Chapter 1

Introduction

1.1 What This Book Is About

When I began graduate study at the University of Wisconsin in 1970, categorical data analysis consisted of chi-square tests for cross-tabulated data, a technique introduced around the turn of the century by the great Karl Pearson. This methodology was viewed with scorn by most of my quantitatively oriented cohorts. It was the province of old fogies who hadn't bothered to learn about REGRESSION ANALYSIS, the new universal tool for social science data analysis. Little did we realize that another revolution was taking place under our noses. By the time I left Wisconsin in 1975, the insanely great new thing was LOGLINEAR ANALYSIS, which made it possible to analyze complicated contingency tables in ways that Karl Pearson never dreamed of. But loglinear analysis was a rather different animal from linear regression and I, for one, never felt entirely comfortable working in the loglinear mode.

In the years after I left Wisconsin, these dissimilar approaches to data analysis came together in the form of LOGIT ANALYSIS, also known as logistic regression analysis. The logit model is essentially a regression model that is tailored to fit a categorical dependent variable. In its most widely used form, the dependent variable is a simple dichotomy, and the independent variables can be either quantitative or categorical. As we shall see, the logit model can be generalized to dependent variables that have more than two categories, both ordered and unordered. Using the method of conditional logit analysis, it can also be extended to handle specialized kinds of data such as discrete-choice applications, matched-

pair analysis, and longitudinal data. Logit models for longitudinal data can also be analyzed with a method called generalized estimating equations.

This book is an introduction to the logit model and its various extensions. Unlike most introductory texts, however, this one is heavily focused on the use of the SAS System to estimate logit and related models. In my judgment, you can't fully understand and appreciate a new statistical technique without carefully considering the practical details of estimation. To accomplish that, it's necessary to choose a particular software system to carry out the computations. Although there are many good statistical packages for doing logit regression, SAS is certainly among the best in terms of the range of estimation methods, available features and options, efficiency and stability of the algorithms, and quality of the documentation. I find I can do almost anything I want to do in SAS and, in the process, I encounter few of the annoying software problems that seem to crop up frequently in other packages.

In addition to the logit model, I also write briefly about two alternatives for binary data, the probit model and the complementary log-log model. The last chapter is about loglinear analysis, a close cousin to logit analysis. Because some kinds of contingency table analysis are awkward to handle with the logit model, the loglinear model can be a useful alternative. I don't pretend to give a comprehensive treatment of loglinear analysis, however. The emphasis is on how to do it with the GENMOD procedure, and to show examples of the type of applications where a loglinear analysis might be particularly useful. The penultimate chapter is about Poisson regression for count data. I've included this topic partly for its intrinsic interest, but also because it's a useful preparation for the loglinear chapter. In PROC GENMOD, loglinear analysis is accomplished by way of Poisson regression.

Besides GENMOD, this book gives extensive coverage to the LOGISTIC procedure. The chapter on multinomial logit analysis focuses on the CATMOD procedure, and the chapters that use conditional logit analysis make heavy use of the PHREG procedure. However, this book is not intended to be a comprehensive guide to these SAS procedures. I discuss only those features that are most widely used, most potentially useful, and most likely to cause problems or confusion. You should always consult the official documentation in the *SAS/STAT User's Guide* or in later updates, such as *SAS/STAT Software: Changes and Enhancements through Release 6.12.*

1.2 What This Book Is Not About

This book does *not* cover a variety of categorical data analysis known as Cochran-Mantel-Haenszel (CMH) statistics, for two reasons. First, I have little expertise on the subject and it would be presumptuous for me to try to teach others. Second, while CMH is widely used in the biomedical sciences, there have been few applications in the social sciences. This disuse is not necessarily a good thing. CMH is a flexible and well-founded approach to categorical data analysis—one that I believe has many potential applications. While it accomplishes many of the same objectives as logit analysis, it is best suited to those situations where the focus is on the relationship between one independent variable and one dependent variable, controlling for a limited set of additional variables. Statistical control is accomplished in a nonparametric fashion by stratification on all possible combinations of the control variables. Consequently, CMH is less vulnerable than logit regression to certain kinds of specification error but at the expense of reduced statistical power. Stokes et al. (1995) give an excellent and thorough introduction to CMH methods as implemented with the FREQ procedure in SAS.

1.3 What You Need to Know

To understand this book, you need to be familiar with multiple linear regression. That means that you should know something about the assumptions of the linear regression model and about estimation of the model via ordinary least squares. Ideally, you should have a substantial amount of practical experience using multiple regression on real data and should feel comfortable interpreting the output from a regression analysis. You should be acquainted with such topics as multicollinearity, residual analysis, variable selection, nonlinearity, interactions, and dummy variables. As part of this knowledge, you must certainly know the basic principles of statistical inference: standard errors, confidence intervals, hypothesis tests, *p*-values, bias, efficiency, and so on. In short, a two-semester sequence in statistics ought to provide the necessary statistical foundation for most people.

I have tried to keep the mathematics at a minimal level throughout the book. Except for one section on maximum likelihood estimation (which can be skipped without loss of continuity), there is no calculus and little use of matrix notation. Nevertheless, to simplify the presentation of regression models, I occasionally use the vector notation

$\boldsymbol{\beta}\mathbf{x} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$. While it would be helpful to have some knowledge of maximum likelihood estimation, it's hardly essential. However, you should know the basic properties of logarithms and exponential functions.

With regard to SAS, the more experience you have with SAS/STAT and the SAS DATA step, the easier it will be to follow the presentation of SAS programs. On the other hand, the programs presented in this book are fairly simple and short, so don't be intimidated if you're just beginning to learn SAS.

1.4 Computing

All the computer input and output displayed in this book was produced by and for Release 6.12 of the SAS System. Occasionally, I point out differences between the syntax of 6.12 and earlier releases. I use the following convention for presenting SAS programs: All SAS keywords are in uppercase. All user-specified variable names and data set names are in lowercase. In the main text, both SAS keywords and user-specified variables are in uppercase. In the output displays, nonessential output lines are often edited out to conserve space.

1.5 References

Most of the topics in this book can be found in any one of several textbooks on categorical data analysis. In preparing this book, I have particularly benefited from consulting Agresti (1990, 1996), Hosmer and Lemeshow (1989), Long (1997), Fienberg (1980), and Everitt (1992). I have generally refrained from giving references for material that is well established in the textbook literature, but I do provide references for any unusual, nonstandard, or controversial claims. I also give references whenever I think the reader might want to pursue additional information or discussion. This book should not be regarded as a substitute for official SAS documentation. As of this writing, the most complete and up-to-date documentation for the LOGISTIC, GENMOD, and PHREG procedures can be found in *SAS/STAT Software: Changes and Enhancements through Release 6.12* (SAS Institute 1996). For PROC CATMOD, consult the *SAS/STAT User's Guide*, volume 1 (SAS Institute 1989).