Statistics for Biology and Health

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

Regression Methods in Biostatistics

Linear, Logistic, Survival, and Repeated Measures Models

With 54 Illustrations

Springer

Eric Vittinghoff
Department of Epidemiology and Biostatistics
University of California
San Francisco, CA 94143
USA
eric@biostat.ucsf.edu

David V. Glidden
Department of Epidemiology and Biostatistics
University of California
San Francisco, CA 94143
USA
dave@biostat.ucsf.edu

Stephen C. Shiboski
Department of Epidemiology and Biostatistics
University of California
San Francisco, CA 94143
USA
steve@biostat.ucsf.edu

Charles E. McCulloch
Department of Epidemiology and Biostatistics
University of California
San Francisco, CA 94143
USA
chuck@biostat.ucsf.edu

Library of Congress Cataloging-in-Publication Data
Regression methods in biostatistics : linear, logistic, survival, and repeated measures models / Eric Vittinghoff ... [et al.].
p. cm. — (Statistics for biology and health)
Includes bibliographical references and index.
ISBN 0-387-20275-7 (alk. paper)
1. Medicine—Research—Statistical methods. 2. Regression analysis. 3. Biometry.
I. Vittinghoff, Eric. II. Series.
R853.S7R44 2004
610′.72′7—dc22 2004056545

ISBN 0-387-20275-7 Printed on acid-free paper.

Printed in the United States of America. (EB)

9 8 7 6 5 4 3 2 1 SPIN 10946190

springeronline.com

For Jessie & Dannie, E.J., Caroline, Erik & Hugo, and J.R.

Preface

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 and regression methods for repeated outcomes have come into widespread use in the medical research literature. Applying these methods and interpreting the results requires 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 datasets. 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 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.

This book grew out of our teaching a two-quarter sequence to post-graduate physicians training for a research career. We thank them for their feedback and patience. Partial support for this came from a K30 grant from the National Institutes of Health awarded to Stephen Hulley, for which we are grateful.

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 Chapter 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. Chapter 5 deals with predictor selection, an issue common to all the multipredictor models covered. In Chapter 6 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 7 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 8 looks at the accommodation of correlated data in both linear and logistic models. Chapter 9 extends Chapter 6, giving an overview of generalized linear models. Finally, Chapter 10 is a brief introduction to the analysis of complex surveys.

The text closes with a summary, Chapter 11, 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
October, 2004

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

Contents

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, 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. 8), what if physicians with more aggressive approaches to back pain also 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 Chapter 9 why a generalized linear model 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.

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 extensions of the logistic model.

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.

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. Related extensions are also required to analyze data from complex surveys.

The various regression modeling approaches, while differing in important statistical details, also share important similarities. Numeric, binary, and cat-

egorical 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 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.

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 for the clearest interpretation of the effects of the aggressiveness of treatment.

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, Chapter 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. We address model checking as well as model selection in considerable detail.

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 Chapter 7 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 Chapter 6, 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. In our view, data analysis profits from considering the estimated effects, using confidence intervals to quantify their precision.

We recommend that readers begin with Chapter 2, on exploratory methods. Since Chapter 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, Chapter 5 covers predictor selection, which is common to the entire family of regression techniques. Chapters 6 and 7 cover regression methods specialized for binary and time-to-event outcomes, while Chapters 8–10 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, Chapter 11 reprises the themes considered in the earlier chapters and is recommended for all readers.

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

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

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 type of anomaly. Data checks revealed missing classifications for patients whose diagnostic score exceeded the 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 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 systolic blood pressure 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. 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.

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 (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 systolic blood pressure (SBP).

Numerical Description

As a first step we obtain basic descriptive statistics for SBP. Table 2.1 gives detailed summary statistics for the systolic blood pressure variable, `sbp`. Several

Table 2.1. Numerical Description of Systolic Blood Pressure

```
. summarize sbp, detail

                        systolic BP
-------------------------------------------------------------
      Percentiles      Smallest
 1%           104            98
 5%           110           100
10%           112           100       Obs                3154
25%           120           100       Sum of Wgt.        3154

50%           126                     Mean           128.6328
                        Largest       Std. Dev.      15.11773
75%           136           210
90%           148           210       Variance       228.5458
95%           156           212       Skewness       1.204397
99%           176           230       Kurtosis       5.792465
```

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 data set, the largest observation is more than six standard deviations above the mean!

Graphical Description

Graphs are often the quickest and most effective way to get a sense of the data. For numerical data, three basic graphs are most useful: the histogram, boxplot, and normal quantile-quantile (or Q-Q) plot. Each is useful for different purposes. The histogram easily conveys information about the location, spread, and shape of the frequency distribution of the data. The boxplot is a schematic identifying key features of the distribution. Finally, the normal quantile-quantile (Q-Q) plot facilitates comparison of the shape of the distribution of the data to a normal (or bell-shaped) distribution.

The histogram displays the frequency of data points falling into various ranges as a bar chart. Fig. 2.1 shows a histogram of the SBP data from WCGS. Generated using an earlier version of Stata, the default histogram uses five intervals and labels axes with the minimum and maximum values only. In this

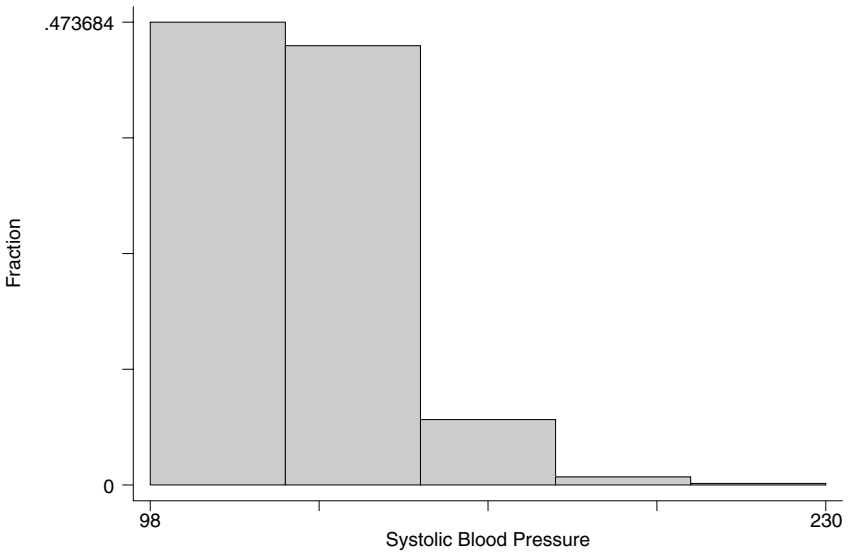

Fig. 2.1. Histogram of the Systolic Blood Pressure Data

figure, we can see that most of the measurements are in the range of about 100 to 150, with a few extreme values around 200. The percentage of observations in the first interval is about 47.4%.

However, this is not a particularly well-constructed histogram. With over 3,000 data points, we can use more intervals to increase the definition of the histogram and avoid grouping the data so coarsely. Using only five intervals, the first two including almost all the data, makes for a loss of information, since we only know the value of the data in those large "bins" to the limits

of the interval (in the case of the first bin, between 98 and 125), and learn nothing about how the data are distributed within those intervals. Also, our preference is to provide more interpretable axis labeling. Fig. 2.2 shows a modified histogram generated using the current version of Stata that provides much better definition as to the shape of the frequency distribution of SBP.

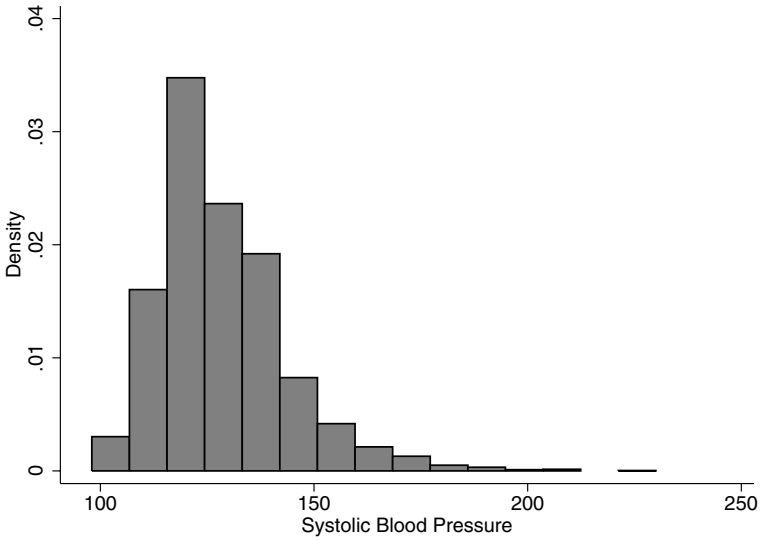

Fig. 2.2. Histogram of the Systolic Blood Pressure Data Using 15 Intervals

The key with a histogram is to use a sufficient number of intervals to define the shape of the distribution clearly and not lose much information, without using so many as to leave gaps, give the histogram a ragged shape, and defeat the goal of summarization. With 3,000 data points, we can afford quite a few bins. A *rough* rule of thumb is to choose the number of bins to be about $1 + 3.3 \log_{10}(n)$, (Sturges, 1926) where n is the sample size (so this would suggest 12 or 13 bins for the WCGS data). More than 20 or so are rarely needed. Fig. 2.2 uses 15 bins and provides a clear definition of the shape as well as a fair bit of detail.

A boxplot represents a compromise between a histogram and a numerical summary. The boxplot in Fig. 2.3 graphically displays information from the summary in Table 2.1, specifically the minimum, maximum, and 25th, 50th (median), and 75th percentiles. This retains many of the advantages of a graphical display while still providing fairly precise numerical summaries. The "box" displays the 25th and 75th percentiles (the lower and upper edges of the box) and the median (the line across the middle of the box). Extending from the box are the "whiskers" (this colorful terminology is due to the

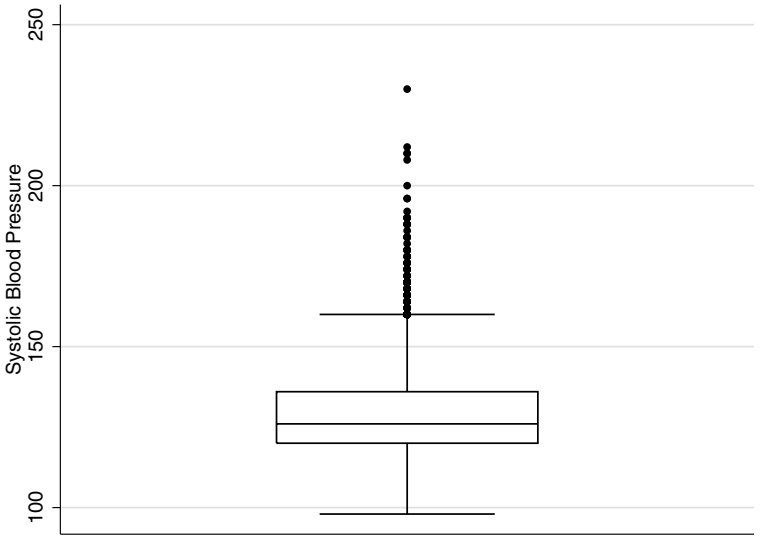

Fig. 2.3. Boxplot of the Systolic Blood Pressure Data

legendary statistician John Tukey, who liked to coin new terms). The bottom whisker extends to the minimum data value, 98, but the maximum is above the upper whisker. This is because Stata uses an algorithm to try to determine if observations are "outliers," that is, values a large distance away from the main portion of the data. Data points considered outliers (they can be in either the upper or lower range of the data) are plotted with symbols and the whisker only extends to the most extreme observation not considered an outlier.

Boxplots convey a wealth of information about the distribution of the variable:

- location, as measured by the median
- spread, as measured by the height of the box (this is called the interquartile range or IQR)
- range of the observations
- presence of outliers
- some information about shape.

This last point bears further explanation. If the median is located toward the bottom of the box, then the data are *right-skewed* toward larger values. That is, the distance between the median and the 75th percentile is greater than that between the median and the 25th percentile. Likewise, right-skewness will be indicated if the upper whisker is longer than the lower whisker or if there are more outliers in the upper range. Both the boxplot and the histogram show evidence for right-skewness in the SBP data. If the

direction of the inequality is reversed (more outliers on the lower end, longer lower whisker, median toward the top of the box), then the distribution is *left-skewed*.

Our final graphical technique, the normal Q-Q plot, is useful for comparing the frequency distribution of the data to a normal distribution. Since it is easy to distinguish lines that are straight from ones that are not, a normal Q-Q plot is constructed so that the data points fall along an approximately straight line when the data are from a normal distribution, and deviate *systematically* from a straight line when the data are from other distributions. Fig. 2.4 shows the Q-Q plot for the SBP data. The line of the data points shows a distinct

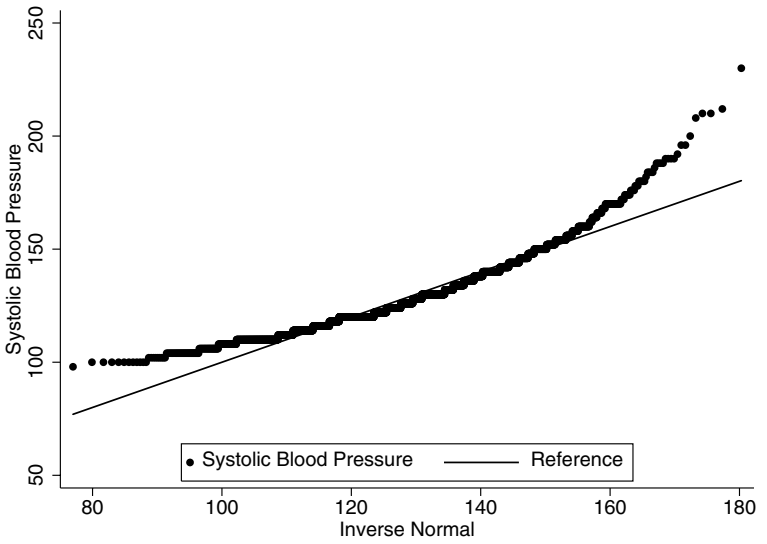

Fig. 2.4. Normal Q-Q Plot of the Systolic Blood Pressure Data

curvature, indicating the data are from a non-normal distribution.

The shape and direction of the curvature can be used to diagnose the deviation from normality. Upward curvature, as in Fig. 2.4, is indicative of right-skewness, while downward curvature is indicative of left-skewness. The other two common patterns are S-shaped. An S-shape as in Fig. 2.5 indicates a *heavy-tailed* distribution, while an S-shape like that in Fig. 2.6 is indicative of a *light-tailed* distribution.

Heavy- and light-tailed are always in reference to a hypothetical normal distribution with the same spread. A heavy-tailed distribution has more observations in the middle of the distribution and way out in the tails, and fewer a modest way from the middle (simply having more in the tails would just mean a larger spread). Light-tailed means the reverse: fewer in the middle and

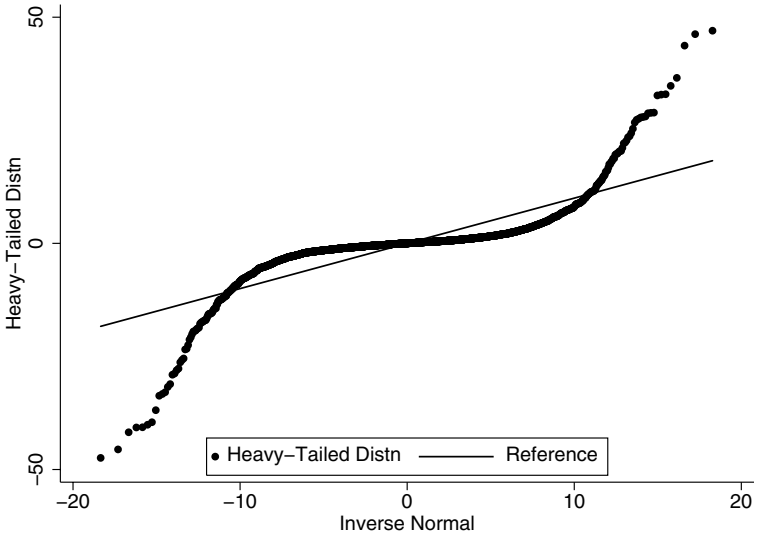

Fig. 2.5. Normal Q-Q Plot of Data From a Heavy-Tailed Distribution

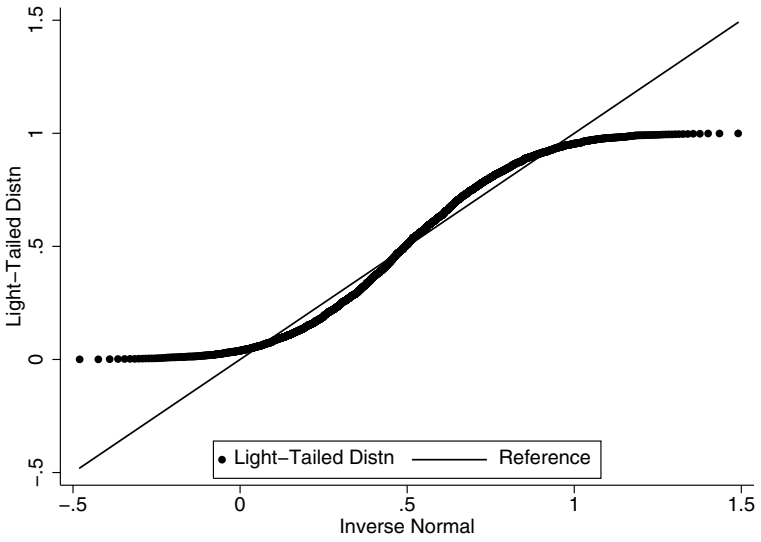

Fig. 2.6. Normal Q-Q plot of Data From a Light-Tailed Distribution

far out tails and more in the mid-range. Heavy-tailed distributions are generally more worrisome than light-tailed since they are more likely to include outliers.

Transformations of Data

A number of the techniques we describe in this book require the assumption of approximate normality or, at least, work better when the data are not highly skewed or heavy-tailed, and do not include extreme outliers. A common method for dealing with these problems is to transform such variables. For example, instead of the measured values of SBP, we might instead use the logarithm of SBP. We first consider why this works and then some of the advantages and disadvantages of transformations.

Transformations affect the distribution of values of a variable because they emphasize differences in a certain range of the data, while de-emphasizing differences in others. Consider a table of transformed values, as displayed in Table 2.2. On the original scale the difference between .01 and .1 is .09, but

Table 2.2. Effect of a $\log_{10}$ Transformation

Value	Difference	$\log_{10}$ value	Difference
0.01	0.09	-2	1
0.1	0.9	-1	1
1	9	0	1
10	90	1	1
100	900	2	1
1000	–	3	–

on the $\log_{10}$ scale, the difference is 1. In contrast, the difference between 100 and 1,000 on the original scale is 900, but this difference is also 1 on the $\log_{10}$ scale. So a log transformation de-emphasizes differences at the upper end of the scale and emphasizes those at the lower end. This holds for the natural log as well as $\log_{10}$ transformation. The effect can readily be seen in Fig. 2.7, which displays histograms of SBP on the original scale and after natural log transformation. The log-transformed data is distinctly less right-skewed, even though some skewness is still evident. Essentially, we are viewing the data on a different scale of measurement.

There are a couple of other reasons to consider transforming variables, as we will see in later sections and chapters: transformations can simplify the relationships between variables (e.g., by making a curvilinear relationship linear), can remove interactions, and can equalize variances across subgroups that previously had unequal variances.

A primary objection to the use of transformations is that they make the data less interpretable. After all, who thinks about medical costs in log dol-

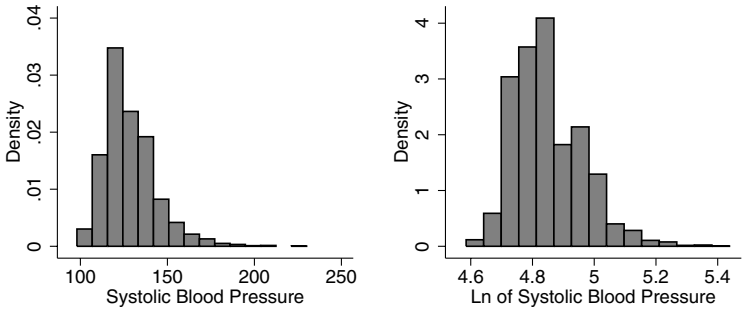

Fig. 2.7. Histograms of Systolic Blood Pressure and Its Natural Logarithm

lars? In situations where there is good reason to stay with the original scale of measurement (e.g., dollars) we may prefer alternatives to transformation including generalized linear models and weighted analyses. Or we may appeal to the robustness of normality-based techniques: many perform extremely well even when used with data exhibiting fairly serious violations of the assumptions.

In other situations, with a bit of work, it is straightforward to express the results on the original scale when the analysis has been conducted on a transformed scale. For example, Sect. 4.7.5 gives the details for log transformations in linear regression.

A compromise when the goal is, for example, to test for differences between two arms in a clinical trial is to plan ahead to present basic descriptive statistics in the original scale, but perform tests on a transformed scale more appropriate for statistical analysis. After all, a difference on the transformed scale is still a difference between the two arms.

Finally we remind the reader that different scales of measurement just take a bit of getting used to: consider pH.

2.3.2 Categorical Variables

Categorical variables require a different approach, since they are less amenable to graphical analyses and because common statistical summaries, such as mean and standard deviation, are inapplicable. Instead we use tabular descriptions. Table 2.3 gives the frequencies, percents, and cumulative percents for each of the behavior pattern categories for the WCGS data. Note that cumulative percentages are really only useful with ordinal categorical data (why?).

When tables are generated by the computer, there is usually little latitude in the details. However, when tables are constructed by hand, thought should be given to their layout; Ehrenberg (1981) is recommended reading. Three

Table 2.3. Frequencies of Behavior Patterns

behavioral pattern (4 level)	Freq.	Percent	Cum.
A1	264	8.37	8.37
A2	1325	42.01	50.38
B3	1216	38.55	88.93
B4	349	11.07	100.00
Total	3154	100.00	

easy-to-follow suggestions from that article are to arrange the categories in a meaningful way (e.g., not alphabetically), report numbers to two effective digits, and to leave a gap every three or four rows to make it easier to read across the table. Table 2.4 illustrates these concepts. With the table arranged

Table 2.4. Characteristics of Top Medical Schools

School	Rank	NIH research ($10 millions)	Tuition ($thousands)	Average MCAT
Harvard	1	68	30	11.1
Johns Hopkins	2	31	29	11.2
Duke	3	16	31	11.6
Penn	4(tie)	33	32	11.7
Washington U.	4(tie)	25	33	12.0
Columbia	6	24	33	11.7
UCSF	7	24	20	11.4
Yale	8	22	30	11.1
Stanford	9(tie)	19	30	11.1
Michigan	9(tie)	20	29	11.0

Source: US News and World Report (http://www.usnews.com, 12/6/01)

in order of the rankings, it is easy to see values that do not follow the pattern predicted by rank, for example, out-of-state tuition.

2.4 Two-Variable Descriptions

Most of the rest of this book is about the relationships among variables. An example from the WCGS is whether behavior pattern is related to systolic blood pressure. In investigating the relationships between variables, it is often useful to distinguish the role that the variables play in an analysis.

2.4.1 Outcome Versus Predictor Variables

A key distinction is whether a variable is being predicted by the remaining variables, or whether it is being used to make the prediction. The variable singled out to be predicted from the remaining variables we will call the *outcome variable*; alternate and interchangeable names are *response variable* or *dependent variable*. The variables used to make the prediction will be called *predictor variables*. Alternate and equivalent terms are *covariates* and *independent variables*. We slightly prefer the outcome/predictor combination, since the term *response* conveys a cause-and-effect interpretation, which may be inappropriate, and *dependent/independent* is confusing with regard to the notion of statistical independence. ("Independent variables do not have to be independent" is a true statement!)

In the WCGS example, we might hypothesize that change in behavior pattern (which is potentially modifiable) might cause change in SBP. This would lead us to consider SBP as the outcome and behavior pattern as the predictor.

2.4.2 Continuous Outcome Variable

As before, it is useful to consider the nature of the outcome and predictor variables in order to choose the appropriate descriptive technique. We begin with continuous outcome variables, first with a continuous predictor and then with a categorical predictor.

Continuous Predictor

When both the predictor and outcome variables are continuous, the typical numerical description is a correlation coefficient and its graphical counterpart is a scatterplot. Again considering the WCGS study, we will investigate the relationship between SBP and weight.

Table 2.5 shows the Stata command and output for the correlation coefficient, while Fig. 2.8 shows a scatterplot. Both the graph and the numerical summary confirm the same thing: there is a weak association between the

Table 2.5. Correlation Coefficient for Systolic Blood Pressure and Weight

```
. correlate sbp weight (obs=3154)

             |      sbp   weight
-------------+------------------
         sbp |   1.0000
      weight |   0.2532   1.0000
```

two variables, as measured by the correlation of 0.25. The graph conveys important additional information. In particular, there are quite a few outliers,