5), whereas the latter consists of the intermediate values of X (2, 3, and 4). Σx^2 in (c) is four times

that of (d): 32 and 8, respectively. As the b's are the same in both sets, ss_{reg} in (c) is four times that of (d). Note also that the standard errors of estimate are very similar in both sets. Although the $s_{y.x}$ is slightly larger in (c), the standard error of b in (c) is .4649, as compared with .9086, which is the standard error of b in (d). This is directly a function of the different Σx^2's in the two sets, leading to a t of 1.61 in (c) and a t of .82 in (d). Also, the proportion of variance accounted for (r^2) in (c) is .30, whereas in (d) it is .06.

Study the example carefully to note the relations and principles I discussed and others I did not discuss.

ASSUMPTIONS

The intelligent and valid application of analytic methods requires knowledge of the rationale, hence the assumptions, behind them. Knowledge and understanding when violations of assumptions lead to serious biases, and when they are of little consequence, are essential to meaningful data analysis. Accordingly, I discuss the assumptions underlying simple linear regression and note some consequences of departures from them.

It is assumed that X, the independent variable, is a fixed variable. What this means is that if the experiment were to be replicated, the same values of X would have to be used. Referring to the numerical example used earlier, this means that the same values of hours of study would have to be used if the experiment were to be replicated.[2]

Inasmuch as the researcher is at liberty to fix the X values in an experimental study, or to select them in a nonexperimental study, the question arises: What are the considerations in determining the X values? Earlier, I showed that the larger Σx^2, the smaller the standard error of the regression coefficient. Therefore, selecting extreme X values optimizes tests of statistical significance. In the limiting case, selecting only two extreme X values maximizes the Σx^2. This, however, forces the regression to be linear even when it is curvilinear along the X continuum. Using only two X values thus precludes the possibility of determining whether the regression departs from linearity (Chapter 13). Decisions about the number of X values, their range, and spacing are to be made in light of substantive interests and theory regarding the process being modeled (Cox, 1958, pp. 138–142; Draper & Smith, 1981, pp. 51–55).

It is further assumed that X is measured without error.

The population means of the Y's at each level of X are assumed to be on a straight line. In other words, the regression of Y on X is assumed to be linear.

Unlike X, Y is a random variable, which means that Y has a range of possible values, each having an associated probability (for discussions of random variables, see Edwards, 1964, Chapter 4; Hays, 1988, pp. 92–106; Winer, 1971, Appendix A). Recall, however, that each Y score (Y_i) is assumed to be composed of a fixed component ($\alpha + \beta X$) and random error (ϵ_i).

The remaining assumptions, which are concerned with the errors, are (1) the mean of errors for each observation, Y_i, over many replications is zero; (2) errors associated with one observation, Y_i, are not correlated with errors associated with any other observation, Y_j; (3) the variance of errors at all values of X is constant, that is, the variance of errors is the same at all levels of X (this property is referred to as *homoscedasticity*, and when the variance of errors differs at

[2] Later in this chapter, I discuss linear regression analysis when Y is a random variable.

different X values, *heteroscedasticity* is indicated); and (4) the errors are assumed to be not correlated with the independent variable, X.

The preceding assumptions are necessary to obtain best linear unbiased estimators (see the discussion earlier in the chapter). For tests of significance, an additional assumption is required, namely that the errors are normally distributed.

Violation of Assumptions

It has been demonstrated that regression analysis is generally robust in the face of departures from assumptions, except for measurement errors and specification errors (for detailed discussions see Bohrnstedt & Carter, 1971; Ezekiel & Fox, 1959; Fox, 1968; Hanushek & Jackson, 1977; Snedecor & Cochran, 1967). Therefore, I comment on these topics only.

Measurement Errors. Measurement errors in the *dependent* variable do not lead to bias in the estimation of the regression coefficient, but they do lead to an increase in the standard error of estimate, thereby weakening tests of statistical significance.

Measurement errors in the *independent* variable lead to underestimation of the regression coefficient. It can be shown that the underestimation is related to the reliability of the measure of the independent variable. Reliability is a complex topic that I cannot discuss here (for different models of reliability and approaches to its estimation, see Nunnally, 1978, Chapters 6 and 7; Pedhazur & Schmelkin, 1991, Chapter 5). For present purposes I will only point out that, broadly speaking, reliability refers to the precision of measurement. Generally symbolized as r_{tt}, reliability can range from .00 to 1.00. The higher the r_{tt}, the more precise the measurement. Now,

$$b = \beta r_{tt} \tag{2.33}$$

where b = the statistic and β = the parameter. Equation (2.33) shows that with perfect reliability of the measure of X (i.e., $r_{tt} = 1.00$), $b = \beta$. When the reliability is less than 1.00, as it almost always is, b underestimates β. When $r_{tt} = .70$, say, there is a 30% underestimation of β. In experimental research, the independent variable is under the control of the experimenter. Consequently, it is reasonable to expect that, with proper care, high reliability of X may be realized.[3] In nonexperimental research, on the other hand, the reliability of the measure of the independent variable tends to be low to moderate (i.e., ranging from about .5 to about .8). This is particularly the case for certain attributes used in such research (e.g., cognitive styles, self–concept, ego strength, attitudes). Thus, bias in estimating the regression coefficient in nonexperimental research may be considerable.

Most researchers seem unaware of the biasing effects of measurement errors.[4] Among those who are aware of such effects, many are complacent about them, presuming their stance to be conservative as it leads to underestimation rather than overestimation of the regression coefficient. However, two things will be noted. First, when one wishes to test whether the regression of Y on X is the same in two groups, say, as in attribute-treatments-interaction (ATI) designs (Chapter 14), conclusions may be seriously in error if there is substantial variation in the reliabilities of the measure of X for the groups under consideration. The same is true for analysis of covariance designs (Chapter 15), and for some designs dealing with test bias (Chapter 14). Second, effects of measurement errors in designs with more than one independent variable (i.e., in multiple regression) are more complex, and the direction of the bias may be in overestimation or underestimation (Chapter 10).

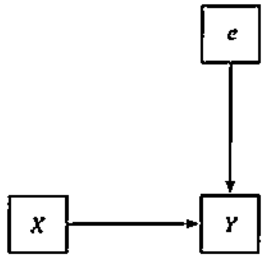
The preceding remarks apply to the effects of random measurement errors. Effects of nonrandom errors are more complex and difficult to trace. In sum, it was no exaggeration on the part of Fleiss and Shrout (1977) when they stated that "effects of measurement errors can become devastating" (p. 1190).

Specification Errors. Broadly, specification errors refer to any errors committed in specifying the model to be tested or to the violation of any of the assumptions that underlie the model. The term is generally used in a narrower sense to refer to errors in model specification (Hanushek & Jackson, 1977, pp. 79–86; Kmenta, 1971, pp. 391–405). The model used (i.e., the regression equation) represents a theoretical conception of the phenomenon under study. When the model is not tenable from a theoretical frame of reference, specification errors are indicated. Among such errors are (1) omission of relevant variables from the equation, (2) inclusion of irrelevant variables in the equation, and (3) specifying that the regression is linear when it is curvilinear.

I discuss specification errors in detail in Chapter 10. At this stage, I will only draw attention to serious biasing effects such errors may have. Earlier, I stated that under errors, e, are subsumed all variables, other than X, that affect the dependent variable, Y. I also stated that e is assumed to be not correlated with X. This situation is depicted in Figure 2.3.

Now, suppose that a relevant variable (or variables) not included in the equation is correlated with X. Since such a variable is subsumed under e, it follows that e and X are correlated, thus violating a crucial assumption (see preceding) and leading to bias in the estimation of the regression coefficient for X. I show the nature of the bias in Chapter 10. For now, suffice it to say that it may be very serious and lead to erroneous conclusions about the effect of X on Y.

The potential for specification errors of the kind just described stems in part from the type of research in which regression analysis is used. Particularly pertinent in this context is the distinction between experimental and nonexperimental research (see Pedhazur & Schmelkin, 1991, Chapters, 10, 12, and 14). In experimental research, subjects are randomly assigned to different levels of X, hence it is reasonable to assume that the effects of all variables, other than X, are



[3] Chatfield (1991) relates an example of an experimenter who faulted the computer program for regression analysis when, contrary to an expectation of an almost perfect fit, it indicated that 10% of the variance of the dependent variable was accounted for. It turned out that the fault was with the manner in which the observations were collected. Replication of the experiment with appropriate controls "resulted in a 99% fit!" (p. 243).

[4] A case in point are researchers who, adopting rules of thumb or "standards" of reliability proposed by various authors (notably Nunnally, 1967, 1978), contend that the reliabilities of their measures that hover around .70 are "acceptable." They then proceed to carry out regression analysis without the slightest hint that it is adversely affected by measurement errors. For a couple of recent examples, see Hobfoll, Shoham, and Ritter (1991, p. 334) and Thomas and Williams (1991,

equally distributed in the various groups. In other words, the assumption about the absence of a relation between X and e is tenable, though not a certainty. The assumption may be highly questionable when the research is nonexperimental. In a very good discussion of the biasing effects of measurement and specification errors, Bohrnstedt and Carter (1971) pointed out that researchers often ignore such errors. "We can only come to the sobering conclusion, then, that many of the published results based on regression analysis . . . are possible distortions of whatever reality may exist" (p. 143). This indictment should serve to alert researchers to possible distortions in their analyses and the need to take steps to avoid them or to cope with them.

DIAGNOSTICS

Diagnostics aimed at affording a better understanding of one's results as well as the detection of possible violations of assumptions and influential observations have grown in sophistication and complexity in recent years. At this stage, I give only a rudimentary introduction to this topic. For more detailed treatments, see Chapters 3 and 10.

Data and Residual Plots

An indispensable approach for a better understanding of one's results and for discerning whether some of the assumptions (e.g., linearity, homoscedasticity) are tenable is the study of data plots (for very good discussions and instructive illustrations, see Anscombe, 1973; Atkinson, 1985; Cleveland & McGill, 1984; du Toit, Steyn, & Stumpf, 1986).

Another very useful approach is the study of residual plots. Probably the simplest and most useful plots are those of the standardized residuals against corresponding X's or predicted Y's (in raw or standardized form). You have, doubtless, encountered standard scores in introductory courses in statistics or measurement. Recall that

$$z = \frac{X - \bar{X}}{s} \tag{2.34}$$

where z = standard score; X = raw score; $\bar{X}$ = mean; and s = standard deviation. As I pointed out earlier, the mean of residuals is zero, and the standard deviation of residuals is the standard error of estimate ($s_{y.x}$). Therefore, to standardize residuals, divide each residual by $s_{y.x}$. Predicted scores (Y') are, of course, obtained through the application of the regression equation.

I use the data of Table 2.2 to illustrate residual plots and to discuss some approaches to studying them. In Table 2.2 the predicted Y's are reported in the column labeled Y', and the residuals are reported under $Y - Y'$. For these data, $s_{y.x}$ = 2.446—see the calculations following (2.26) and (2.27). I divided the residuals by 2.446 to obtain standardized residuals, which I plotted against the predicted Y's in Figure 2.4.

Several things are being sought when studying plots like that of Figure 2.4. First, do the points appear to scatter randomly about the line originating from the mean of the residuals, depicting what appears to be a rectangle? Figure 2.5 illustrates departure from such a scatter, sug-

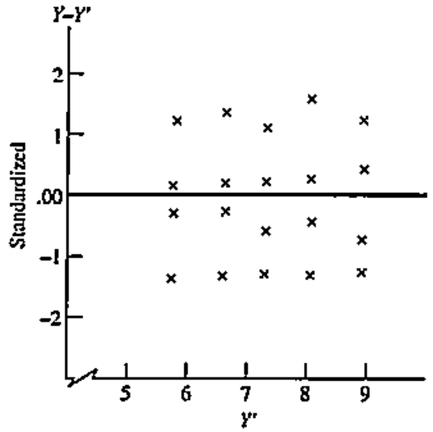


Figure 2.4

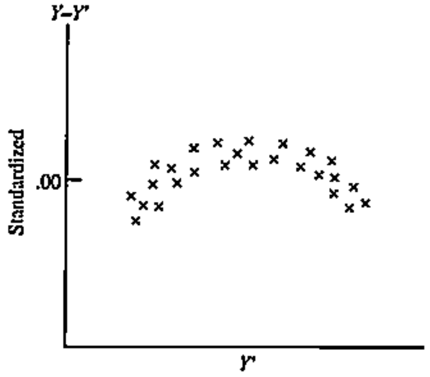


Figure 2.5

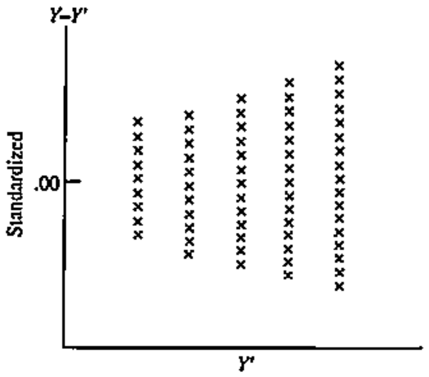


Figure 2.6

Second, are the points scattered evenly about the line originating from the mean of the residuals? If they are not, as exemplified in Figure 2.6, heteroscedasticity is indicated.

Third, are there outliers? I discuss outliers in Chapter 3. For now I will only point out that ob-

REGRESSION ANALYSIS WHEN X IS A RANDOM VARIABLE

Thus far, my discussion was limited to designs in which X is fixed. As is well known, however, in much of behavioral research, the researcher does not, or cannot, fix X. Instead, a sample is drawn from a defined population, and measures of X and Y are obtained. Thus, both X and Y are random variables. It was shown (e.g., Fox, 1984, pp. 61–63; Kmenta, 1971, pp. 297–304; Snedecor & Cochran, 1967, pp. 149–150) that when the other assumptions are reasonably met, particularly the assumption that X and e are not correlated, least-squares estimators and tests of significance presented earlier apply equally to the situation when both X and Y are random variables.

When both variables are random, the researcher may choose to study the regression of Y on X, or the regression of X on Y. The equation for the regression of X on Y is

$$X' = a + bY \tag{2.35}$$

where X' is the predicted X. The formulas for a and b are

$$b = \frac{\Sigma xy}{\Sigma y^2} \tag{2.36}$$

$$a = \overline{X} - b\overline{Y} \tag{2.37}$$

Compare (2.36) and (2.37) with the corresponding formulas for the regression of Y on X—(2.7) and (2.8)—and note the similarities and the differences.

Generally, subscripts are used to distinguish between the constants of the two equations. Thus, for example, b_{yx} is used to denote the regression coefficient for the regression of Y on X, whereas b_{xy} is used to denote the regression coefficient for the regression of X on Y. When there is no ambiguity about the designations of the independent variable and the dependent variable, it is convenient to dispose of the use of subscripts—a practice I followed in preceding sections of this chapter.

THE CORRELATION MODEL

Unlike the regression model, in the correlation model no distinction is made between an independent and a dependent variable. Instead, the nature (i.e., positive or negative) and degree of relation between two variables is sought.

Although the concept of covariance, which I discussed earlier in this chapter, is useful, it is difficult to interpret because its magnitude is affected by the specific scales used. For example, in studying the covariance between height and weight, one might express height in inches, say, and weight in ounces. If, instead, one were to express height in feet and weight in pounds, the underlying relation between the two variables would, of course, not change but the value of the covariance would change. This problem may be overcome by using the correlation coefficient:

$$\rho = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{2.38}$$

where ρ (rho) = population correlation coefficient; Σxy = sum of cross products; σ_{xy} = covariance of X and Y; and σ_x, σ_y = standard deviation of X and Y, respectively.

Earlier, I pointed out that dividing a deviation score by the standard deviation yields a standard score—see (2.34). Inspection of (2.38), particularly the first term on the right, reveals that the scores on X and Y are standardized. To make this more explicit, I express ρ in standard score form,

$$\rho = \frac{\Sigma z_x z_y}{N} \tag{2.39}$$

from which it can be seen clearly that the correlation coefficient is a covariance of standard scores, therefore not affected by the specific units used to measure X and Y. It can be shown that the maximum value of ρ is $|1.00|$. $\rho = +1.00$ indicates a perfect positive correlation, whereas $\rho = -1.00$ indicates a perfect negative correlation. $\rho = .00$ indicates no *linear* relation between X and Y. The closer ρ is to $|1.00|$, the stronger is the relation between X and Y. Also, the correlation coefficient is a symmetric index: $\rho_{xy} = \rho_{yx}$.

The sample correlation, r, is

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{2.40}$$

where s_{xy} = sample covariance; and s_x, s_y = sample standard deviations of X and Y, respectively. Of various formulas for the calculation of r, two that are particularly easy to use are

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \tag{2.41}$$

where Σxy = sum of the products; and Σx^2, Σy^2 = sums of squares of X and Y, respectively, and

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \tag{2.42}$$

where all the terms are expressed in raw scores. Formula (2.42) is particularly useful for calculations by hand, or with the aid of a calculator.

Recall that in the regression model Y is a random variable assumed to be normally distributed, whereas X is fixed, its values being determined by the researcher. In the correlation model, both X and Y are random variables assumed to follow a bivariate normal distribution. That is, the joint distribution of the two variables is assumed to be normal. The assumptions about homoscedasticity and about the residuals, which I discussed earlier in this chapter, apply also to the correlation model.

Although r and r^2 enter regression calculations, r is irrelevant in the regression model. Therefore, interpreting r as indicating the linear relation between X and Y is inappropriate. Look back at Figure 2.2 and note that the b's are the same in the four sets of data, but the r's range from a low of .24 to a high of .73. Careful study of the figure and the calculations associated with it will reveal that r changes as a function of scatter of points about the regression line and the variability of X. The greater the scatter, other things equal, the lower r is. The smaller the variability of X, other things equal, the lower r is. As I said earlier, in regression analysis the researcher may increase the variability of X at will, thereby increasing r. There is nothing wrong in doing this provided one does not interpret r as the sample estimate of the linear correlation between two random variables.

Earlier, I showed that r^2 is a meaningful term in regression analysis, indicating the proportion of variance of Y accounted for by X. Moreover, $1 - r^2$ is closely related to the variance of estimate and the standard error of estimate. Also, I expressed the residual sum of squares as

$$ss_{\text{res}} = (1 - r^2)\Sigma y^2 \tag{2.43}$$

and the variance of estimate as

$$s^2_{y.x} = \frac{(1 - r^2)\Sigma y^2}{N - 2} \tag{2.44}$$

From (2.43) and (2.44) it can be seen that when $1 - r^2$ is zero, ss_{res} and $s_{y.x}$ are zero. In other words, no error is committed. This, of course, happens when $r^2 = 1.00$, indicating that all the variance is due to regression. The larger r^2 is, the smaller is $1 - r^2$ (the proportion of variance due to error). It is this use of r^2 that is legitimate and meaningful in regression analysis, and not the use of its square root (i.e., r) as an indicator of the linear correlation between two random variables.

The regression model is most directly and intimately related to the primary goals of scientific inquiry: explanation and prediction of phenomena. When a scientist wishes, for instance, to state the expected changes in Y because of manipulations of, or changes in, X, it is the regression coefficient, b, that provides this information. Because of the greater potency of the regression model, some writers (e.g., Blalock, 1968; Tukey, 1954) argued that it be used whenever possible and that the correlation model be used only when the former cannot be applied. Tukey (1954), who referred to himself as a member of the "informal society for the suppression of the correlation coefficient" (p. 38), advanced strong arguments against its use. He maintained that "It is an enemy of generalization, a focuser on the 'here and now' to the exclusion of the 'there and then'" (Tukey, 1969, p. 89). Only bad reasons came to Tukey's mind when he pondered the appeal the correlation coefficient holds for behavioral researchers.

> Given two perfectly meaningless variables, one is reminded of their meaninglessness when a regression coefficient is given, since one wonders how to interpret its value. A correlation coefficient is less likely to bring up the unpleasant truth—we *think* we know what $r = .7$ means. *Do we?* How often? Sweeping things under the rug is the enemy of good data analysis. Often, using the correlation coefficient is "sweeping under the rug" with a vengeance.

Expressing the same point of view, though in a less impassioned tone, Fisher (1958) stated, "The regression coefficients are of interest and scientific importance in many classes of data where the correlation coefficient, if used at all, is an artificial concept of no real utility" (p. 129).

These are, admittedly, postures with which some writers may disagree. The important point, however, is that you keep in mind the differences between the regression and the correlation models and apply the one most suited for the given research problem.

For further discussions of the distinction between the two models, see Binder (1959), Ezekiel and Fox (1959, pp. 279–280), Fox (1968, pp. 167–190, 211–223), Kendall (1951), and Warren (1971). Although I deal occasionally with the correlation model, my primary concern in this book is with the regression model.

CONCLUDING REMARKS

In this chapter, I introduced elements of simple linear regression analysis. Along with the

In addition, I discussed assumptions underlying the regression model and pointed out that measurement errors in the independent variable and specification errors deserve special attention because they may lead to serious distortions of results of regression analysis. Finally, I discussed the distinction between the regression and correlation models.

STUDY SUGGESTIONS

1. You will do well to study simple regression analysis from a standard text. The following two sources are excellent—and quite different: Hays (1988, Chapter 13) and Snedecor and Cochran (1967, Chapter 6). Although these two chapters are somewhat more difficult than certain other treatments, they are both worth the effort.
2. Here are X and Y scores (the second, third, and fourth pairs of columns are continuations of the first pair of columns):

X	Y	X	Y	X	Y	X	Y
2	2	4	4	4	3	9	9
2	1	5	7	3	3	10	6
1	1	5	6	6	6	9	6
1	1	7	7	6	6	4	9
3	5	6	8	8	10	4	10

Calculate the following:
(a) Means, sums of squares and cross products, standard deviations, and the correlation between X and Y.
(b) Regression equation of Y on X.
(c) Regression and residual sum of squares.
(d) F ratio for the test of significance of the regression of Y on X, using the sums of squares (i.e., ss_{reg} and ss_{res}) and using r^2_{xy}.
(e) Variance of estimate and the standard error of estimate.
(f) Standard error of the regression coefficient.
(g) t ratio for the test of the regression coefficient. What should the square of the t equal? (That is, what statistic calculated above should it equal?)

Using the regression equation, calculate the following:
(h) Each person's predicted score, Y', on the basis of the X's.
(i) The sum of the predicted scores and their mean.
(j) The residuals, $(Y - Y')$; their sum, $\Sigma(Y - Y')$, and the sum of the squared residuals, $\Sigma(Y - Y')^2$.
(k) Plot the data, the regression line, and the standardized residuals against the predicted scores.

3. Following are summary data from a study: $N = 200$; $\bar{X} = 60$; $\bar{Y} = 100$; $s_x = 6$; $s_y = 9$; $r_{xy} = .7$. Calculate the following:
(a) Sum of squares for X and Y and the sum of products.
(b) Proportion of variance of Y accounted for by X.
(c) Regression of Y on X.
(d) Regression sum of squares.
(e) Residual sum of squares.
(f) F ratio for the test of significance of the regression of Y on X.

ANSWERS

2. (a) $\bar{X} = 4.95$; $\bar{Y} = 5.50$; $\Sigma x^2 = 134.95$; $\Sigma y^2 = 165.00$; $\Sigma xy = 100.50$; $s_x = 2.6651$; $s_y = 2.9469$; $r_{xy} = .6735$
(b) $Y' = 1.81363 + .74472X$
(c) $ss_{\text{reg}} = 74.84439$; $ss_{\text{res}} = 90.15561$
(d) $F = 14.94$, with 1 and 18 *df*
(e) $s^2_{y.x} = 5.00865$; $s_{y.x} = 2.23800$
(f) $s_b = .19265$
(g) $t = 3.87$ with 18 *df*; $t^2 = F$ obtained in (d).
(h) $Y'_1 = 3.30307 \ldots$; $Y'_{20} = 4.79251$
(i) $\Sigma Y' = 110.00 = \Sigma Y$; $\bar{Y}' = 5.50 = \bar{Y}$
(j) $Y_1 - Y'_1 = -1.30307 \ldots$; $Y_{20} - Y'_{20} = 5.20749$;
$\Sigma(Y - Y') = 0$; $\Sigma(Y - Y')^2 = 90.15561$

3. (a) $\Sigma x^2 = (N-1)s_x^2 = (199)(36) = 7164;$
$\Sigma y^2 = (N-1)s_y^2 = 16119;$
$\Sigma xy = (r_{xy}s_x s_y)(\mathrm{N}-1) = 7522.2;$
(b) $.49 = r_{xy}^2$
(c) $Y' = 37 + 1.05X$
(d) $ss_{reg} = 7898.31$
(e) $ss_{res} = 8220.69$
(f) $F = 190.24$ with 1 and 198 *df*

CHAPTER

3

Regression Diagnostics

Merits of most regression diagnostics can be especially appreciated in multiple regression analysis (i.e., analysis with more than one independent variable). Some diagnostics are applicable only in this case. Further, familiarity with matrix algebra and analysis by computer are essential for the application and understanding of most diagnostics. Nevertheless, a rudimentary introduction in the context of simple regression analysis should prove helpful because the calculations involved are relatively simple, requiring neither matrix operations nor computer analysis. After introducing computer programs and basic notions of matrix algebra (Chapters 4 and 5), I elaborate and expand on some topics I introduce here.[1] The present introduction is organized under two main headings: "Outliers" and "Influence Analysis."[2]

OUTLIERS

As the name implies, an outlier is a data point distinct or deviant from the rest of the data. Of factors that may give rise to outliers, diverse errors come readily to mind. Thus, an outlier may be a result of a recording or an input error, measurement errors, the malfunctioning of an instrument, or inappropriate instructions in the administration of a treatment, to name but some. Detecting errors and correcting them, or discarding subjects when errors in their scores are not correctable, are the recommended strategies in such instances.

Outliers may occur in the absence of errors. In essence, these are "true" outliers, as contrasted with "false" ones arising from errors of the kind I discussed in the preceding paragraph. It is outliers not due to discernable errors that are of interest for what they may reveal, among other things, about (1) the model being tested, (2) the possible violation of assumptions, and (3) observations that have undue influence on the results.[3] This is probably what Kruskal (1988) had in mind when he asserted that "investigation of the mechanism for outlying may be far more important than the original study that led to the outlier" (p. 929).

[1] An advanced review of topics presented in this chapter is given by Chatterjee and Hadi (1986a) and is followed by comments by some leading authorities. See also Hoaglin (1992) for a very good explication of diagnostics.

[2] I do not present here diagnostic approaches addressed to issues of collinearity (see Chapter 10), as they are only relevant for the case of multiple regression analysis.

[3] As I explain in the next section, an influential observation is a special case of an outlier.

Individuals with a unique attribute, or a unique combination of attributes, may react uniquely to a treatment making them stand out from the rest of the group. Discovery of such occurrences may lead to new insights into the phenomenon under study and to the designing of research to explore and extend such insights.

DETECTION OF OUTLIERS

Procedures for the detection of outliers rely almost exclusively on the detection of extreme residuals, so much so that the two are used interchangeably by some authors and researchers. Using the outlier concept in a broader sense of a deviant case, it is possible for it to be associated with a small residual, even one equal to zero. Such outliers may become evident when studying influence analysis—a topic I present in the next section. In what follows, I present three approaches to the detection of outliers based on residual analysis: (1) standardized residuals, (2) studentized residuals, and (3) studentized deleted residuals.

Standardized Residuals (ZRESID)

I introduced standardized residuals in Chapter 2—see (2.34) and the discussion related to it. Various authors have suggested that standardized residuals greater than 2 in absolute value (i.e., $z > |2.0|$) be scrutinized. Notice that large standardized residuals serve to alert the researcher to study them; *not* to automatically designate the points in question as outliers. As in most other matters, what counts is informed judgment. The same is true of studentized and studentized deleted residuals, which I discuss later in the chapter.

To illustrate the calculation of the various indices presented here, I will use data from the numerical example I introduced in Chapter 2 (Table 2.1). For convenience, I repeat the data from Table 2.1 in the first two columns of Table 3.1. Also repeated in the table, in the column labeled RESID, are residuals I took from Table 2.2.

As an example, I will calculate the standardized residual for the last subject in Table 3.1. This subject's residual is –2.80. For the data under consideration, $s_{y.x} = 2.446$ (see Chapter 2, for calculations). Dividing the residual by 2.446 yields a standardized residual of –1.1447.

Standardized residuals for the rest of the subjects, reported in Table 3.1 in the column labeled ZRESID, were similarly calculated. As you can see, none of the standardized residuals is greater than $|2.0|$. Had standardized residuals been used for detection of outliers, it would have been plausible to conclude that there are no outliers in the data under consideration.

Studentized Residuals (SRESID)

Calculation of standardized residuals is based on the generally untenable assumption that all residuals have the same variance. To avoid making this assumption, it is suggested that SRESIDs be used instead. This is accomplished by dividing each residual by its estimated standard deviation, which for simple regression analysis is

$$s_{e_i} = s_{y.x}\sqrt{1-\left[\frac{1}{N}+\frac{(X_i-\bar{X})^2}{\Sigma x^2}\right]} \qquad (3.1)$$

Y-Y'

Table 3.1 Residual Analysis for Data of Table 2.1

X	*Y*	*RESID*	*ZRESID*	*SRESID*	*SDRESID*
1	3	–2.80	–1.1447	–1.2416	–1.2618
1	5	–.80	–.3271	–.3547	–.3460
1	6	.20	.0818	.0887	.0862
1	9	3.20	1.3082	1.4190	1.4632
2	4	–2.55	–1.0425	–1.0839	–1.0895
2	6	–.55	–.2248	–.2338	–.2275
2	7	.45	.1840	.1913	.1861
2	10	3.45	1.4104	1.4665	1.5188
3	4	–3.30	–1.3491	–1.3841	–1.4230
3	6	–1.30	–.5315	–.5453	–.5343
3	8	.70	.2862	.2936	.2860
3	10	2.70	1.1038	1.1325	1.1420
4	5	–3.05	–1.2469	–1.2965	–1.3232
4	7	–1.05	–.4293	–.4463	–.4362
4	9	.95	.3884	.4038	.3942
4	12	3.95	1.6148	1.6790	1.7768
5	7	–1.80	–.7359	–.7982	–.7898
5	10	1.20	.4906	.5321	.5212
5	12	3.20	1.3082	1.4190	1.4633
5	6	–2.80	–1.1447	–1.2416	–1.2618

NOTE: *X* and *Y* were taken from Table 2.1.
RESID = residual (taken from Table 2.2)
ZRESID = standardized residual
SRESID = studentized residual
SDRESID = studentized deleted residual
See text for explanations.

the radical. Examine the latter and notice that the more X_i deviates from the mean of X, the smaller the standard error of the residual; hence the larger the studentized residual. As I show in the "Influence Analysis" section of this chapter, the term in the brackets (i.e., that subtracted from 1) is referred to as *leverage* and is symbolized as h_i.[4]

For illustrative purposes, I will apply (3.1) to the last subject of Table 3.1. For the data of Table 3.1, $\bar{X} = 3.00$ and $\Sigma x^2 = 40$ (see Chapter 2 for calculations). Hence,

$$s_{e_i} = 2.446\sqrt{1-\left[\frac{1}{20}+\frac{(5-3.0)^2}{40}\right]} = 2.2551$$

Dividing the residual (–2.80) by its standard deviation (2.2551), SRESID for the last subject is –1.2416. Note that subjects having the same *X* have an identical standard error of residual. For example, the standard error of the residual for the last four subjects is 2.2551. Dividing these subjects' residuals by 2.2551 yields their SRESIDs. Studentized residuals for all the subjects in the example under consideration are reported in Table 3.1 under SRESID.

[4] The *h* stands for the so-called hat matrix, and *i* refers to the *i*th diagonal element of this matrix. If you are unfamiliar with matrix terminology, don't worry about it. I explain it in subsequent chapters (especially Chapter 6). I introduced

When the assumptions of the model are reasonably met, SRESIDs follow a t distribution with $N - k - 1$ *df*, where N = sample size, k = number of independent variables. For the present example, $df = 18$ (20 − 1 − 1). It should be noted that the t's are not independent. This, however, is not a serious drawback, as the usefulness of the t's lies not so much in their use for tests of significance of residuals but as indicators of relatively large residuals whose associated observations deserve scrutiny.

The SRESIDs I discussed thus far are referred to by some authors (e.g., Cook & Weisberg, 1982, pp. 18–20) as "internally studentized residuals," to distinguish them from "externally studentized residuals." The distinction stems from the fact that $s_{y..x}$ used in the calculation of internally studentized residuals is based on the data for *all* the subjects, whereas in the case of externally studentized residuals $s_{y..x}$ is calculated after excluding the individual whose studentized residual is being sought (see the next section).

Studentized Deleted Residuals (SDRESID)

The standard error of SDRESID is calculated in a manner similar to (3.1), except that the standard error of estimate is based on data from which the subject whose studentized deleted residual is being sought was excluded. The reasoning behind this approach is that to the extent that a given point constitutes an outlier, its retention in the analysis would lead to upward bias in the standard error of estimate ($s_{y..x}$), thereby running the risk of failing to identify it as an outlier. Accordingly, the standard error of a deleted residual is defined as

$$s_{e(i)} = s_{y..x(i)}\sqrt{1-\left[\frac{1}{N}+\frac{(X_i-\overline{X})^2}{\Sigma x^2}\right]} \tag{3.2}$$

where $s_{e(i)}$ = standard error of residual for individual i, who has been excluded from the analysis; and $s_{y..x(i)}$ = standard error of estimate based on data from which i was excluded. Dividing i's residual by this standard error yields a SDRESID, which, as I stated previously, is also called an externally studentized residual.

For illustrative purposes, I will calculate SDRESID for the last subject of Table 3.1. This requires that the subject in question be deleted and a regression analysis be done to obtain the standard error of estimate.[5] Without showing the calculations, the standard error of estimate based on the data from which the last subject was deleted (i.e., an analysis based on the first 19 subjects) is 2.407. Applying (3.2),

$$s_{e(i)} = 2.407\sqrt{1-\left[\frac{1}{20}+\frac{(5-3.0)^2}{40}\right]} = 2.2190$$

Dividing the last subject's residual (−2.80) by this standard error yields a SDRESID of −1.2618.

As you can see, application of (3.2) for all the subjects would entail 20 regression analyses, in each of which one subject is deleted. Fortunately, formulas obviating the need for such laboriously repetitious calculations are available.[6] Following are two alternative approaches to the calculation of SDRESID based on results of an analysis in which all the subjects were included.

$$SDRESID_{(i)} = e_i\sqrt{\frac{N-k-2}{ss_{res}(1-h_i)-e_i^2}} \tag{3.3}$$

[5] Later, I give formulas that obviate the need to do a regression analysis from which the subject in question was excluded.

where $\text{SDRESID}_{(i)}$ = studentized deleted residual for subject i; e_i = residual for subject i; N = sample size; k = number of independent variables; ss_{res} = residual sum of squares from the analysis in which *all* the subjects were included; and $h_i = 1/N + (X_i - \overline{X})^2/\Sigma x^2$—see (3.1) and Footnote 4.

Using (3.3), I will calculate SDRESID for the last subject in the present example (using the data from Table 3.1). Recall that $N = 20$, and $k = 1$. From earlier calculations, $e_{20} = -2.80$; the mean of $X = 3.0$; $ss_{res} = 107.70$. Hence,

$$SDRESID_{(20)} = -2.80\sqrt{\frac{20-1-2}{107.70\,(1-.15)-(-2.80)^2}} = -1.2618$$

which agrees with the value I obtained previously. Similarly, I calculated SDRESIDs for the rest of the subjects. I reported them in Table 3.1 under SDRESID.[7]

Having calculated studentized residuals—as I did earlier and reported under SRESID in Table 3.1—SDRESIDs can also be calculated as follows:

$$SDRESID_{(i)} = SRESID_i\sqrt{\frac{N-k-2}{N-k-1-SRESID_i^2}} \tag{3.4}$$

where all the terms were defined earlier.

Using (3.4), I will calculate SDRESID for the last subject of Table 3.1. From earlier calculations (see also Table 3.1), $\text{SRESID}_{20} = -1.2416$. Hence,

$$SDRESID_{(20)} = -1.2416\sqrt{\frac{20-1-2}{20-1-1-(-1.2416)^2}} = -1.2619$$

which is, within rounding, the same value I obtained earlier.

The SDRESID is distributed as a t distribution with $N - k - 2$ *df*. As with ZRESID and SRESID, it is generally used not for tests of significance but for identifying large residuals, alerting the user to examine the observations associated with them.

INFLUENCE ANALYSIS

Although it has been recognized for some time that certain observations have greater influence on regression estimates than others, it is only in recent years that various procedures were developed for identifying influential observations. In their seminal work on influence analysis, Belsley, Kuh, and Welsch (1980) defined an influential observation as

> one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates (coefficients, standard errors, t-values, etc.) than is the case for most of the other observations. (p. 11)

As I illustrate later in this chapter, an outlier (see the preceding section) is not necessarily an influential observation. Rather, "an influential case is a special kind of outlier" (Bollen & Jackman, 1985, p. 512). As with outliers, greater appreciation of the role played by influential observations can be gained in the context of multiple regression analysis. Nevertheless, I introduce this topic here for the same reasons I introduced outliers earlier, namely, in simple regression

[7] For the present data, SDRESIDs differ little from SRESIDs. In the next section, I give an example where the two differ considerably.

analysis the calculations are simple, requiring neither matrix operations nor computer analysis. Later in the text (especially Chapter 6), I show generalizations to multiple regression analysis of indices presented here.

LEVERAGE

As the name implies, an observation's undue influence may be likened to the action of a lever providing increased power to pull the regression line, say, in a certain direction. In simple regression analysis, leverage can be calculated as follows:

$$h_i = \frac{1}{N} + \frac{(X-\bar{X})^2}{\Sigma x^2} \tag{3.5}$$

As I pointed out earlier—see (3.1) and the discussion related to it—h_i refers to the ith diagonal element of the so-called hat matrix (see Chapter 6). Before applying (3.5) to the numerical example under consideration, I will list several of its properties.

1. Leverage is a function solely of scores on the independent variable(s). Thus, as I show in the next section, a case that may be influential by virtue of its status on the dependent variable will not be detected as such on the basis of its leverage.
2. Other things equal, the larger the deviation of X_i from the mean of X, the larger the leverage. Notice that leverage is at a minimum $(1/N)$ when X_i is equal to the mean of X.
3. The maximum value of leverage is 1.
4. The average leverage for a set of scores is equal to $(k + 1)/N$, where k is the number of independent variables.

In light of these properties of leverage, Hoaglin and Welsch (1978, p. 18) suggested that, as a rule of thumb, $h_i > 2(k + 1)/N$ be considered high (but see Velleman & Welsch, 1981, pp. 234–235, for a revision of this rule of thumb in light of N and the number of independent variables). Later in this chapter, I comment on rules of thumb in general and specifically for the detection of outliers and influential observations and will therefore say no more about this topic here.

For illustrative purposes, I will calculate h_{20} (leverage for the last subject of the data in Table 3.1). Recalling that $N = 20$, $X_{20} = 5$, $\bar{X} = 3$, $\Sigma x^2 = 40$,

$$h_{20} = \frac{1}{20} + \frac{(5-3)^2}{40} = .15$$

Leverage for subjects having the same X is, of course, identical. Leverages for the data of Table 3.1 are given in column (1) of Table 3.2, from which you will note that all are relatively small, none exceeding the criterion suggested earlier.

To give you a feel for an observation with high leverage, and how such an observation might affect regression estimates, assume for the last case of the data in Table 3.1 that $X = 15$ instead of 5. This may be a consequence of a recording error or it may truly be this person's score on the independent variable. Be that as it may, after the change, the mean of X is 3.5, and $\Sigma x^2 = 175.00$ (you may wish to do these calculations as an exercise). Applying now (3.5), leverage for the changed case is .81 (recall that maximum leverage is 1.0).

Table 3.2 Influence Analysis for Data of Table 3.1

(1) *h* *Leverage*	*(2)* *Cook's* *D*	*(3)* *a* *DFBETA*	*(4)* *b* *DFBETA*	*(5)* *a* *DFBETAS*	*(6)* *b* *DFBETAS*
.15	.13602	–.65882	.16471	–.52199	.43281
.15	.01110	–.18824	.04706	–.14311	.11866
.15	.00069	.04706	–.01176	.03566	–.02957
.15	.17766	.75294	–.18824	.60530	–.50189
.07	.04763	–.34459	.06892	–.27003	.17912
.07	.00222	–.07432	.01486	–.05640	.03741
.07	.00148	.06081	–.01216	.04612	–.03059
.07	.08719	.46622	–.09324	.37642	–.24969
.05	.05042	–.17368	.00000	–.13920	.00000
.05	.00782	–.06842	.00000	–.05227	.00000
.05	.00227	.03684	.00000	.02798	.00000
.05	.03375	.14211	.00000	.11171	.00000
.07	.06814	.08243	–.08243	.06559	–.21754
.07	.00808	.02838	–.02838	.02162	–.07171
.07	.00661	–.02568	.02568	–.01954	.06481
.07	.11429	–.10676	.10676	–.08807	.29210
.15	.05621	.21176	–.10588	.16335	–.27089
.15	.02498	–.14118	.07059	–.10781	.17878
.15	.17766	–.37647	.18824	–.30265	.50189
.15	.13602	.32941	–.16471	.26099	–.43281

NOTE: The data, originally presented in Table 2.1, were repeated in Table 3.1. I discuss Column (2) under Cook's *D* and Columns (3) through (6) under DFBETA. a = intercept.

Using the data in Table 3.1, change X for the last case to 15, and do a regression analysis. You will find that

$$Y' = 6.96 + .10X$$

In Chapter 2—see the calculations following (2.9)—the regression equation for the original data was shown to be

$$Y' = 5.05 + .75X$$

Notice the considerable influence the change in one of the X's has on both the intercept and the regression coefficient (incidentally, r^2 for these data is .013, as compared with .173 for the original data). Assuming one could rule out errors (e.g., of recording, measurement, see the earlier discussion of this point), one would have to come to grips with this finding. Issues concerning conclusions that might be reached, and actions that might be taken, are complex. At this stage, I will give only a couple of examples.

Recall that I introduced the numerical example under consideration in Chapter 2 in the context of an experiment. Assume that the researcher had intentionally exposed the last subject to $X = 15$ (though it is unlikely that only one subject would be used). A possible explanation for the undue influence of this case might be that the regression of Y on X is curvilinear rather than linear. That is, the last case seems to change a linear trend to a curvilinear one (*but see the caveats that follow;* note also that I present curvilinear regression analysis in Chapter 13).

Assume now that the data of Table 3.1 were collected in a nonexperimental study and that errors of recording, measurement, and the like were ruled out as an explanation for the last person's X score being so deviant (i.e., 15). One would scrutinize attributes of this person in an attempt to discern what it is that makes him or her different from the rest of the subjects. As an admittedly unrealistic example, suppose that it turns out that the last subject is male, whereas the rest are females. This would raise the possibility that the status of males on X is considerably higher than that of females. Further, that the regression of Y on X among females differs from that among males (I present comparison of regression equations for different groups in Chapter 14).

Caveats. *Do not place too much faith in speculations such as the preceding.* Needless to say, one case does not a trend make. At best, influential observations should serve as clues. Whatever the circumstances of the study, and whatever the researcher's speculations about the findings, two things should be borne in mind.

1. Before accepting the findings, it is necessary to ascertain that they are replicable in newly designed studies. Referring to the first illustration given above, this would entail, among other things, exposure of more than one person to the condition of $X = 15$. Moreover, it would be worthwhile to also use intermediate values of X (i.e., between 5 and 15) so as to be in a position to ascertain not only whether the regression is curvilinear, but also the nature of the trend (e.g., quadratic or cubic; see Chapter 13). Similarly, the second illustration would entail, among other things, the use of more than one male.
2. Theoretical considerations should play the paramount role in attempts to explain the findings.

Although, as I stated previously, leverage is a property of the scores on the independent variable, the extent and nature of the influence a score with high leverage has on regression estimates depend also on the Y score with which it is linked. To illustrate this point, I will introduce a different change in the data under consideration. Instead of changing the last X to 15 (as I did previously), I will change the one before the last (i.e., the 19th subject) to 15.

Leverage for this score is, of course, the same as that I obtained above when I changed the last X to 15 (i.e., .81). However, the regression equation for these data differs from that I obtained when I changed the last X to 15. When I changed the last X to 15, the regression equation was

$$Y' = 6.96 + .10X$$

Changing the X for the 19th subject to 15 results in the following regression equation:

$$Y' = 5.76 + .44X$$

Thus, the impact of scores with the same leverage may differ, depending on the dependent-variable score with which they are paired. You may find it helpful to see why this is so by plotting the two data sets and drawing the regression line for each. Also, if you did the regression calculations, you would find that $r^2 = .260$ when the score for the 19th subject is changed to 15, as contrasted with $r^2 = .013$ when the score for the 20th subject is changed to 15. Finally, the residual and its associated transformations (e.g., standardized) are smaller for the second than for the first change:

	X	*Y*	*Y′*	*Y − Y′*	*ZRESID*	*SRESID*	*SDRESID*
20th subject	15	6	8.4171	−2.4171	−.9045	−2.0520	−2.2785
19th subject	15	12	12.3600	−.3600	−.1556	−.3531	−.3443

Based on residual analysis, the 20th case might be deemed an outlier, whereas the 19th would not be deemed thus.

COOK'S D

Earlier, I pointed out that leverage cannot detect an influential observation whose influence is due to its status on the dependent variable. By contrast, Cook's (1977, 1979) D (distance) measure is designed to identify an influential observation whose influence is due to its status on the independent variable(s), the dependent variable, or both.

$$D_i = \left[\frac{SRESID_i^2}{k+1}\right]\left[\frac{h_i}{1-h_i}\right] \tag{3.6}$$

where SRESID = studentized residual (see the "Outliers" section presented earlier in this chapter); h_i = leverage (see the preceding); and k = number of independent variables. Examine (3.6) and notice that D will be large when SRESID is large, leverage is large, or both.

For illustrative purposes, I will calculate D for the last case of Table 3.1. $\text{SRESID}_{20} = -1.2416$ (see Table 3.1); $h_{20} = .15$ (see Table 3.2); and $k = 1$. Hence,

$$D_{20} = \left[\frac{-1.2416^2}{1+1}\right]\left[\frac{.15}{1-.15}\right] = .1360$$

D's for the rest of the data of Table 3.1 are given in column (2) of Table 3.2.

Approximate tests of significance for Cook's D are given in Cook (1977, 1979) and Weisberg (1980, pp. 108–109). For diagnostic purposes, however, it would suffice to look for relatively large D values, that is, one would look for relatively large gaps between D for a given observation and D's for the rest of the data. Based on our knowledge about the residuals and leverage for the data of Table 3.1, it is not surprising that all the D's are relatively small, indicating the absence of influential observations.

It will be instructive to illustrate a situation in which leverage is relatively small, implying that the observation is not influential, whereas Cook's D is relatively large, implying that the converse is true. To this end, change the last observation so that $Y = 26$. As X is *unchanged* (i.e., 5), the leverage for the last case is .15, as I obtained earlier. Calculate the regression equation, SRESID, and Cook's D for the last case. Following are some of the results you will obtain:

$$Y' = 3.05 + 1.75X$$

$$SRESID_{20} = 3.5665; \; h_{20} = .15; \; k = 1$$

Notice the changes in the parameter estimates resulting from the change in the Y score for the 20th subject.[8] Applying (3.6),

$$D_{20} = \left[\frac{3.5665^2}{1+1}\right]\left[\frac{.15}{1-.15}\right] = 1.122$$

If you were to calculate D's for the rest of the data, you would find that they range from .000 to .128. Clearly, there is a considerable gap between D_{20} and the rest of the D's. To reiterate, sole reliance on leverage would lead to the conclusion that the 20th observation is not influential, whereas the converse conclusion would be reached based on the D.

[8] Earlier, I pointed out that SRESID (studentized residual) and SDRESID (studentized deleted residual) may differ con-

I would like to make two points about my presentation of influence analysis thus far.

1. My presentation proceeded backward, so to speak. That is, I examined consequences of a change in an X or Y score on regression estimates. Consistent with the definition of an influential observation (see the preceding), a more meaningful approach would be to study changes in parameter estimates that would occur because of deleting a given observation.
2. Leverage and Cook's D are global indices, signifying that an observation may be influential, but not revealing the effects it may have on specific parameter estimates.

I now turn to an approach aimed at identifying effects on specific parameter estimates that would result from the deletion of a given observation.

DFBETA

DFBETA$_{j(i)}$ indicates the change in j (intercept or regression coefficient) as a consequence of deleting subject i.[9] As my concern here is with simple regression analysis—consisting of two parameter estimates—it will be convenient to use the following notation: DFBETA$_{a(i)}$ will refer to the change in the intercept (a) when subject i is deleted, whereas DFBETA$_{b(i)}$ will refer to the change in the regression coefficient (b) when subject i is deleted.

To calculate DFBETA for a given observation, then, delete it, recalculate the regression equation, and note changes in parameter estimates that have occurred. For illustrative purposes, delete the last observation in the data of Table 3.1 and calculate the regression equation. You will find it to be

$$Y' = 4.72 + .91X$$

Recall that the regression equation based on all the data is

$$Y' = 5.05 + .75X$$

Hence, DFBETA$_{a(20)}$ = .33 (5.05 − 4.72), and DFBETA$_{b(20)}$ = −.16 (.75 − .91). Later, I address the issue of what is to be considered a large DFBETA, hence identifying an influential observation.

The preceding approach to the calculation of DFBETAs is extremely laborious, requiring the calculation of as many regression analyses as there are subjects (20 for the example under consideration). Fortunately, an alternative approach based on results obtained from a single regression analysis in which all the data are used is available. The formula for DFBETA for a is

$$DFBETA_{a(i)} = a - a(i) = \left[\left(\frac{\Sigma X^2}{N\Sigma X^2 - (\Sigma X)^2}\right) + \left(\frac{-\Sigma X}{N\Sigma X^2 - (\Sigma X)^2}\right)X_i\right]\frac{e_i}{1 - h_i} \qquad (3.7)$$

where N = number of cases; ΣX^2 = sum of squared raw scores; ΣX = sum of raw scores; $(\Sigma X)^2$ = square of the sum of raw scores; e_i = residual for subject i; and h_i = leverage for subject i. Earlier, I calculated all the preceding terms. The relevant sum and sum of squares (see Table 2.1 and the presentation related to it) are

$$\Sigma X = 60 \qquad \Sigma X^2 = 220$$

N = 20. Residuals are given in Table 3.1, and leverages in Table 3.2.

For illustrative purposes, I will apply (3.7) to the last (20th) case, to determine the change in a that would result from its deletion.

$$DFBETA_{a(20)} = a - a(20) = \left[\left(\frac{220}{(20)(220) - (60)^2}\right) + \left(\frac{-60}{(20)(220) - (60)^2}\right)5\right]\frac{-2.8}{1 - .15} = .32941$$

which agrees with the result I obtained earlier.

The formula for DFBETA for b is

$$DFBETA_{b(i)} = b - b(i) = \left[\left(\frac{-\Sigma X}{N\Sigma X^2 - (\Sigma X)^2}\right) + \left(\frac{N}{N\Sigma X^2 - (\Sigma X)^2}\right)X_i\right]\frac{e_i}{1 - h_i} \qquad (3.8)$$

where the terms are as defined under (3.7). Using the results given in connection with the application of (3.7),

$$DFBETA_{b(20)} = b - b(20) = \left[\left(\frac{-60}{(20)(220) - (60)^2}\right) + \left(\frac{20}{(20)(220) - (60)^2}\right)5\right]\frac{-2.8}{1 - .15} = -.16471$$

which agrees with the value I obtained earlier.

To repeat, DFBETAs indicate the change in the intercept and the regression coefficient(s) resulting from the deletion of a given subject. Clearly, having calculated DFBETAs, calculation of the regression equation that would be obtained as a result of the deletion of a given subject is straightforward. Using, as an example, the DFBETAs I calculated for the last subject (.33 and −.16 for a and b, respectively), and recalling that the regression equation based on all the data is $Y' = 5.05 + .75X$,

$$a = 5.05 - .33 = 4.72$$

$$b = .75 - (-.16) = .91$$

Above, I obtained the same values when I did a regression analysis based on all subjects but the last one.

Using (3.7) and (3.8), I calculated DFBETAs for all the subjects. They are given in columns (3) and (4) of Table 3.2.

Standardized DFBETA

What constitutes a large DFBETA? There is no easy answer to this question, as it hinges on the interpretation of regression coefficients—a topic that will occupy us in several subsequent chapters. For now, I will only point out that the size of the regression coefficient (hence a change in it) is affected by the scale of measurement used. For example, using feet instead of inches to measure X will yield a regression coefficient 12 times larger than one obtained for inches, though the nature of the regression of Y on X will, of course, not change.[10]

In light of the preceding, it was suggested that DFBETA be standardized, which for a is accomplished as follows:

[9] DF is supposed to stand for the difference between the estimated statistic with and without a given case. I said "supposed," as initially the prefix for another statistic suggested by the originators of this approach (Belsley et al., 1980) was DI, as in DIFFITS, which was then changed to DFFITS and later to DFITS (see Welsch, 1986, p. 403). Chatterjee and Hadi (1986b) complained about the "computer-speak (à la Orwell)," saying, "We aesthetically rebel against DFFIT, DFBETA, etc., and have attempted to replace them by the last name of the authors according to a venerable statistical tradition" (p. 416). Their hope that "this approach proves attractive to the statistical community" (p. 416) has not mate-

[10] It is for this reason that some researchers prefer to interpret standardized regression coefficients or beta weights—a topic I discuss in detail in Chapters 4 and 10.

$$DFBETAS_{a(i)} = \frac{DFBETA_{a(i)}}{\sqrt{MSR_{(i)}\left[\frac{\Sigma X^2}{N\Sigma X^2 - (\Sigma X)^2}\right]}} \qquad (3.9)$$

where DFBETAS = standardized DFBETA;[11] and $MSR_{(i)}$ = mean square residual when subject *i* is deleted. The rest of the terms were defined earlier.

If, as I suggested earlier, you did a regression analysis in which the last subject was deleted, you would find that $MSR_{(20)} = 5.79273$. Hence,

$$DFBETAS_{a(20)} = \frac{.32941}{\sqrt{5.79273\left[\frac{220}{(20)(220) - (60)^2}\right]}} = .26099$$

The formula for standardizing DFBETA for *b* is

$$DFBETAS_{b(i)} = \frac{DFBETA_{b(i)}}{\sqrt{MSR_{(i)}\left[\frac{N}{N\Sigma X^2 - (\Sigma X)^2}\right]}} \qquad (3.10)$$

Applying (3.10) to the 20th case,

$$DFBETAS_{b(20)} = \frac{-.16471}{\sqrt{5.79273\left[\frac{20}{(20)(220) - (60)^2}\right]}} = -.43282$$

Notice that $MSR_{(i)}$ in the denominator of (3.9) and (3.10) is based on an analysis in which a given subject is deleted. Hence, as many regression analyses as there are subjects would be required to calculate DFBETAS for all of them. To avoid this, $MSR_{(i)}$ can be calculated as follows:

$$MSR_{(i)} = \frac{ss_{res} - \frac{(e_i)^2}{1 - h_i}}{N - k - 1 - 1} \qquad (3.11)$$

For comparative purposes, I will apply (3.11) to the 20th subject. $e_{(20)} = -2.8$ (see Table 3.1); $h_{(20)} = .15$ (see Table 3.2); $ss_{res} = 107.70$ (see earlier calculations). $N = 20$ and $k = 1$. Therefore,

$$MSR_{(20)} = \frac{107.70 - \frac{(-2.8)^2}{1 - .15}}{20 - 1 - 1 - 1} = 5.79273$$

which agrees with the value I obtained earlier. Using (3.7) through (3.11), I calculated DFBETAS for all the subjects in the example under consideration (i.e., Table 3.1) and reported the results in columns (5) and (6) of Table 3.2.

In line with the recommendation that DFBETAS (standardized) be used instead of DFBETA (nonstandardized) for interpretive purposes (see preceding), criteria for what is to be considered a "large" DFBETAS have been proposed. Not surprisingly, there is no consensus on this point. Following are some examples of cutoffs that have been proposed.

Belsley et al. (1980) suggested, "as a first approximation," an "*absolute cutoff*" of 2 (p. 28). They went on to suggest that, because DFBETAS is affected by sample size, $2/\sqrt{n}$ serve as a "*size-adjusted cutoff*" (p. 28), when small samples are used. Neter, Wasserman, and Kutner (1989, p. 403), on the other hand, recommended that $2/\sqrt{n}$ serve as a cutoff for "large data sets," whereas 1 serve as a cutoff for "small to medium-size data sets." Finally, Mason, Gunst, and Hess (1989, p. 520) proposed $3/\sqrt{n}$ as a general cutoff.

Recalling that $N = 20$ for the example under consideration, following Belsley et al., the size-adjusted cutoff is .45, whereas following Mason et al. the cutoff is .67. Examine columns (5) and (6) of Table 3.2 and notice that a few of the DFBETASs are slightly larger than the size-adjusted cutoff proposed by Belsley et al. and that none meet the criteria proposed by Neter et al. or Mason et al. In sum, it is safe to assume that most researchers would conclude that none of the DFBETASs in the numerical example under consideration are "large."

Before I comment generally on criteria and rules of thumb, I will use an additional example to illustrate: (1) the value of DFBETA in pinpointing changes occurring as a result of the deletion of a subject and (2) that an outlier does *not* necessarily signify that the observation in question is influential. To this end, let us introduce yet another change in the data of Table 3.1. This time, change the *Y* for the first subject in the group whose $X = 3$ (i.e., the ninth subject) to 14 (instead of 4). Calculate the regression equation. In addition, for this subject, calculate (1) ZRESID, SRESID, and SDRESID; (2) leverage and Cook's *D;* (3) DFBETA (nonstandardized) and DFBETAS (standardized). Following are results you will obtain:

$$Y' = 5.55 + .75X$$

For the ninth subject,

(1)	*ZRESID*	*SRESID*	*SDRESID*
	2.2498	2.3082	2.6735
(2)	*Leverage*	*Cook's D*	
	.050	.140	
(3)	*DFBETA*	*DFBETAS*	
	a: .32632 *b*: .00000	.26153 .00000	

Beginning with the residual, note that the observation under consideration would probably be identified as an outlier, especially when it is compared with those for the rest of the data. For example, the next largest SDRESID is −1.3718.

Turning to leverage, it is clear that it is small. The same is true of *D*. If you were to calculate the *D*'s for the rest of the data, you would find that they range from .000 to .149. Clearly, the *D* for the ninth subject is not out of line from the rest of the *D*'s, leading to the conclusion that the ninth observation is not influential. Here, then, is an example where an observation that might be identified as an outlier would not be deemed as influential.

Examine now the DFBETA and DFBETAS and note that the deletion of the ninth subject will result in an intercept change from 5.55 to 5.22 (i.e., 5.55 − .32632). The regression coefficient will, however, *not* change as a result of the deletion of the ninth subject. Thus, the regression equation based on the data from which the ninth subject was deleted would be[12]

$$Y' = 5.22 + .75X$$

[11] [illegible] DFBETAS, although something like STDFBETA would be less

[12] [illegible]

It will be instructive to concentrate first on the interpretation of a change in *a*. Recall that *a* indicates the point at which the regression line intercepts the *Y* ordinate when $X = 0$. Stated differently, it is the predicted *Y* when $X = 0$. In many areas of behavioral sciences $X = 0$ is of little or no substantive meaning. Suffice it to think of *X* as a measure of mental ability, achievement, depression, and the like, to see why this is so. Therefore, even if the change in *a* was much larger than the one obtained earlier, and even if it was deemed to be large based on some criterion, it is conceivable that it would be judged not meaningful. This is not to say that one would ignore the extreme residual that would be associated with the observation in question. But this matter need not concern us here, as I addressed it earlier.

What is, however, most revealing in the present example—indeed my reason for presenting it—is the absence of change in the regression coefficient (*b*) as a result of deleting the ninth subject.[13] Thus, even if based on other indices (e.g., *D*), one was inclined to consider the ninth observation as influential, it is conceivable that focusing on the change in *b*, one would deem it not influential.

CRITERIA AND RULES OF THUMB

Dependence on criteria and rules of thumb in the conduct of behavioral research is so prevalent that it requires no documentation. The ubiquity of such practices is exemplified by conventions followed in connection with statistical tests of significance (e.g., Type I and Type II errors, effect size).[14]

Authors who propose criteria and rules of thumb do so, in my opinion, with the best of intentions to assist their readers to develop a "feel" for the indices in question. Notably, most stress the need for caution in resorting to criteria they propose and attempt to impress upon the reader that they are not meant to serve as substitutes for informed judgment. For instance, preceding their proposed criteria for influential observations, Belsley et al. (1980) cautioned:

> As with all empirical procedures, this question is ultimately answered by judgment and intuition in choosing reasonable cutoffs most suitable for the problem at hand, guided whenever possible by statistical theory. (p. 27)

Unfortunately, many researchers not only ignore the cautions, but also misinterpret, even misrepresent recommended guidelines.[15] Drawing attention to difficulties in interpreting outliers, Johnson (1985) bemoans the practice of treating methods for detecting them as a "technological fix," prompting "many investigators . . . to believe that statistical procedures will sort a data set into the 'good guys' and the 'bad guys'" (p. 958).

Perusal of published research reveals that many authors flaunt criteria with an air of finality and certainty. The allure of a criterion adorned by references to authorities in the field is apparently so potent as to dazzle even referees and editors of professional journals. Deleterious consequences of this practice cannot be overestimated. The most pernicious effect of this practice is that it seems to absolve the researcher of the responsibility of making an informed interpretation and decision—actions unimaginable without thorough knowledge of the research area, an understanding of statistical and design principles, and, above all, hard thinking.

The paramount role of knowledge and judgment in deciding what is an influential observation, say, may be discerned from the last example I gave earlier. Recall that it concerned a situation in which the deletion of an observation resulted in a change in *a* (intercept), but not in *b* (regression coefficient). Clearly, a researcher whose aim is to interpret *b* only would not deem an observation influential, regardless of the effect its deletion would have on *a*.

In sum, beware of being beguiled by criteria and rules of thumb. It is only in light of various aspects of the study (e.g., cost, duration, consequences, generalizability), as well as theoretical and analytic considerations, that you can hope to arrive at meaningful statements about its findings.

A Numerical Example

Before considering remedies, I present another numerical example designed to illustrate the potential hazards of neglecting to examine one's data and of failing to apply regression diagnostics. The example is reported in Part (*a*) of Table 3.3. Included in the table are summary statistics and results of tests of statistical significance.[16] As I used a similar format in Chapter 2 (see Table 2.3), I will not explain the terms.

Table 3.3 Two Data Sets

	(*a*)		(*b*)	
	X	*Y*	*X*	*Y*
	2	2	2	2
	3	3	3	3
	3	1	3	1
	4	1	4	1
	4	3	4	3
	5	2	5	2
	8	8		
N:	7		6	
M:	4.14	2.86	3.50	2.00
s:	1.95	2.41	1.05	.89
r_2:	.67		.00	
a:	−1.34		2.00	
b:	1.01		.00	
ss_{reg}:	23.43		.00	
ss_{res}:	11.42		4.00	
F:	10.25 (1,5)		.00 (1,4)	
t:	3.20 (5)		.00 (4)	
p:	.02		1.00	

[13] I suggest that you experiment by introducing other changes in *Y* for the same subject (e.g., make it 24, 30, or 40), and reanalyze the data. For the suggested changes, you will find the DFBETAS$_a$ becomes increasingly larger (.6623, .9027, and 1.3034, respectively), but the *b* is unchanged. Incidentally, the same will hold true if you changed any of the *Y*'s whose *X* scores are equal to the mean of *X*. The main point is that when *a* is not substantively meaningful, neither is a change in it, whatever its size.

[14] For examples relating to measurement models, see Bollen and Lennox (1991); for examples relating to adoption of "standards" of reliability of measures, see Pedhazur and Schmelkin (1991, pp. 109–110).

Examine Part (*a*) of Table 3.3 and note that, assuming $\alpha = .05$ was selected, the regression of Y on X is statistically significant. In the absence of diagnostics, one would be inclined to conclude, among other things, that (1) about 67% of the variance in Y is accounted for by X and (2) the expected change in Y associated with a unit change in X is 1.01.

In what follows, I will scrutinize the role of the last subject in these results. The residual and some of its transformations for this subject are as follows:

RESID	*ZRESID*	*SRESID*	*SDRESID*
1.2375	.8178	1.8026	2.7249

Inspection of ZRESID and SRESID would lead to the conclusion that there is nothing distinctive about this subject, although SDRESID might raise doubt about such a conclusion.

Here now are diagnostic indices associated with the last subject:

H_7	D_7	$DFBETA_{a(7)}$	$DFBETA_{b(7)}$	$DFBETAS_{a(7)}$	$DFBETAS_{b(7)}$
.79	6.25	−3.3375	1.0125	−3.5303	4.8407

Clearly, this is an influential observation. To appreciate how influential it is, I will use DFBETAs (unstandardized) to calculate the regression equation based on the first six subjects (i.e., deleting the seventh subject).

$$a = -1.34 - (-3.34) = 2.00$$

$$b = 1.01 - 1.01 = 0$$

These statistics are reported also in Part (*b*) of Table 3.3, which consists of results of a regression analysis based on the first six subjects of Part (*a*).

The most important thing to note is that in the absence of the seventh subject, the regression of Y on X is zero ($b = 0$). At the risk of being redundant, it is noteworthy that the statistically significant and, what appeared to be, the strong regression of Y on X was due to the inclusion of a single subject.

Note that, consistent with (2.10) and the discussion related to it, when $b = 0$, the intercept (a) is equal to the mean of the dependent variable.

REMEDIES

Awareness of the existence of a problem is, needless to say, a prerequisite for attempts to do something about it. More than a decade ago, Belsley et al. (1980) observed that "[i]t is increasingly the case that the data employed in regression analysis, and on which the results are conditioned, are given only the most cursory examination for their suitability" (p. 2). Remarkable increases in availability of computers and reliance on technicians (euphemistically referred to as "consultants") to analyze one's data have greatly exacerbated this predicament.

The larger the project, the greater the likelihood for data analysis "chores" to be relegated to assistants, and the lesser the likelihood for principal investigators to examine their data. Consequently, many a researcher is unaware that "dramatic" or "puzzling" findings may be due to one or more influential observations, or that a relation they treat as linear is curvilinear, to give but two

Suggested Remedies

Difficulties in selecting from among indices of influential observations, and of designating observations as influential, pale in comparison to those arising concerning action to be taken when influential observations are detected. Earlier, I pointed out that when it is determined that an observation in question is due to error, the action that needs to be taken is relatively uncomplicated. It is when errors are ruled out that complications abound, as the decision regarding action to be taken is predicated on a host of theoretical and analytic considerations (e.g., model, subjects, settings). What follows is not an exhaustive presentation of remedies but a broad sketch of some, along with relevant references.

Probably the first thing that comes to mind is to delete the influential observation(s) and reanalyze the data. Nevertheless, in light of norms against "fudging" data and "dishonesty" in data analysis, the tendency to refrain from doing this is strong. I concur strongly with Judd and McClelland's (1989) cogent argument that when an influential observation(s) affects the results, it is "misleading . . . to pretend" that this is not so.

> Somehow, however, in the social sciences the reporting of results with outliers included has come to be viewed as the "honest" thing to do and the reporting of results with outliers removed is sometimes unfortunately viewed as "cheating." Although there is no doubt that techniques for outlier identification and removal can be abused, we think it far more honest to omit outliers from the analysis with the explicit admission in the report that there are some observations which we do not understand and to report a good model for those observations which we do understand. If that is not acceptable, then separate analyses, with and without the outliers included, ought to be reported so that the reader can make his or her own decision about the adequacy of the models. *To ignore outliers by failing to detect and report them is dishonest and misleading.* (pp. 231–232; see also, Fox, 1991, p. 76)

I believe that, in addition to reporting results of analyses with and without influential observations, sufficient information ought to be given (or made available on request) so that readers who so desire may reanalyze the data.

Deletion of influential observations is by no means the only suggested course of action. Among others, a transformation of one or more variables may reduce the impact of influential observations (for discussions of transformations and their role in data analysis see, among others, Atkinson, 1985, Chapters 6–9; Fox, 1984, Chapter 3; Judd & McClelland, 1989, Chapter 16; Stoto & Emerson, 1983).

Another approach is to subject the data to a robust regression method (for a review of-four such methods, see Huynh, 1982; see also, Neter et al., 1989, pp. 405–407; Rousseeuw & Leroy, 1987).

CONCLUDING REMARKS

I hope that this chapter served to alert you to the importance of scrutinizing data and using regression diagnostics. In subsequent chapters, I extend and elaborate on concepts I introduced in this chapter.

In Chapter 4, which is devoted to computers and computer programs, I will use several computer programs to reanalyze some of the numerical examples I presented in Chapter 2 and in the present chapter.

CONCLUDING REMARKS

Except for specialized programs (e.g., EQS, LISREL), which I use in specific chapters, I will use the packages I introduced in this chapter throughout the book. When using any of these packages, I will follow the format and conventions I presented in this chapter (e.g., commentaries on input and output). However, my commentaries on input and output will address primarily the topics under consideration. Consequently, if you have difficulties in running the examples given in subsequent chapters, or if you are puzzled by some aspects of the input, output, commentaries, and the like, you may find it useful to return to this chapter. To reiterate: study the manual(s) of the package(s) you are using, and refer to it when in doubt or at a loss.

In most instances, a single independent variable is probably not sufficient for a satisfactory, not to mention thorough, explanation of the complex phenomena that are the subject matter of behavioral and social sciences. As a rule, a dependent variable is affected by multiple independent variables. It is to the study of simultaneous effects of independent variables on a dependent variable that I now turn. In Chapter 5, I discuss analysis and interpretation with two independent variables, whereas in Chapter 6 I present a generalization to any number of independent variables.

CHAPTER

5

Elements of Multiple Regression Analysis: Two Independent Variables

In this chapter, I extend regression theory and analysis to the case of two independent variables. Although the concepts I introduce apply equally to multiple regression analysis with any number of independent variables, the decided advantage of limiting this introduction to two independent variables is in the relative simplicity of the calculations entailed. Not having to engage in, or follow, complex calculations will, I hope, enable you to concentrate on the meaning of the concepts I present. Generalization to more than two independent variables is straightforward, although it involves complex calculations that are best handled through matrix algebra (see Chapter 6).

After introducing basic ideas of multiple regression, I present and analyze in detail a numerical example with two independent variables. As in Chapter 2, I carry out all the calculations by hand so that you may better grasp the meaning of the terms presented. Among topics I discuss in the context of the analysis are squared multiple correlation, regression coefficients, statistical tests of significance, and the relative importance of variables. I conclude the chapter with computer analyses of the numerical example I analyzed by hand, in the context of which I extend ideas of regression diagnostics to the case of multiple regression analysis.

BASIC IDEAS

In Chapter 2, I gave the sample linear regression equation for a design with one independent variable as (2.6). I repeat this equation with a new number. (For your convenience, I periodically resort to this practice of repeating equations with new numbers attached to them.)

$$Y = a + bX + e \tag{5.1}$$

where Y = raw score on the dependent variable; a = intercept; b = regression coefficient; X = raw score on the independent variable; and e = error, or residual.

Equation (5.1) can be extended to any number of independent variables or X's:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e \tag{5.2}$$

where $b_1, b_2, \ldots, b_k$ are regression coefficients associated with the independent variables $X_1, X_2, \ldots, X_k$ and e is the error, or residual. As in simple linear regression (see Chapter 2), a solution is

(Σe^2) is minimized. This, it will be recalled, is referred to as the *principle of least squares*, according to which the independent variables are differentially weighted so that the sum of the squared errors of prediction is minimized or that prediction is optimized.

The prediction equation in multiple regression analysis is

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k \tag{5.3}$$

where Y' = predicted Y score. All other terms are as defined under (5.2). One of the main calculation problems of multiple regression is to solve for the b's in (5.3). With only two independent variables, the problem is not difficult, as I show later in this chapter. With more than two X's, however, it is considerably more difficult, and reliance on matrix operations becomes essential. To reiterate: the principles and interpretations I present in connection with two independent variables apply equally to designs with any number of independent variables.

In Chapter 2, I presented and analyzed data from an experiment with one independent variable. Among other things, I pointed out that r^2 (squared correlation between the independent and the dependent variable) indicates the proportion of variance accounted for by the independent variable. Also, of course, $1 - r^2$ is the proportion of variance not accounted for, or error. To minimize errors, or optimize explanation, more than one independent variable may be used. Assuming two independent variables, X_1 and X_2, are used, one would calculate $R^2_{y.x_1x_2}$ where R^2 = squared multiple correlation of Y (the dependent variable, which is placed before the dot) with X_1 and X_2 (the independent variables, which are placed after the dot). To avoid cumbersome subscript notation, I will identify the dependent variable as Y, and the independent variables by numbers only. Thus,

$$R^2_{y.x_1x_2} = R^2_{y.12} \qquad r^2_{y.x_1} = r^2_{y.1} \qquad r^2_{x_1x_2} = r^2_{12}$$

$R^2_{y.12}$ indicates the proportion of variance of Y accounted for by X_1 and X_2.

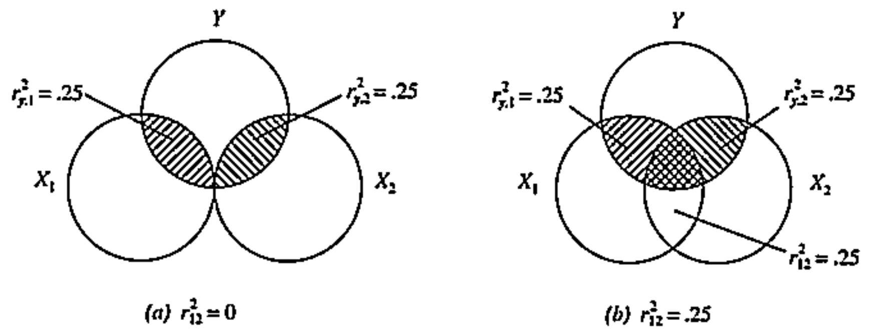
As I discussed in Chapter 2, regression analysis may be applied in different designs (e.g., experimental, quasi-experimental, and nonexperimental; see Pedhazur & Schmelkin, 1991, Chapters 12–14, for detailed discussions of such designs and references). In various subsequent chapters, I discuss application of regression analysis in specific designs. For present purposes, it will suffice to point out that an important property of a well-designed and well-executed experiment is that the independent variables are not correlated. For the case of two independent variables, this means that $r_{12} = .00$. Under such circumstances, calculation of R^2 is simple and straightforward:

$$R^2_{y.12} = r^2_{y1} + r^2_{y2} \qquad (\text{when } r_{12} = 0)$$

Each r^2 indicates the proportion of variance accounted for by a given independent variable.[1] Calculations of other regression statistics (e.g., the regression equation) are equally simple.

In quasi-experimental and nonexperimental designs, the independent variables are almost always correlated. For the case of two independent variables, or two predictors, this means that $r_{12} \neq .00$. The nonzero correlation indicates that the two independent variables, or predictors, provide a certain amount of redundant information, which has to be taken into account when calculating multiple regression statistics.

These ideas can perhaps be clarified by Figure 5.1, where each set of circles represents the variance of a Y variable and two X variables, X_1 and X_2. The set on the left, labeled (a), is a simple situation where $r_{y1} = .50$, $r_{y2} = .50$, and $r_{12} = 0$. Squaring the correlation of X_1 and X_2 with Y and adding them $[(.50)^2 + (.50)^2 = .50]$, the proportion of variance of Y accounted for by both X_1 and X_2 is obtained, or $R^2_{y.12} = .50$.



(a) $r^2_{12} = 0$

(b) $r^2_{12} = .25$

Figure 5.1

But now study the situation in (b). The sum of r^2_{y1} and r^2_{y2} is *not* equal to $R^2_{y.12}$ because r_{12} is *not* equal to 0. (The degree of correlation between two variables is expressed by the amount of overlap of the circles.[2]) The hatched areas of overlap represent the variances common to pairs of depicted variables. The one doubly hatched area represents that part of the variance of Y that is common to the X_1 and X_2 variables. Or, it is part of r^2_{y1}, it is part of r^2_{y2}, and it is part of r^2_{12}. Therefore, to calculate that part of Y that is determined by X_1 *and* X_2, it is necessary to subtract this doubly hatched overlapping part so that it will not be counted twice.

Careful study of Figure 5.1 and the relations it depicts should help you grasp the principle I stated earlier. Look at the right-hand side of the figure. To explain or predict more of Y, so to speak, it is necessary to find other variables whose variance circles will intersect the Y circle and, at the same time, not intersect each other, or at least minimally intersect each other.

A Numerical Example

I purposely use an example in which the two independent variables are correlated, as it is the more general case under which the special case of $r_{12} = 0$ is subsumed. It is the case of correlated independent variables that poses so many of the interpretational problems that will occupy us not only in this chapter but also in subsequent chapters.

Suppose we have the reading achievement, verbal aptitude, and achievement motivation scores on 20 eighth-grade pupils. (There will, of course, usually be many more than 20 subjects.) We want to calculate the regression of Y, reading achievement, on both verbal aptitude and achievement motivation. But since verbal aptitude and achievement motivation are correlated, it is necessary to take the correlation into account when studying the regression of reading achievement on both variables.

Calculation of Basic Statistics

Assume that scores for the 20 pupils are as given in Table 5.1. To do a regression analysis, a number of statistics have to be calculated. The sums, means, and the sums of squares of raw scores on the three sets of scores are given in the three lines directly below the table. In addition,

[2] Although the figure is useful for pedagogical purposes, it is not always possible to depict complex relations among vari-

Table 5.1 Illustrative Data: Reading Achievement (Y), Verbal Aptitude (X_1), and Achievement Motivation (X_2)

	Y	X_1	X_2	Y'	$Y-Y'=e$
	2	1	3	2.0097	−.0097
	4	2	5	3.8981	.1019
	4	1	3	2.0097	1.9903
	1	1	4	2.6016	−1.6016
	5	3	6	5.1947	−.1947
	4	4	5	5.3074	−1.3074
	7	5	6	6.6040	.3960
	9	5	7	7.1959	1.8041
	7	7	8	9.1971	−2.1971
	8	6	4	6.1248	1.8752
	5	4	3	4.1236	.8764
	2	3	4	4.0109	−2.0109
	8	6	6	7.3086	.6914
	6	6	7	7.9005	−1.9005
	10	8	7	9.3098	.6902
	9	9	6	9.4226	−.4226
	3	2	6	4.4900	−1.4900
	6	6	5	6.7167	−.7167
	7	4	6	5.8993	1.1007
	10	4	9	7.6750	2.3250
Σ:	117	87	110	117	0
M:	5.85	4.35	5.50		
SS:	825	481	658		$\Sigma e^2 = 38.9469$

NOTE: SS = sum of squared raw scores.

the following statistics will be needed: the deviation sums of squares for the three variables, their deviation cross products, and their standard deviations. They are calculated as follows:

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 825 - \frac{(117)^2}{20} = 825 - 684.45 = 140.55$$

$$\Sigma x_1^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{N} = 481 - \frac{(87)^2}{20} = 481 - 378.45 = 102.55$$

$$\Sigma x_2^2 = \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{N} = 658 - \frac{(110)^2}{20} = 658 - 605.00 = 53.00$$

$$\Sigma x_1 y = \Sigma X_1 Y - \frac{(\Sigma X_1)(\Sigma Y)}{N} = 604 - \frac{(87)(117)}{20} = 604 - 508.95 = 95.05$$

$$\Sigma x_2 y = \Sigma X_2 Y - \frac{(\Sigma X_2)(\Sigma Y)}{N} = 702 - \frac{(110)(117)}{20} = 702 - 643.50 = 58.50$$

$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - \frac{(\Sigma X_1)(\Sigma X_2)}{N} = 517 - \frac{(87)(110)}{20} = 517 - 478.50 = 38.50$$

$$s_y = \sqrt{\frac{\Sigma y^2}{N-1}} = \sqrt{\frac{140.55}{20-1}} = 2.720$$

$$s_{x_1} = \sqrt{\frac{\Sigma x_1^2}{N-1}} = \sqrt{\frac{102.55}{20-1}} = 2.323$$

$$s_{x_2} = \sqrt{\frac{\Sigma x_2^2}{N-1}} = \sqrt{\frac{53.00}{20-1}} = 1.670$$

For visual convenience, I pulled together in Table 5.2 the results of the previous calculations. Below the principal diagonal of the matrix, I included the correlations between the variables (.792, .678, and .522), as I will need them later.

Table 5.2 Deviation Sums of Squares and Cross Products, Correlation Coefficients, and Standard Deviations for the Data in Table 5.1

	y	x_1	x_2
y	140.55	95.05	58.50
x_1	*.792*	102.55	38.50
x_2	*.678*	*.522*	53.00
s	2.720	2.323	1.670

NOTE: The tabled values are as follows: the first line is comprised, successively, of Σy^2, the deviation sum of squares of Y; the cross product of the deviations of X_1 and Y, or $\Sigma x_1 y$; and finally $\Sigma x_2 y$. The entries in the second and third lines, on the diagonal or above, are Σx_1^2, $\Sigma x_1 x_2$, and (in the lower right corner) Σx_2^2. The italicized entries *below* the diagonal are the correlation coefficients. The standard deviations are given in the last line.

There is more than one way to calculate the essential statistics of multiple regression analysis. Ultimately, I will cover several approaches. Now, however, I concentrate on calculations that use sums of squares, because they have the virtue of being additive and intuitively comprehensible.

Reasons for the Calculations

Before proceeding with the calculations, it will be useful to review why we are doing all this. First, we want to calculate the constants (a, b_1, and b_2) of the regression equation $Y' = a + b_1X_1 + b_2X_2$, so that we can, if we wish, use the X's of individuals and predict their Y's (Y'). This means, in the present example, that if we have scores of individuals on verbal aptitude and achievement motivation, we can insert them into the equation and obtain Y' values, or predicted reading achievement scores.

Second, we want to know the proportion of variance "accounted for," that is $R^2_{y.12}$. In other words, we want to know how much of the total variance of Y, reading achievement, is due to its regression on the X's, on verbal aptitude and achievement motivation.

Third, we wish to test the results for statistical significance so that we could state, for instance, whether the regression of Y on the X's is statistically significant, or whether each regression coefficient, b, in the regression equation is statistically different from zero.

Finally, we wish to determine the relative importance of the different X's in explaining Y. We wish to know, in this case, the relative importance of X_1 and X_2, verbal aptitude and achievement motivation, in explaining verbal achievement. As you will see, this is the most difficult question to answer. In this chapter, I deal briefly with the complexities attendant with such questions and some alternative approaches to answering them. A deeper knowledge of multiple regression analysis is necessary for fully comprehending the diverse approaches and the complexities of each. In succeeding chapters, I broaden and deepen the scope of the different uses and interpretations of multiple regression analysis.

Calculation of Regression Statistics

The calculation of the b's of the regression equation is done rather mechanically with formulas for two X variables. They are

$$b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$
$$b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2} \tag{5.4}$$

Taking relevant values from Table 5.2 and substituting them in the formulas, I calculate the b's:

$$b_1 = \frac{(53.00)(95.05) - (38.50)(58.50)}{(102.55)(53.00) - (38.50)^2} = \frac{5037.65 - 2252.25}{5435.15 - 1482.25} = \frac{2785.40}{3952.90} = .7046$$

$$b_2 = \frac{(102.55)(58.50) - (38.50)(95.05)}{(102.55)(53.00) - (38.50)^2} = \frac{5999.175 - 3659.425}{5435.15 - 1482.25} = \frac{2339.75}{3952.90} = .5919$$

The formula for a is

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 \tag{5.5}$$

Substituting relevant values yields

$$a = 5.85 - (.7046)(4.35) - (.5919)(5.50) = -.4705$$

The regression equation can now be written with the calculated values of a and the b's

$$Y' = -.4705 + .7046X_1 + .5919X_2$$

As examples of the use of the equation in prediction, I calculate predicted Y's for the first and the last subjects of Table 5.1:

$$Y' = -.4705 + (.7046)(1) + (.5919)(3) = 2.0098$$

$$Y' = -.4705 + (.7046)(4) + (.5919)(9) = 7.6750$$

The *observed* Y's are $Y_1 = 2$ and $Y_{20} = 10$. The residuals, $e = Y - Y'$, are

$$e_1 = 2 - 2.0098 = -.0098$$

$$e_{20} = 10 - 7.6750 = 2.3250$$

The predicted Y's and the residuals are given in the last two columns of Table 5.1.[3] Recall that the a and the b's of the regression equation were calculated to satisfy the least-squares principle, that is, to minimize the squares of errors of prediction. Squaring each of the residuals and adding them, as I did in Chapter 2, $\Sigma e^2 = 38.9469$. (Note that $\Sigma e = 0$.) As I showed earlier, this can be symbolized Σy^2_{res} or ss_{res}. In short, the residual sum of squares expresses that portion of the total Y sum of squares, Σy^2, that is *not* due to regression. In the following, I show that the residual

sum of squares can be more readily calculated. I went through the lengthy calculations to show what this sum of squares consists of.

The regression sum of squares is calculated with the following general formula:

$$ss_{\text{reg}} = b_1\Sigma x_1 y + \ldots + b_k\Sigma x_k y \tag{5.6}$$

where k = the number of X, or independent variables. In the case of two X variables, $k = 2$, the formula reduces to

$$ss_{\text{reg}} = b_1\Sigma x_1 y + b_2\Sigma x_2 y \tag{5.7}$$

Taking the b values I calculated in the preceding and the deviation sums of cross products from Table 5.2, and substituting in (5.7),

$$ss_{\text{reg}} = (.7046)(95.05) + (.5919)(58.50) = 101.60$$

This is the portion of the total sum of squares of Y, or Σy^2, that is due to the regression of Y on the two X's. Note that the total sum of squares is 140.55 (from Table 5.2). In Chapter 2—see (2.17)—I showed that

$$ss_{\text{res}} = \Sigma y^2 - ss_{\text{reg}} \tag{5.8}$$

Therefore,

$$ss_{\text{res}} = 140.55 - 101.60 = 38.95$$

which is, within rounding, the same as the value I obtained through the lengthy calculations of Table 5.1.

Alternative Calculations

To reinforce and broaden your understanding of multiple regression analysis, I will calculate regression statistics by other methods, in which correlation coefficients are used. Before presenting the alternative calculations, it is necessary to digress briefly to discuss the distinction between standardized and unstandardized regression coefficients and how they are related to each other.

Regression Weights: *b* and β

Earlier, I used b as a symbol for the statistic and β as a symbol for the parameter. There is, however, another way in which these symbols are frequently used: b is the unstandardized regression coefficient, and β is the standardized regression coefficient (see the following). Unfortunately, there is no consistency in notation. For example, some authors use b^* as the symbol for the standardized regression coefficient, others use $\hat{\beta}$ as the symbol for the estimator of β (the unstandardized coefficient) and $\hat{\beta}^*$ as the symbol for the standardized coefficient. Although the use of the different symbols, as exemplified here, is meant to avoid confusion, I believe that they are unnecessarily cumbersome and may therefore lead to greater confusion. Adding to the potential confusion is another usage of β as a symbol for Type II error (see Hays, 1988, p. 261; Pedhazur & Schmelkin, 1991, p. 206). *Henceforth, I will use* b *as the symbol for the sample unstandardized regression coefficient and* β *as the symbol for the sample standardized coefficient.* Occasion-

[3] [illegible] given here and those given in Table 5.1 are due to rounding

When raw scores are used, as I did until now, b's are calculated and applied to the X's (raw scores) in the regression equation. If, however, the Y and X scores were standardized (i.e., converted to z scores), β's would be calculated and applied to z's in the regression equation. For simple regression, the regression equation in which standard scores are used is

$$z'_y = \beta z_x \tag{5.9}$$

where z'_y = predicted standard score of Y; β = standardized regression coefficient; and z_x = standard score of X. As in the case of b, β is interpreted as the expected change in Y associated with a unit change in X. But because the standard deviation of z scores is equal to 1.00, a unit change in X, when it has been standardized, refers to a change of one standard deviation in X. Later in this chapter, I discuss distinctions in the use and interpretation of b's and β's.

With one independent variable, the formula for the calculation of β is

$$\beta = \frac{\Sigma z_x z_y}{\Sigma z_x^2} \tag{5.10}$$

In Chapter 2, I showed that b is calculated as follows:

$$b = \frac{\Sigma xy}{\Sigma x^2} \tag{5.11}$$

Note the similarity between (5.10) and (5.11). Whereas sum of cross products and sum of squares of standard scores are used in the former, the latter requires the deviation sum of cross products and sum of squares. It is, however, not necessary to carry out the calculations indicated in (5.10) as b and β are related as follows:

$$\beta = b\frac{s_x}{s_y} \qquad b = \beta\frac{s_y}{s_x} \tag{5.12}$$

where β = standardized regression coefficient; b = unstandardized regression coefficient; and s_x, s_y = standard deviations of X and Y, respectively. Substituting (5.11) and the formulas for the standard deviations of X and Y in (5.12), we obtain

$$\beta = b\frac{s_x}{s_y} = \frac{\Sigma xy\sqrt{\Sigma x^2}\sqrt{N-1}}{\Sigma x^2\sqrt{N-1}\sqrt{\Sigma y^2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = r_{xy} \tag{5.13}$$

Note that with one independent variable $\beta = r_{xy}$. Also, when using standard scores, the intercept, a, is zero. The reason for this is readily seen when you recall that the mean of z scores is zero. Therefore,

$$a = \bar{Y} - \beta\bar{X} = 0 - \beta 0 = 0$$

For two independent variables, X_1 and X_2, the regression equation with standard scores is

$$z'_y = \beta_1 z_1 + \beta_2 z_2 \tag{5.14}$$

where β_1 and β_2 are standardized regression coefficients; z_1 and z_2 are standard scores on X_1 and X_2, respectively. The formulas for calculating the β's when two independent variables are used are

Note that when the independent variables are not correlated (i.e., $r_{12} = 0$), $\beta_1 = r_{y1}$ and $\beta_2 = r_{y2}$ as is the case in simple linear regression. This is true for any number of independent variables: when the independent variables are not intercorrelated, β for a given independent variable is equal to the correlation coefficient (r) of that variable with the dependent variable.

The correlations for the data of Table 5.1 (see Table 5.2) are $r_{y1} = .792$; $r_{y2} = .678$; and $r_{12} = .522$.

$$\beta_1 = \frac{.792 - (.678)(.522)}{1 - .522^2} = .602$$

$$\beta_2 = \frac{.678 - (.792)(.522)}{1 - .522^2} = .364$$

The regression equation, in standard scores, for the data of Table 5.1 is

$$z'_y = .602z_1 + .364z_2$$

Having calculated the β's, the corresponding b's, the unstandardized regression coefficients, can be calculated as follows:

$$b_j = \beta_j\frac{s_y}{s_j} \tag{5.16}$$

where b = unstandardized regression coefficient; j = 1, 2; β = standardized regression coefficient; and s_y and s_j are, respectively, standard deviations of Y and X_j. For the data of Table 5.1,

$$s_y = 2.720 \qquad s_1 = 2.323 \qquad s_2 = 1.670$$

$$b_1 = .602\,\frac{2.720}{2.323} = .705 \qquad b_2 = .364\,\frac{2.720}{1.670} = .593$$

which are, within rounding errors, the same values I obtained earlier. Once the b's are calculated, a can be calculated by (5.5).

SQUARED MULTIPLE CORRELATION COEFFICIENT

In Chapter 2, I showed that the ratio of ss_{reg} to the total sum of squares, Σy^2, equals the squared correlation coefficient between the independent and the dependent variable. The same is true for the case of multiple independent variables, except that the ratio equals the squared multiple correlation:

$$R^2 = \frac{ss_{\text{reg}}}{\Sigma y^2} \tag{5.17}$$

R^2, then, indicates the proportion of variance of the dependent variable accounted for by the independent variables. Using the sums of squares I calculated earlier,

$$R^2 = \frac{101.60}{140.55} = .723$$

About 72% of the variance in reading achievement is accounted for by verbal aptitude and achievement motivation.

Another way of viewing R^2 is to note that it is the squared correlation of Y (observed Y's) and Y' (predicted Y's), which are of course a linear combination of the X's.

The values of (5.18) can be calculated from the Y and Y' columns of Table 5.1. We already have $\Sigma y^2 = 140.55$. The comparable value of $\Sigma y'^2$ is calculated as follows:

$$\Sigma y'^2 = \Sigma Y'^2 - \frac{(\Sigma Y')^2}{N} = 786.0525 - \frac{(117)^2}{20} = 101.60$$

The sum of the deviation cross products is:

$$\Sigma yy' = \Sigma YY' - \frac{(\Sigma Y)(\Sigma Y')}{N} = 786.0528 - \frac{(117)(117)}{20} = 101.60$$

Note that $\Sigma y'^2 = \Sigma yy'$. Substituting in (5.18)

$$R^2 = \frac{(101.60)^2}{(140.55)(101.60)} = .723$$

The positive square root of R^2 gives R. Unlike r, which can take positive as well as negative values, R may vary from .00 to 1.00. For the data of Table 5.1

$$R_{y.12} = \sqrt{.723} = .85$$

or

$$R_{y.12} = r_{yy'} = \frac{\Sigma yy'}{\sqrt{\Sigma y^2}\sqrt{\Sigma y'^2}} = \frac{101.60}{\sqrt{140.55}\sqrt{101.60}} = .85$$

For completeness of presentation, I calculated R, even though it may be irrelevant in regression analysis. As I explained in the case of simple linear regression (see Chapter 2), r^2, not r, is the meaningful term in regression analysis; likewise R^2, not R, is the meaningful term in multiple regression analysis.

Calculation of Squared Multiple Correlation Coefficient

There are various formulas for the calculation of R^2. Following is one in which β's and r's are used:

$$R^2_{y.12} = \beta_1 r_{y1} + \beta_2 r_{y2} \tag{5.19}$$

For the present data,

$$R^2_{y.12} = (.602)(.792) + (.364)(.678) = .724$$

which is, within rounding, the same value I obtained in the lengthier calculations.

Yet another formula for the calculation of R^2 can be obtained by substituting (5.15) in (5.19):

$$R^2_{y.12} = \frac{r^2_{y1} + r^2_{y2} - 2r_{y1}r_{y2}r_{12}}{1 - r^2_{12}} \tag{5.20}$$

Formula (5.20) is very simple and very useful for the calculation of R^2 with two independent variables. All one needs are the three r's. Note that when the correlation between the independent variables is zero (i.e., $r_{12} = 0$), then (5.20) reduces to $R^2_{y.12} = r^2_{y1} + r^2_{y2}$, as was noted earlier.

For the data of Table 5.1,

$$R^2_{y.12} = \frac{.792^2 + .678^2 - 2(.792)(.678)(.522)}{1 - .522^2} = .723$$

Again, this is the same as the value I obtained earlier.

TESTS OF SIGNIFICANCE AND INTERPRETATIONS

I discussed the role of tests of significance and the assumptions underlying them in Chapter 2 and will therefore not address these issues here. Of several tests of significance that may be applied to results of multiple regression analysis, I present three here: (1) test of R^2, (2) tests of regression coefficients, and (3) tests of increments in the proportion of variance accounted for by a given variable.

Test of R^2

The test of R^2 proceeds as the test of r^2, which I presented in Chapter 2.

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} \tag{5.21}$$

with k and $N-k-1$ df. k = number of independent variables and N = sample size. For the data of Table 5.1, $R^2_{y.12} = .723$, $N = 20$.

$$F = \frac{.723/2}{(1-.723)/(20-2-1)} = \frac{.3615}{.0163} = 22.18$$

with 2 and 17 df, $p < .01$. One can, of course, calculate F using the appropriate sums of squares. The formula is

$$F = \frac{ss_{\text{reg}}/df_{\text{reg}}}{ss_{\text{res}}/df_{\text{res}}} \tag{5.22}$$

The degrees of freedom associated with ss_{reg} are $k = 2$, the number of independent variables. The degrees of freedom associated with ss_{res} are $N - k - 1 = 17$. Earlier I calculated $ss_{\text{reg}} = 101.60$ and $ss_{\text{res}} = 38.95$. Therefore,

$$F = \frac{101.60/2}{38.95/17} = \frac{50.80}{2.29} = 22.18$$

This agrees with the F I got when I tested R^2.

Whether one uses (5.21) or (5.22) is a matter of taste, as the same test is being performed. The identity of the two tests can be seen when it is noted that $ss_{\text{reg}} = R^2\Sigma y^2$ and $ss_{\text{res}} = (1-R^2)\Sigma y^2$. Substituting these equivalencies in (5.22), Σy^2 can be canceled from the numerator and the denominator, yielding (5.21).

Based on the R^2, one would conclude that verbal aptitude and achievement motivation account for about 72% of the variance in reading achievement and that this finding is statistically significant at the .01 level.

The test of R^2 indicates whether the regression of Y on the independent variables taken together is statistically significant. Stated differently, testing R^2 is tantamount to testing whether at least one regression coefficient differs from zero. Failure to reject the null hypothesis leads to the conclusion that all the regression coefficients do not differ significantly from zero (I discuss this point in the following section).

When, however, one wishes to determine whether the effect of a given variable is significantly different from zero, it is the regression coefficient, b, associated with it that is tested.

Tests of Regression Coefficients

Each b in a multiple regression equation indicates the expected change in Y associated with a unit change in the independent variable under consideration while controlling for, or holding constant, the effects of the other independent variables. Accordingly, the b's are called partial regression coefficients or partial slopes. To avoid cumbersome notation, I omitted certain subscripts from the multiple regression equations I presented thus far. Inserting the relevant subscripts, the regression equation with two independent variables, for example, is written thus:

$$Y' = a + b_{y1.2}X_1 + b_{y2.1}X_2 \tag{5.23}$$

where $b_{y1.2}$ and $b_{y2.1}$ are partial regression coefficients. Each of these b's is referred to as a first-order partial regression coefficient, the order pertaining to the number of variables that are held constant, or partialed. With two independent variables, each b is of a first order because one variable is partialed in each case. With three independent variables, there are three second-order partial coefficients, as two variables are partialed in the calculation of each b. With k independent variables, the order of each b is $k - 1$. I will not use the notation of (5.23), as it is clear which variables are partialed.

In Chapter 2, I stated that dividing a b by its standard error yields a t ratio. The same is true in multiple regression analysis, where each b has a standard error associated with it. The standard error of b_1, for instance, when there are k independent variables, is

$$s_{b_{y1.2...k}} = \sqrt{\frac{s^2_{y.12...k}}{\Sigma x_1^2(1 - R^2_{1.2...k})}} \tag{5.24}$$

where $s_{b_{y1.2...k}}$ = standard error of b_1; $s^2_{y.12...k}$ = variance of estimate; Σx_1^2 = sum of squares of X_1; and $R^2_{1.2...k}$ = squared multiple correlation of X_1, treated as a dependent variable, with X_2 to X_k as the independent variables. All other b's should be similarly treated. For the case of two independent variables,

$$s_{b_{y1.2}} = \sqrt{\frac{s^2_{y.12}}{\Sigma x_1^2(1 - r^2_{12})}} \qquad s_{b_{y2.1}} = \sqrt{\frac{s^2_{y.12}}{\Sigma x_2^2(1 - r^2_{12})}} \tag{5.25}$$

The denominator of (5.24), or (5.25), reveals important aspects of tests of significance of b's, namely the effects of the correlations among the independent variables on the standard errors of the b's. I discuss the topic of high intercorrelations among independent variables in detail in Chapter 10 (see the "Collinearity" section). Here, I will only point out that the higher the intercorrelation among the independent variables, the larger the standard errors of the b's. It therefore follows that when the independent variables are highly intercorrelated, it may turn out that none of the b's is statistically significant when each is tested separately. Note, on the other hand, that when, for example, $r_{12} = 0$, the denominator of (5.25) reduces to Σx^2, as when a single independent variable is used—see (2.28) and the discussion related to it. These properties of s_b underscore the virtue of designing studies in which the independent variables are not correlated among themselves, as can be done in experimental research.

Test of R² versus Test of b. Before illustrating calculations of standard errors of b's and their use in tests of significance, I elaborate on the distinction between the test of R^2 and the test of a given b in a multiple regression equation. I said earlier that the test of R^2 is tantamount to testing all the b's simultaneously. When testing a given b for significance, the question addressed is whether *it differs from zero while controlling for the effects of the other independent variables.*

Failure to distinguish between the purposes of the two tests has led some researchers to maintain that they might lead to contradictory or puzzling conclusions. For example, R^2 may be statistically significant, leading to the conclusion that at least one regression coefficient is statistically significant. Yet when each regression coefficient is tested separately, it may turn out that *none* is statistically significant. Earlier I alluded to a possible reason for such an occurrence—when the independent variables are highly intercorrelated the standard errors of the b's are relatively large. As long as the different questions addressed by the test of R^2 and by the test of a b are borne in mind, there should be no reason for puzzlement about seemingly contradictory results they may yield.[4]

Tests of b's for the Present Example. Recall that

$$s^2_{y.12} = \frac{ss_{\text{res}}}{N - k - 1} \tag{5.26}$$

where $s^2_{y.12}$ = variance of estimate; ss_{res} = residual sum of squares; N = sample size; and k = number of independent variables. For the present example (see calculations earlier in the chapter),

$$ss_{\text{res}} = 38.95$$

$$s^2_{y.12} = \frac{38.95}{20 - 2 - 1} = 2.29$$

$$\Sigma x_1^2 = 102.55 \qquad \Sigma x_2^2 = 53.00$$

$$b_1 = .7046 \qquad b_2 = .5919 \qquad r_{12} = .522$$

$$s_{b_1} = \sqrt{\frac{2.29}{102.55(1 - .522^2)}} = .1752$$

$$t_{b_1} = \frac{b_1}{s_{b_1}} = \frac{.7046}{.1752} = 4.02$$

with 17 *df* (*df* associated with the variance of estimate: $N - k - 1$), $p < .05$.

$$s_{b_2} = \sqrt{\frac{2.29}{53.00(1 - .522^2)}} = .2437$$

$$t_{b_2} = \frac{b_2}{s_{b_2}} = \frac{.5919}{.2437} = 2.43$$

with 17 *df*, $p < .05$. Assuming that α (level of significance) = .05 was selected, it would be concluded that the effects of both independent variables on the dependent variable are statistically significant. *I remind you not to overlook the important distinction between statistically significant and substantively meaningful findings* (see Chapter 2).

Earlier in this chapter, I distinguished between b (unstandardized regression coefficient) and β (standardized regression coefficient) and pointed out that the former is applied to raw scores, whereas the latter is applied to standard scores. I do not show how to test β's inasmuch as the t (or F) for a given b is the same as the one that would be obtained when its corresponding β is tested (see formulas for getting β from b, and vice versa, earlier in this chapter). In short, testing a b is tantamount to testing its corresponding β.

In Chapter 2, I discussed several factors that affect the precision of regression statistics. One of these factors is the variability of X as reflected by Σx^2. The effect of the variability of X in the present example may be noted from the two standard errors of the b's, given previously. Except for Σx^2, all other terms are identical in both standard errors. Since $\Sigma x_1^2 = 102.55$ and $\Sigma x_2^2 = 53.00$, the standard error of b_1 is smaller than the standard error of b_2. Other things equal, division by a smaller standard error will, of course, yield a larger t ratio.

Confidence Intervals. In Chapter 2, I explained the idea and benefits of setting confidence intervals around the regression coefficient in simple regression analysis. The same approach is taken in multiple regression analysis, namely,

$$b \pm t_{(\alpha/2,\, df)} s_b$$

where t is the tabled t ratio at $\alpha/2$ with df associated with the variance of estimate, or the mean square residual; and s_b is the standard error of the b. For the present example, $df = 17$. Assuming it is desired to set the 95% confidence interval for the present example, the tabled t at .05/2 (.025) with 17 df is 2.11 (see table of t in statistics books, or take $\sqrt{F}$ with 1 and 17 df from Appendix B).

Using the results obtained in the preceding, 95% confidence intervals for the first and second regression coefficients, respectively, are

$$.7046 \pm (2.11)(.1752) = .3349 \text{ and } 1.0743$$

$$.5919 \pm (2.11)(.2437) = .0777 \text{ and } 1.1061$$

Note that, as expected based on the tests of significance of the two b's (see preceding), the confidence intervals do not include zero. Also, as expected based on the standard error of the b's, the confidence interval for b_1 is narrower than that for b_2.

Testing Increments in Proportion of Variance Accounted For. In multiple regression analysis, an increment in the proportion of variance accounted for by a given variable, or a set of variables, can be tested. I discuss this approach in detail later in the text (notably Chapter 9). For now, I introduce some rudimentary ideas about this topic. The test for an increment in the proportion of variance accounted for is given by

$$F = \frac{(R^2_{y.12\ldots k_1} - R^2_{y.12\ldots k_2})/(k_1 - k_2)}{(1 - R^2_{y.12\ldots k_1})/(N - k_1 - 1)} \tag{5.27}$$

where $R^2_{y.12\ldots k_1}$ = squared multiple correlation coefficient for the regression of Y on k_1 variables (the larger coefficient, referred to as the full model); $R^2_{y.12\ldots k_2}$ = squared multiple correlation for the regression of Y on k_2 variables, where k_2 = the smaller set of variables selected from among those of k_1 (referred to as the restricted model); and N = sample size. The F ratio has $k_1 - k_2$ df for the numerator and $N - k_1 - 1$ df for the denominator. Formula (5.27) could also be used to test increments in regression sum of squares. The test is, of course, identical, as a regression sum of squares is a product of a proportion of variance multiplied by the total sum of squares; for example,

$$ss_{\text{reg}(12\ldots k_1)} = (R^2_{y.12\ldots k_1})\Sigma y^2$$

For the example under consideration, (5.27) can be used to test the increment due to X_2 (i.e.,

These are, respectively,

$$F = \frac{(R^2_{y.12} - R^2_{y.1})/(2 - 1)}{(1 - R^2_{y.12})/(N - 2 - 1)}$$

and

$$F = \frac{(R^2_{y.12} - R^2_{y.2})/(2 - 1)}{(1 - R^2_{y.12})/(N - 2 - 1)}$$

Earlier, I calculated $r_{y1} = R_{y.1} = .792$ and $r_{y2} = R_{y.2} = .678$. Thus, X_1 by itself accounts for about 63% ($.792^2$) of the variance of Y, and X_2 accounts for about 46% ($.678^2$) of the variance of Y. Together, the two variables account for about 72% of the variance ($R^2_{y.12} = .723$).

Testing the increment due to X_2,

$$F = \frac{(.723 - .6273)/(2 - 1)}{(1 - .723)/(20 - 2 - 1)} = \frac{.0957}{.0163} = 5.87$$

with 1 and 17 df, $p < .05$. Although, as I noted earlier, X_2 by itself accounts for about 46% of the variance of Y, its increment to the accounting of variance over X_1 is about 10% (.0957). This is to be expected, as some of the information X_1 and X_2 provide is redundant ($r_{12} = .522$). Clearly, the larger the correlation between the two variables, the smaller the increment in the proportion of variance accounted by either. Testing now the increment due to X_1,

$$F = \frac{(.723 - .4597)/(2 - 1)}{(1 - .723)/(20 - 2 - 1)} = \frac{.2633}{.0163} = 16.15$$

with 1 and 17 df, $p < .05$. Recall that by itself, X_1 accounted for about 62% of the variance of Y. Again, the reduction from 62% to 26% (.2633) reflects the correlation between X_1 and X_2.

For now, I will make two points about this procedure. One, testing the increment in proportion of variance accounted for by a single variable is equivalent to the test of the b associated with the variable. From earlier calculations, $b_1 = .7046$ with $t = 4.02$ (17 df) and $b_2 = .5919$ with $t = 2.43$ (17 df). Recall that $t = \sqrt{F}$, when F has one degree of freedom for the numerator. Using the above two F ratios, $\sqrt{16.15} = 4.01$ and $\sqrt{5.87} = 2.42$. To repeat, a test of a b is equivalent to a test of the increment in proportion of variance that is due to the variable with which it is associated. Two, the increment in the proportion of variance accounted for by a given variable (or by a set of variables) may be considerably different from the proportion of variance it accounts by itself, the difference being a function of the correlations of the variable with the other variables in the equation.

RELATIVE IMPORTANCE OF VARIABLES

Researchers use diverse approaches aimed at determining the relative importance of the independent variables under study. This is an extremely complex topic that I discuss later in the text (see, in particular, Chapters 9 and 10). Here, I comment briefly on the use of regression coefficients and increments in proportion of variance accounted for as indices of the relative importance of variables.

b's and β's

The magnitude of b is affected, in part, by the scale of measurement used to measure the variable with which the b is associated. Assume, for example, a simple linear regression in which Y is length of objects measured in feet. Suppose that one were to express X in inches instead of feet. The nature of the regression of Y on X will, of course, not change, nor will the test of significance of the b. The magnitude of the b, however, will change considerably. In the present case, the b associated with X when measured in inches will be one-twelfth of the b when X is measured in feet. This should alert you to two things: (1) a relatively large b may be neither substantively meaningful nor statistically significant, whereas a relatively small b may be both meaningful and statistically significant, and (2) sizes of b's should *not* be used to infer the relative importance of the variables with which they are associated (see Chapter 10).

Incidentally, because b's are affected by the scales being used, it is necessary to carry out their calculations to several decimal places. For a given scale, the b may, for example, be .0003 and yet be substantively meaningful and statistically significant. Had one solved to two decimal places, this b would have been declared to equal zero (I give some such numerical examples in subsequent chapters). In general, it is suggested that calculations of regression analysis be carried out to as many decimal places as is feasible. Further rounding may be done at the end of the calculations.

Because of the incomparability of b's, researchers who wish to speak of the relative importance of variables resort to comparisons among β's, as they are based on standard scores. In the numerical example analyzed previously, $\beta_1 = .602$ and $\beta_2 = .364$. Thus one may wish to conclude that the effect of X_1 is more than 1.5 times as great as the effect of X_2. Broadly speaking, such an interpretation is legitimate, but it is not free of problems because the β's are affected, among other things, by the variability of the variable with which they are associated. Recall that $\beta = r$ in simple linear regression. In Chapter 2, I showed that while r, hence β, varied widely as a function of the variability of X, b remained constant (see Table 2.3 and the discussion related to it). The same principle operates in multiple regression analysis (for a discussion of this point, and numerical examples, see Chapter 10). My aim in the present discussion is only to alert you to the need for caution when comparing magnitudes of β's for the purpose of arriving at conclusions about the relative importance of variables. I postpone discussions of other issues regarding the interpretation of regression coefficients until after I give a more thorough presentation of multiple regression analysis.

Increment in Proportion of Variance Accounted For

I cannot discuss issues concerning the use of the increment in the proportion of variance accounted for by an independent variable as an indication of its relative importance without addressing the broader problem of variance partitioning—a topic to which I devote Chapter 9 in its entirety. All I will say here is that when the independent variables are intercorrelated, the proportion of variance incremented by a variable depends, among other things, on its point of entry into the regression analysis. Thus, when all the correlations among the variables are positive, the later the point of entry of a variable, the smaller the proportion of variance it is shown to account for in the dependent variable. Questions will undoubtedly come to your mind: How, then, does one determine the order of entry of the variables? Is there a "correct" order? As I show in Chapters 8 and 9, attempts to answer such questions are closely related to considerations of the theory that has generated the research and its focus—that is, explanation or prediction.

I can well imagine your sense of frustration at the lack of definitive answers to questions about the relative importance of variables. Yet, attempts to come to grips with such questions

must be postponed until after multiple regression analysis has been explored in greater detail and depth. Only then will it become evident that there is more than one answer to such questions and that the ambiguity of some situations is not entirely resolvable.

COMPUTER ANALYSIS

Using computer packages I introduced in Chapter 4, I will analyze the numerical example I analyzed in preceding sections (Table 5.1).[5] First, I will present a detailed analysis through SPSS. Then, I will give listings of inputs for the other packages (BMDP, MINITAB, and SAS) and brief excerpts of output. Replicate my analysis using one or more of the computer programs available to you, and compare your output with mine. As I pointed out in Chapter 4, I introduce and discuss substantive issues in the context of commentaries on output. Therefore, *study the output and my commentaries on it, even when the specific program that generated the output is of no interest or is irrelevant to you.*

SPSS

Input

```
TITLE TABLE 5.1.   TWO INDEPENDENT VARIABLES.
DATA LIST FREE/Y,X1,X2.
BEGIN DATA
2 1 3
4 2 5     [first two subjects]
. . .
7 4 6     [last two subjects]
10 4 9
END DATA
LIST.
REGRESSION VAR Y,X1,X2/DES ALL/STAT ALL/DEP=Y/
  ENTER X1/ENTER X2/
  RESIDUALS OUTLIERS (COOK LEVER)/
  SCATTERPLOT (Y,X1)(Y,X2)(X1,X2)(*RES,*PRE)
  (*RES,X1)(*RES,X2)/
  SAVE COOK(COOK) LEVER(LEVER) DFBETA SDBETA/PARTIALPLOT/
  DEP Y/ENTER X2/ENTER X1.
PLOT /HSIZE=40/VSIZE=12/
  PLOT COOK LEVER WITH X1.
LIST COOK TO SDB2_1/FORMAT=NUMBERED.
TITLE TABLE 5.1. LAST CASE DELETED.
N 19.
REGRESSION VAR Y,X1,X2/DES ALL/STAT ALL/DEP=Y/
  ENTER X1/ENTER X2.
```

[5] If necessary, refer to Chapter 4 for a general orientation to the packages and to my practice in presenting and commenting on input and output.

Commentary

As I pointed out in Chapter 4, italicized comments in brackets (e.g., *[first two subjects]*), are *not* part of either the input or the output. As I gave an orientation to SPSS, with special emphasis on the use of REGRESSION, in Chapter 4, I will comment only on aspects of the input relevant to the analysis under consideration.

Examine the second line of the REGRESSION procedure and notice the two ENTER statements. Their effect is to enter the independent variables sequentially (X1 followed by X2), thereby enabling one to see how much of the variance of Y is accounted for by X1 and how much X2 adds over and above X1. Notice also that, several lines later, I reversed the order of entry of the variables (i.e., ENTER X2/ENTER X1). Entering variables, or blocks of variables, sequentially—often called hierarchical regression analysis—is probably the most misunderstood and abused approach in applications of multiple regression analysis. *Here, I use it solely to replicate my earlier analyses in this chapter.* In subsequent chapters (especially Chapters 9 and 10), I discuss hierarchical regression analysis in detail.

RESIDUALS Subcommand. Instead of specifying CASEWISE ALL, as in Chapter 4, I specified OUTLIERS for Cook's *D* and Leverage. As a result, the ten cases with the largest values for each of these indices will be listed. This approach, which I use here for illustrative purposes, is particularly useful when analyzing a large data set.

I will explain PARTIALPLOT when I reproduce output generated by it. For explanations of all other REGRESSION subcommands, see Chapter 4. The end of the REGRESSION procedure is signified by the period after X1 (see line preceding PLOT).

I then invoke two procedures: PLOT and LIST. In the former, I use HSIZE and VSIZE to specify the number of columns (horizontal) and number of rows (vertical) to be used in the plots. I call for the plotting of Cook's *D* and leverage against X1.

Following the LIST procedure, I invoke again the REGRESSION procedure. As I indicated in the TITLE, in this analysis I exclude the last subject. In my commentaries on the output generated by it, I explain why I do this analysis. Here it is convenient to exclude the last subject by specifying N 19. Various other approaches to subject selection (e.g., SELECT IF) are more useful in other circumstances.

Output

```
           Mean      Std Dev      Variance
Y          5.850      2.720         7.397
X1         4.350      2.323         5.397
X2         5.500      1.670         2.789

N of Cases =     20

Correlation, Covariance, Cross-Product:
              Y             X1            X2
Y           1.000          .792          .678
            7.397         5.003         3.079
          140.550        95.050        58.500

X1           .792         1.000          .522
            5.003         5.397         2.026
```

```
X2           .678          .522         1.000
            3.079         2.026         2.789
           58.500        38.500        53.000
```

Commentary

I explained layout of this type of output in Chapter 4. To recapitulate briefly, though, each set of three rows in the second portion of the output is composed of correlations (first row), covariances or variances (second row), and deviation sum of cross products or sum of squares (third row). Following are some examples: (1) .792 (first row, second column) is the correlation between Y and X1; (2) 5.397 (fifth row, second column) is the variance of X1; (3) 38.500 (sixth row, third column) is the deviation sum of cross products of X1 and X2; and (4) 53.00 (ninth row, third column) is the deviation sum of squares of X2. Other terms are treated similarly.

Output

```
Equation Number 1     Dependent Variable..     Y
Variable(s) Entered on Step Number·1..     X1

Multiple R        .79171                                                  Analysis of Variance
R Square          .62681   R Square Change     .62681                  DF   Sum of Squares    Mean Square
                           F Change          30.23314   Regression      1      88.09851        88.09851
Standard Error   1.70704   Signif F Change     .0000    Residual       18      52.45149         2.91397

                                                        F =     30.23314     Signif F =     .0000

------------------ Variables in the Equation ------------------   ------------------ Variables not in the Equation ------------------

Variable          B          Beta        T        Sig T          Variable      Beta In      Partial        T       Sig T

X1             .926865     .791715     5.498      .0000          X2            .363476      .507417      2.428     .0266
(Constant)    1.818137                 2.199      .0412
```

Commentary

Notwithstanding some of the nomenclature (e.g., Multiple R), these results refer to a simple regression analysis, as only one independent variable (X1) was entered at this step. Thus, for example, .79171 is the Pearson correlation between Y and X1. I explained procedures for tests of significance in simple regression analysis in Chapter 2 and will therefore not comment on them here.

When only some of the independent variables are entered into the analysis, SPSS reports statistics for Variables in the Equation and Variables not in the Equation. I reproduced here *excerpts* from this output. Examine first the Variables in the Equation. B refers to the unstandardized regression coefficient (i.e., *b*). Constant refers to the intercept (*a*). Thus, the regression equation for Y on X1 is

$$Y' = 1.82 + .93X_1$$

Recall that dividing *b* by its standard error (not reproduced here) yields a *t* ratio: 5.498, with $N - k - 1$ *df* (18, in the present example) the numerator and $N - k - 1$ df for the denominator. Accordingly, $5.498^2 = 30.228$, which is, within rounding, the same as the F reported under the Analysis of Variance portion of the output. Signif F refers to the probability of obtaining an F ratio of this size or larger, given that the null hypothesis is true. As SPSS reports this probability to four decimal places, one would conclude with $p < .0001$ that $b = 0$. Stated differently, the null hypothesis that $b = 0$ would be rejected at the indicated α level.

The foregoing should be not construed as an attempt to explain the logic of tests of significance. All I meant to do is give a rough explanation of the output. I assume that you are familiar with the major issues and controversies concerning statistical tests of significance (e.g., Type I and Type II errors, effect size, sample size). I reviewed briefly this topic in Chapter 2 (see "Tests of Significance"), where I also gave references for further study.

Beta in the preceding output refers to the standardized regression coefficient (β). Earlier in this chapter—see (5.10)–(5.13) and the discussion related to them—I pointed out, among other things, that in an equation with one independent variable, $\beta = r$. Examine the preceding output and notice that Beta for the variable in the equation is equal to Multiple R which, as I pointed out above, is r in the present case.

Look now at the variables *not* in the equation. "Beta In" means the Beta that would be obtained when X2 is added to the analysis. In addition, the t ratio (2.428) with its associated probability (.0266) for this Beta are reported (see also next excerpt of output). Again, my purpose here is solely to acquaint you with the output, *not to imply that a decision whether or not to add the variable, and how to interpret the results of such an addition, are simple and straightforward.* Earlier, I commented briefly on this topic and pointed out that I will discuss it in detail in Chapter 10.

Later, I explain the information reported under Partial.

Output

```
Variable(s) Entered on Step Number 2..      X2

                                                             Analysis of Variance
Multiple R        .85023                                     DF    Sum of Squares     Mean Square
R Square          .72290    R Square Change    .09609
                            F Change          5.89475   Regression   2     101.60329     50.80164
Standard Error   1.51360    Signif F Change    .0266    Residual    17      38.94671      2.29098

                                                        F =    22.17461     Signif F =    .0000

------------------------------------------ Variables in the Equation ------------------------------------------

Variable          B         SE B     95% Confdnce Intrvl B      Beta     Correl   Part Cor   Partial     T     Sig T

X1             .704647    .175263     .334874    1.074420    .601900   .791715   .513306   .698143   4.021   .0009
X2             .591907    .243793     .077549    1.106265    .363476   .677801   .309976   .507417   2.428   .0266
 (Constant)   -.470705   1.194154   -2.990149    2.048739                                             -.394   .6984
```

Commentary

I trust that you will have no difficulty identifying and interpreting most of this output, as I gave the same results earlier (see my hand calculations) and commented on them. Therefore, I limit my comments here to several specific issues.

R Square Change refers to the change in the squared multiple correlation as a result of

.72290 − .62681); and the F Change (5.89475) is the F ratio for the test of significance of this increment—see (5.27) and the discussion related to it, where I obtained the same F, within rounding.

Earlier, I reported, and commented on, the regression equation, the confidence intervals of the b's, and the Beta coefficients. Note that Beta for X2 and its associated T ratio are the same as those reported in the preceding segment of output under Variables not in the Equation.

Correl is the Pearson correlation between the dependent variable and the independent variable on the given line. Compare with my earlier calculations and with the first segment of the output given in the preceding.

I discuss Part Cor(relation), also referred to as semipartial correlation, in Chapter 7. For now, I will only point out that the squared part correlation is equal to the proportion of variance that will be added (incremented) when the variable with which it is associated is entered last in the analysis. Thus, $.309976^2 = .09609$, which is the same as the R Square Change reported above when X2 is entered after X1. Earlier I showed that X1 by itself (first step of output) accounts for .62681 (or about 63%) of the variance of Y. The squared Part Cor for X1 ($.513306^2 = .26348$) indicates that if X1 were entered after X2 it would add about 26% to the variance accounted for (see output reported in the following). The marked difference between the proportion of variance X1 accounts for when it enters first or second is, of course, due to its correlation with X2. Had the correlation between the two variables been zero, X2 (and X1) would have accounted for the same proportion of variance (equal to the squared correlation with the dependent variable), regardless of their point of entry into the analysis. This important point is part of the general topic of variance partitioning, which I present in Chapter 9.

Partial is the partial correlation. In Chapter 7, I give a detailed discussion of partial correlation and its relation to part correlation.

Output

```
                                   Summary table
                                   -------------
Step   Variable    MultR    Rsq     F(Eqn)   SigF   RsqCh     FCh     SigCh
  1    In:  X1     .7917   .6268    30.233   .000   .6268    30.233    .000
  2    In:  X2     .8502   .7229    22.175   .000   .0961     5.895    .027
```

Commentary

The preceding is an excerpt from the summary table. As you can see, this is a handy summary from which the results of the analysis can be gleaned at a glance. This being a summary, it goes without saying that the information given has been reported earlier. I suggest that you examine elements of this table in conjunction with output given earlier or with output you obtained from your run.

At each step, the program reports which variable(s) was entered (In)[6] and its effect. F(Eqn) is the F ratio for the test of the b's for the variables that are in the equation up to and including the step in question (equivalently, it is the test of Rsq[uared] up to that point). Thus, 30.233 is the test of the b for X1, or the test of .6268 (Rsq), whereas 22.175 is the test of the b's for X1 *and* X2, or the test of .7229 (Rsq). In contrast, FCh(ange) is the F ratio for the increment in the proportion of variance accounted for by the variable(s) in question at the given point—Rsq(uared)Ch(ange). In

[6] [illegible]

view of the fact that X1 is the first variable to enter, Rsq and RsqCh are equal, as are F(Eqn) and FCh. In contrast, at the second step F(Eqn) differs from FCh, as the former refers to the test of Rsq due to X1 *and* X2 and the latter refers to the increment in the proportion of variance due to X2 when it is entered after X1.

Output

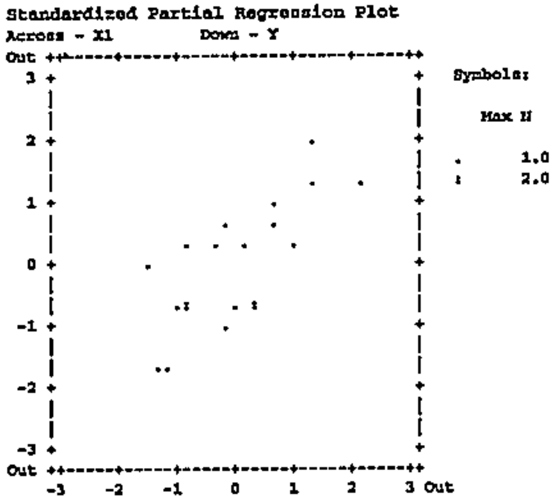
```
     Outliers -- Cook's Distance                 Outliers -- Leverage
Case #            *COOK D              Case #                *LEVER
  20               .84038                20                   .34331
   9               .18391                16                   .25111
   3               .17941                 3                   .14947
  10               .15172                 1                   .14947
   4               .08533                11                   .14674
  17               .06915                10                   .14309
  12               .06682                15                   .13035
  14               .06225                 9                   .12725
   8               .05496                 4                   .11096
  11               .03408                17                   .10342
```

Commentary

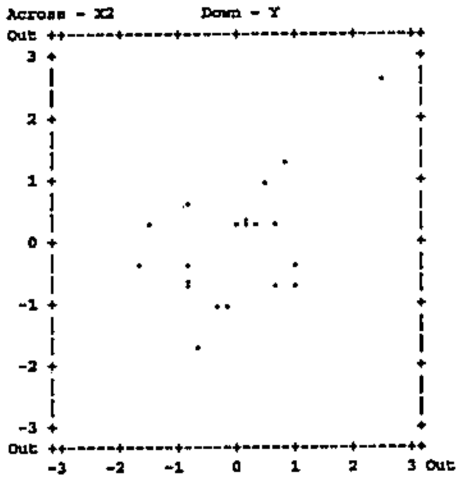
SPSS reports the indices in descending order of magnitude, making it easy to identify extreme cases. As I discussed Cook's D and leverage in Chapter 4, I will only point out that case #20 (the last subject) differs from the others, particularly on Cook's D.

Output

Standardized Partial Regression Plot
Across - X1   Down - Y



Standardized Partial Regression Plot
Across - X2   Down - Y



Commentary

Although I called for additional plots (see Input), in the interest of space I did not reproduce them. If, as I suggested earlier, you ran the data, then you may want to study the other plots. You will find that the data do not seem to contain serious irregularities.

My comments here are limited to the Partial Regression Plots, reproduced previously.[7] I will explain the meaning and benefits of such plots in the context of the current example. In partial regression plots, residuals, *not* original data, are plotted. For example, referring to the plot on the left, residuals of Y are plotted against residuals of X1. What this means is that the dependent variable and each independent variable are, in turn, regressed on all remaining variables (X2 in the present case) to obtain the residuals, which are then plotted (SPSS does this with standardized residuals).[8]

The reasoning behind this approach is that residuals of a given variable are *not* correlated with the independent variable(s), or predictor(s), used to obtain them.[9] Thus, residuals of Y are not correlated with the variable used to generate them (X2 in the example in question). Similarly, residuals of X1 are not correlated with X2. Another way of stating this is that X2 was partialed from Y and X1, hence the residuals of these two variables have nothing in common with X2. Now, if the residuals of X1 are shown to be correlated with the residuals of Y, then this correlation is independent of X2. In other words, Y and X1 are correlated after controlling for or partialing X2, thereby indicating that X1 would enhance the prediction of Y, after the contribution of X2 has been taken into account. This is why some authors (e.g., Cook & Weisberg, 1982, pp. 44–50) refer to the plots under consideration as "added variable plots." Others (e.g., Belsley et al., 1980, p. 30) prefer the term "partial-regression leverage plot." And, "at the risk of further confusion," Welsch (1986, p. 403) proposed "adjusted partial residual plot."

Examine now the two partial regression plots given earlier, and notice that after partialing X2 from Y and X1 (plot on the left) there is a fairly clear linear relation between the residuals, indicating that adding X1 after X2 would lead to a noticeable improvement in predicting Y. When, however, X1 is partialed from Y and X2 (plot on the right), the relation between the residuals seems to be primarily due to a single influential subject, namely case #20 (the point in the upper right-hand corner; recall that this case had a relatively large Cook's D; see also the following plots of Cook's D and leverage with X1). Notice that the scatter of remaining points appears almost random. I will return to this issue in my commentary of the analysis from which I removed the last subject.

Output

```
        Equation Number 2      Dependent Variable..     Y
        Variable(s) Entered on Step Number 1..          X2

Multiple R        .67780                                              Analysis of Variance
R Square          .45941   R Square Change     .45941                   DF   Sum of Squares    Mean Square
                           F Change          15.29725   Regression       1        64.57075       64.57075
Standard Error   2.05452   Signif F Change     .0010    Residual        18        75.97925        4.22107

                                                        F =      15.29725      Signif F =    .0010
```

[7] For discussions of other diagnostic plots, see the references that follow.

[8] For comparative purposes, I use MINITAB, later on, to generate residuals (unstandardized) and plot them.

```
---------------------- Variables in the Equation ----------------------      ---------- Variables not in the Equation ----------

Variable          B            Beta          T         Sig T             Variable        Beta In        T          Sig T

X2             1.103774       .677801      3.911      .0010              X1              .601900       4.021       .0009
 (Constant)    -.220755                    -.136      .8930

Variable(s) Entered on Step Number 2..          X1

Multiple R        .85023                                                    Analysis of Variance
R Square          .72290    R Square Change       .26348                        DF     Sum of Squares      Mean Square
                            F Change            16.16447     Regression          2       101.60329          50.80164
Standard Error   1.51360    Signif F Change        .0009     Residual           17        38.94671           2.29098

                                                             F =      22.17461        Signif F =     .0000

------------------------------------------------ Variables in the Equation ------------------------------------------------

Variable           B          SE B        95% Confdnce Intrvl B          Beta        . T        Sig T      Part Cor

X1          .   .704647     .175263     .334874       1.074420        .601900      4.021      .0009      .513306
X2              .591907     .243793     .077549       1.106265        .363476      2.428      .0266      .309976
 (Constant)    -.470705    1.194154   -2.990149       2.048739                     -.394      .6984

                                           Summary table
                                           -------------

Step      Variable       MultR        Rsq        F(Eqn)       SigF        RsqCh        FCh        SigCh
  2       In:  X2        .6778       .4594       15.297       .001        .4594       15.297       .001
  1       In:  X1        .8502       .7229       22.175       .000        .2635       16.164       .001
```

Commentary

The preceding are excerpts from the second analysis in which I reversed the order of entry of the variables (see my commentary on the input). That is, I entered X2 first, and then I entered X1. As expected, when X2 enters first, the proportion of variance it accounts for is equal to the square of its correlation with Y (i.e., .46; see R Square at the first step or in the summary table). In contrast, when this variable entered second it accounted for only about 10% of the variance (see the earlier analysis). In the present analysis, the proportion of variance X1 accounts for, over and above that accounted by X2, is about 26% (see R Square Change in the second step or in the summary table). Contrast this with the 63% of the variance accounted for by this variable when it entered first. Note, however, that the overall proportion of variance accounted for, that is, R^2 of Y with all the independent variables (two, in the present example), is the same (.7229), regardless of the order in which the variables are entered. In other words, the order of entry of the variables in the analysis does *not* affect the overall R^2, but rather what portion of it is attributed to each variable.

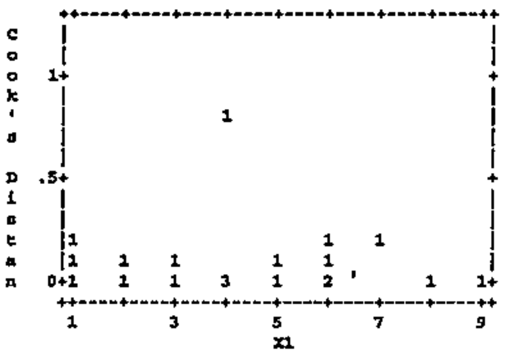
Earlier I pointed out that the squared part correlation indicates the proportion of variance that the variable with which it is associated accounts for when the variable is entered last in the analysis. For X2, $.309976^2 = .09609$, the proportion of variance it accounted for when it entered second (compare with R Square Change in the second step of the first analysis). For X1, $.513306^2 = .26348$, the proportion of variance it accounted for when it entered second (compare with R Square Change in the second step of the current analysis).

To reiterate: when, as in the present example, the independent variables, or predictors, are correlated, the proportion of variance a given variable is shown to account for will vary, depending on its point of entry into the analysis. As I pointed out earlier, I discuss this topic in Chapter 9.

Earlier in this chapter, I discussed tests of regression coefficients and of increments in proportion of variance accounted for by a given variable. For present purposes, therefore, I will only remind you that each *b* indicates the expected change in the dependent variable associated with a unit change in the variable in question, while controlling for the remaining independent variables (this is why the *b*'s are referred to as partial regression coefficients). It follows that a test of a *b* is equivalent to the test of the proportion of variance the variable with which it is associated accounts for when it is entered last in the analysis. Thus, T^2 for each *b* in the preceding output is equal to the *F* for the corresponding R Square Change. For the test of the *b* associated with X2, $2.428^2 = 5.895$, which is the same as the *F* ratio for R Square Change in the second step of the first analysis. For the test of the *b* associated with X1, $4.021^2 = 16.168$, which is the same as the *F* ratio for R Square Change in the second step of the current analysis. In light of the preceding it should come as no surprise that, when all the variables are in the equation, tests of the *b*'s are the same regardless of the order in which the variables were entered. Compare the output given here with that given earlier, when X1 entered first. Lest you think that I am belaboring the obvious, I suggest that you peruse Chapter 10 where I give research examples of confusion and misconceptions about tests of regression coefficients.

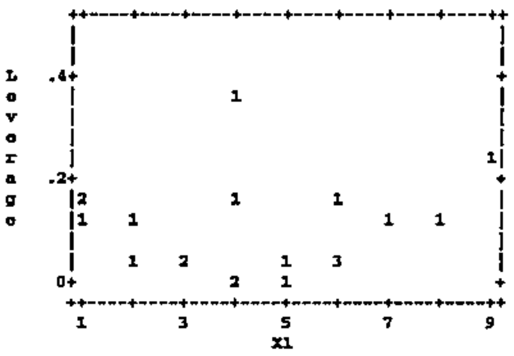
Output

Plot of COOK with X1



20 cases plotted.

Plot of LEVER with X1



20 cases plotted.

Commentary

I generated the preceding by the PLOT procedure (see Input). Examine the plots and notice that on both indices, but especially on Cook's D, one subject whose X1 = 4 is set apart from all the rest. If you looked back at the output for the outliers, you would see that it is Case #20—the last subject in the input file (see also, the next excerpt of output).

Output

```
Case#  COOK    LEVER   DFB0_1    DFB1_1   DFB2_1   SDB0_1    SDB1_1   SDB2_1
  1    .00000  .14947  -.00382   .00025   .00039   -.00311   .00137   .00155
  2    .00020  .05764   .01271  -.00304   .00113    .01033  -.01684   .00451
  .      .       .        .        .        .         .        .        .
 19    .01232  .01154   .00177  -.01122   .01921    .00146  -.06314   .07775
 20    .84038  .34331  -1.14755 -.14862   .36103  -1.06158  -.93675  1.63591
```

Commentary

I introduced DFBETA and DFBETAS (standardized) in Chapter 3 for the case of simple linear regression. To recapitulate: DFBETA$_{j(i)}$ indicates the change in j (intercept or regression coefficient) as a consequence of deleting subject i. In Chapter 4, I reproduced, and commented on, computer output of such indices for simple linear regression. If necessary, review relevant sections in the aforementioned chapters.

In multiple regression analysis, DFBETAs are calculated for each independent variable, as well as for the intercept. In the above output, DFB refers to DFBETA whereas SDB refers to DFBETAS (standardized; see Chapter 3 for an explanation of the nomenclature used). For example, DFB1_1 is DFBETA associated with X1, whereas SDB2_1 is DEFBETAS associated with X2.

The results given earlier, notably Cook's D, led me to focus on the last subject (case #20). If I were inclined to delete this subject, then the output given above would tell me the nature of the changes that would ensue. For convenience, I repeat the regression equation for the entire group, calculated earlier in this chapter and also reported in the SPSS output:

$$Y' = -.470705 + .704647X_1 + .591907X_2$$

Using the DFBETAs associated with the last subject, the regression equation that would be obtained if this subject were deleted is

$$Y' = .67685 + .85327X_1 + .23088X_2$$

Note that deletion of the last subject would result in a slight increase in b_1 and a considerable decrease in b_2, reinforcing the notions gathered from the examination of the partial regression plots, namely that the effect of X_2 is largely due to a single influential subject (#20). In the next excerpt of the output, I show that when this subject is deleted, b_2 is statistically not significant at conventional levels.

Output

```
TABLE 5.1. LAST CASE DELETED
Equation Number 1     Dependent Variable. .     Y
Variable(s) Entered on Step Number 1. .         X1

Multiple R       .86250                                          Analysis of Variance
R Square         .74391   R Square Change     .74391               DF   Sum of Squares   Mean Square
                          F Change          49.38255  Regression    1       91.07008      91.07008
Standard Error  1.35800   Signif F Change     .0000   Residual     17       31.35098       1.84418
```

```
------------------------- Variables in the Equation -------------------------

Variable            B          SE B          T        Sig T

X1               .942960     .134186      7.027      .0000
 (Constant)     1.512333     .663829      2.278      .0359

Variable(s) Entered on Step Number 2. .    X2

Multiple R       .86870                                          Analysis of Variance
R Square         .75464   R Square Change   .01073                 DF   Sum of Squares   Mean Square
                          F Change          .69986   Regression     2       92.38394      46.19197
Standard Error  1.37015   Signif F Change   .4152    Residual      16       30.03711       1.87732

                                                     F =     24.60528     Signif F =     .0000

------------------------- Variables in the Equation -------------------------

Variable            B          SE B          T        Sig T

X1               .853265     .172699      4.941      .0001
X2               .230881     .275983       .837      .4152
 (Constant)      .676841    1.202495       .563      .5813
```

Commentary

Except for the fact that the last case was deleted, the format of the output here is the same as that I reproduced earlier. Accordingly, I focus on the second step only. Notice that when entered after X1, X2 accounts for about 1% of the variance of Y and that this increment is statistically not significant at conventional levels of significance (see R Square Change, F Change, and Signif F Change). Examine now the Variables in the Equation and notice that (1) the regression equation is the same as I calculated above, using DFBETAs for the last case; and (2) the regression coefficient for X2 is statistically not significant. The latter piece of information is, of course, the same as obtained from R Square Change and the F Change (as I explained earlier, $T^2 = F$).

Incidentally, if you replicated this analysis and called also for a residual analysis, you would find that, for instance, all values of Cook's D are relatively small, the largest being .20 for subject #7. In sum, then, as expected from the diagnostic results based on the analysis of the data for all the subjects, the effect of X2 is drastically diminished when the last case is deleted.

Having found in this analysis that b_2 is statistically not significant, the idea of deleting X_2 from the analysis suggests itself. I discuss considerations in making such a decision in subsequent chapters (e.g., Chapter 10). Here, I will only point out that if I decided to delete X_2, then I would interpret the simple regression equation obtained in step 1 above.

Referring to the substantive example I used when I introduced these data (see the discussion of Table 5.1 presented earlier in this chapter), I would, for example, conclude that the expected change in reading achievement associated with a unit change in verbal aptitude is .85. Further, that achievement motivation seems to have no effect on reading achievement. *Recall, however, that the data are fictitious.*

Before turning to the other computer programs, I would like to suggest that you carry out additional analyses that will, I believe, help you better understand the material I presented thus far. In particular, I recommend that you use SPSS to (1) regress Y on X1 and save the residuals, la-

Y.X1 on X2.X1; and (4) plot Y.X1 against X2.X1. Following is an example of control statements to carry out the suggested analyses.

Input

```
REGRESSION VAR Y,X1,X2/DES ALL/STAT ALL/
   DEP Y/ENTER X1/SAVE RESID(Y.X1)/
   DEP X2/ENTER X1/SAVE RESID(X2.X1).
REGRESSION VAR Y.X1 X2.X1/DES ALL/
   DEP Y.X1/ENTER X2.X1.
PLOT VSIZE=15/HSIZE=50/VERTICAL='Y.X1'/
   HORIZONTAL='X2.X1'/
   PLOT Y.X1 WITH X2.X1.
```

Commentary

In the preceding, I did *not* include control statements for reading in the data, as they are the same as those I used earlier. Note that, with slight adjustments, the preceding can be incorporated in the input file I used earlier. Whichever way you run the preceding, study the results and compare them with those given earlier.[10] Among other things, you will find that *b* for the regression of Y.X1 on X2.X1 is .5919, which is the same as the *b* for X2 when Y was regressed on X1 and X2 earlier in this section. Also, the residual sum of squares in the suggested analysis is 38.9467, which is the same as that reported earlier when Y was regressed on X1 and X2. Finally, a comparison of the plot from the suggested analysis with the corresponding partial regression plot from the earlier analysis should also prove instructive, especially as standardized residuals were used in the earlier analysis.

BMDP

Input

```
/PROBLEM TITLE IS 'TABLE 5.1'.
/INPUT VARIABLES ARE 3. FORMAT IS FREE. FILE IS 'T51.DAT'.
/VARIABLE NAMES ARE Y,X1,X2.
/REGRESS DEPEND IS Y. INDEP=X1,X2. LEVEL=0,1,2.
/PRINT COVA. CORR. DATA.
 DIAG=HATDIAG,RESIDUAL,STRESID,DELRESID,DSTRESID,COOK.
/PLOT RESID. XVAR=X1,X1. YVAR=COOK,HATDIAG.
 SIZE=44,12.
/END
/END
/PROBLEM TITLE IS 'TABLE 5.1. LAST CASE OMITTED'.
/INPUT
/REGRESS SETNAMES=BOTH. BOTH=X1,X2.
/TRANSFORM OMIT=20.
/END
```

Commentary

This input is for program 2R. See Chapter 4 for an orientation to BMDP and to basic elements of 2R.

INPUT. For illustrative purposes, I placed the data in an external file (see FILE IS 'T51.DAT'). This is particularly useful when, as in the present case, multiple problems are processed (see Dixon, 1992, Vol. 1, Chapter 9).

REGRESS. LEVEL is used to specify the order in which the variables are to enter into the equation. The zero is for the dependent variable, 1 is for X1, and 2 is for X2. Accordingly, X1 will enter first.

PLOT. I called for residual plots and plots of COOK (Cook's D) and HATDIAG (diagonal of the hat matrix; in Chapter 3, I pointed out that this is another term for leverage) against X1.

As I stated earlier, the present example is composed of two problems. For conventions for running multiple problems, see Dixon (1992, Vol. 1, Chapter 9). In the second problem, I reanalyzed the data after deleting the last subject. I used TRANSFORMATION OMIT=20 to omit case #20. To enter X1 and X2 together in the analysis, I used SETNAMES (see Dixon, 1992, Vol. 1, pp. 408–409).

Output

```
BMDP2R -- STEPWISE REGRESSION
TABLE 5.1
STEP NO.     1
----------------------------
VARIABLE ENTERED     2  X1

MULTIPLE R              0.7917
MULTIPLE R-SQUARE       0.6268
STD. ERROR OF EST.      1.7070

ANALYSIS OF VARIANCE

                  SUM OF SQUARES      DF      MEAN SQUARE       F RATIO
   REGRESSION       88.098540          1       88.09854          30.23
   RESIDUAL         52.451460         18       2.913970

                         VARIABLES IN EQUATION FOR Y

                                                STD. ERROR         STD REG
    VARIABLE            COEFFICIENT             OF COEFF           COEFF
 (Y-INTERCEPT                 1.81814 )
X1              2         0.92687                0.1686            0.792

STEP NO.     2
----------------------------
VARIABLE ENTERED     3  X2

MULTIPLE R              0.8502
MULTIPLE R-SQUARE
```

```
ANALYSIS OF VARIANCE
                   SUM OF SQUARES      DF     MEAN SQUARE     F RATIO
 REGRESSION          101.60330          2      50.80165        22.17
 RESIDUAL            38.946690         17      2.290982

                       VARIABLES IN EQUATION FOR Y

                                      STD. ERROR    STD REG        F
   VARIABLE         COEFFICIENT        OF COEFF      COEFF    TO REMOVE
(Y-INTERCEPT           -0.47071 )
X1          2        0.70465            0.1753       0.602        16.16
X2          3        0.59191            0.2438       0.363         5.89

SUMMARY TABLE

STEP      VARIABLE           MULTIPLE          CHANGE       F TO
 NO.       ENTERED         R         RSQ       IN RSQ       ENTER
  1         2 X1         0.7917    0.6268      0.6268       30.23
  2         3 X2         0.8502    0.7229      0.0961        5.89
```

```
       ...+....+....+....+....+....+....+....+....+...        ...+....+....+....+....+....+....+....+....+...
C      -                                             -  H     -                                             -
O      -                                             -  A     -                                             -
O  1.0 +                                             +  T  .4 +                 1                           +
K      -                                             -  D     -                                             -
       -                 1                           -  I     -                                          1  -
       -                                             -  A     -                                             -
       -                                             -  G     -                                             -
   .50 +                                             +     .2 +  2              1         1         1       +
       -                                             -        -  1    1                        1            -
       -                                             -        -       1                   1                 -
       -  1                        1    1            -        -            2    1    1    2                 -
       -  1    1    1         1    1                 -        -                 1    1                      -
   0.0 +  1    1    1    3    1    2         1    1  +     0. +                                             +
       ...+....+....+....+....+....+....+....+....+...        ...+....+....+....+....+....+....+....+....+...
          1         3         5         7         9              1         3         5         7         9
               2         4         6         8                        2         4         6         8

                              X1                                                     X1
```

```
2R TABLE 5.1.    LAST CASE OMITTED

NUMBER OF CASES READ ..................   20
  CASES WITH USE SET TO ZERO ...........   1
    REMAINING NUMBER OF CASES ........    19

                 VARIABLES IN EQUATION FOR Y

                                           STD. ERROR        STD REG
   VARIABLE            COEFFICIENT          OF COEFF          COEFF
 (Y-INTERCEPT              0.67684 )
 SET BOTH
 X1          2            0.85327             0.1727          0.780
 X2          3            0.23088             0.2760          0.132
```

Commentary

In line with my earlier statement, I reproduced only brief excerpts of the output. If you are using this program, compare your output with the SPSS output I reproduced in the preceding section. When in doubt, reread my commentaries on the SPSS output and on my hand calculations. Here, I comment only on F TO REMOVE and F TO ENTER. These terms are used primarily in stepwise regression analysis—an approach I discuss in detail in Chapter 8. Although 2R was designed for stepwise regression analysis, it can be used for other types of analyses (see Chapter 4).

F TO REMOVE is essentially a test of the regression coefficient with which it is associated. Thus, 16.16 is for the test of b_1, and 5.89 is for the test of b_2. Each of these F ratios has 1 and $N-k-1$ df (17, in the present example) or df for the mean square residuals (*MSR;* see output). Recall that when F has 1 df for the numerator, $t^2 = F$. For the present example, the corresponding t's are 4.02 and 2.43, respectively, which are the values I obtained earlier in this chapter (see "Tests of Regression Coefficients"; see also the SPSS output in the preceding section).

F TO ENTER is essentially a test of the R^2 change associated with a given variable. Thus, 30.23 is for the test of proportion of variance accounted for by X1 (.6268), whereas 5.89 is for the test of the proportion of variance incremented by X2 (0.0961). I obtained these values several times earlier. See, for example, RsqCh and FCh in the Summary Table of the SPSS output in the preceding section.

Finally, recall that the test of a b is equivalent to the test of the proportion of variance accounted for by a variable when it is entered last in the analysis. This is why F TO REMOVE and the F TO ENTER for X2 are identical (5.89).

MINITAB

Input

```
GMACRO
T51
OUTFILE='T51.MIN';
  NOTERM.
NOTE TABLE 5.1
READ C1-C3;
  FILE 'T51.DAT'.          [reading data from external file]
NAME C1 'Y' C2 'X1' C3 'X2'
ECHO
DESCRIBE C1-C3
CORRELATION C1-C3
COVARIANCE C1-C3 M1
PRINT M1
NOTE M1 IS A COVARIANCE MATRIX
BRIEF 3                    [maximum output]
REGRESS C1 2 C2-C3 C4-C5;
  RESIDUALS C6;
  HI C7;
  COOKD C8
```

```
NAME C5 'FITS' C6 'RESIDS' C7 'LEVER' C8 'COOKD'
PRINT C4-C8
GSTD
WIDTH 50
PLOT C6 C5
PLOT C7 C2
PLOT C8 C2
REGRESS C1 1 C2;          [regress Y on X1]
  RESIDUALS C9.           [put residuals in C9]
REGRESS C3 1 C2;          [regress X2 on X1]
  RESIDUALS C10.          [put residuals in C10]
NAME C9 'Y.X1' C10 'X2.X1'
PRINT C9-C10
PLOT C9 C10               [partial regression plot]
DELETE 20 C1-C3           [delete case number 20. MINITAB, 1995a, p. 6–2]
PRINT C1-C3
NOTE CASE NUMBER 20 DELETED
REGRESS C1 2 C2-C3 C4-C5
ENDMACRO
```

Commentary

For an orientation to MINITAB, see Chapter 4, where I commented on many of the commands I use here. As MINITAB does not have an option for partial regression plots, I show how this can be accomplished with relative ease by generating the relevant residuals and plotting them.

Output

```
The regression equation is
Y = -0.47 + 0.705 X1 + 0.592 X2

Predictor        Coef         Stdev        t-ratio        p
Constant       -0.471         1.194         -0.39       0.698
X1              0.7046        0.1753         4.02       0.001
X2              0.5919        0.2438         2.43       0.027

s = 1.514     R-sq = 72.3%

Analysis of Variance

SOURCE        DF        SS          MS          F          p
Regression     2      101.603     50.802      22.17      0.000
Error         17       38.947      2.291
Total         19      140.550

SOURCE        DF      SEQ SS
X1             1       88.099
X2             1       13.505
```

Commentary

SEQ SS = sequential sum of squares, that is, the sum of squares incremented by a given variable at its point of entry into the analysis. Thus, 88.099 is the sum of squares of Y accounted for by X1, and 13.505 is the sum of squares *incremented* by X2 (i.e., what X2 accounts for over and above X1). Thus, SEQ SS constitute a hierarchical regression analysis.

To transform SEQ SS to proportions of variance accounted for, divide each by the total sum of squares: 140.550. Thus, 88.099/140.550 = .627 and 13.505/140.550 = .096. The sum of these two proportions is, of course, $R^2_{y.12}$. I obtained the preceding values several times earlier (see "Testing Increments in Proportion of Variance Accounted For," earlier in this chapter; see also relevant SPSS and BMDP output and commentaries).

Output

```
 RESIDS           LEVER          COOKD
-0.00966         0.199474       0.000004
 0.10187         0.107642       0.000204      [first two subjects]
    .               .              .
 1.10067         0.061537       0.012316      [last two subjects]
 2.32495         0.393306       0.840381
```

```
          -
          -                     *
      0.36+
          -
LEVER     -                                              *
          -
          -
      0.24+
          -
          -      2              *         *         *
          -      *                             *
          -           *
      0.12+                               *
          -           *    2         *    *
          -                     *         *
          -                     *    *
          -
            +---------+---------+---------+---------+---------X1
```

```
         -
     0.90+
         -                     *
COOKD    -
         -
         -
     0.60+
         -
         -
         -
         -
     0.30+
         -
         -      *                        *    *
         -
         -      *    *    *    *    *    *
     0.00+      *    *    *    2    *    2         *    *
           +---------+---------+---------+---------+---------X1
          0.0       2.0       4.0       6.0       8.0
```

```
ROW        Y.X1        X2.X1        [residuals of Y and X2]

 I       –0.74500     –1.24232
 2        0.32813      0.38225      [first two subjects]
 .           .            .
19        1.47440      0.63140      [last two subjects]
20        4.47440      3.63140
```

```
         -                                              *
Y.X1     -
         -
         -
      2.5+                              *
         -
         -             *           *
         -                      *
         -       *            *  **   *
      0.0+
         -     *       *                *
         -             *
         -              *    2       *   *
         -
     -2.5+              *
         -
         -
           +---------+---------+---------+---------+---------X2.X1
         -3.0      -1.5       0.0       1.5       3.0

                    [partial regression plot]
```

Commentary

Although the terminology and the layout of the preceding output differ somewhat from those of SPSS and BMDP, you should have no difficulties in understanding it. If necessary, reread relevant comments on output from the aforementioned programs, notably SPSS.

SAS

Input

```
TITLE 'TABLE 5.1';
DATA T51;
  INPUT Y X1 X2;
  CARDS;
2 1 3
4 2 5        [first two subjects]
. . .
7 4 6        [last two subjects]
10 4 9
;
PROC PRINT;
PROC REG;
  MODEL Y=X1 X2/ALL R INFLUENCE PARTIAL;
RUN;
TITLE 'SECOND RUN AFTER DELETING CASES WITH COOK D >.5';
  REWEIGHT COOKD. >.5;
  PRINT;
RUN;
```

Commentary

For an orientation to SAS and its PROC REG, see Chapter 4. Here, I will only point out that for illustrative purposes I use the REWEIGHT command with the condition that cases whose Cook's D is greater than .5 be excluded from the analysis (see SAS Institute Inc., 1990a, Vol. 2, pp. 1381–1384, for a detailed discussion of REWEIGHT). Given the values of Cook's D for the present data, this will result in the exclusion of the last subject, thus yielding results similar to those I obtained from analyses with the other computer programs. PRINT calls for the printing of results of this second analysis.

Output

TABLE 5.1

```
                              Analysis of Variance

                            Sum of          Mean
Source           DF        Squares         Square          F Value        Prob>F

Model             2      101.60329       50.80164           22.175        0.0001
Error            17       38.94671        2.29098
C Total          19      140.55000
[C Total = Corrected Total Sum of Squares]

              Root MSE           1.51360        R-square          0.7229
              Dep Mean           5.85000

                              Parameter Estimates

                                                                                                  Squared      Squared
              Parameter     Standard     T for H0:                                 Standardized   Semi-partial Semi-part
Variable  DF  Estimate      Error        Parameter=0  Prob >|T|  Type I SS  Type II SS  Estimate   Corr Type I  Corr Type

INTERCEP  1   -0.470705   1.19415386      -0.394     0.6984                                            .            .
X1        1    0.704647   0.17526329       4.021     0.0009   88.098513  37.032535  0.60189973   0.62681261   0.2634830
X2        1    0.591907   0.24379278       2.428     0.0266   13.504777  13.504777  0.36347633   0.09608522   0.0960852
```

Commentary

You should have no difficulty with most of the preceding, especially if you compare it with output I presented earlier from other packages. Accordingly, I comment only on Type I and II SS and their corresponding squared semipartial correlations.

For a general discussion of the two types of sums of squares, see SAS Institute Inc. (1990a, Vol. 1, pp. 115–117). For present purposes I will point out that Type I SS are sequential sums of squares, which I explained earlier in connection with MINITAB output. To reiterate, however, the first value (88.0985) is the sum of squares accounted for by X1, whereas the second value (13.505) is the sum of squares incremented by X2. Squared Semi-partial Corr(elations) Type I are corresponding proportions accounted for sequentially, or in a hierarchical analysis. They are equal to each Type I SS divided by the total sum of squares (e.g., 88.0985/140.55 = .6268, for the value associated with X1). I calculated the same values in my commentaries on MINITAB output.

Type II SS is the sum of squares a variable accounts for when it enters last in the analysis, that is, after having been adjusted for the remaining independent variables. This is why some authors refer to this type of sum of squares as the unique sum of squares, and to the corresponding Squared Semi-partial correlation Type II as the unique proportion of variance accounted for by the variable in question. Thus, when X1 enters last, the increment in sum of squares due to it is 37.0325. Dividing this value by the total sum of squares yields the Squared Semi-partial Corr Type II (37.0325/140.55 = .2635; compare with the previous output). Similarly, the sum of squares accounted for uniquely by X2 (i.e., when it enters last) is 13.5048, and the corresponding Squared Semi-partial Corr Type II is 13.5048/140.55 = .0961 (compare with the previous output). What is labeled here Semi-partial Corr Type II is labeled Part Cor in SPSS (note that SAS reports the square of these indices). Clearly, *only when the independent variables, or predictors, are not correlated will the sum of the unique regression sums of squares be equal to the overall regression sum of squares. The same is true of the sum of the unique proportions of variance accounted for, which will be*

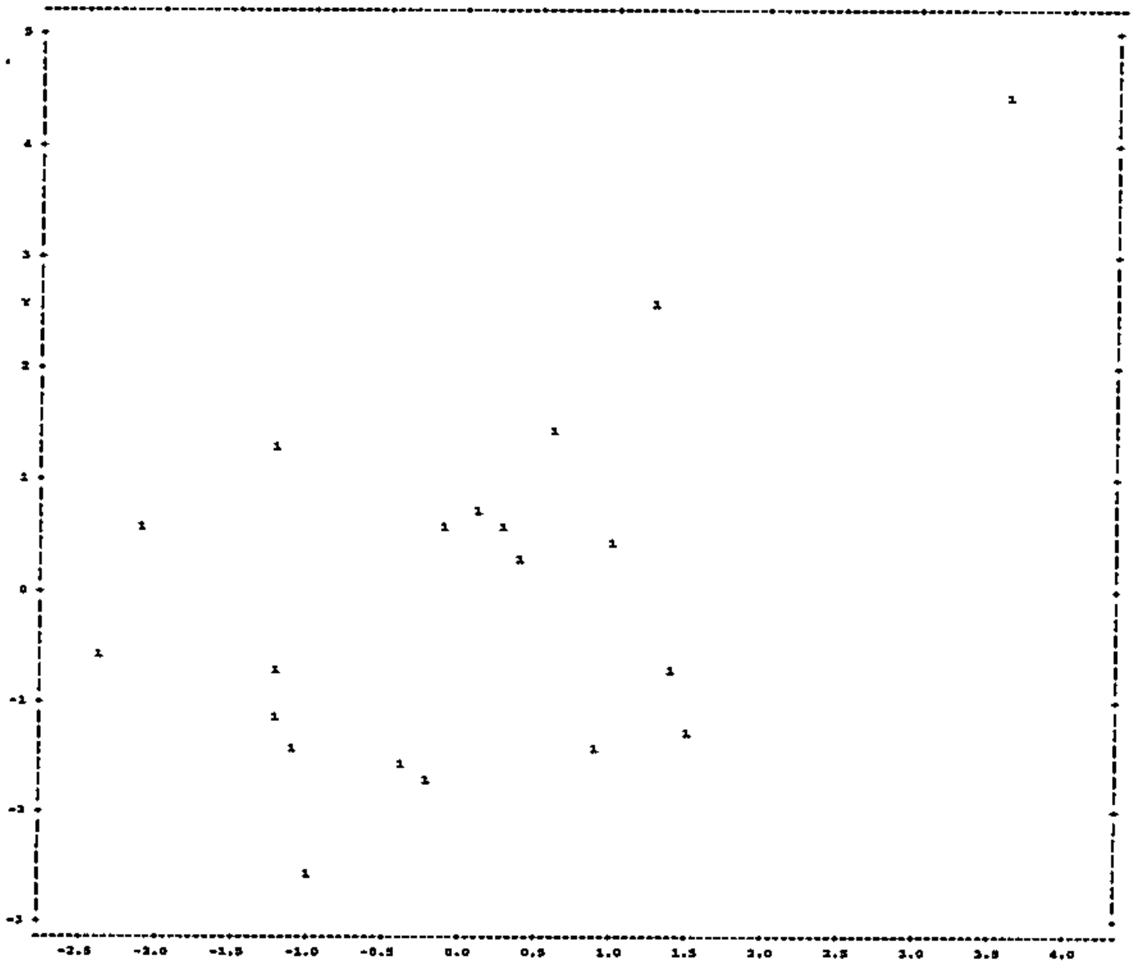
You are probably wondering how one arrives at a decision when to use Type I SS and when to use Type II SS (or the corresponding Squared Semi-partial Correlations), and how they are interpreted substantively. I discuss this complex topic in detail in Chapter 9.

Output

```
            Cook's        Hat Diag      INTERCEP        X1          X2
Obs           D              H          Dfbetas      Dfbetas     Dfbetas

  1         0.000         0.1995        -0.0031       0.0014      0.0015
  2         0.000         0.1076         0.0103      -0.0168      0.0045
  .           .              .              .            .           .
 19         0.012         0.0615         0.0015      -0.0631      0.0777
 20         0.840         0.3933        -1.0616      -0.9367      1.6359
```

Partial Regression Residual Plot



SECOND RUN AFTER DELETING CASES WITH COOK D > .5

```
                                                                                                   Squared        Squared
                 Parameter    Standard      T for H0:                                    Standardized  Semi-partial   Semi-parti
Variable   DF    Estimate     Error         Parameter=0  Prob > |T|  Type I SS  Type II SS  Estimate     Corr Type I    Corr Type I

INTERCEP    1    0.676841     1.20249520    0.563        0.5813                                            .              .
X1          1    0.853265     0.17269862    4.941        0.0001     91.070076  45.827718   0.78045967   0.74390862     0.3743450
X2          1    0.230881     0.27598330    0.837        0.4152      1.313865   1.313865   0.13214835   0.01073235     0.01073235
```

Commentary

My comments on the preceding excerpts will be brief, as I obtained similar results several times earlier.

Hat Diag H. In Chapter 4, I explained that this is another term for leverage. I pointed out that SAS reports standardized DFBETAs only. By contrast, SPSS reports unstandardized as well as standardized values. What was labeled earlier partial regression plots, is labeled by SAS Partial Regression Residual Plots.

The last segment of the output is for the analysis in which the last subject was omitted. Compare with outputs from other packages for the same analysis (given earlier).

CONCLUDING REMARKS

I hope that by now you understand the basic principles of multiple regression analysis. Although I used only two independent variables, I presented enough of the subject to lay the foundations for the use of multiple regression analysis in scientific research. A severe danger in studying a subject like multiple regression, however, is that of becoming so preoccupied with formulas, numbers, and number manipulations that you lose sight of the larger purposes. Becoming engrossed in techniques, one runs the risk of ending up as their servant rather than their master. While it was necessary to go through a good deal of number and symbol manipulations, this poses the real threat of losing one's way. It is therefore important to pause and take stock of why we are doing what we are doing.

In Chapter 1, I said that multiple regression analysis may be used for two major purposes: explanation and prediction. To draw the lines clearly though crassly, if we were interested only in prediction, we might be satisfied with selecting a set of predictors that optimize R^2, and with using the regression equation for the predictors thus selected to predict individuals' performance on the criterion of interest. Success in high school or college as predicted by certain tests is a classic case. In much of the research on school success, the interest has been on prediction of this criterion, however defined. One need not probe too deeply into the whys of success in college; one wants mainly to be able to predict successfully, and this is, of course, no mean achievement. As I show in Chapter 8, which is devoted solely to the use of multiple regression analysis for prediction, various approaches are available to achieve this goal.

In much of behavioral research, however, prediction, successful or not, is not enough. We want to know why; we want to explain phenomena. This is the main goal of science. We want to explain, for instance, phenomena such as problem solving, achievement, creativity, aggression, prejudice, or job satisfaction. When the goal is explanation, the focus shifts to the interpretation of the regression equation. We want to know the magnitudes of the effects of independent variables on the dependent variable as they are reflected by the regression coefficients. But should we use unstandardized (b) or standardized (β) regression coefficients? In this chapter, I could offer only a glimpse at the answer to this question. In Part 2 of this book, I elaborate on applications of multiple regression analysis for explanatory purposes in various research designs. For example, in Chapters 9 and 10 I discuss applications of multiple regression in nonexperimental research, whereas in Chapters 11 and 12 I discuss such applications primarily in experimental research.

In sum, then, Chapters 2 through 5 were designed to set the stage for the study of analytic and technical problems encountered in the use of elements of multiple regression analysis in predictive and explanatory scientific research. Mastery of the technical aspects of an analytic approach is a prerequisite for its valid application and an important antidote against misapplications. Needless to say, the study and mastery of the technical aspects of research are necessary but not sufficient conditions for the solution of research problems.

Finally, recall that in the beginning of this chapter I said that the presentation was limited to two independent variables because of the ease this affords in calculating and discussing elements of multiple regression analysis. Although the meaning of elements of multiple regression analysis with more than two independent variables is the same as for the case of two independent variables, their calculations are best accomplished by the use of matrix algebra, a topic I present in Chapter 6.

STUDY SUGGESTIONS

1. Use the following illustrative data for the calculations indicated as follows. The second set of three columns is merely a continuation of the first set. I suggest that you do all the calculations by hand with the aid of a calculator. You may wish to set up tables like the ones I presented in this chapter to keep things orderly.

X_1	X_2	Y	X_1	X_2	Y
2	5	2	4	3	3
2	4	1	3	6	3
1	5	1	6	9	6
1	3	1	6	8	6
3	6	5	8	9	10
4	4	4	9	6	9
5	6	7	10	4	6
5	4	6	9	5	6
7	3	7	4	8	9
6	3	8	4	9	10

Calculate the following:

(a) Means, standard deviations, sums of squares and cross products, and the three r's.

(b) Regression equation of Y on X_1 and X_2.

(c) ss_{reg}, ss_{res}, $R^2_{y.12}$, F, $s^2_{y.12}$.

(d) t ratios for the two regression coefficients.

(e) Increment in the proportion of variance accounted for by X_2, over and above X_1, and the F ratio for the test of this increment. To what test statistic, calculated earlier, should this F be equal?

(f) Increment in the proportion of variance accounted for by X_1, over and above X_2, and the F ratio for the test of this increment. To what test statistic, calculated earlier, should this F be equal?

(g) Using the regression equation, calculate the predicted Y's, $Y - Y'$ (the residuals), $\Sigma(Y - Y')$, and $\Sigma(Y - Y')^2$.

(h) Calculate the squared correlation between Y' and Y. To what statistic, calculated earlier, should this correlation be equal?

2. Using a computer program, analyze the data given in the previous exercise. Compare results with those obtained in the previous exercise.

   (a) For all subjects, obtain the following: standardized residuals (ZRESID), studentized residuals (SRESID), studentized deleted residuals (SDRESID), leverage (h), Cook's D, and DFBETAs (raw and standardized).