Preface

Linear regression analysis is by far the most popular analytical method in the social and behavioral sciences, not to mention other fields like medicine and public health. Everyone is exposed to regression analysis in some form early on who undertakes scientific training, although sometimes that exposure takes a disguised form. Even the most basic statistical procedures taught to students in the sciences—the *t*-test and analysis of variance (ANOVA), for instance—are really just forms of regression analysis. After mastering these topics, students are often introduced to multiple regression analysis as if it is something new and designed for a wholly different type of problem than what they were exposed to in their first course. This book shows how regression analysis, ANOVA, and the independent groups *t*-test are one and the same. But we go far beyond drawing the parallels between these methods, knowing that in order for you to advance your own study in more advanced statistical methods, you need a solid background in the fundamentals of linear modeling. This book attempts to give you that background, while facilitating your understanding using a conversational writing tone, minimizing the mathematics as much as possible, and focusing on application and implementation using statistical software.

Although our intention was to deliver an introductory treatment of regression analysis theory and application, we think even the seasoned researcher and user of regression analysis will find him- or herself learning something new in each chapter. Indeed, with repeated readings of this book we predict you will come to appreciate the glory of linear modeling just as we have, and maybe even develop the kind of passion for the topic that we developed and hope we have successfully conveyed to you.

Regression analysis is conducted with computer software, and you have many good programs to choose from. We emphasize three commercial packages that are heavily used in the social and behavioral sciences: IBM SPSS Statistics (referred to throughout the book simply as "SPSS"), SAS, and STATA. A fourth program, R, is given some treatment in one of the appendices. But this book is about the concepts and application of regression analysis and is not written as a how-to guide to using your software. We assume that you already have at least some exposure to one of these programs, some working experience entering and manipulating data, and perhaps a book on your program available or a local expert to guide you as needed. That said, we do provide relevant commands for each of these programs for the key analyses and uses of regression analysis presented in these pages, using different fonts and shades of gray to most clearly distinguish them from each other. Your program's reference manual or user's guide, or your course instructor, can help you fine-tune and tailor the commands we provide to extract other information from the analysis that you may need one day.

In this rest of this preface, we provide a nonexhaustive summary of the contents of the book, chapter by chapter, to give you a sense of what you can expect to learn about in the pages that follow.

Overview of the Book

Chapter 1 introduces the book by focusing on the concept of "accounting for something" when interpreting research results, and how a failure to account for various explanations for an association between two variables renders that association ambiguous in meaning and interpretation. Two examples are offered in this first chapter, where the relationship between two variables changes after accounting for the relationship between these two variables and a third—a *covariate*. These examples are used to introduce the concept of *statistical control*, which is a major theme of the book. We discuss how the linear model, as a general analytic framework, can be used to account for covariates in a flexible, versatile manner for many types of data problems that a researcher confronts.

Chapters 2 and 3 are perhaps the core of the book, and everything that follows builds on the material in these two chapters. Chapter 2 introduces the concept of a *conditional mean* and how the ordinary least squares criterion used in regression analysis for defining the best-fitting model yields a model of conditional means by minimizing the sum of the squared residuals. After illustrating some simple computations, which are then replicated using regression routines in SPSS, SAS, and STATA, distinctions are drawn between the correlation coefficient and the regression coefficient as

related measures of association sensitive to different things (such as scale of measurement and restriction in range). Because the residual plays such an important role in the derivation of measures of partial association in the next chapter, considerable attention is paid in Chapter 2 to the properties of residuals and how residuals are interpreted.

Chapter 3 lays the foundation for an understanding of statistical control by illustrating again (as in Chapter 1, but this time using all continuous variables) how a failure to account for covariates can lead to misleading results about the true relationship between an independent and dependent variable. Using this example, the partialing process is described, focusing on how the residuals in a regression analysis can be thought of as a new measure—a variable that has been cleansed of its relationships with the other variables in the model. We show how the partial regression coefficient as well as other measures of partial association, such as the partial and semipartial correlation, can be thought of as measures of association between residuals. After showing how these measures are constructed and interpreted without using multiple regression, we illustrate how multiple regression analysis yields these measures without the hassle of having to generate residuals yourself. Considerable attention is given in this chapter to the meaning and interpretation of various measures of partial association, including the sometimes confusing difference between the semipartial and partial correlation. Venn diagrams are introduced at this stage as useful heuristics for thinking about shared and partial association and keeping straight the distinction between semipartial and partial correlation.

In many books, you find the topic of statistical inference addressed first in the simple regression model, before additional regressors and measures of partial association are introduced. With this approach, much of the same material gets repeated when models with more than one predictor are illustrated later. Our approach in this book is different and manifested in Chapter 4. Rather than discussing inference in the single and multiple regressor case as separate inferential problems in Chapters 2 and 3, we introduce inference in Chapter 4 more generally for any model regardless of the number of variables in the model. There are at least two advantages to this approach of waiting until a bit later in the book to discuss inference. First, it allows us to emphasize the mechanics and theory of regression analysis in the first few chapters while staying purely in the realm of description of association between variables with or without statistical control. Only after these concepts have been introduced and the reader has developed some comfort with the ideas of regression analysis do we then add the burden that can go with the abstraction of generalization, populations, degrees of freedom, tolerance and collinearity, and so forth. Second, with this approach, we need to cover the theory and mechanics of inference

only once, noting that a model with only a single regressor is just a special case of the more general theory and mathematics of statistical inference in regression analysis.

We return to the uses and theory of multiple regression in Chapter 5, first by showing that a dichotomous regressor can be used in a model and that, when used alone, the result is a model equivalent to the independent groups *t*-test with which readers are likely familiar. But unlike the independent groups *t*-test, additional variables are easily added to a regression model when the goal is to compare groups when holding one or more covariates constant (variables that can be dichotomous or numerical in any combination). We also discuss the phenomenon of regression to the mean, how regression analysis handles it, and the advantages of regression analysis using pretest measurements rather than difference scores when a variable is measured more than once and interest is in change over time. Also addressed in this chapter are measures and inference about partial association for sets of variables. This topic is particularly important later in the book, where an understanding of variable sets is critical to understanding how to form inferences about the effect of multicategorical variables on a dependent variable as well as testing interaction between regressors.

In Chapter 6 we take a step away from the mechanics of regression analysis to address the general topic of cause and effect. Experimentation is seen by most researchers as the gold-standard design for research motivated by a desire to establish cause–effect relationships. But fans of experimentation don't always appreciate the limitations of the randomized experiment or the strengths of statistical control as an alternative. Ultimately, experimentation and statistical control have their own sets of strengths and weaknesses. We take the position in this chapter that statistical control through regression analysis and randomized experimentation complement each other rather than compete. Although data analysis can only go so far in establishing cause–effect, statistical control through regression analysis and the randomized experiment can be used in tandem to strengthen the claims that one can make about cause–effect from a data analysis. But when random assignment is not possible or the data are already collected using a different design, regression analysis gives a means for the researcher to entertain and rule out at least some explanations for an association that compete with a cause–effect interpretation.

Emphasis in the first six chapters is on the regression coefficient and its derivatives. Chapter 7 is dedicated to the use of regression analysis as a prediction system, where focus is less on the regression coefficients and more on the multiple correlation *R* and how accurately a model generates estimates of the dependent variable in currently available or future data. Though no doubt this use of regression analysis is less common, an understanding of the subtle and sometimes complex issues that come up when

using regression analysis to make predictions is important. In this chapter we make the distinction between how well a sample model predicts the dependent variable in the sample, how well the "population model" predicts the dependent variable in the population, and how well a sample model predicts the dependent variable in the population. The latter is quantified with *shrunken R*, and we discuss some ways of estimating it. We also address mechanical methods of model construction, best known as *stepwise regression*, including the pitfalls of relinquishing control of model construction to an algorithm. Even if you don't anticipate using regression analysis as a prediction system, the section in this chapter on predictor variable configurations is worth reading, because complementarity, redundancy, and suppression are phenomena that, though introduced here in the context of prediction, do have relevance when using regression for causal analysis as well.

Chapter 8 is on the topic of variable importance. Researchers have an understandable impulse to want to describe relationships in terms that convey in one way or another the *size* of the effect they have quantified. It is tempting to rely on rules of thumb circulating in the empirical literature and statistics books for what constitutes a small versus a big effect using concepts such as the proportion of variance that an independent variable explains in the dependent variable. But establishing the size of a variable's effect or its importance is far more complex than this. For example, small effects can be important, and big effects for variables that can't be manipulated or changed have limited applied value. Furthermore, as discussed in this chapter, there is reason to be skeptical of the use of squared measures of correlations, which researchers often use, as measures of effect size. In this chapter we describe various quantitative, value-free measures of effect size, including our attraction to the semipartial correlation relative to competitors such as the standardized regression coefficient. We also provide an overview of dominance analysis as an approach to ordering the contribution of variables in explaining variation in the dependent variable.

In Chapters 9 and 10 we address how to include multicategorical variables in a regression analysis. Chapter 9 focuses on the most common means of including a categorical variable with three or more categories in a regression model through the use of *indicator* or *dummy* coding. An important take-home message from this chapter is that regression analysis can duplicate anything that can be done with a traditional single-factor one-way ANOVA or ANCOVA. With the principles of interpretation of regression coefficients and inference mastered, the reader will expand his or her understanding in Chapter 10, where we cover other systems for coding groups, including Helmert, effect, and sequential coding. In both of these chapters we also discuss contrasts between means either with or without control, including pairwise comparisons between means and

more complex contrasts that can be represented as a linear combination of means.

In the classroom, we have found that after covering multicategorical regressors, students invariably bring up the so-called *multiple test problem*, because students who have been exposed to ANOVA prior to taking a regression course often learn about Type I error inflation in the context of comparing three or more means. So Chapter 11 discusses the multiple test problem, and we offer our perspective on it. We emphasize that the problem of multiple testing surfaces any time one conducts more than one hypothesis test, whether that is done in the context of comparing means or when using any linear model that is the topic of this book. Rather than describing a litany of approaches invented for pairwise comparisons between means, we focus almost exclusively on the Bonferroni method (and a few variants) as a simple, easy-to-use, and flexible approach. Although this method is conservative, we take the position that its advantages outweigh its conservatism most of the time. We also offer our own philosophy of the multiple test problem and discuss how one has to be thoughtful rather than mindless when deciding when and how to compensate for multiple hypothesis tests in the inference process. This includes contemplating such things as the logical independence of the hypotheses, how well established the research area is, and the interest value of various hypotheses being conducted.

By the time you get to Chapter 12, the versatility of linear regression analysis will be readily apparent. By the end of Chapter 12 on nonlinearity, any remaining doubters will be convinced. We show in this chapter how *linear* regression analysis can be used to model *nonlinear* relationships. We start with polynomial regression, which largely serves as a reminder to the reader what he or she probably learned in secondary school about *functions*. But once these old lessons are combined with the idea of minimizing residuals through the least squares criterion, it seems almost obvious that linear regression analysis can and should be able to model curves. We then describe linear spline regression, which is a means of connecting straight lines at joints so as to approximate complex curves that aren't always captured well by polynomials. With the principles of linear spline regression covered, we then merge polynomial and spline regression into polynomial spline regression, which allows the analyst to model very complex curvilinear relationships without ever leaving the comfort of a linear regression analysis program. Finally, it is in this chapter that we discuss various transformations, which have a variety of uses in regression analysis including making nonlinear relationships more linear, which can have its advantages in some circumstances.

Up to this point in the book, one variable's effect on a dependent variable, as expressed by a measure of partial association such as the partial regression coefficient, is fixed to be independent of any other regressor.

This changes in Chapters 13 and 14, where we discuss *interaction*, also called *moderation*. Chapter 13 introduces the fundamentals by illustrating the flexibility that can be added to a regression model by including a cross-product of two variables in a model. Doing so allows one variable's effect—the focal predictor—to be a linear function of a second variable—the moderator. We show how this approach can be used with focal predictors and moderators that are numerical, dichotomous, or multicategorical in any combination. In Chapter 14 we formalize the linear nature of the relationship between focal predictor and moderator and how a function can be constructed, allowing you to estimate one variable's effect on the dependent variable, knowing the value of the moderator. We also address the exercise of *probing* an interaction and discuss a variety of approaches, including the appealing but less widely known Johnson–Neyman technique. We end this section by discussing various complications and myths in the study and analysis of interactions, including how nonlinearity and interaction can masquerade as each other, and why a valid test for interaction does not require that variables be centered before a cross-product term is computed, although centering may improve the interpretation of the coefficients of the linear terms in the cross-product.

Moderation is easily confused with *mediation*, the topic of Chapter 15. Whereas moderation focuses on estimating and understanding the boundary conditions or contingencies of an effect—when an effect exists and when it is large versus small—mediation addresses the question how an effect operates. Using regression analysis, we illustrate how one variable's effect in a regression model can be partitioned into direct and indirect components. The indirect effect of a variable quantifies the result of a causal chain of events in which an independent variable is presumed to affect an intermediate *mediator* variable, which in turn affects the dependent variable. We describe the regression algebra of path analysis first in a simple model with only a single mediator before extending it to more complex models involving more than one mediator. After discussing inference about direct and indirect effects, we dedicate considerable space to various controversies and extensions of mediation analysis, including cause–effect, models with multicategorical independent variables, nonlinear effects, and combining moderation and mediation analysis.

Under the topic of "irregularities," Chapter 16 is dedicated to regression diagnostics and testing regression assumptions. Some may feel these important topics are placed later in the sequence of chapters than they should be, but our decision was deliberate. We feel it is important to focus on the general concepts, uses, and remarkable flexibility of regression analysis before worrying about the things that can go wrong. In this chapter we describe various diagnostic statistics—measures of *leverage*, *distance*, and *influence*—that analysts can use to find problems in their data

or analysis (such as clerical errors in data entry) and identify cases that might be causing distortions or other difficulties in the analysis, whether they take the form of violating assumptions or producing results that are markedly different than they would be if the case were excluded from the analysis entirely. We also describe the assumptions of regression analysis more formally than we have elsewhere and offer some approaches to testing the assumptions, as well as alternative methods one can employ if one is worried about the effects of assumption violations.

Chapters 17 and 18 close the book by addressing various additional complexities and problems not addressed in Chapter 16, as well as numerous extensions of linear regression analysis. Chapter 17 focuses on power and precision of estimation. Though we do not dedicate space to how to conduct a power analysis (whole books on this topic exist, as does software to do the computations), we do dissect the formula for the standard error of a regression coefficient and describe the factors that influence its size. This shows the reader how to increase power when necessary. Also in Chapter 17 is the topic of measurement error and the effects it has on power and the validity of a hypothesis test, as well as a discussion of other miscellaneous problems such as missing data, collinearity and singularity, and rounding error. Chapter 18 closes the book with an introduction to logistic regression, which is the natural next step in one's learning about linear models. After this brief introduction to modeling dichotomous dependent variables, we point the reader to resources where one can learn about other extensions to the linear model, such as models of ordinal or count dependent variables, time series and survival analysis, structural equation modeling, and multilevel modeling.

Appendices aren't usually much worth discussing in the precis of a book such as this, but other than Appendix C, which contains various obligatory statistical tables, a few of ours are worthy of mention. Although all the analyses can be described in this book with regression analysis and in a few cases perhaps a bit of hand computation, Appendix A describes and documents the RLM macro for SPSS and SAS written for this book and referenced in a few places elsewhere in the book that makes some of the analyses considerably easier. RLM is not intended to replace your preferred program's regression routine, though it can do many ordinary regression functions. But RLM has some features not found in software off the shelf that facilitates some of the computations required for estimating and probing interactions, implementing the Johnson–Neyman technique, dominance analysis, linear spline regression, and the Bonferroni correction to the largest *t*-residual for testing regression assumptions, among a few other things. RLM can be downloaded from this book's web page at *www.afhayes.com*. Appendix B is for more advanced readers who are interested in the matrix algebra behind basic regression computations. Finally, Appendix D

addresses regression analysis with R, a freely available open-source computing platform that has been growing in popularity. Though this quick introduction will not make you an expert on regression analysis with R, it should get you started and position you for additional reading about R on your own.

To the Instructor

Instructors will find that our precis above combined with the Contents provides a thorough overview of the topics we cover in this book. But we highlight some of its strengths and unique features below:

- Repeated references to syntax for regression analysis in three statistical packages: SPSS, SAS, and STATA. Introduction of the R statistical language for regression analysis in an appendix.
- Introduction of regression through the concept of statistical control of covariates, including discussions of the relative advantages of statistical and experimental control in section 1.1 and Chapter 6.
- Differences between simple regression and correlation coefficients in their uses and properties; see section 2.3.
- When to use partial, semipartial, and simple correlations, or standardized and unstandardized regression coefficients; see sections 3.3 and 3.4.
- Is collinearity really a serious problem? See section 4.7.1.
- Truly understanding regression to the mean; see section 5.2.
- Using regression for prediction. Why the familiar "adjusted" multiple correlation overestimates the accuracy of a sample regression equation; see section 7.2.
- When should a mechanical regression prediction replace expert judgment in making decisions about real people? See sections 7.1 and 7.5.
- Assessing the relative importance of the variables in a model; see Chapter 8.
- Should correlations be squared when assessing relative importance? See section 8.2.
- Sequential, Helmert, and effect coding for multicategorical variables; see Chapter 10.
- A different view of the multiple test problem. Why should we correct for some tests, but not correct for all tests in the entire history of science? See Chapter 11.
- Fitting curves with polynomial, spline, and polynomial spline regression; see Chapter 12.
- Advanced techniques for probing interactions; see Chapter 14.

Acknowledgments

Writing a book is a team effort, and many have contributed in one way or another to this one, including various reviewers, students, colleagues, and family members. C. Deborah Laughton, Seymour Weingarten, Judith Grauman, Katherine Sommer, Jeannie Tang, Martin Coleman, and others at The Guilford Press have been professional and supportive at various phases while also cheering us on. They make book writing enjoyable and worth doing often. Amanda Montoya and Cindy Gunthrie provided editing and readability advice and offered a reader's perspective that helped to improve the book. Todd Little, the editor of Guilford's Methodology in the Social Sciences series, was an enthusiastic supporter of this book from the very beginning. Scott C. Roesch and Chris Oshima reviewed the manuscript prior to publication and made various suggestions, most of which we incorporated into the final draft. And our families, and in particular our wives, Betsy and Carole, deserve much credit for their support and also tolerating the divided attention that often comes with writing a book of any kind, but especially one of this size and scope.

RICHARD B. DARLINGTON
Ithaca, New York

ANDREW F. HAYES
Columbus, Ohio

1

Statistical Control and Linear Models

Researchers routinely ask questions about the relationship between an independent variable and a dependent variable in a research study. In experimental studies, relationships observed between a manipulated independent variable and a measured dependent variable are fairly easy to interpret. But in many studies, experimental control in the form of random assignment is not possible. Absent experimental or some form of procedural control, relationships between variables can be difficult to interpret but can be made more interpretable through *statistical control*. After discussing the need for statistical control, this chapter overviews the linear model—widely used throughout the social sciences, health and medical fields, business and marketing, and countless other disciplines. Linear modeling has many uses, among them being a means of implementing statistical control.

1.1 Statistical Control

1.1.1 The Need for Control

If you have ever described a piece of research to a friend, it was probably not very long before you were asked a question like "But did the researchers account for this?" If the research found a difference between the average salaries of men and women in a particular industry, did it account for differences in years of employment? If the research found differences among several ethnic groups in attitudes toward social welfare spending, did it account for income differences among the groups? If the research found that males who hold relatively higher-status jobs are seen as less physically attractive by females than are males in lower-status jobs, did it account for age differences among men who differ in status?

All these studies concern the relationship between an *independent variable* and a *dependent variable*. The study on salary differences concerns the

relationship between the independent variable of sex and the dependent variable of salary. The study on welfare spending concerns the relationship between the independent variable of ethnicity and the dependent variable of attitude. The study on perceived male attractiveness concerns the relationship between the independent variable of status and the dependent variable of perceived attractiveness. In each case, there is a need to account for, in some way, a third variable; this third variable is called a *covariate*. The covariates for the three studies are, respectively, years of employment, income, and age.

Suppose you wanted to study these three relationships without worrying about covariates. You may be familiar with three very different statistical methods for analyzing these three problems. You may have studied the *t*-test for testing questions like the sex difference in salaries, analysis of variance (also known as "ANOVA") for questions like the difference in average attitude among several ethnic groups, and the Pearson or rank-order correlation for questions like the relationship between status and perceived attractiveness. These three methods are all similar in that they can all be used to test the relationship between an independent variable and a dependent variable; they differ primarily in the type of independent variable used. For sex differences in salary you could use the *t*-test because the independent variable—sex—is *dichotomous*; there are two categories—male and female. In the example on welfare spending, you could use analysis of variance because the independent variable of ethnicity is *multicategorical*, since there are several categories rather than just two—the various ethnic groups in the study. You could use a correlation coefficient for the example about perceived attractiveness because status is *numerical*—a more or less continuous dimension from high status to low status. But for our purposes, the differences among these three variable types are relatively minor. You should begin thinking of problems like these as basically similar, as this book presents the *linear model* as a single method that can be applied to all of these problems and many others with fairly minor variations in the method.

1.1.2 Five Methods of Control

The layperson's notion of "accounting for" something in a study is a colloquial expression for what scientists refer to as *controlling for* that something. Suppose you want to know whether driver training courses help students pass driving tests. One problem is that the students who take a driver training course may differ in some way before taking the course from those

who do not take the course. If that thing they differ on is related to test performance, then any differences in test performance may be due to that thing rather than the training course itself. This needs to be accounted for or "controlled" in some fashion in order to determine whether the course helps students pass the test. Or perhaps in a particular town, some testers may be easier than others. The driving schools may know which testers are easiest and encourage their students to take their tests when they know those testers are on duty. So the standards being used to evaluate a student driver during the test may be systematically different for students who take the driver training course relative to those who do not. This also needs to be controlled in some fashion.

You might control the problem caused by preexisting difference between those who do and do not take the course by using a list of applicants for driving courses, randomly choosing which of the applicants is allowed to take the course, and using the rejected applicants as the control group. That way you know that students are likely to be equal on all things that might be related to performance on the test before the course begins. This is *random assignment on the independent variable*. Or, if you find that more women take the course than men, you might construct a sample that is half female and half male for both the trained and untrained groups by discarding some of the women in the available data. This is control by *exclusion of cases*.

You might control the problem of differential testing standards by training testers to make them apply uniform evaluation standards; that would be *manipulation of covariates*. Or you might control that problem by randomly altering the schedule different testers work, so that nobody would know which testers are on duty at a particular moment. That would not be random assignment on the independent variable, since you have not determined which applicants take the course; rather, it would be *other types of randomization*. This includes randomly assigning which of two or more forms of the dependent variable you use, choosing stimuli from a population of stimuli (e.g., in a psycholinguistics study, all common English adjectives), and manipulating the order of presentation of stimuli.

All these methods except exclusion of cases are types of *experimental control* since they all require you to manipulate the situation in some way rather than merely observe it. But these methods are often impractical or impossible. For instance, you might not be allowed to decide which students take the driving course or to train testers or alter their schedules. Or, if a covariate is worker seniority, as in one of our earlier examples, you cannot manipulate the covariate by telling workers how long to keep

their jobs. In the same example, the independent variable is sex, and you cannot randomly decide that a particular worker will be male or female the way you can decide whether the worker will be in the experimental or control condition of an experiment. Even when experimental control is possible, the very exertion of control often intrudes the investigator into the situation in a way that disturbs participants or alters results; ethologists and anthropologists are especially sensitive to such issues. Experimental control may be difficult even in laboratory studies on animals. Researchers may not be able to control how long a rat looks at a stimulus, but they are able to measure looking time.

Control by exclusion of cases avoids these difficulties, because you are manipulating data rather than participants. But this method lowers sample size, and thus lowers the precision of estimates and the power of hypothesis tests.

A fifth method of controlling covariates—statistical control—is one of the main topics of this book. It avoids the disadvantages of the previous four methods. No manipulation of participants or conditions is required, and no data are excluded. Several terms mean the same thing: to control a covariate statistically means the same as to *adjust for* it or to *correct for* it, or to *hold constant* or to *partial out* the covariate.

Statistical control has limitations. Scientists may disagree on what variables need to be controlled—an investigator who has controlled age, income, and ethnicity may be criticized for failing to control education and family size. And because covariates must be measured to be controlled, they will be controlled inaccurately if they are measured inaccurately. We return to these and other problems in Chapters 6 and 17. But because control of some covariates is almost always needed, and because the other four methods of control are so limited, statistical control is widely recognized as one of the most important statistical tools in the empiricist's toolbox.

1.1.3 Examples of Statistical Control

The nature of statistical control can be illustrated by a simple fictitious example, though the precise methods used in this example are not those we emphasize later. In Holly City, 130 children attended a city-subsidized preschool program and 130 others did not. Later, all 260 children took a "school readiness test" on entering first grade. Of the 130 preschool children, only 60 scored above the median on the test; of the other 130 children, 70 scored above the median. In other words, the preschool children scored worse on the test than the others. These results are shown in the "Total"

TABLE 1.1. Test Scores, Socioeconomic Status, and Preschool Attendance in Holly City

	Raw frequencies								
	Middle-class			Working-class			Total		
	A	B	Total	A	B	Total	A	B	Total
Preschool	30	10	40	30	60	90	60	70	130
Other	60	30	90	10	30	40	70	60	130

TABLE 1.2. Socioeconomic Status and Preschool Attendance in Holly City

	Percentage scoring above the median		
	Middle-class	Working-class	Total
Preschool	75	33	46
Other	67	25	54

section of Table 1.1; A and B refer to scoring above and below the test median, respectively.

But when the children are divided into "middle-class" and "working-class," the results are as shown on the left and center of Table 1.1. We see that of the 40 middle-class children attending preschool, 30, or 75%, scored above the median. There were 90 middle-class children not attending preschool, and 60, or 67%, of them scored above the median. These values of 75 and 67% are shown on the left in Table 1.2. Similar calculations based on the working-class and total tables yield the other figures in Table 1.2. This table shows clearly that within each level of socioeconomic status (SES), the preschool children outperform the other children, even though they appear to do worse when you ignore socioeconomic status (SES). We have *held constant* or *controlled* or *partialed out* the covariate of SES.

When we perform a similar analysis for nearby Ivy City, we find the results in Table 1.3. When we inspect the total percentages, preschool appears to have a positive effect. But when we look within each SES group, no effect is found. Thus, the "total" tables overstate the effect of

preschool in Ivy City and understate it in Holly City. In these examples the independent variable is preschool attendance and the dependent variable is test score. In Holly City, we found a negative simple relationship between these two variables (those attending preschool scored lower on the test) but a positive *partial* relationship (a term more formally defined later) when SES was controlled. In Ivy City, we found a positive simple relationship but no partial relationship.

By examining the data more carefully, we can see what caused these paradoxical results, known as *Simpson's paradox* (for a discussion of this and related phenomena, see Tu, Gunnel, & Gilthorpe, 2008). In Holly City, the 130 children attending preschool included 90 working-class children and 40 middle-class children, so 69% of the preschool attenders were working-class. But the 130 nonpreschool children included 90 middle-class children and 40 working-class children, so this group was only 31% working-class. Thus, the test scores of the preschool group were lowered by the disproportionate number of working-class children in that group. This might have occurred if city-subsidized preschool programs had been established primarily in poorer neighborhoods. But in Ivy City this difference was in the opposite direction: The preschool group was 75% middle-class, while the nonpreschool group was only 25% middle-class; thus, the test scores of the preschool group were raised by the disproportionate number of middle-class children. This might have occurred if parents had to pay for their children to attend preschool. In both cities the effects of preschool were seen more clearly by controlling for or holding constant SES.

All three variables in this example were dichotomous—they had just two levels each. The independent variable of preschool attendance had two levels we called "preschool" and "other." The dependent variable of test score was dichotomized into those above and below the median. The covariate of SES was also dichotomized. Such dichotomization is rarely if ever something you would want do in practice (as discussed later in section 5.1.6). Fortunately, with the methods described in this book, such categorization is not necessary. Any or all of the variables in this problem could have been numerically scaled. Test scores might have ranged from 0 to 100, and SES might have been measured on a scale with very many points on a continuum. Even preschool attendance might have been numerical, such as if we measured the exact number of days each child had attended preschool. Changing some or all variables from dichotomous to numerical would change the details of the analysis, but in its underlying logic the problem would remain the same.

TABLE 1.3. Socioeconomic Status and Preschool Attendance in Ivy City

	Raw frequencies								
	Middle-class			Working-class			Total		
	A	B	Total	A	B	Total	A	B	Total
Preschool	90	30	120	10	30	40	100	60	160
Other	30	10	40	30	90	120	60	100	100

	Percentage scoring above the median		
	Middle-class	Working-class	Total
Preschool	75	25	62
Other	75	25	38

Consider now a problem in which the dependent variable is numerical. At Swamp College, the dean calculated that among professors and other instructional staff under 30 years of age, the average salary among males was $81,000 and the average salary among females was only $69,000. To see whether this difference might be attributed to different proportions of men and women who have completed the Ph.D., the dean made up the table given here as Table 1.4.

If the dean had hoped that different rates of completion of the Ph.D. would explain the $12,000 difference between men and women in average salary, that hope was frustrated. We see that men had completed the Ph.D. *less* often than women: 10 of 40 men, versus 15 of 30 women. The first column of the table shows that among instructors with a Ph.D., the mean difference in salaries between men and women is $15,000. The second column shows the same difference of $15,000 among instructors with no Ph.D. Therefore, in this artificial example, controlling for completion of the Ph.D. does not lower the difference between the mean salaries of men and women, but rather *raises* it from $12,000 to $15,000.

This example differs from the preschool example in its mechanical details; we are dealing with means rather than frequencies and proportions. But the underlying logic is the same. In the present case, the independent variable is sex, the dependent variable is salary, and the covariate is

TABLE 1.4. Average Salaries at Swamp College

	Ph.D. completed		
	Yes	No	Total
Men	\$90,000 $n = 10$	\$78,000 $n = 30$	\$81,000 $n = 40$
Women	\$75,000 $n = 15$	\$63,000 $n = 15$	\$69,000 $n = 30$

educational level. Again, the partial relationship differs from the simple relationship, though this time both the simple and partial relationships have the same sign, meaning that men make more than women, with or without controlling for education.

1.2 An Overview of Linear Models

The examples presented in section 1.1.3 are so simple that you may be wondering why a whole book is needed to discuss statistical control. But when the covariate is numerical, it may be that no two participants in a study have the same measurement on the covariate and so we cannot construct tables like those in the two earlier examples. And we may want to control many covariates at once; the dean might want to simultaneously control teaching ratings and other covariates as well as completion of the Ph.D. Also, we need methods for inference about partial relationships such as hypothesis testing procedures and confidence intervals. Linear modeling, the topic of this book, offers a means of accomplishing all of these things and many others.

This book presents the fundamentals of linear modeling in the form of *linear regression analysis*. A linear regression analysis yields a mathematical equation—a *linear model*—that estimates a dependent variable Y from a set of predictor variables or *regressors* X. Such a linear model in its most general form looks like

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e \tag{1.1}$$

Each regressor in a linear model is given a numerical weight—the b next to each X in equation 1.1—called its *regression coefficient*, *regression slope*, or simply its *regression weight* that determines how much the equation uses values on that variable to produce an estimate of Y. These regression weights are derived by an algorithm that produces a mathematical equation or *model* for Y that best fits the data, using some kind of criterion for defining "best." In this book, we focus on linear modeling using the *least squares* criterion.

Linear modeling has many uses, among them being the process of statistical control introduced conceptually in the prior section. Linear modeling is widely used throughout the behavioral sciences, medical research and public health, business and marketing, and countless other fields. It is safe to say that one really cannot progress far in one's development as a scientist without a solid understanding of linear modeling. Most universities offer at least one and typically several courses on linear regression analysis. Indeed, it is so important that many if not most academic departments whose faculty use the scientific method regularly offer their own version of a course on linear modeling in one form or another.

The basic linear model method imposes six requirements:

1. As in any statistical analysis, there must be a set of "participants," "cases," or "units." In most every example and application in this book, the data come from people, so we use the term "participant" frequently. But case, unit, and participant can be thought of as synonymous and we use all three of these terms.

2. Each of these participants must have values or measurements on two or more variables, each of which is numerical, dichotomous, or multicategorical. Thus, the raw data for the analysis form a rectangular data matrix with participants in the rows and variables in the columns.

3. Each variable must be represented by a single column of numbers. For instance, the dichotomy of sex can be represented by letting the number 1 represent male and 0 represent female, so that the sexes of 100 people could be represented by a column of 100 numbers, each 0 or 1. A multicategorical variable with, say, five categories can be represented by a column of numbers, each 1, 2, 3, 4, or 5. For both dichotomous and multicategorical variables, the numbers representing categories are mere codes and are arbitrary. They carry no meaning about quantity and can be exchanged with any other set

of numbers without changing the results of the analysis so long as proper coding methods are used. And of course a numerical variable such as age can be represented by a column of ages.

4. Each analysis must have just one dependent variable, though it may have several independent variables and several covariates.

5. The dependent variable must be numerical. A numerical variable is something like age or income with interval properties, such that values can be meaningfully averaged.

6. Statistical inference from linear models often requires several additional assumptions that are described elsewhere in this book, such as in section 4.1.2 and Chapter 16.

Within these conditions, linear models are flexible in many ways:

1. A variable might be a natural property of a participant, such as age or sex, or might be a property manipulated in an experiment, such as which of two or more experimental conditions into which the participant is placed through a random assignment procedure. Manipulated variables are typically categorical but may be numerical, such as the number of hours of practice at a task participants are given or the number of acts of violence on television a person is exposed to during an experiment.

2. You may choose to conduct a series of analyses from the same rectangular data matrix, and the same variable might be a dependent variable in one analysis and an independent variable or covariate in another. For instance, if the matrix includes the variables age, sex, years of education, and salary, one analysis may examine years of education as a function of age and sex, while another analysis examines salary as a function of age, sex, and education.

3. As explained more fully in section 3.1.2, the distinction between independent variables and covariates may be fuzzy since linear modeling programs make no distinction between the two. The program computes a measure of the relationship between the dependent variable and every other variable in the analysis while controlling statistically for all remaining variables, including both covariates and other independent variables. Independent variables are those whose relationship to the dependent variable you wish to discuss or are the focus of your study, while covariates are other variables you wish to

control or otherwise include in the model for some other purpose. Thus, the distinction between the two determines how you describe the results of the analysis but is not used in writing the computer commands that specify the analysis or the underlying mathematics.

4. Each independent variable or covariate may be dichotomous, multicategorical, or numerical. All three variable types may occur in the same problem. For instance, if we studied salary in a professional firm as a function of sex, ethnicity, and age while controlling for seniority, citizenship (American or not), and type of college degree (business, arts, engineering, etc.), we would have one independent variable and one covariate from each of the three scale types.

5. The independent variables and covariates may all be intercorrelated, as they are likely to be in all these examples. In fact, the need to control a covariate typically arises because it correlates with one or more independent variables or the dependent variable or both.

6. In addition to correlating with each other, the independent variables and covariates may *interact* in affecting the dependent variable. For instance, age or sex might have a larger or smaller effect on salary for American citizens than for noncitizens. Interaction is explained in detail in Chapters 13 and 14.

7. Despite the names "linear regression" and "linear model," these methods can easily be extended to a great variety of problems involving curvilinear relations between variables. For example, physical strength is curvilinearly related to age, peaking in the 20s. But a linear model could be used to study the relationship between age and strength or even to estimate the age at which strength peaks. We discuss how in Chapter 12.

8. The assumptions required for statistical inference are not extremely limiting. There are a number of ways around the limits imposed by those assumptions.

There are many statistical methods that are just linear models in disguise, or closely related to linear regression analysis. For example, ANOVA, which you may already be familiar with, can be thought of as a particular subset of linear models designed early in the 20th century, well before computers were around. Mostly this meant using only categorical independent variables, no covariates, and equal cell frequencies if there were two or

more independent variables. When a problem does meet the narrow requirements of ANOVA, linear models and analysis of variance give the same answers. Thus, ANOVA is just a special subset of the linear model method. As shown in various locations throughout this book, ANOVA, *t*-tests on differences between means, tests on Pearson correlations—things you likely have already been exposed to—can all be thought of as special simple cases of the general linear model, and can all be executed with a program that can estimate a linear model.

Logistic regression, probit regression, and multilevel modeling are close relatives of linear regression analysis. In logistic and probit regression, the dependent variable can be dichotomous or ordinal, such as whether a person succeeds or fails at a task, acts or does not act in a particular way in some situation, or dislikes, feels neutral, or likes a stimulus. Multilevel modeling is used when the data exhibit a "nested" structure, such as when different subsets of the participants in a study share something such as the neighborhood or housing development they live in or the building in a city they work in. But you cannot fruitfully study these methods until you have mastered linear models, since a great many concepts used in these methods are introduced in connection with linear models.

1.2.1 What You Should Know Already

This book assumes a working familiarity with the concepts of means and standard deviations, correlation coefficients, distributions, samples and populations, random sampling, sampling distributions, standardized variables, null hypotheses, standard errors, statistical significance, power, confidence intervals, one-tailed and two-tailed tests, summation, subscripts, and similar basic statistical terms and concepts. It refers occasionally to basic statistical methods including *t*-tests, ANOVA, and factorial analysis of variance. It is not assumed that you remember the mechanics of these methods in detail, but some sections of this book will be easier if you understand the uses of these methods.

1.2.2 Statistical Software for Linear Modeling and Statistical Control

In most research applications, statistical control is undertaken not by looking at simple association in subsets of the data, as in the two examples presented earlier, but through *mathematical equating* or *partialing*. This process is conducted automatically through linear regression analysis and will

be described starting in Chapter 3. Suffice it to say now that statistical control is usually accomplished by computer software. Only the simplest linear models are practical without the aid of a computer.

Fortunately, most statistical packages that researchers have access to in one way or another include routines that conduct linear regression analysis. There are many statistical packages that can conduct regression analysis; examples include SPSS, SAS, SYSTAT, Minitab, and STATA, and most are available for Windows and MacOS. These are all commercial programs and can be quite expensive. Fortunately most universities purchase licenses for one or more of these programs that provide free or low-cost access to its faculty, staff, and students. Over the last decade, a freely available statistical language and program called R has become quite popular. It also has procedures built in that conduct the kind of analyses described in this book. R can be downloaded at no charge from *www.r-project.org*.

This book is about the principles of linear modeling, not about using software that implements the methods we describe. These principles are not software specific. We assume you already have some working familiarity with at least one statistics program capable of doing the types of analyses described in this book. In many chapters we include code for SPSS, SAS, or STATA that generates output pertinent to the analyses described. In Appendix B we offer a brief primer on the use of R for linear regression analysis. We chose to emphasize SPSS, SAS, and STATA because these programs are arguably most readily available and widely used by researchers in the social sciences, medical and health fields, business and marketing, and elsewhere. But you will not become an expert on the use of any of these programs by reading this book. It is no substitute for the documentation, a book dedicated to specific software packages, or a local expert who can guide you on its use.

SAS and R require the user to write *syntax* or *code* instructing the software which analysis is desired, which variables are playing the roles of independent and dependent variable, what options to produce in output, and so forth. SPSS is often chosen by beginners or adopted by instructors of introductory statistics classes because it has a friendly menu-based interface that allows the user to select various analyses and options by pointing and clicking on the screen rather than by typing instructions. STATA has a similar interface, though most users instruct STATA using code. For consistency, and because we believe that ultimately researchers need to be familiar with how to write code for their chosen program, we offer SPSS syntax rather than point-and-click instructions. Consult a local SPSS expert

for guidance on how to type in and execute syntax in SPSS if you are not already familiar with this.

A convention we follow in this book is to use different text colors and background for code corresponding to different programs. For SPSS code, we use white text in a black box. For example, SPSS code to produce a scatterplot, such as the one found in the beginning of Chapter 2, would appear in this book as

```
graph/scatterplot plays with points.
```

For SAS, the code will be set in black text in a white box. Thus, the corresponding SAS code to produce this scatterplot would look like

```
proc sgscatter data=golf;plot points*plays;run;
```

STATA code will appear as white text in a gray box. So corresponding STATA code would would appear as

```
twoway (scatter points plays)
```

Data files we use are archived on the web page for this book at *www.afhayes.com* in the native data file formats of SPSS and STATA as well as text files, along with code to produce corresponding data files in SAS. Throughout this book, when we provide computer instructions, we assume you already have a data file available for analysis and know how to open or generate data files in your chosen software. Therefore, we do not provide code or instructions for how to do so. We refer to data files by name using CAPITAL letters. Variables in those data files we refer to in the text using an *`italicized courier`* font. When we otherwise refer to computer code within the body of the text of this book rather than set in boxes as described above, we use the **`boldface courier`** font.

1.2.3 About Formulas

If you glance through this book, you will see many algebraic formulas. If formulas frighten you, relax. You can master the material in this book without memorizing any formulas. Most of the formulas in this book, or variations closely related to these formulas, are applied by computer programs. They are provided here merely so you will know what the program is doing. Many other formulas are for relatively uncommon problems. Still other formulas are so simple that they merely express concepts you can easily put into words. For instance, in Chapter 2 we define a deviation

score x_i as the difference between a raw score X_i and the sample mean $\overline{X}$, and another formula defines a variance Var(X) as the mean of the squared deviation scores. In this book there are only a few formulas that you can expect to ever have to actually apply to your own data by hand, without the assistance of a computer. That said, understanding the formulas is sometimes a good way of learning what the formulas represent conceptually as well as mathematically.

1.2.4 On Symbolic Representations

If you were to lay three or four statistical methods books side by side and open to chapters on a common topic, you would find a considerable lack of consistency in the symbols the authors use to refer to the same concepts. Although there would be some overlap—for instance, the use of the Greek letter mu (μ) to refer to a population mean and the Roman letter r as a reference to a correlation coefficient are both nearly universal—it would be hard to generalize your learning from one book to another if you relied entirely on symbolic representations of ideas used by one author. And two people who learn statistics from different books, who are taught by different instructors who use particular (and perhaps idiosyncratic) symbols to communicate ideas, or who come from different fields may appear to both outsiders and insiders to be speaking different languages even when talking about the same thing.

The Greek letter "beta" (β) and Roman letter b or B are examples. Some use β to refer to a population regression weight, whereas others use this symbol to refer to a sample regression weight. Still others might use β when talking about a standardized regression weight. SPSS labels some regression coefficients in its output with the word "beta" and others with B, and it is not uncommon to hear people talk about the "beta weights" or simply the "betas" in a regression model. You might be asked by someone if you aren't clear in a presentation whether you are presenting "bees or betas" from your model. The questioner may be asking whether you are reporting standardized or unstandardized regression coefficients, though others in the audience may not understand the meaning of this question as phrased if they were trained elsewhere or used a different book. Others restrict the use of Roman letters such as b or B to refer to estimates from a sample and reserve Greek symbols for parameters. But not all Greek letters refer to parameters in scientific discourse. For instance, people commonly report Cronbach's α calculated in a sample of participants who filled out

some kind of measurement instrument used to measure personality or attitude.

Ultimately, symbols are arbitrary. We can use whatever symbols we want to communicate ideas so long as we communicate what those symbols refer to when using them for the first time. Because there are relatively few conventions in the literature and books on linear modeling, we do not attempt to follow any of them. We introduce various symbols we use along the way—symbols that may at times be idiosyncratic to this book—but we always communicate what those symbols refer to unless there is a very strong convention in existence. As a reader, your job is to avoid assuming that one symbol that we use has the same meaning as this symbol when used by others. Such an assumption will inevitably result in confusion in your mind at some point.

1.3 Chapter Summary

Association between two variables X and Y can be difficult to interpret or obscured when a third variable Z is related to both X and Y. Researchers have a variety of procedural tricks they can employ to deal with such circumstances either prior to or following data collection, such as random assignment to X or other forms of randomization, case deletion, and maintaining strict control over various aspects of the research design and its administration. When none of these are possible, as is often the case, statistical control is an option and can render relationships easier to interpret and less susceptible to competing explanations. Linear modeling is one of the more frequently used procedures for statistical control by behavioral scientists, business and marketing researchers, investigators in health and medical fields, and many other disciplines. This book introduces the linear model in the form of linear regression analysis not only as a means of implementation of statistical control but also as a general and flexible tool that can be used for a variety of data-analytic tasks.

Guilford Publications
370 Seventh Avenue, Suite 1200
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com