Preface

Categorical data, whether categories are nominal or ordinal, consist of multinomial responses along with suitable covariates from a large number of independent individuals, whereas longitudinal categorical data consist of similar responses and covariates collected repeatedly from the same individuals over a small period of time. In the latter case, the covariates may be time dependent but they are always fixed and known. Also it may happen in this case that the longitudinal data are not available for the whole duration of the study from a small percentage of individuals. However, this book concentrates on complete longitudinal multinomial data analysis by developing various parametric correlation models for repeated multinomial responses. These correlation models are relatively new and they are developed by generalizing the correlation models for longitudinal binary data [Sutradhar (2011, Chap. 7), *Dynamic Mixed Models for Familial Longitudinal Data,* Springer, New York]. More specifically, this book uses dynamic models to relate repeated multinomial responses which is quite different than the existing books where longitudinal categorical data are analyzed either marginally at a given time point (equivalent to assume independence among repeated responses) or by using the so-called working correlations based GEE (generalized estimating equation) approach that cannot be trusted for the same reasons found for the longitudinal binary (two category) cases [Sutradhar (2011, Sect. 7.3.6)]. Furthermore, in the categorical data analysis, whether it is a cross-sectional or longitudinal study, it may happen in some situations that responses from individuals are collected on more than one response variable. This type of studies is referred to as the bivariate or multivariate categorial data analysis. On top of univariate categorical data analysis, this book also deals with such multivariate cases, especially bivariate models are developed under both cross-sectional and longitudinal setups. In the cross-sectional setup, bivariate multinomial correlations are developed through common individual random effect shared by both responses, and in the longitudinal setup, bivariate structural and longitudinal correlations are developed using dynamic models conditional on the random effects.

As far as the main results are concerned, whether it is a cross-sectional or longitudinal study, it is of interest to examine the distribution of the respondents (based on their given responses) under the categories. In longitudinal studies, the possible

change in distribution pattern over time is examined after taking the correlations of the repeated multinomial responses into account. All these are done by fitting a suitable univariate multinomial probability model in the cross-sectional setup and correlated multinomial probability model in the longitudinal setup. Also these model fittings are first done for the cases where there is no covariate information from the individuals. In the presence of covariates, the distribution pattern may also depend on them, and it becomes important to examine the dependence of response categories on the covariates. Remark that in many existing books, covariates are treated as response variables and contingency tables are generated between response variable and the covariates, and then a full multinomial or equivalently a suitable log linear model is fitted to the joint cell counts. This approach lacks theoretical justification mainly because the covariates are usually fixed and known and hence the Poisson mean rates for joint cells should not be constructed using association parameters between covariates and responses. This book avoids such confusions and emphasizes on regression analysis all through to understand the dependence of the response(s) on the covariates.

The book is written primarily for the graduate students and researchers in statistics, biostatistics, and social sciences, among other applied statistics research areas. However, the univariate categorical data analysis discussed in Chap. 2 under cross-sectional setup, and in Chap. 3 under longitudinal setup with time independent (stationary) covariates, is written for undergraduate students as well. These two chapters containing cross-sectional and longitudinal multinomial models, and corresponding inference methodologies, would serve as the theoretical foundation of the book. The theoretical results in these chapters have also been illustrated by analyzing various biomedical or social science data from real life. As a whole, the book contains six chapters. Chapter 4 contains univariate longitudinal categorical data analysis with time dependent (non-stationary) covariates, and Chaps. 5 and 6 are devoted to bivariate categorical data analysis in cross-sectional and longitudinal setup, respectively. The book is technically rigorous. More specifically, this is the first book in longitudinal categorical data analysis with high level technical details for developments of both correlation models and inference procedures, which are complemented in many places with real life data analysis illustrations. Thus, the book is comprehensive in scope and treatment, suitable for a graduate course and further theoretical and/or applied research involving cross-sectional as well as longitudinal categorical data. In the same token, a part of the book with first three chapters is suitable for an undergraduate course in statistics and social sciences. Because the computational formulas all through the book are well developed, it is expected that the students and researchers with reasonably good computational background should have no problems in exploiting them (formulas) for data analysis.

The primary purpose of this book is to present ideas for developing correlation models for longitudinal categorical data, and obtaining consistent and efficient estimates for the parameters of such models. Nevertheless, in Chaps. 2 and 5, we consider categorical data analysis in cross-sectional setup for univariate and bivariate responses, respectively. For the analysis of univariate categorical data in

Chap. 2, multinomial logit models are fitted irrespective of the situations whether the data contain any covariates or not. To be specific, in the absence of covariates, the distribution of the respondents under selected categories is computed by fitting multinomial logit model. In the presence of categorical covariates, similar distribution pattern is computed but under different levels of the covariate, by fitting product multinomial models. This is done first for one covariate with suitable levels and then for two covariates with unequal number of levels. Both nominal and ordinal categories are considered for the response variable but covariate categories are always nominal. Remark that in the presence of covariates, it is of primary interest to examine the dependence of response variable on the covariates, and hence product multinomial models are exploited by using a multinomial model at a given level of the covariate. Also, as opposed to the so-called log linear models, the multinomial logit models are chosen for two main reasons. First, the extension of log linear model from the cross-sectional setup to the longitudinal setup appears to be difficult whereas the primary objective of the book is to deal with longitudinal categorical data. Second, even in the cross-sectional setup with bivariate categorical responses, the so-called odds ratio (or association) parameters based Poisson rates for joint cells yield complicated marginal probabilities for the purpose of interpretation. In this book, this problem is avoided by using an alternative random effects based mixed model to reflect the correlation of the two variables but such models are developed as an extension of univariate multinomial models from cross-sectional setup. With regard to inferences, the likelihood function based on product multinomial distributions is maximized for the case when univariate response categories are nominal. For the inferences for ordinal categorical data, the well-known weighted least square method is used. Also, two new approaches, namely a binary mapping based GQL (generalized quasi-likelihood) and pseudo-likelihood approaches, are developed. The asymptotic covariances of such estimators are also computed.

Chapter 3 deals with longitudinal categorical data analysis. A new parametric correlation model is developed by relating the present and past multinomial responses. More specifically, conditional probabilities are modeled using such dynamic relationships. Both linear and non-linear type models are considered for these dynamic relationships based conditional probabilities. The models are referred to as the linear dynamic conditional multinomial probability (LDCMP) and multinomial dynamic logit (MDL) models, respectively. These models have pedagogical virtue of reducing to the longitudinal binary cases. Nevertheless, for simplicity, we discuss the linear dynamic conditional binary probability (LDCBP) and binary dynamic logit (BDL) models in the beginning of the chapter, followed by detailed discussion on LDCMP and MDL models. Both covariate free and stationary covariate cases are considered. As far as the inferences for longitudinal binary data are concerned, the book uses the GQL and likelihood approaches, similar to those in Sutradhar (2011, Chap. 7), but the formulas in the present case are simplified in terms of transitional counts. The models are then fitted to a longitudinal Asthma data set as an illustration. Next, the inferences for the covariate free LDCMP model are developed by exploiting both GQL and likelihood approaches; however, for simplicity, only likelihood approach is discussed for the covariate free MDL model.

In the presence of stationary covariates, the LDCMP and MDL regression models are fitted using the likelihood approach. As an illustration, the well-known Three Miles Island Stress Level (TMISL) data are reanalyzed in this book by fitting the LDCMP and MDL regression models through likelihood approach. Furthermore, correlation models for ordinal longitudinal multinomial data are developed and the models are fitted through a binary mapping based pseudo-likelihood approach.

Chapter 4 is devoted to theoretical developments of correlation models for longitudinal multinomial data with non-stationary covariates, whereas similar models were introduced in Chap. 3 for the cases with stationary covariates. As opposed to the stationary case, it is not sensible to construct contingency tables at a given level of the covariates in the non-stationary case. This is because the covariate levels are also likely to change over time in the non-stationary longitudinal setup. Consequently, no attempt is made to simplify the model and inference formulas in terms of transitional counts. Two non-stationary models developed in this chapter are referred to as the NSLDCMP (non-stationary LDCMP) and NSMDL (non-stationary MDL) models. Likelihood inferences are employed to fit both models. The chapter also contains discussions on some of the existing models where odds ratios (equivalent to correlations) are estimated using certain "working" log linear type working models. The advantages and drawbacks of this type of "working" correlation models are also highlighted.

Chapters 2 through 4 were confined to the analysis of univariate longitudinal categorical data. In practice, there are, however, situations where more than one response variables are recorded from an individual over a small period of time. For example, to understand how diabetes may affect retinopathy, it is important to analyze retinopathy status of both left and right eyes of an individual. In this problem, it may be of interest to study the effects of associated covariates on both categorical responses, where these responses at a given point of time are structurally correlated as they are taken from the same individual. In Chap. 5, this type of bivariate correlations is modeled through a common individual random effect shared by both response variables, but the modeling is confined, for simplicity, to the cross-sectional setup. Bivariate longitudinal correlation models are discussed in Chap. 6. For inferences for the bivariate mixed model in Chap. 5, we have developed a likelihood approach where a binomial approximation to the normal distribution of random effects is used to construct the likelihood estimating equations for the desired parameters. Chapter 5 also contains a bivariate normal type linear conditional model, but for multinomial response variables. A GQL estimation approach is used for the inferences. The fitting of the bivariate normal model is illustrated by reanalyzing the well-known WESDR (Wisconsin Epidemiologic Study of Diabetic Retinopathy) data.

In Chap. 6, correlation models for longitudinal bivariate categorical data are developed. This is done by using a dynamic model for each multinomial variables conditional on the common random effect shared by both variables. Theoretical details are provided for both model development and inferences through a GQL estimation approach. The bivariate models discussed in Chaps. 5 and 6 may be

extended to the multivariate multinomial setup, which is, however, beyond the scope of the present book. The incomplete longitudinal multinomial data analysis is also beyond the scope of the present book.

St. John's, Newfoundland, Canada Brajendra C. Sutradhar

http://www.springer.com/978-1-4939-2136-2

Longitudinal Categorical Data Analysis
Sutradhar, B.C.
2014, XVIII, 369 p., Hardcover
ISBN: 978-1-4939-2136-2