Preface

This book has been written with several guiding principles in mind. First, the focus is on marginal models for *categorical* data. This is partly done because marginal models for continuous data are better known, statistically better developed, and more widely used in practice — although not always under the name of marginal models — than the corresponding models for categorical data (as shown in the last chapter of this book). But the main reason is that we are convinced that a large part of the data used in the social and behavioral sciences are categorical and should be treated as such. Categorical data refer either to discrete characteristics that are categorical by nature (e.g., religious denominations, number of children) or result from a categorical measurement process (e.g., using response categories such as *Yes* versus *No*, *Agree* versus *Neither Agree nor Disagree* versus *Disagree*). Treating categorical data as realizations of continuous variables and forcing them somehow into models for continuous data always implies making untestable assumptions about the measurement process and about underlying continuous distributions. In general, an explicit a priori model about the specific way in which the respondents 'translate' the continuous information into a choice among discrete categories must be postulated, and the underlying latent variable must be assumed to have a specific form, e.g., a normal distribution. Violation of these largely untestable assumptions can seriously distort the conclusions of an analysis. Therefore, it is very fortunate that during the last few decades enormous progress has been made in the development of models that take the categorical nature of the data as given without having to make strong untestable assumptions. This monograph will apply and extend these insights into the area of marginal modeling.

Our second guiding principle seems an obvious one: *marginal models* are needed to answer important research questions. However, we state this principle explicitly because there is some controversy about it. This monograph will elaborate our position extensively (throughout the book and more explicitly in the next and last chapters), but a concise illustration of it runs as follows. We have two variables A and B, which are measurements of the characteristic *Political Preference* at two instances, e.g., political preference at time one (variable A) and two (B) for the same sample, or the political preferences of husband (A) and wife (B) for a sample of married

couples. To answer questions like *Are there any net differences between time one and two?* or *Are there any overall differences between the opinions of husbands and wives?*, one needs marginal modeling. Because the marginal distributions of *A* and *B* do not come from independent samples, standard estimation and testing techniques involving independent observations cannot be used. Comparison of the marginal distributions, taking the dependencies of the observations into account, is the specific purpose of marginal modeling. This approach is often referred to as 'unconditional' or 'population-averaged' and contrasted with the conditional or subject-specific approach. In the latter one, the scores on *B* are regressed on *A* (or vice versa) and the interest lies in the (expected) individual scores on *B*, conditional upon and as a function of the scores on *A*. In this sense, the interest in conditional analyses lies explicitly in the nature of the dependencies in the data. In terms of longitudinal studies, one focuses on the gross changes and transitions from *A* to *B*. The unconditional and the conditional analyses will generally show different outcomes. It does not make much sense to argue about what approach is better — they just answer different research questions.

An important and widely used alternative way of handling dependencies in the data is by means of random coefficient models, also referred to as multilevel or hierarchical models. The (dis)similarities between the marginal and the random coefficient approaches may be rather complex. Just a few main remarks will be made here. In the next and especially the last chapter, a more extensive discussion will be presented. In marginal modeling, the marginal distributions of *A* and *B* are being compared, taking the dependencies in the data into account in the estimation and testing procedures, without making any restrictive assumptions about the nature of the dependencies. Many researchers tend to use random coefficient models for essentially the same purposes. In terms of our little example, one can test and estimate the differences between *A* and *B* while introducing random intercepts for each individual (in the longitudinal example) or for each couple (in the husband-wife example), thus explicitly taking the dependencies in the data into account. Generally (and certainly for loglinear models), the outcomes of random coefficient models and corresponding marginal models will not be the same. In the random coefficient model for the example above, one investigates the differences between the marginal distributions of *A* and *B* conditioning on subject (or couple) differences. In this sense, random coefficient models belong to the class of conditional or subject-specific models. Further, when introducing random coefficients, one introduces certain assumptions about the distribution of the random coefficient and about the nature of the dependencies in the data, assumptions that usually constrain the dependencies in the data and that may or may not be correct. Finally, it is usually not so easy and straightforward to constrain random-effect models in such a way that these constraints yield the intended marginal restrictions by means of which the intended hypotheses about the marginal distributions can be tested.

The estimation method that will be used in this book is *Maximum Likelihood* (ML) estimation, directly building upon the work of Lang and Agresti (Lang, 1996a; Lang & Agresti, 1994). ML estimation for marginal modeling has many advantages, but ML estimates are sometimes difficult to obtain. The algorithms proposed here at

least partly overcome these problems. In an alternative general approach towards the testing and estimation of marginal models, weighted least squares (WLS) procedures are used. This approach is often called the GSK method, named after its developers Grizzle, Starmer and Koch (Grizzle, Starmer, & Koch, 1969; Landis & Koch, 1979). In general, WLS estimates are computationally simpler to obtain than ML estimates, but have some statistical disadvantages. To overcome some of the computational difficulties of ML estimation without the disadvantages of WLS, Liang and Zeger developed a quasi-likelihood procedure — the GEE (Generalized Estimating Equations) method — for marginal-modeling purposes (Liang & Zeger, 1986; Diggle, Heagerty, Liang, & Zeger, 2002). GEE provides consistent parameter estimates, but faces problems regarding the efficiency and accuracy of the estimated standard errors. A more extensive comparison between ML on the one hand, and WLS and GEE on the other hand will be discussed in the last chapter. A third alternative to ML might be Bayesian inference. Despite the important recent developments in Bayesian methods, at least for categorical data analysis, its accomplishments and promises for marginal modeling of categorical data are still too unclear to treat them here. It would also go beyond the intended scope of this book.

This book has been written for *social and behavioral scientists* with a good background in social science statistics and research methods. Familiarity with basic loglinear modeling and some basic principles of matrix algebra is needed to understand the contents of this book. Nevertheless, the emphasis is on an intuitive understanding of the methodology of marginal modeling and on applications and research examples. Parts that are statistically more difficult are indicated by ***: they are included to provide a deeper insight into the statistical background but are not necessary to follow the main argument.

The real world examples presented in this book are on the book's website:

www.cmm.st

In addition, our Mathematica and R programmes for fitting marginal models can be found there, as well as the user-friendly code needed to fit the models discussed in the book. This is further discussed in Chapter 7, along with a presentation of some other user-friendly programs for marginal modeling.

In the first chapter, we will explain the basic concepts of marginal modeling. Because loglinear models form the basic tools of categorical data analysis, loglinear marginal models will be discussed in Chapter 2. However, not all interesting research questions involving marginal modeling can be answered within the loglinear framework. Therefore, in Chapter 3 it will be shown how to estimate and test nonloglinear marginal models. The methods explained in Chapters 2 and 3 will then be applied in Chapter 4 to investigate changes over time using longitudinal data. Data resulting from repeated measurements on the same subjects probably form the most important field for the application of marginal models. In Chapter 5, marginal modeling is related to causal modeling. For many decades now, Structural Equation Modeling (SEM) has formed an important and standard part of the researcher's tool kit, and it has also been well developed for categorical data. It is shown in Chapter 5 that there are many useful connections between SEM and marginal modeling for the

analysis of cross-sectional or longitudinal data. The use of marginal models for the analysis of (quasi-)experimental data is another important topic in Chapter 5. In all analyses in Chapters 2 through 5, the observed data are treated as given, in the sense that no questions are asked regarding their reliability and validity. All the analyses are manifest-level analyses only. Marginal models involving latent variables are the topic of Chapter 6. In the final Chapter 7, a number of important conclusions, discussions and extensions will be discussed: marginal models for continuous data, alternative estimation methods, comparisons of marginal models to random and fixed-effect models, some specific applications, possible future developments, and very importantly, software and the contents of the book's website.

The very origins of this book lie in the Ph.D. thesis *Marginal Models for Categorical Data* (Bergsma, 1997), written by the first and supervised by the second (as co-promotor) and third author (as promotor). Each of us has written in one form or another all lines, sections, and chapters of this book, and we are all three responsible for its merits and shortcomings, but at the same time we acknowledge the fundamental work done in the Ph.D. thesis. Grants from the Netherlands Organization for Scientific Research NWO have made the Ph.D. project possible, as well as a subsequent postdoc project (NWO grant 400-20-001P) that contributed enormously to the birth of this monograph.

Finally, we would like to thank several people and acknowledge their often critical but always constructive, helpful, and important contributions. Jeroen Vermunt, Joe Lang, Antonio Forcina and Tamas Rudas contributed in many ways, in discussions, in answering questions, in working together both on the aforementioned Ph.D. thesis and on this book. John Gelissen, Steffen Kühnel, J. Scott Long, Ruud Luijkx, and Michael E. Sobel commented on (large) parts of the manuscript. Andries van der Ark contributed to the R routines for fitting the models. Matthijs Kalmijn provided us with the Dutch NKPS data (Netherlands Kinship Panel; http://www.nkps.nl) used in Chapters 2, 5, and 6. Marrie Bekker allowed us to use the data on body satisfaction which are analyzed in Chapters 2 and 3. Finally, we thank Bettina Hoeppner for providing us with the smoking cessation data from the Cancer Prevention Research Center, University of Rhode Island (Kingston, RI), used in Chapter 5 (and we will not insult the readers' intelligence by assuming that they might think the aforementioned people are in any way responsible for the book's errors and shortcomings).

If it were not for the patience and enduring support of John Kimmel, editor at Springer Verlag, these pages would never have appeared between covers.

Wicher Bergsma, Marcel Croon, Jacques Hagenaars