Preface

"Please, sir, I'd like some more."
— Oliver Twist

This book introduces students to statistical modeling beyond what they learn in an introductory course. We assume that students have successfully completed a Stat 101 college course or an AP Statistics course. Building on basic concepts and methods learned in that course, we empower students to analyze richer datasets that include more variables and address a broader range of research questions.

Guiding Principles

Principles that have guided the development of this book include:

- **Modeling as a unifying theme**. Students will analyze many types of data structures with a wide variety of purposes throughout this course. These purposes include making predictions, understanding relationships, and assessing differences. The data structures include various numbers of variables and different kinds of variables in both explanatory and response roles. The unifying theme that connects all of these data structures and analysis purposes is *statistical modeling*. The idea of constructing statistical models is introduced at the very beginning, in a setting that students encountered in their Stat 101 course. This modeling focus continues throughout the course as students encounter new and increasingly more complicated scenarios.

  Basic principles of statistical modeling that apply in all settings, such as the importance of checking model conditions by analyzing residuals with graphical and numerical, are emphasized throughout. Although it's not feasible in this course to prepare students for all possible contingencies that they might encounter when fitting models, we want students to recognize when a model has substantial faults. Throughout the book, we offer two general approaches for analyzing data when model conditions are not satisfied: data transformations and computer-intensive methods such as bootstrapping and randomization tests.

  Students will go beyond their Stat 101 experience by learning to develop and apply models with both quantitative and categorical response variables, with both quantitative and categorical explanatory variables, and with multiple explanatory variables.

- **Modeling as an interactive process**. Students will discover that the practice of statistical modeling involves applying an interactive process. We employ a four-step process in all statistical modeling: *Choose* a form for the model, *fit* the model to the data, *assess* how well the model describes the data, and *use* the model to address the question of interest.

  As students gain more and more facility with the interplay between data and models, they will find that this modeling process is not as linear as it might appear. They will learn how to apply their developing judgment about statistical modeling. This development of judgment, and the growing realization that statistical modeling is as much an art as a science, are more ways in which this second course is likely to differ from students' Stat 101 experiences.

- **Modeling of real, rich datasets**. Students will encounter real and rich datasets throughout this course. Analyzing and drawing conclusions from real data are crucial for preparing students to use statistical modeling in their professional lives. Using real data to address genuine research questions also helps to motivate students to study statistics. The richness stems not only from interesting contexts in a variety of disciplines, but also from the multivariable nature of most datasets.

  This multivariable dimension is an important aspect of how this course builds on what students learned in Stat 101 and prepares them to analyze data that they will see in our modern world that is so permeated with data.

Prerequisites

We assume that students using this book have successfully completed an introductory statistics course (Stat 101), including statistical inference for comparing two proportions and for comparing two means. No further mathematical prerequisites are needed to learn the material in this book. Some material on data transformations and logistic regression assumes that students are able to understand and work with exponential and logarithmic functions.

Overlap with Stat 101

We recognize that Stat 101 courses differ with regard to coverage of topics, so we expect that students come to this course with different backgrounds and levels of experience. We also realize that having studied material in Stat 101 does not ensure that students have mastered or can readily use those ideas in a second course. To help all students make a smooth transition to this course, we recommend introducing the idea of statistical modeling while presenting some material that students are likely to have studied in their first course. Chapter 0 reminds students of basic statistical terminology and also uses the familiar two-sample t-test as a way to illustrate the approach of specifying, estimating, and testing a statistical model. Chapters 1 and 2 lead students through specification, fit, assessment, and inference for simple linear models with a single quantitative predictor. Some topics in these chapters (for example, inference for the slope of a regression line) may be familiar to students from their first course, but most likely not in the more formal setting of a linear model that we present here. A thorough introduction of the formal linear model and related ideas in the "simple" one-predictor setting makes it easier to move to datasets with multiple

predictors in Chapter 3. For a class of students with strong backgrounds, an instructor may choose to move more quickly through the first chapters, treating that material mostly as review to help students get "up to speed."

Organization of Chapters

After completing this course, students should be able to work with statistical models where the response variable is either quantitative or categorical and where explanatory/predictor variables are quantitative or categorical (or with both kinds of predictors). Chapters are grouped to consider models based on the type of response and type of predictors.

Chapter 0: Introduction. We remind students about basic statistical terminology and present our four-step process for constructing statistical models in the context of a two-sample t-test.

Unit A (Chapters 1–4): Linear regression models. These four chapters develop and examine statistical models for a quantitative response variable, first with one quantitative predictor and then with multiple predictors of both quantitative and categorical types.

Unit B (Chapters 5–8): Analysis of variance models. These four chapters also consider models for a quantitative response variable, but specifically with categorical explanatory variables/factors. We start with a single factor (one-way ANOVA) and then move to models that consider multiple factors. We follow this with an overview of experimental design issues.

Unit C (Chapters 9–11): Logistic regression models. These three chapters introduce models for a binary response variable with either quantitative or categorical predictors.

These three units follow a similar structure:

- Each unit begins by considering the "simple" case with a single predictor/factor (Chapters 1–2 for Unit A, 5 for Unit B, 9 for Unit C). This helps students become familiar with the basic ideas for that type of model (linear regression, analysis of variance, or logistic regression) in a relatively straightforward setting where graphical visualizations are most feasible.
- The next chapter of the unit (Chapters 3, 6, 10) extends these ideas to models with multiple predictors/factors.
- Each unit then presents a chapter of additional topics that extend ideas discussed earlier (Chapters 4, 7, 11). For example, Section 1.5 gives a brief and informal introduction to outliers and influential points in linear regression models. Topic 4.3 covers these ideas in more depth, introducing more formal methods to measure leverage and influence and to detect outliers. The topics in these chapters are relatively independent and so allow for considerable flexibility in choosing among the additional topics.
- Unit B also has a chapter providing an overview of experimental design issues (Chapter 8).

Flexibility within and between Units

The units and chapters are arranged to promote flexibility regarding order and depth in which topics are covered. Within a unit, some instructors may choose to "splice" in an additional topic when related ideas are first introduced. For example, Section 5.4 in the first ANOVA chapter introduces techniques for conducting pairwise comparisons with one-way ANOVA using Fisher's LSD method. Instructors who prefer a more thorough discussion of pairwise comparison issues at this point, including alternate techniques such as the Bonferroni adjustment or Tukey's HSD method, can proceed to present those ideas from Section 7.2. Other instructors might want to move immediately to two-way ANOVA in Chapter 6 and then study pairwise procedures later.

Instructors can also adjust the order of topics between the units. For example, some might prefer to consider logistic regression models (Unit C) before studying ANOVA models (Unit B). Others might choose to study all three types of models in the "simple setting" (Chapters 1–2, 5, 9), and then return to consider each type of model with multiple predictors. One could also move to the ANOVA material in Unit B directly after starting with a "review" of the two-sample t-test for means in Unit 0, then proceed to the material on regression.

Technology

Modern statistical software is essential for doing statistical modeling. We assume that students will use statistical software for fitting and assessing the statistical models presented in this book. We include output from both Minitab and R throughout the book, but we do not include specific software commands or instructions. Our goal is to allow students to focus on understanding statistical concepts, developing facility with statistical modeling, and interpreting statistical output while reading the text. Toward these ends, we want to avoid the distractions that often arise when discussing or implementing specific software instructions. This choice allows instructors to use other statistical software packages (e.g., SAS, SPSS, DataDesk, JMP, etc.).

Exercises

Developing skills of statistical modeling requires considerable practice working with real data. Homework exercises are an important component of this book. Exercises appear at the end of each chapter, except for the "Additional Topics" chapters that have exercises after each independent topic. These exercises are grouped into four categories:

- **Conceptual exercises.** These questions are brief and require minimal (if any) calculations. They give students practice with applying basic terminology and assess students' understanding of concepts introduced in the chapter.
- **Guided exercises.** These exercises ask students to perform various stages of a modeling analysis process by providing specific prompts for the individual steps.
- **Open-ended exercises.** These exercises ask for more complete analyses and reporting of conclusions, without much or any step-by-step direction.
- **Supplemental exercises.** Topics for these exercises go somewhat beyond the scope of the material covered in the chapter.

To the Student

In your introductory statistics course you saw many facets of statistics but you probably did little if any work with the formal concept of a statistical model. To us, modeling is a very important part of statistics. In this book, we develop statistical models, building on ideas you encountered in your introductory course. We start by reviewing some topics from Stat 101 but adding the lens of modeling as a way to view ideas. Then we expand our view as we develop more complicated models.

You will find a thread running through the book:

- Choose a type of model.
- Fit the model to data.
- Assess the fit and make any needed changes.
- Use the fitted model to understand the data and the population from which they came.

We hope that the Choose, Fit, Assess, Use quartet helps you develop a systematic approach to analyzing data.

Modern statistical modeling involves quite a bit of computing. Fortunately, good software exists that enables flexible model fitting and easy comparisons of competing models. We hope that by the end of your Stat2 course, you will be comfortable using software to fit models that allow for deep understanding of complex problems.

Stat2 Book Companion Web site at www.whfreeman.com/stat2 provides a range of resources.
Available for instructors only:
- Instructor's Manual
- Instructor's Solutions Manual
- Sample Tests and Quizzes
- Lecture PowerPoint Slides

Available for students:
- Datasets (in Excel, Minitab, R, .csv, and .txt formats)

Each new copy of Stat2 is packaged with an access code students can use to access premium resources via the Book Companion Web site. These resources include:
- Student Solutions Manual
- R Companion Manual
- Minitab Companion Manual

Acknowledgments

We are grateful for the assistance of a great number of people in writing Stat2.

First, we thank all the reviewers and classroom testers listed at the end of this section. This group of people gave us valuable advice, without which we would have not progressed far from early drafts of our book.