Preface

Our Purpose

This book is intended to be used for a one-semester course in "applied regression" that focuses on concepts and applications to be taught to juniors, seniors, or first-year graduate students. Its prerequisites are: (1) one course in statistical methods, (2) proficiency in high school mathematics, and (3) some familiarity with computers. Of course, it is always useful for students to have more mathematics, statistics, or computing than the minimal prerequisites.

Our Approach

There are two kinds of statistics books—theory books and methods books. Theory books generally start with a discussion of probability and populations (probability distributions) and then discuss sampling, whereas books on statistical methods tend to start directly with data and do not discuss populations thoroughly. We believe that the theoretical approach is sound and that high school mathematics is sufficient to understand populations as they relate to many applications. Consequently we first introduce populations and then discuss sampling, focusing all the while on the concepts underlying statistical inference.

Our Emphasis

We have endeavored to give an accurate and clear account of linear regression with an emphasis on *prediction.* We have deemphasized some traditional concepts such as correlation and tests of hypotheses and significance. We stress standard deviations of prediction errors (when using unbiased predictors), rather

than correlations, to judge the adequacy of regression functions, and we use confidence intervals in place of tests. This change in emphasis is more than superficial because it influences the way practical questions are formulated and answered.

Use of Finite Population Models

Although most of the *mathematical theory* of regression is derived using infinite population models, we feel that it is easier for students to understand many concepts—such as random sampling, unbiasedness, sampling distributions, and the relative frequency interpretation of confidence intervals and tests—by using *finite population* models. Moreover, most populations we come across in real problems happen to be finite populations. For this reason, we attempt to explain important concepts using finite populations. On the other hand, the theory based on infinite models is essentially valid for topics covered in this book because real populations under study usually consist of a large number of items, and so infinite population models serve as good approximations.

The Organization

Chapter 1 is a review of what is normally covered in a first course in statistical methods plus some material about matrices, functional notation, and the multivariate Gaussian (normal) population. We suggest that Chapter 1 not be omitted even if students have had one or more courses in statistics because it includes some notation, terminology, and specific prerequisites for the rest of the book. Chapter 2 introduces many of the fundamental concepts underlying regression. Chapter 3 contains a detailed discussion of straight line regression. Chapter 4 gives an in-depth study of multiple linear regression. Chapter 5 deals with diagnostic procedures in regression analysis. Chapters 6 and 7 treat special topics that cover several important practical applications of linear regression. Chapter 8 discusses some procedures, including a distribution-free method, that are valid when traditional assumptions for straight line regression are not appropriate. Finally, Chapter 9 gives a brief discussion of nonlinear regression.

The first four chapters form the core of the book. Thereafter, any or all of the sections in Chapters 5 through 9 can be studied in any order since each section in these chapters depends on only the material in Chapters 1 through 4.

Computing

It is difficult, and practically speaking impossible, to do all of the computing required in multiple regression without the use of some software computing package, yet we did not want details about computing software or hardware to interfere with the flow of the material. Thus we decided to avoid discussion of

computing issues in the textbook and to make laboratory manuals available where computing is discussed. The textbook is self-contained and is not dependent on knowledge of any computing package. We include relevant outputs from MINITAB and/or SAS whenever the required computing is, to the say the least, tedious. As a result, students can learn the material without any statistical computing package, although we strongly advise using a suitable one.

We have written two laboratory manuals—one for MINITAB and one for SAS. We chose MINITAB because it is an easy statistical package for students to learn without detracting from the main subject. We also chose SAS because many graduate students know and use this package or want to learn it because it is widely used in government, industry, and business. The laboratory manuals contain instructions for using MINITAB and SAS to solve problems and exercises in the textbook. This arrangement offers students the following choices:

1 For those who wish to use MINITAB, a laboratory manual explaining MINITAB commands needed for regression is available with the textbook.
2 For those who wish to use SAS, a laboratory manual explaining basic SAS commands needed for regression is similarly available with the textbook.
3 For those who do not wish to use computers, the textbook is self-sufficient and they need not concern themselves with the laboratory assignments contained in the laboratory manuals.

The Data Sets

Practically every data set used in the book appears "real," and many are based on real studies. The problem descriptions associated with the data sets in the book do not require knowledge of any particular field, such as physics, chemistry, biology, or engineering. The data sets involve common and well-known topics such as insurance premiums, grade point averages, final and midterm exams, electric bills, grocery costs, strength of plastic containers, professors' salaries, and atmospheric pollutants.

Examples, Problems, and Exercises

In the examples, problems, and exercises we have aimed to ask and answer questions that investigators may encounter in the course of an investigation and not questions that are of interest to statisticians only. In particular, we have tried to explain how to formulate practical questions about population parameters in the statistical model being considered.

There are more than 100 examples (over half of them worked out in detail), and there are 14 tasks and more than 85 data sets throughout the book to illustrate each new procedure as it is presented. Problems at the end of each section help students determine whether they understand the material covered in that section. Exercises at the end of each chapter cover the material in the entire chapter. Answers to selected problems and exercises are given in an appendix.

The Mathematics

We believe that students can learn, understand, and use the techniques of regression correctly and effectively in their applied work without a great deal of mathematics and theoretical statistics. In many statistics courses, students prove theorems without really understanding the statistical concepts the theorems imply, and in some cases this can be an impediment to developing their intuition about statistical methods. Consequently this book requires very little mathematics and no theorems are proven. For many procedures, we give a heuristic explanation to help students understand the underlying ideas. Instructors who wish to teach theoretical (mathematical) regression techniques can supplement this book with appropriate mathematical procedures. This should be much easier than using a mathematical text and supplementing it with appropriate methodological procedures that include data sets, computing techniques, etc.

The Notation

Some students may find the notation difficult and troublesome, but adequate notation is required to understand many of the concepts and procedures. We have simplified the necessary notation as much as possible, and we believe that it will actually help clarify many of the concepts if students study it carefully.

Recommended Coverage

One way to cover the material in this book for a one-semester course that meets three hours per week is as follows:

1 The material in Chapter 1 can be covered rather quickly but deliberately because it is supposed to be a review.

2 Chapter 2, which is an introduction to the concept of regression, can also be covered rather quickly.

3 Chapters 3 and 4 should be covered in depth. However, the instructor may elect to spend very little time on *regression analysis when there are measurement errors* in Section 3.10 and *lack-of-fit analysis* in Section 4.11.

4 After Chapters 1 through 4 have been completed, instructors can teach any section in the remainder of the book. Some instructors may elect to cover the topics in one of the ways described below:

- **a** Selection of variables, growth curves, and nonlinear regression, along with other topics.
- **b** Selection of variables, tolerance intervals, and prediction intervals, etc.
- **c** Selection of variables, nonlinear regression, spline regression, etc.

Noteworthy Features

Some additional noteworthy features of the book are:

1. An exceptional range of applications is presented in Chapters 6 and 7—tolerance intervals, calibration, regulation, intersection of lines, variable selection, growth curves, maximum and minimum of a quadratic, spline regression, and many others.
2. For each and every technique discussed, we stress the assumptions needed for that technique to be valid.
3. We point out that no valid point estimates or confidence intervals exist for certain population parameters unless sampling is carried out by a prescribed method. This relationship between the sampling method and the availability of valid inference procedures is often ignored in many textbooks.
4. We downplay tests of hypotheses and tests of significance but emphasize the use of confidence intervals for making practical decisions from data because confidence intervals are more informative than tests for making these decisions. Tests are used by investigators to assess "statistical significance," but confidence intervals can be used to help assess "practical importance."
5. We use word problems, called *tasks*, to help students see how regression can be used to answer practical questions.
6. We provide an alternative approach for assessing lack-of-fit of prediction models using confidence intervals rather than traditional tests.
7. We stress the fact that correlation must be used with caution in real problems and that it is not an adequate measure of regression functions.
8. We use standard deviations, rather than variances, because standard deviations are easier to interpret in applied problems.
9. All the data sets used in the book are available on a data disk that accompanies each laboratory manual.
10. In the laboratory manuals we show how MINITAB and SAS can be used to analyze the data sets and solve problems and exercises in the textbook. For every topic in regression that is discussed in the book, we discuss a computer program that can be used for computations.
11. For problems that cannot be solved using the built-in commands in MINITAB or SAS, we have written programs for MINITAB and SAS. These programs are called *macros*, and the laboratory manuals explain how to use them. These macros are available on the data disk. They are intended for only the problems discussed in this book, and they are not necessarily valid for other, more complex, problems.
12. We make recommendations as to which procedures should be used in applied problems.
13. We have included several conversations between a statistician and an investigator to help clarify certain concepts.

It is our hope that this book will serve the needs of students of statistics and researchers who wish to have an in-depth understanding of the concepts underlying regression analysis and alternative ways of formulating questions of interest in practical applications.

To Jeanne, our children and grandchildren
for the good times. F. G.

To my grandparents, my parents, all my teachers, and to
Pam, Matthew, Kristin, Kevin, and Geoffrey. H. I.

Far better an approximate answer to the *right* question, which is often vague, than an exact answer to the *wrong* question, which can always be made precise."

John W. Tukey, *Ann. Math. Stat.*, vol. 33, p. 13, 1962.