Statistics for Biology and Health

Regression Methods in Biostatistics

Linear, Logistic, Survival, and Repeated Measures Models

Second Edition

Springer

Statistics for Biology and Health

For further volumes:
http://www.springer.com/series/2848

Eric Vittinghoff • David V. Glidden
Stephen C. Shiboski • Charles E. McCulloch

Regression Methods in Biostatistics

Linear, Logistic, Survival, and Repeated Measures Models

Second edition

Springer

Eric Vittinghoff
Department of Epidemiology
and Biostatistics
University of California, San Francisco
Parnassas Ave. 500
94143 San Francisco California
MU-420 West
USA

David V. Glidden
Department of Epidemiology
and Biostatistics
University of California, San Francisco
Parnassas Ave. 500
94143 San Francisco California
MU-420 West
USA

Stephen C. Shiboski
Department of Epidemiology
and Biostatistics
University of California, San Francisco
Parnassas Ave. 500
94143 San Francisco California
MU-420 West
USA

Prof. Charles E. McCulloch
Department of Epidemiology
and Biostatistics
University of California, San Francisco
Berry 185
94107 San Francisco California
Suite 5700
USA

ISSN 1431-8776
ISBN 978-1-4614-1352-3 e-ISBN 978-1-4614-1353-0
DOI 10.1007/978-1-4614-1353-0
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011945441

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

For Rupert & Jean; Kay & Minerva;
Caroline, Erik & Hugo; and J.R.

Preface

In the second edition of *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, we have substantially revised and expanded the core chapters of the first edition, and added two new chapters. The first of these, Chap. 9, on strengthening causal inference, introduces potential outcomes, average causal effects, and two primary methods for estimating these effects, what we call *potential outcomes estimation* and inverse probability weighting. It also covers propensity scores in detail, then more briefly discusses time-dependent exposures, controlled and natural direct effects, instrumental variables, and principal stratification. The second, Chap. 11, on missing data, explains why this is a problem, classifies missingness by mechanism, and discusses the shortcomings of some simple approaches. Its focus is on three primary approaches for dealing with missing data: maximum likelihood estimation, multiple imputation, and inverse weighting, and lays out in detail when each of these approaches is most appropriate.

Among the core chapters of the first edition, Chap. 5, on logistic regression, has substantial new sections on models for ordinal and multinomial outcomes, as well as exact logistic regression. Chapter 6, on survival analysis, has an in-depth new section on competing risks, as well as new coverage of interval censoring and left truncation. Chapter 7, on repeated measures analysis, introduces recently developed methods for distinguishing between- and within-cluster effects, and for estimating the effects of fixed and time-dependent covariates (TDCs) on change. Chapter 8, on generalized linear models, adds coverage of negative binomial as well as zero-inflated and zero-truncated models for counts. Chapters 4–8 all now cover restricted cubic splines, take a new approach to mediation, and provide methods for sample size, power, and detectable effect calculation. Chapter 10, on predictor selection, has expanded coverage of developing and assessing models for prediction, as well as a new section on *directed acyclic graphs*. Our summary in Chap. 13 includes a new discussion of multiple comparisons and updated coverage of software packages. All Stata examples have been updated. As before, Stata, SAS, and Excel datasets and Stata do-files for most examples are provided on the website for the book, `http://www.biostat.ucsf.edu/vgsm`. We also posted implementations of analyses for time-dependent exposures too complicated for inclusion in the text.

At UCSF, we have used the first edition for a two-quarter course on regression methods for clinical researchers and epidemiologists, the first quarter covering linear and logistic models and predictor selection, and the second covering survival and repeated measures analysis. The new chapter on strengthening causal inference is the basis of new quarter-long course, and the new missing data chapter will play an important role in a more advanced quarter-long course next year. The new breadth of coverage of the second edition should make it more widely useful in year-long biostatistics courses for students like ours, MPH students, and for masters-level courses in biostatistics.

Finally, we gratefully acknowledge the very important contributions made by Professors Joseph Hogan of Brown University, Michael Hudgens of the University of North Carolina, Barbara McKnight of the University of Washington, and Maya Peterson of the University of California, Berkeley, who generously provided detailed, insightful reviews of the two new chapters. Any remaining errors and shortcomings are of course entirely ours.

San Francisco, CA, USA

Eric Vittinghoff
David V. Glidden
Stephen C. Shiboski
Charles E. McCulloch

Preface to the First Edition

The primary biostatistical tools in modern medical research are single-outcome, multiple-predictor methods: multiple linear regression for continuous outcomes, logistic regression for binary outcomes, and the Cox proportional hazards model for time-to-event outcomes. More recently, generalized linear models (GLMs) and regression methods for repeated outcomes have come into widespread use in the medical research literature. Applying these methods and interpreting the results require some introduction. However, introductory statistics courses have no time to spend on such topics and hence they are often relegated to a third or fourth course in a sequence. Books tend to have either very brief coverage or to be treatments of a single topic and more theoretical than the typical researcher wants or needs.

Our goal in writing this book was to provide an accessible introduction to multipredictor methods, emphasizing their proper use and interpretation. We feel strongly that this can only be accomplished by illustrating the techniques using a variety of real data sets. We have incorporated as little theory as feasible. Further, we have tried to keep the book relatively short and to the point. Our hope in doing so is that the important issues and similarities between the methods, rather than their differences, will come through. We hope this book will be attractive to medical researchers needing familiarity with these methods and to students studying statistics who would like to see them applied to real data. The methods we describe are, of course, the same as those used in a variety of fields, so non-medical readers will find this book useful if they can extrapolate from the predominantly medical examples.

A prerequisite for the book is a good first course in statistics or biostatistics or an understanding of the basic tools: paired and independent samples t-tests, simple linear regression and one-way analysis of variance (ANOVA), contingency tables and χ^2 (chi-square) analyses, Kaplan–Meier curves, and the logrank test.

We also think it is important for researchers to know how to interpret the output of a modern statistical package. Accordingly, we illustrate a number of the analyses with output from the Stata statistics package. There are a number of other packages that can perform these analyses, but we have chosen this one because of its accessibility and widespread use in biostatistics and epidemiology.

We begin the book with a chapter introducing our viewpoint and style of presentation and the big picture as to the use of multipredictor methods. Chapter 2 presents descriptive numerical and graphical techniques for multipredictor settings and emphasizes choice of technique based on the nature of the variables. Chapter 3 briefly reviews the statistical methods we consider prerequisites for the book.

We then make the transition in Chap. 4 to multipredictor regression methods, beginning with the linear regression model. This chapter also covers confounding, mediation, interaction, and model checking in the most detail. In Chap. 5, we turn to binary outcomes and the logistic model, noting the similarities to the linear model. Ties to simpler, contingency table methods are also noted. Chapter 6 covers survival outcomes, giving clear indications as to why such techniques are necessary, but again emphasizing similarities in model building and interpretation with the previous chapters. Chapter 7 looks at the accommodation of correlated data in both linear and logistic models. Chapter 8 extends Chap. 5, giving an overview of GLMs.

In the second edition, new sections of Chaps. 4–8 deal with pooled and exact logistic regression (Chap. 5), competing risks (Chap. 6), and time-varying predictors and separating between and within cluster information (Chap. 7). Chapters 4–8, also now conclude with short sections on calculating sample size, power, and minimum detectable effects.

The next three chapters, two of them new in the second edition, cover broader issues. Chapter 9 looks more closely at making causal inferences, using the models discussed in Chaps. 4–8, as well as alternatives including propensity scores and instrumental variables. Chapter 10 deals with predictor selection, with expanded treatment of methods for prediction problems. Chapter 11 considers missing data and methods for dealing with it, including maximum likelihood models, multiple imputation, and complete case analysis, the problematic default.

Finally, Chap. 12 is a brief introduction to the analysis of complex surveys. The text closes with a summary, Chap. 13, attempting to put each of the previous chapters in context. Too often it is hard to see the forest for the trees of each of the individual methods. Our goal in this final chapter is to provide guidance as to how to choose among the methods presented in the book and also to realize when they will not suffice and other techniques need to be considered.

San Francisco, CA, USA

Eric Vittinghoff
David V. Glidden
Stephen C. Shiboski
Charles E. McCulloch

Contents

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 *semi-parametric* 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

Chapter 2
Exploratory and Descriptive Methods

Before beginning any sort of statistical analysis, it is imperative to take a preliminary look at the data with three main goals in mind: first, to check for errors and anomalies; second, to understand the distribution of each of the variables on its own; and third, to begin to understand the nature and strength of relationships among variables. Errors should, of course, be corrected, since even a small percentage of erroneous data values can drastically influence the results. Understanding the distribution of the variables, especially the outcomes, is crucial to choosing the appropriate multipredictor regression method. Finally, understanding the nature and strength of relationships is the first step in building a more formal statistical model from which to draw conclusions.

2.1 Data Checking

Procedures for data checking should be implemented before data entry begins, to head off future headaches. Many data entry programs have the capability to screen for egregious errors, including values that are out the expected range or of the wrong "type." If this is not possible, then we recommend regular checking for data problems as the database is constructed.

Here are two examples we have encountered recently. First, some values of a variable defined as a proportion were inadvertently entered as percentages (i.e., 100 times larger than they should have been). Although they made up less than 3% of the values, the analysis was completely invalidated. Fortunately, this simple error was easily corrected once discovered. A second example involved patients with a heart anomaly. Those whose diagnostic score was poor enough (i.e., exceeded a numerical threshold) were to be classified according to the type of anomaly. Data checks revealed missing classifications for patients whose diagnostic score exceeded the

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_2,

threshold, as well as classifications for patients whose score did not, complicating planned analyses. Had the data been screened as they were collected, this problem with study procedures could have been avoided.

2.2 Types of Data

The proper description of data depends on the nature of the measurement. The key distinction for statistical analysis is between numerical and categorical variables. The number of diagnostic tests ordered is a numerical variable, while the gender of a person is categorical. Systolic blood pressure (SBP) is numerical, whereas the type of surgery is categorical.

A secondary but sometimes important distinction within numerical variables is whether the variable can take on a whole continuum or just a discrete set of values. So SBP would be continuous, while number of diagnostic tests ordered would be discrete. Cost of a hospitalization would be continuous, whereas number of mice able to successfully navigate a maze would be discrete. More generally,

> *Definition*: A numerical variable taking on a continuum of values is called *continuous* and one that only takes on a discrete set of values is called *discrete*.

A secondary distinction sometimes made with regard to categorical variables is whether the categories are ordered or unordered. So, for example, categories of annual household income (<\$20,000, \$20,000–\$40,000, \$40,000–\$100,000, >\$100,000) would be ordered, while marital status (single, married, divorced, widowed) would be unordered. More exactly,

> *Definition*: A categorical variable is *ordinal* if the categories can be logically ordered from smallest to largest in a sense meaningful for the question at hand (we need to rule out silly orders like alphabetical); otherwise it is unordered or *nominal*.

Some overlap between types is possible. For example, we may break a numerical variable (such as exact annual income in dollars and cents) into ranges or categories. Conversely, we may treat a categorical variable as a numerical score, for example, by assigning values one to five to the ordinal responses Poor, Fair, Good, Very Good, and Excellent.

Most of the analysis methods we will describe for numerical scores (e.g., linear regression or t-tests) have interpretations based on average scores. So assigning scores to a categorical variable is effective if average scores are readily interpretable. This may well be the case for scoring the categories Poor through Excellent as 1 through 5: an average value of 3.5 is between Good and Very Good. It might be a less effective strategy ordinal categorical variables such as the modified Rankin Scale, a scale used to assess disability following a stroke. For that scale, 0 represents no symptoms, 1 and 2 slight disability, 3 and 4 moderate disability, 5 severe disability, and 6 is dead. Consider two sets of three patients, the first set with scores of 0, 6, and 6 and the second with scores of 4, 4, and 4. Both have averages of 4, but the first set would generally be considered as having worse outcomes since two of the patients died. In such a case, summarizing with the average, and hence treating the variable as numeric, may not be appropriate.

In the following sections, we present each of the descriptive and exploratory methods according to the types of variables involved.

2.3 One-Variable Descriptions

We begin by describing techniques useful for examining a single variable at a time. These are useful for uncovering mistakes or extreme values in the data and for assessing distributional shape.

2.3.1 Numerical Variables

We can describe the distribution of numerical variables using either numerical or graphical techniques.

2.3.1.1 Example: Systolic Blood Pressure

The western collaborative group study (WCGS) was a large epidemiological study designed to investigate the association between the "type A" behavior pattern and coronary heart disease (CHD) (Rosenman et al. 1964). We will revisit this study later in the book, focusing on the primary outcome, but for now we want to explore the distribution of SBP.

2.3.1.2 Numerical Description

As a first step, we obtain basic descriptive statistics for SBP. Table 2.1 gives detailed summary statistics for the SBP variable, `sbp`. Several features of the output are worth consideration. The largest and smallest values should be scanned for outlying or incorrect values, and the mean (or median) and standard deviation should be assessed as general measures of the location and spread of the data. Secondary features are the skewness and kurtosis, though these are usually more easily assessed by the graphical means described in the next section. Another assessment of skewness is a large difference between the mean and median. In *right-skewed* data, the mean is quite a bit larger than the median, while in *left-skewed* data, the mean is much smaller than the median. Of note, in this dataset, the largest observation is more than six standard deviations above the mean!