

Chapter 13

Summary

13.1 Introduction

Our goal in writing this book was to provide researchers and students with a practical guide to the analysis of data from research studies focusing on the relationship between outcomes and multiple predictor variables. Through our experience as coinvestigators and instructors at the University of California, San Francisco, we have observed that students and researchers from many fields can benefit greatly from being able to conduct their own data analyses. Mastering these skills promotes better study designs, clearer and more informative papers and presentations, and more focused and productive interactions with professional statisticians concerning more advanced topics.

Despite the fundamentally mathematical foundations of statistics, the prerequisites needed to acquire adequate data analysis skills are surprisingly nontechnical. Perhaps the most important one is critical thinking. As is true with many technical fields, the key ideas underlying the methods presented here become much clearer when applied in actual data analyses. All of them are characterized by a common structure that mirrors the majority of research questions arising in clinical research: the relationship between an outcome and measured explanatory variables.

In this chapter, we provide a brief review of the general approach to data analysis developed in this book, and provide guidance on how to use it as a resource to address particular analytical issues. We also briefly discuss a number of topics relevant to investigators undertaking their own data analyses, including development of analysis plans and finding help with technical questions. Finally, we discuss briefly some advanced topics that are not covered extensively in this book, and represent areas of current research that are relevant to many modern applications of regression methods.

13.2 Selecting Appropriate Statistical Methods

Selection of the right statistical tool to apply in addressing a research question is not always easy. Despite a number of unsuccessful attempts to use concepts from artificial intelligence in the development of algorithms to automate this process, common sense and experience remain most important for choosing an appropriate analysis method. In this section, we provide some general guidelines on selecting statistical methods, with references to appropriate chapters and sections in the book. In keeping with our overall theme, we assume that the research question and available data involve investigating the relationship between a specified outcome and one or multiple measured predictor variables.

The first step in most data analyses is to define clearly the candidate outcome and predictor variable(s) and choose an appropriate analytic approach. As described in Sect. 1.1, outcomes can generally be classified as being either numeric (e.g., measured characteristics such as cholesterol level or body weight) or categorical (e.g., disease status indicators). Table 13.1 uses this classification to distinguish the main types of outcomes considered in the book (that subsume the majority considered in health research applications), along with the standard regression approaches for each, and the chapters in which they are discussed. Clearly many outcomes do not fit cleanly into the categories provided in the table. For example, the severity score in the back pain example introduced in Chap. 1 could be considered as either continuous or as a categorical variable with ordinal categories. In many such cases, the decision of how to consider such variables for the purpose of analysis will be driven by practicality (e.g., available software) and/or convention. In cases where multiple approaches are available, it is often a good idea to try more than one to insure that results are not sensitive to the choice.

Although the type of outcome usually dictates the choice of which regression model to consider, further consideration of how the outcome is observed and measured is necessary before settling on an analysis approach. A fundamental consideration is whether individual outcomes can be viewed as independent or not. Examples of studies with independent outcomes include diagnosis of CHD in participants in the WCGS study (used for examples in Chaps. 2–5) and baseline glucose levels in women participating in the HERS study (Sect. 4.2). Dependence between outcomes can arise in a number of ways detailed in Chap. 7. These include repeated measures of outcomes measured in the same individuals, or outcomes

Table 13.1 Outcome, regression model, and chapter reference

Outcome classification	Outcome type	Regression model	Chapter reference
Numerical	Continuous	Linear	4
	Count	Poisson model	8
	Time-to-event	Proportional hazards	6
Categorical	Binary	Logistic	5
	Ordinal	Proportional odds	5
	Nominal	Polytomous logistic	5

on different individuals that are associated via a shared environment or genetic relationship (e.g., disease outcomes among members of the same family). Examples include repeated measures of fat content of feces (Sect. 7.1) and birthweights of first- and last-born infants from the same mothers (Sect. 7.3). As described in Chap. 7, most of the regression approaches for independent outcomes have direct analogs applicable in the dependent outcome setting.

In addition to dependence between individual outcomes, it is also important to consider how individuals were selected for inclusion in the study being analyzed. Although for many studies, it is reasonable to assume that study participants in a defined population had equal chances of being selected, in some cases these chances are controlled by the investigator to obtain a sample with desired properties. Examples include case-control studies for binary outcomes and complex sample surveys. As illustrated in Sect. 5.3 and Chap. 12, regression methods for such studies generally involve minor modifications of techniques applicable for independent samples.

Finally, we want to stress that despite the large number of outcome types and corresponding approaches to regression modeling covered here, the tools used for model fitting and evaluation are quite similar in most cases. Key concepts and techniques in model construction and interpretation such as accounting for confounding, mediation, and interaction and non-linearity are shared across approaches as well. Experience with regression modeling for different types of outcomes and study designs will surely reinforce these points.

13.3 Planning and Executing a Data Analysis

Data analyses are usually complex and benefit from careful planning in order to proceed in a timely and organized fashion. In our experience, few analyses are limited to straightforward application of textbook procedures. Invariably, technical questions arise related to data structure and/or quality, application of particular techniques, use of software programs, and interpretation of results. In this section, we provide some advice on several topics related to conducting an efficient analysis.

13.3.1 *Analysis Plans*

Before beginning a data analysis, it is useful to formulate a plan for how the work will proceed. For randomized controlled trials, analysis plans are generally specified in advance by the study protocol. For observational and clinical studies, preliminary plans are often formulated at the proposal stage. However, even when existing plans are not available to guide analyses, a clear outline of the important issues and tasks

can aid in organizing the process. A detailed plan should include a summary of the study design, statements of the research hypotheses, descriptions of each stage of analysis, and clear procedures for record-keeping, data distribution, and security.

13.3.2 Choice of Software

Fortunately, there are a number of excellent software packages available that implement the majority of techniques discussed here. Although we have used Stata in our examples, SAS, S-PLUS, and SPSS all provide commercial alternatives that offer many of the same facilities and run on a variety of computer platforms and operating systems. Also, the R language for statistical computing and graphics ([R Development Core Team 2004](#)) is freely available and includes most of the procedures presented here. Finally, there are a number of special-purpose programs providing methods not well-represented in the major packages, including StatXact and LogXact (exact inference for contingency tables and logistic regression), and SUDAAN (analysis of data from complex surveys). Frequently, multiple programs will be used for a given analysis. For example, SAS may be used in preparation of analysis data sets, and specific analyses conducted in Stata or R. Fortunately, there are programs such as StatTransfer that translate data sets between common formats used by different analysis packages, preserving important features such as variable labels and formats.

13.3.3 Data Preparation

Perhaps the single most time consuming phase of any data analysis is preparation of analysis-ready data sets from source data. Source data frequently reside in relational databases or proprietary formats and must be exported and re-formatted for specific analyses. Since particular analytic procedures rely on specific data structures and variable definitions, sufficient time and resources should be allocated for proper preparation and checking of analysis data sets prior to conducting statistical analyses.

13.3.4 Record Keeping and Reproducibility of Results

An important part of a complete data analysis includes keeping files of relevant commands and procedures used in each of the stages above. Adding comments and explanatory text to programs and keeping text files outlining the analysis procedures

and cataloging the important files are very useful in this regard. This information should be kept in an identifiable place (preferably organized with other project-specific materials) and backed up in a secure location for disaster recovery.

Because a typical data analysis involves a large number of steps, having all files necessary to recreate results from source data can save work for revision of research publications, and is critical in demonstrating that the results are reproducible. The merits of making this material, including source data, public are a topic of current debate in the scientific literature. See [Sedransk et al. \(2010\)](#) for a discussion of relevant issues from the statistician's perspective.

13.3.5 Data Security

Records from research studies often contain sensitive patient information and must be protected from unauthorized access. Although studies generally have data security measures in place to protect primary data sources, data analyses often involve creation of multiple datasets that may be distributed between investigators. As a general rule, it is a good practice to keep analysis datasets physically separate from source data, with any variables that can be linked to participant identities removed. Make sure that all analysis and data distribution procedures conform to current government, institutional, and study-specific guidelines on data security and protected health information.

13.3.6 Consulting a Statistician

As we have noted frequently in the text, there are many instances where analysis issues arise that do not fall in the neat categories typical of many of the examples. Complex sampling schemes, extensive missing data, unusual patterns of censoring, misclassification in measured outcomes and predictors, causal inferences in longitudinal observational studies subject to time-dependent confounding—all are examples of situations where standard methods and attendant assumptions may not apply without modification. Being able to recognize these circumstances is an important step in addressing these issues. When faced with an analysis problem that appears to fall outside of the range of techniques covered here, having access to a professional statistician is a valuable resource. For investigators at research institutions, the best way to insure the availability of sound statistical support is to include a statistician as a consultant or coinvestigator in proposals. Participating in courses or workshops on specialized statistical methods is another way to gain access to expert advice on advanced topics.

13.3.7 Use of Internet Resources

The Internet provides a vast and very valuable resource to assist in selection of statistical methods and planning data analyses. Frequently, answers to questions about particular applications and methods can quickly be found via a search using one of the available Web search engines. Unfortunately, even judicious searches often yield too many results to review completely. Also, the relevance of returned results is frequently influenced by factors completely unrelated to their scientific value. For these reasons, beginning with searches of established research resources such as the PubMed interface to the MEDLINE index and the Current Index to Statistics will often yield more focused searches. Many educational institutions and private companies provide free online access to electronic scientific journals. Also, statistical software sites frequently have online documentation and message lists that can provide useful information on the use of particular methods. Finally, message boards related to particular software programs and academic interests can frequently be a good way to get answers to analysis questions. Of course, unless the qualifications of individuals posting are known, blindly following advice can be dangerous.

13.4 Further Notes and References

13.4.1 Multiple Hypothesis Tests

The majority of the examples and applications considered in this book can be characterized by single outcome variables and their relationship to one or multiple predictors. While these are representative of many of the research questions that arise in epidemiological and medical research, we have largely ignored issues that arise when analyses include testing multiple hypotheses. These can arise in many contexts, including genomic studies that seek to identify important predictors of a primary disease outcome from a potentially very large pool of candidates, and in clinical studies investigating the effect of a treatment on multiple disease outcomes. The primary concern in these examples is the inflation of type-I error resulting from the occurrence of false-positive results arising from multiple hypothesis tests. Valid inferences in these situations generally involve adjustment of P -values from individual tests to control family-wise error rate (FER) to desired levels.

Consider a study of the use of gene expression data in the classification of two types of acute leukemia (myeloid and lymphoblastic) (Golub et al. 1999). RNA from bone marrow samples from 38 patients (27 lymphoblastic and 11 myeloid) was hybridized to oligonucleotide microarrays, each containing probes for 6,817 genes. The research questions centered on the use of genes as predictors for leukemia type. Although some form of binary regression model relating the disease outcome to predictors is clearly appropriate in this example, the fact that the number of

candidate predictors greatly outnumber the observations, and that the correlation between predictors may be quite complex (reflecting functional relationships between genes) raises a number of difficult computational and inferential issues. Clearly, an analysis that screened for candidate genes via independent hypothesis tests of each would potentially yield many false-positive results if the type-I error was fixed at the conventional 5% level.

Conventional procedures for controlling FER such as the Bonferroni correction outlined in Sects. 3.1.5 and 4.3.4 may be quite stringent in this example, resulting in significance levels that may rule out even associations of interest as potential false-positive results. These concerns have led to development of multiple testing procedures designed to control the *false discovery rate* (FDR), defined as the number of false-positives relative to the total number of positives, rather than focus solely on the former. In the example, the choice of an FDR of 5% implies that on average, 5% of genes selected as positively associated with the leukemia outcome would represent false-positive results. This approach generally results in improved power relative to procedures designed to minimize EER, at the expense of an increased likelihood of type-I errors. We refer readers to the seminal papers by [Benjamini and Hochberg \(1995\)](#) and [Storey \(2002\)](#) for further information about FDR procedures.

Multiple testing problems also arise in studies involving the effect of a predictor of interest on multiple outcomes. For example, randomized trials may consider more than one primary outcome in addition to a number of secondary outcomes. This is common in fields such as psychiatry, where treatments may influence a number of behavioral characteristics, many of which are related. Similar issues arise in *subgroup analyses*, which repeat the primary outcome in groups of individuals defined by enrollment characteristics in an effort to identify factors that may influence treatment efficacy. They are also a concern in safety analyses, in which rates of occurrence of adverse events are compared between arms. In all these situations, conventional hypothesis testing with no adjustment for multiple testing can lead to potentially misleading conclusions about results. Results from the Bonferroni adjustment in these situations is expected to be fairly conservative, both because it ignores correlations between outcomes and also gives equal weight to each. Alternative procedures tailored to prespecified ordering of hypotheses about primary and secondary endpoints are sometimes appropriate. These issues are discussed further in [Dmitrienko et al. \(2009\)](#) and [Piantadosi \(2005\)](#).

As discussed in Sect. 10.3, multiple comparison issues are also a concern in regression analyses targeting the relationship between an outcome and multiple predictors, where the primary goal is to identify important predictors and characterize their relationship to the outcome rather than construct a model that provides accurate outcome prediction. In this case, use of formal adjustment methods is debatable.

13.4.2 Statistical Learning

In Sect. 10.1, we considered the application of regression methods in developing clinical prediction models. These problems are typically characterized by using a

potentially large collection of predictor variables to develop a regression model for predicting individual patient outcomes with the aim of minimizing prediction error. Regression methods represent just one approach in a large class of *statistical learning* methods for such addressing such problems. Many of these methods are computationally intensive, and depart radically from the familiar additive linear structure familiar from the models presented here. We refer readers to [Hastie et al. \(2009\)](#) for a book-length overview of some modern approaches being applied in this area.