Chapter 1
Introduction

The book describes a family of statistical techniques that we call *multipredictor* regression modeling. This family is useful in situations where there are multiple measured factors (also called predictors, covariates, or independent variables) to be related to a single outcome (also called the response or dependent variable). The applications of these techniques are diverse, including those where we are interested in prediction, isolating the effect of a single predictor, or understanding multiple predictors. We begin with an example.

1.1 Example: Treatment of Back Pain

Korff et al. (1994) studied the success of various approaches to treatment for back pain. Some physicians treat back pain more aggressively, with prescription pain medication and extended bed rest, while others recommend an earlier resumption of activity and manage pain with over-the-counter medications. The investigators classified the aggressiveness of a sample of 44 physicians in treating back pain as low, medium, or high, and then followed 1,071 of their back pain patients for two years. In the analysis, the classification of treatment aggressiveness was related to patient outcomes, including cost, activity limitation, pain intensity, and time to resumption of full activity.

The primary focus of the study was on a single categorical predictor, the aggressiveness of treatment. Thus for a continuous outcome like cost, we might think of an analysis of variance (ANOVA), while for a categorical outcome we might consider a contingency table analysis and a χ^2-test. However, these simple analyses would be incorrect at the very least because they would fail to recognize that multiple patients were *clustered* within physician practice and that there were *repeated outcome measures* on patients.

Looking beyond the clustering and repeated measures (which are covered in Chap. 7), what if physicians with more aggressive approaches to back pain also

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_1,

tended to have older patients? If older patients recover more slowly (regardless of treatment), then even if differences in treatment aggressiveness have no effect, the age imbalance would nonetheless make for poorer outcomes in the patients of physicians in the high-aggressiveness category. Hence, it would be misleading to judge the effect of treatment aggressiveness without correcting for the imbalances between the physician groups in patient age and, potentially, other prognostic factors—that is, to judge without *controlling for confounding*. This can be accomplished using a model which relates study outcomes to age and other prognostic factors as well as the aggressiveness of treatment. In a sense, multipredictor regression analysis allows us to examine the effect of treatment aggressiveness while *holding the other factors constant*.

1.2 The Family of Multipredictor Regression Methods

Multipredictor regression modeling is a family of methods for relating multiple predictors to an outcome, with each member of the family suitable for a different type of outcome. The cost outcome, for example, is a numerical measure and for our purposes can be taken as *continuous*. This outcome could be analyzed using the linear regression model, though we also show in Chap. 8 why a *generalized linear model* (GLM) might be a better choice.

Perhaps the simplest outcome in the back pain study is the yes/no indicator of moderate-to-severe activity limitation; a subject's activities are limited by back pain or not. Such a categorical variable is termed *binary* because it can only take on two values. This type of outcome is analyzed using the logistic regression model, presented in Chap. 5.

In contrast, pain intensity was measured on a scale of ten equally spaced values. The variable is numerical and could be treated as continuous, although there were many tied values. Alternatively, it could be analyzed as a categorical variable, with the different values treated as ordered categories, using the proportional-odds or continuation-ratio models, both extensions of the logistic model and briefly covered in Chap. 5.

Another potential outcome might be time to resumption of full activity. This variable is also continuous, but what if a patient had not yet resumed full activity at the end of the follow-up period of two years? Then the time to resumption of full activity would only be known to exceed two years. When outcomes are known only to be greater than a given value (like two years), the variable is said to be *right-censored*—a common feature of time-to-event data. This type of outcome can be analyzed using the Cox proportional hazards model, the primary topic of Chap. 6.

Furthermore, in the back pain example, study outcomes were measured on groups, or clusters, of patients with the same physician, and on multiple occasions for each patient. To analyze such *hierarchical* or *longitudinal* outcomes, we need to use extensions of the basic family of regression modeling techniques suitable for


repeated measures data, described in Chap. 7. Related extensions are also required to analyze data from complex surveys, briefly covered in Chap. 12.

The various regression modeling approaches, while differing in important statistical details, also share important similarities. Numeric, binary, and categorical predictors are accommodated by all members of the family, and are handled in a similar way: on some scale, the systematic part of the outcome is modeled as a linear function of the predictor values and corresponding *regression coefficients*. The different techniques all yield estimates of these coefficients that summarize the results of the analysis and have important statistical properties in common. This leads to unified methods for selecting predictors and modeling their effects, as well as for making inferences to the population represented in the sample. Finally, all the models can be applied to the same broad classes of practical questions involving multiple predictors.

1.3 Motivation for Multipredictor Regression

Multipredictor regression can be a powerful tool for addressing three important practical questions. These questions, which provide the framework for our discussion of predictor selection in Chap. 10, include *prediction, isolating the effect of a single predictor,* and *understanding multiple predictors*.

1.3.1 Prediction

How can we identify which patients with back pain will have moderate-to-severe limitation of activity? Multipredictor regression is a powerful and general tool for using multiple measured predictors to make useful predictions for future observations. In this example, the outcome is binary and thus a multipredictor logistic regression model could be used to estimate the predicted probability of limitation for any possible combination of the observed predictors. These estimates could then be used to classify patients as likely to experience limitation or not. Similarly, if our interest was future costs, a continuous variable, we could use a linear regression model to predict the costs associated with new observations characterized by various values of the predictors. In developing models for this purpose, we need to avoid *over-fitting*, and to *validate* their predictiveness in actual practice.

1.3.2 Isolating the Effect of a Single Predictor

In settings where multiple, related predictors contribute to study outcomes, it will be important to consider multiple predictors even when a single predictor is of interest. In the von Korff study, the primary predictor of interest was how

aggressively a physician treated back pain. But incorporation of other predictors was necessary to minimize *confounding*, so that we could plausibly consider a causal interpretation of the estimated effects of the aggressiveness of treatment. Estimating causal effects from observational data is difficult, and sometimes requires special methods, including *potential outcomes estimation* and *propensity scores*. These approaches depend on the assumption that there are no unmeasured confounders. Causal estimation using *instrumental variables* depends on different but equally stringent assumptions. We consider these specialized methods in Chap. 9.

1.3.3 Understanding Multiple Predictors

Multipredictor regression can also be used when our aim is to identify multiple independent predictors of a study outcome—independent in the sense that they appear to have an effect over and above other measured variables. Especially in this context, we may need to consider other complexities of how predictors jointly influence the outcome. For example, the effect of injuries on activity limitation may in part operate through their effect on pain; in this view, pain *mediates* the effect of injury and should not be adjusted for, at least initially. Alternatively, suppose that among patients with mild or moderate pain, younger age predicts more rapid recovery, but among those with severe pain, age makes little difference. The effects of both age and pain severity will both potentially be misrepresented if this *interaction* is not taken into account. Fortunately, all the multipredictor regression methods discussed in this book easily handle interactions, as well as mediation and confounding, using essentially identical techniques. Though certainly not foolproof, multipredictor models are well suited to examining the complexities of how multiple predictors are associated with an outcome of interest.

1.4 Guide to the Book

This text attempts to provide practical guidance for regression analysis. We interweave real data examples from the biomedical literature in the hope of capturing the reader's interest and making the statistics as easy to grasp as possible. Theoretical details are kept to a minimum, since it is usually not necessary to understand the theory to use these methods appropriately. We avoid formulas and keep mathematical notation to a minimum, instead emphasizing selection of appropriate methods and careful interpretation of the results.

This book grew out a two-quarter sequence in multipredictor methods for physicians beginning a career in clinical research, with a focus on techniques appropriate to their research projects. For these students, mathematical explication is an ineffective way to teach these methods. Hence our reliance on real-world examples and heuristic explanations.

Our students take the course in the second quarter of their research training. A beginning course in biostatistics is assumed and some understanding of epidemiologic concepts is clearly helpful. However, Chap. 3 presents a review of topics from a first biostatistics course, and we explain epidemiologic concepts in some detail throughout the book.

Although theoretical details are minimized, we do discuss techniques of practical utility that some would consider advanced. We treat extensions of basic multipredictor methods for repeated measures and hierarchical data, for data arising from complex surveys, and for the broader class of *generalized linear models*, of which logistic regression is the most familiar example. In addition, we consider alternative approaches to estimating the causal effects of an exposure or treatment from observational data, including *propensity scores* and *instrumental variables*. We address model checking as well as model selection in considerable detail, including specialized methods for avoiding over-fitting in selecting prediction models. And we consider how missing data arise, and the conditions under which maximum likelihood methods for repeated measures as well as multiple imputation of the missing values can successfully deal with it.

The orientation of this book is to *parametric* methods, in which the systematic part of the model is a simple function of the predictors, and substantial assumptions are made about the distribution of the outcome. In our view, parametric methods are usually flexible and robust enough, and we show how model adequacy can be checked. The Cox proportional hazards model covered in Chap. 6 is a *semiparametric* method which makes few assumptions about an important component of the systematic part of the model, but retains most of the efficiency and many of the advantages of fully parametric models. *Generalized additive models*, briefly reviewed in Chap. 5, go an additional step in this direction. However, fully *nonparametric* regression methods in our view entail losses in efficiency and ease of interpretation which make them less useful to researchers. We do recommend a popular bivariate nonparametric regression method, LOWESS, but only for exploratory data analysis.

Our approach is also to encourage exploratory data analysis as well as thoughtful interpretation of results. We discourage focusing solely on P-values, which have an important place in statistics but also important limitations. In particular, P-values measure the strength of the evidence for an effect, but not its size. Furthermore, they can be misleading when data-driven model selection has been carried out. In our view, data analysis profits from considering the estimated effects, using confidence intervals (CIs) to quantify their precision. In prediction problems, P-values are a poor guide to *prediction error*, the proper focus of interest, and over-reliance of them can lead to over-fitting.

We recommend that readers begin with Chap. 2, on exploratory methods. Since Chap. 3 is largely a review, students may want to focus only on unfamiliar material. Chapter 4, on multipredictor regression methods for continuous outcomes, introduces most of the important themes of the book, which are then revisited in later chapters, and so is essential reading. Similarly, Chap. 9 covers causal inference, Chap. 10 addresses predictor selection, and Chap. 11 deals with missing data, all

topics common to the entire family of regression techniques. Chapters 5 and 6 cover regression methods specialized for binary and time-to-event outcomes, while Chaps. 7, 8, and 12 cover extensions of these methods for repeated measures, counts, and other special types of outcomes, and complex surveys. Readers may want to study these chapters as the need arises. Finally, Chap. 13 reprises the themes considered in the earlier chapters and is recommended for all readers.

For interested readers, Stata code and selected datasets used in examples and problems, plus errata, are posted on the website for this book:

http://www.biostat.ucsf.edu/vgsm