Preface

Discrete familial data consist of count or binary responses along with suitable covariates from the members of a large number of independent families, whereas discrete longitudinal data consist of similar responses and covariates collected repeatedly over a small period of time from a large number of independent individuals. As the statistical modelling of correlation structures especially for the discrete longitudinal data has not been easy, many researchers over the last two decades have used either certain 'working' models or mixed (familial) models for the analysis of discrete longitudinal data. Many books are also written reflecting these 'working' or mixed models based research. This book, however, presents a clear difference between the modelling of familial and longitudinal data. Parametric or semiparametric mixed models are used to analyze familial data, whereas parametric dynamic models are exploited to analyze the longitudinal data. Consequently, dynamic mixed models are used to analyze combined familial longitudinal data. Basic properties of the models are discussed in detail. As far as the inferences are concerned, various types of consistent estimators are considered, including simple ones based on method of moments, quasi-likelihood, and weighted least squares, and more efficient ones such as generalized quasi-likelihood estimators which account for the underlying familial and/or longitudinal correlation structure of the data. Special care is given to the mathematical derivation of the estimating equations.

The book is written for readers with a background knowledge of mathematics and statistics at the advanced undergraduate level. As a whole, the book contains eleven chapters including Chapters 2 and 3 on linear fixed and mixed models (for continuous data) with autocorrelated errors. The remaining chapters are also presented in a systematic fashion covering mixed models, longitudinal models, longitudinal mixed models, and familial longitudinal models, both for count and binary data. Furthermore, in almost every chapter, the inference methodologies have been illustrated by analyzing biomedical or econometric data from real life. Thus, the book is comprehensive in scope and treatment, suitable for a graduate course and further theoretical and/or applied research involving familial and longitudinal data.

Familial models for discrete count or binary data are generally known as the generalized linear mixed models (GLMMs). There is a long history on inferences in GLMMs with single or multiple random effects. In this GLMMs setup, the correlations among the responses under a family are clearly generated through the common random effects shared by the family members. However, as opposed to the GLMMs setup, it has not been easy to model the longitudinal correlations in generalized linear longitudinal models (GLLMs) setup. Chapter 1 provides an overview on difficulties and remedies with regard to (1) the consistent and efficient estimation in the GLMMs setup, and (2) the modelling of longitudinal correlations and subsequently efficient estimation of the parameters in GLLMs.

The primary purpose of this book is to present ideas for developing correlation models for discrete familial and/or longitudinal data, and obtaining consistent and efficient estimates for the parameters of such models. Nevertheless, in Chapter 2, we consider a clustered linear regression model with autocorrelated errors. There are two main reasons to deal with such linear models with autocorrelated errors. First, in

practice, one may also need to analyze the continuous longitudinal data. Secondly, the knowledge of autocorrelation models for continuous repeated data should be helpful to distinguish them from similar autocorrelation models for discrete repeated data. Several estimation techniques, namely the method of moments (MM), ordinary least squares (OLS), and generalized least squares (GLS) methods are discussed. An overview on the relative efficiency performances of these approaches is also presented.

In Chapter 3, a linear mixed effects model with autocorrelated errors is considered for the analysis of clustered correlated continuous data, where the repeated responses in a cluster are also assumed to be influenced by a random cluster effect. A generalized quasi-likelihood (GQL) method, similar to but different from the GLS method, is used for the inferences in such a mixed effects model. The relative performance of this GQL approach to the so-called generalized method of moments (GMM), used mainly in the econometrics literature, is also discussed in the same chapter.

When the responses from the members of a given family are counts, and they are influenced by the same random family effect in addition to the covariates, they are routinely analyzed by fitting a familial model (i.e., GLMM) for count data. In this setup, the familial correlations among the responses of the members of the same family become the function of the regression parameters (effects of the covariates on the count responses) as well as the variance of the random effects. However, obtaining consistent and efficient estimates especially for the variance of the random effects has been proven to be difficult. With regard to this estimation issue, Chapter 4 discusses the advantages and the drawbacks of the existing highly competitive approaches, namely the method of moments, penalized quasi-likelihood (PQL), hierarchical likelihood (HL), and a generalized quasi-likelihood. The relatively new GQL approach appears to perform the best among these approaches, in obtaining consistent and efficient estimates for both regression parameters and the variance of the random effects (also known as the overdispersion parameter). This is demonstrated for the GLMMs for Poisson distribution based count data, first with single– and then with two-dimensional random effects in the linear predictor of the familial model. The aforementioned estimation approaches are discussed in detail in the parametric setup under the assumption that the random effects follow a Gaussian distribution. The estimation in the semiparametric and nonparametric set up is also discussed in brief.

Chapter 5 deals with familial models for binary data. These models are similar but different from those for count data discussed in Chapter 4. The difference lies in the fact that conditional on the random family effect, the distribution of the response of a member is assumed to follow the log-linear based Poisson distribution in the count data setup, whereas in the familial models for binary data, the response of a member is assumed to follow the so-called linear logistic model based binary distribution. This makes the computation of the unconditional likelihood and moments of the data more complicated under the binary set up as compared to the count data setup. A binomial approximation as well as a simulation approach is discussed to tackle this difficulty of integration over the distribution of the random effect to

obtain unconditional likelihood or moments of the binary responses under a given family. Formulas for unconditional moments up to order four are clearly outlined for the purpose of obtaining the MM and GQL estimates for both regression and the overdispersion parameters.

In the longitudinal setup, the repeated responses collected from the same individual over a small period of time become correlated due to the influence of time itself. Thus, it is not reasonable to model these correlations through the common random effect of the individual. This becomes much clearer when it is understood that in some situations, conditional on the random effect, the repeated responses can be correlated. It has not, however, been easy to model the correlations of the repeated discrete such as count or binary responses. One of the main reasons for this is that unlike in the linear regression setup (Chapters 2 and 3), the correlations for the discrete data depend on the time-dependent covariates associated with the repeated responses. In fact, the modelling of the correlations for discrete data, even if the covariates are time independent, has also not been easy. Over the last two decades, many existing studies, consequently, have used arbitrary 'working' correlations structure to obtain efficient regression estimates as compared to the moment or least squares estimates. This is, however, known by now that this type of 'working' correlations model based estimates [usually referred to as the generalized estimating equations (GEE) based estimates] may be less efficient than the simpler moment or least squares estimates. Chapter 6 deals with a class of autocorrelation models constructed based on certain dynamic relationships among repeated count responses. When covariates are time independent, in this approach, it is not necessary to identify the true correlation structure for the purpose of estimation of the regression coefficients. A GQL approach is used which always produces consistent and highly efficient regression estimates, especially as compared to the moment or independence assumption based estimates. The modelling for correlations when covariates are time dependent is also discussed in detail. In order to use the GQL estimation approach, this chapter also demonstrates how to identify the true correlation structure of the data when it is assumed that the true model belongs to an autocorrelations class.

Similar to Chapter 6, Chapter 7 deals with dynamic models and various inference techniques including the GQL approach for the analysis of repeated binary data collected from a large number of independent individuals. Note that the correlated binary models based on linear dynamic conditional probabilities (LDCP) are quite different from those dynamic models discussed in Chapter 6 for the repeated count data. Furthermore, for the cases where it is appropriate to consider that the means and variances of repeated binary responses over time may maintain a recursive relationship, Chapter 7 provides a discussion on the inferences for such data by fitting a binary dynamic logit (BDL) model.

Chapter 8 develops a longitudinal mixed model for count data as a generalization of the longitudinal fixed effects model for count data discussed in Chapter 6. This generalization arises in practice because of the fact that if the response of an individual at a given time is influenced by the associated covariates as well as a random effect of the individual, then this random effect will remain the same throughout

the data collection period over time. In such a situation, conditional on the random effect, the repeated responses will be influenced by the associated time dependent covariates as well as by time as a stochastic factor. Thus, conditional on the random effect, the repeated count responses will follow a dynamic model for count data as in Chapter 6. Note that unconditional correlations, consequently, will be affected by both the variance of the random effects as well as the correlation index parameter from the dynamic model. This extended correlation structure has been exploited to obtain the consistent and efficient GQL estimates for the regression parameters, as well as a consistent GQL estimate for the variance of the random effects.

By the same token as that of Chapter 8, Chapter 9 deals with various longitudinal mixed models for binary data. These models are developed based on the assumption that conditional on the individual's random effect, the repeated binary responses either follow the LDCP or BDL models as in Chapter 7. Conditional on the random effects, a binary dynamic probit (BDP) model is also considered. This generalized model is referred to as the binary dynamic mixed probit (BDMP) model. In general, the GQL estimation approach is used for the inferences. The GMM and maximum likelihood (ML) estimation approaches are also discussed.

Chapter 10 is devoted to the inferences in familial longitudinal models for count data. These models are developed by combining the familial models for count data discussed in Chapter 4 and the longitudinal models (GLLMs) for count data discussed in Chapter 6. The combined model has been referred to as the GLLMM (generalized linear longitudinal mixed model). In this setup, the count responses are two-way correlated, familial correlations occur due to the same random family effect shared by the members of a given family, and the longitudinal correlations arise due to the possible dynamic relationship among the repeated responses of a given member of the family. These two-way correlations are taken into account to develop the GQL estimating equations for the regression effects and variance component for the random family effects, and the moment estimating equation for the longitudinal correlation index parameter.

Chapter 11 discusses the inferences in GLLMMs for binary data. A variety of longitudinal correlation models is considered, whereas the familial correlations are developed through the introduction of the random family effects only. The GQL approach is discussed in detail for the estimation of the parameters of the models. Because the likelihood estimation is manageable when longitudinal correlations are introduced through dynamic logit models, this chapter, similar to Chapter 9, discusses the ML estimation as well. As a further generalization, two-dimensional random family effects are also considered in the dynamic logit relationship based familial longitudinal models. Both GQL and ML approaches are given for the estimation of the parameters of such multidimensional random effects based familial longitudinal models.

http://www.springer.com/978-1-4419-8341-1

Dynamic Mixed Models for Familial Longitudinal Data
Sutradhar, B.C.
2011, XVIII, 494 p., Hardcover
ISBN: 978-1-4419-8341-1