SOME CONTRIBUTIONS
IN
LONGITUDINAL DATA ANALYSIS

Thesis submitted to the

UNIVERSITY OF CALICUT

for the award of the degree of

DOCTOR OF PHILOSOPHY

IN

STATISTICS

under the Faculty of Science

by

RADHAKRISHNAN NAIR K

under the guidance of

Prof.(Dr.) M. MANOHARAN

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALICUT
KERALA, 673 635
INDIA

November 2012

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALICUT

Dr. M.MANOHARAN
Professor

Calicut University,
16th November 2012.

CERTIFICATE

Certified that the work presented in this thesis entitled 'SOME CONTRIBUTIONS IN LONGITUDINAL DATA ANLAYSIS', submitted to the University of Calicut for the award of the Degree of Doctor of Philosophy in Statistics, is a bonafide work done by Mr. Radhakrishnan Nair K, under my guidance in the Department of Statistics, University of Calicut and that this work has not been included in any other thesis submitted previously for the award of any degree, diploma, associate ship, fellowship.

Prof. (Dr.) M. Manoharan

Supervising Guide

DECLARATION

I hereby declare that the matter embodied in this thesis entitled 'SOME CONTRIBUTIONS IN LONGITUDINAL DATA ANLAYSIS', submitted to the University of Calicut for the award of the Degree of Doctor of Philosophy in Statistics, is based on the original work done by me under the guidance and supervision of Prof. (Dr.) M. Manoharan, Department of Statistics, University of Calicut, and has not been previously formed the basis for the award of any degree, diploma, associateship, fellowship *etc.* of this university or any other university or institution.

University of Calicut

28 October 2012 Radhakrishnan Nair K

ACKNOWLEDGEMENTS

It gives me immense pleasure to put in words my sincere gratitude to one and all who extended their helping hands towards me in my work.

I am greatly indebted to my Guide Prof. (Dr.) M. Manoharan, Professor, Department of Statistics, University of Calicut, for his affectionate as well as scholarly guidance, generous help, and encouragement through out my work, without which my pursuit would not have been materialized.

I am indebted to Prof. (Dr.) M. Manoharan, Head of the Department, and the former Heads of the Department Dr. C Chandran and Dr. K Jayakumar for providing the necessary facilities to complete my research work. I would like to extend my heartfelt thanks to Dr. N. Raju, Professor, Department of Statistics of Calicut University, for the inspiration, encouragement and technical help.

I am grateful to the Principal and the Manager of Nehru Arts & Science College, Kanhangad, for deputing me to undertake this study.

I also thank the UGC for the sanction of the financial support for this research work by way of granting me teacher fellowship under FDP.

My sincere thanks are also due to Sri. Shanthakumar M V, Librarian and the non - teaching staff of the Department of Statistics, Calicut University for their help and co-operation.

To all research scholars and students of the Department of Statistics of Calicut University, for their help, discussions and suggestions during this research.

I also thank my wife and children for their love, patience, and support and for the inconveniences they have borne because of my preoccupations.

Radhakrishnan Nair K

SOME CONTRIBUTIONS

IN

LONGITUDINAL DATA ANALYSIS

Table of Contents

Chapter 1

Introduction

The main concern of the thesis is on the statistical models and methods useful in the analysis of longitudinal data, that are routinely collected in a broad range of applications, including agricultural and life sciences, medical and public health research, and physical science and engineering. In this set up, data in the form of repeated measurements on the same unit over time are used. The characteristics of this data is that the same response is measured repeatedly on each unit. This type of the data structure will be the focus of the thesis.

In addition to the usual kinds of questions being asked, such as how the mean response differs across treatments, we are more curious to know how change in mean response over time differs and other issues concerning the relationship between response and time. All these warrants, an appropriate representation of the situation in terms of a statistical model that signifies the way in which the data were collected in order to address these questions. Needless to say special methods of analysis are

required in this study to complement the models.

In almost every discipline there is increased awareness of the importance of longitudinal studies for understanding change over time and the factors that influence change. This has led to a steady growth in the availability of longitudinal data, often arising from relatively complex study designs. The analysis of longitudinal data continues to pose many interesting methodological challenges and is likely to do so for the foreseeable future. The objective of this research work is to highlight the current state of the art of longitudinal data analysis, to unravel the potentials of the longitudinal data analysis to find answers to challenging questions posed in various disciplines and to provide a glimpse of future directions.

Longitudinal studies represent one of the major strategies employed in research in various areas nowadays. The main advantage of longitudinal studies is that they can distinguish changes over time within individuals (longitudinal effects) from differences between subjects at the start of the study (base-line characteristics; cross-sectional effects). The statistical literature on the analysis of repeated measurements is based on the paradigm of multivariate regression. But longitudinal studies, typically, have unbalanced designs, time varying coefficients, missing data and other characteristics that make standard multivariate procedures inapplicable.

The type of problems in longitudinal study differs from that usually treated as time series in the statistical literature to the extent that in the latter, we have, in general, a single sample unit evaluated at many instants while in the former, we usually deal with several sample units observed at a few instants. Thus they differ from classical time series data in consisting of a large number of short series, one

from each subject, rather than a single long series. The distinctive characteristic of longitudinal data is the ordered dimension along which data are collected. For further details of the characterisation of longitudinal studies and their relation with other forms of data collection one may refer to Goldstein (1979), Duncan and Kalton (1987), Crowder and Hand (1990), Lindsey (1993), Jones (1993), Baltagi (1995), Vonesh and Chinchilli (1997) and Diggle *et al* (2003), among others.

In cross sectional studies a single observation (at a specified instant) of the response variable is recorded for each element of the sample. If two or more observations (at different instances) of the response variable is recorded for each observation of the sample, the study is known as longitudinal. That is, in longitudinal studies, each subject gives rise to a vector of measurements, the measurements representing the same physical quantity at different observation times. For example, we might measure a subject's blood pressure on each of five successive days. Longitudinal data therefore combine elements of multivariate and time series data. However, they differ from classical multivariate data in that the time series aspect of the data typically imparts a much more highly structured pattern of interdependence among measurements than for standard multivariate data sets.

Longitudinal studies are of particular interest for investigating global or individual changes along time. In the first place, they allow for the observation of the response variable under uniform exposure of the sample units to different covariates. Secondly, longitudinal designs provide information about individual variation of the levels of the response variable. Finally, some parameters of the underlying statistical models may be more efficiently estimated under longitudinal data collection schemes

than under cross sectional schemes with the same number of observations.

The major advantage of longitudinal study is its capacity to separate what in the context of population studies are called cohort and age effects. The longitudinal study can distinguish changes over time within individuals(ageing effects) from differences among people in their baseline levels(cohort effects), whereas a cross-sectional study cannot. In some studies, a third timescale, the period, or calendar date of measurement, is also important. Any two of age, period and cohort determine the third. Analyses which must consider all three scales require external assumptions which unfortunately are difficult to validate. (See Mason and Feinberg (1985) for details).

Longitudinal designs are superior to cross sectional designs in several ways. First, it offers investigators the opportunity for controlled and uniform measurement of exposure history and other factors related to outcome. Hence the relevant information is more reliably quantified. Second, longitudinal designs provide information about individual patterns of change. Finally, longitudinal designs can provide more efficient estimators of some parameters than cross sectional designs with the same number and pattern of measurement.

The main disadvantage associated to longitudinal studies is related to cost, for in many instances they require efforts to assure that the sample units be observed at designated instants and in others, they need long observation periods. Technical difficulties due to comparatively complex statistical models may also constitute a problem.

1.1 Advances in Longitudinal Data Analysis: A Historical Perspective

There have been remarkable developments in statistical methodology for longitudinal data analysis in the last few decades. As a result statisticians and empirical researchers now have access to an increasingly sophisticated toolbox of methods. However, there has been a lag between the recent developments that have appeared in the statistical journals and their widespread application to substantive problems. One reason for these advances to be somewhat slow to move into the mainstream is their limited implementation in widely available standard computer software. Recently, however, the introduction of new programs for analyzing multivariate and longitudinal data has made many of these methods far more accessible to statisticians and empirical researchers alike. Also, in the present scenario, in which statistical software is constantly evolving, we can anticipate that many of the more recent advances will soon be implemented. Thus, the outlook is bright that modern methods for longitudinal analysis will be applied more widely and across a broader spectrum of disciplines.

In this Section, we look back with an historical perspective and review many of the key advances that have been made, especially in the past four decades. Our review will be somewhat selective, and the main goal is to highlight the important and enduring developments in the methodology. We collocate the seminal works that turned out to be the landmarks in the headway of longitudinal data analysis. The review will set the stage for the remaining chapters of this thesis, where the focus is

on the current state of the art of longitudinal data analysis.

For a few years, most of the statistical methodological developments for the analysis of longitudinal data were directed towards univariate continuous response variables, with special attention to those with an underlying Gaussian distribution. Due to the nice behaviour of the multivariate Gaussian distribution and its analytic characteristics, there has been very sharp focus on research on models under this distribution, for such a long time after its first introduction by Wishart(1938). As a result a rich variety of models for Gaussian data is available. The classical papers of Box (1950), Geisser and Green House (1958), Potthoff and Roy (1964), Rao (1959, 1965), Elston and Grizzle (1962) and Grizzle and Allen (1969) are examples of the early efforts, which developed into a fruitful research area in the past four decades. Grizzle and Allen (1969) suggest an interesting computational procedure to implement Rao's (1959) technique of covariance adjustment. This technique is further discussed in Baksalary *et al* (1978), Kenward (1985), Calinski and Caussinus (1989) and is summarised in Kshirsagar and Smith (1995).

The use of random effects models for longitudinal data analysis was first considered by Rao (1959) and Elston and Grizzle (1962), in the context of growth curves. A generalization was considered in Graybill (1976) and its actual form was presented by Laird and Ware (1982), based on the idea of Harville (1977). Laird and Ware (1982) present an excellent discussion of restricted maximum likelihood estimation (REML) methods for estimation and proposed the use of EM algorithm to fit a class of linear mixed effects models appropriate for the analysis of repeated measurements. Jennrich and Schluchter (1986) propose a variety of alternative algorithms, including Fisher-

scoring and Newton-Raphson algorithms. Winer (1971) provides some application of repeated measures by ANOVA.

Ware and Liang (1996) provide an interesting historical perspective on the development of statistical methods for the analysis of longitudinal data. The foundations for the repeated measures analysis of variance can be found in the seminal monograph by Fisher (1925) and in the method of analysing split-plot experiments proposed by Yates (1935) and Scheffe (1959).

There is increased evidence about the value of using Bayesian predictive inference for finite population parameters, especially for difficult problems such as estimation. Bayesian estimation for longitudinal data models has been considered by Lee and Geisser (1972), Giesser (1970, 1980), Rao (1987), Lee (1988) and Kass and Steffey (1989). The Bayesian models allow us to pass along both the qualitative form of the longitudinal model and the information about the parameters in the model, conveyed by the prior distribution and likelihood. Bayesian nonparametric methods have been proposed for population models to accommodate population heterogeneity and to relax distributional assumptions and restrictive models.

Henderson *et al* (2000) propose a shared latent process for jointly modelling Gaussian longitudinal data and event time data that incorporated autocorrelation through a latent process. Li, Shao and Palta (2005) propose a latent variable measurement error model for the error structure and implement it in a linear mixed model, wherein the estimation procedure is similar to regression calibration but involves a distributional assumption for the latent variable through an example concerning sleep-disordered breathing (SDB) in the Wisconsin Sleep Cohort Study (WSCS).

More recently, statistical methods based on generalized linear models have been considered for the analysis of data with underlying distributions in the exponential family, although they still fall behind the need. Generalized linear models are used for discrete data as well.

1.1.1 Early origins of linear models for longitudinal data analysis

The analysis of change is a fundamental component of so many research endeavours in almost every discipline. Many of the earliest statistical methods for the analysis of change were based on the analysis of variance (ANOVA) paradigm, as originally developed by R. A. Fisher. One of the earliest methods proposed for analyzing longitudinal data was a mixed-effects ANOVA, with a single random subject effect. The inclusion of a random subject effect induced positive correlation among the repeated measurements on the same subject. We use the terms subjects and individuals interchangeably to refer to the participants in a longitudinal study. Interestingly, it was the British astronomer George Biddel Airy who laid the foundations for the linear mixed-model formulation (Airy (1861)), before it was put on a more formal theoretical footing in the seminal work of R. A. Fisher (see, for example, Fisher (1918, 1925). Airy's work on a model for errors of observation in astronomy predated Fishers more systematic study of related issues within the ANOVA paradigm (e.g., Fisher's (1921, 1925) writings on the intraclass correlation). Scheffe (1956) provides a fascinating discussion of the early contributions of 19^{th} century astronomers to the development of the theory of random-effects models. As such, it can be argued that statistical

methods for the analysis of longitudinal data, in common with classical linear regression and the method of least squares, have their earliest origins in the field of astronomy.

The mixed-effects ANOVA model has a long history of use for analyzing longitudinal data, where it is often referred to as the univariate repeated-measures ANOVA. Statisticians recognized that a longitudinal data structure, with N individuals and n repeated measurements, has striking similarities to data collected in a randomized block design, or the closely related split-plot design. So it seemed natural to apply ANOVA methods developed for these designs (e.g., Yates (1935), Scheffe (1959)) to the repeated-measures data collected from longitudinal studies. In doing so, the individuals in the study are regarded as the blocks or main plots. The univariate repeated-measures ANOVA model can be written as

$$Y_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + b_i + e_{ij}, \quad i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, n,$$

where Y_{ij} is the outcome of interest, $\boldsymbol{X}_{ij}$ is a design vector, $\boldsymbol{\beta}$ is a vector of regression parameters, $b_i \sim N(0, \sigma_b^2)$, and $e_{ij} \sim N(0, \sigma_e^2)$. In this model, the blocks or plot effects are regarded as random rather than fixed effects. The random effect, b_i, represents an aggregation of all the unobserved or unmeasured factors that make individuals respond differently. The inclusion a single, individual-specific random effect induces a positive correlation among the repeated measurements, although be it with the following highly restrictive "compound symmetry" structure for the covariance: constant variance $\mathrm{Var}(Y_{ij}) = \sigma_b^2 + \sigma_e^2$ and constant covariance $\mathrm{Cov}(Y_{ij}, Y_{ik}) = \sigma_b^2$.

The univariate repeated-measures ANOVA model can be considered a forerunner of more versatile regression models for longitudinal data. For balanced data, estimates of variance components could be readily obtained in closed form by equating ANOVA mean squares to their expectations; sometime later, Henderson (1963) developed a related approach for unbalanced data. So, from an historical perspective, an undoubted appeal of the repeated measures ANOVA was that it was one of the few models that could realistically be fit to longitudinal data at a time when computing was in its infancy. This explains why, in those days, the key issue perceived to arise with incomplete data was lack of balance.

The repeated-measures multivariate analysis of variance (MANOVA) is a related approach for the analysis of longitudinal data with an equally long history, requiring somewhat more advanced computations. While the univariate repeated-measures ANOVA is conceptualized as a model for a single response variable, allowing for positive correlation among the repeated measures on the same individual *via* the inclusion of a random subject effect, MANOVA is a model for multivariate responses. As originally developed, MANOVA was intended for the simultaneous analysis of a single measure of a multivariate vector of substantively distinct response variables. In contrast, while longitudinal data are multivariate, the vector of responses is commensurate, being repeated measures of the same response variable over time. So, although MANOVA was developed for multiple, but distinct, response variables, statisticians recognized that such data share a common feature with longitudinal data, namely, that they are correlated. This led to the development of a very specific variant of MANOVA, known as repeated-measures analysis by MANOVA (or sometimes referred to as multivariate repeated-measures ANOVA).

A special case of the repeated-measures analysis by MANOVA is a general approach known as profile analysis (Box (1950), see also Geisser and Greenhouse (1958) and Greenhouse and Geisser (1959)). It proceeds by constructing a set of derived variables, based on a linear combination of the original sequence of repeated measures, and using relevant subsets of these to address questions about longitudinal change and its relation to between subject factors. These derived variables provide information about the mean level of the response, averaged over all measurement occasions, and also about change in the response over time. For the most part, the primary interest in a longitudinal analysis is in the analysis of the latter derived variables. The multiple derived variables representing the effects of measurement occasions are then analyzed by MANOVA.

Box (1950) provided one of the earliest descriptions of this approach, proposing the construction of derived variables that represent polynomial contrasts of the measurement occasions. A closely related work can be found in Danford *et al* (1960), Geisser (1963), Potthoff and Roy (1964), Cole and Grizzle (1966), and Grizzle and Allen (1969). Alternative transformations can be used, as the MANOVA test statistics are invariant to how change over time is characterized in the transformation of the original repeated measures. Although the MANOVA approach is computationally more demanding than the univariate repeated-measures ANOVA, an appealing feature of the method is that it allows assumptions on the structure of the covariance among repeated measures to be relaxed. In standard applications of the method, no explicit structure is assumed for the covariance among repeated measures (other than homogeneity of covariance across different individuals).

1.1.2 Linear mixed-effects model for longitudinal data

The linear mixed-effects model is probably the most widely used method for analyzing longitudinal data. Although the early development of mixed-effects models for hierarchical or clustered data can be traced back to the ANOVA paradigm (see, for example, Scheffe (1959)) and to the seminal paper by Harville (1977), their usefulness for analyzing longitudinal data, especially in the life sciences, was highlighted in the 1980's in a widely cited paper by Laird and Ware (1982). Goldstein (1979) is often seen as the counterpart for the humanities. The idea of allowing certain regression coefficients to vary randomly across individuals was also a recurring theme in the early contributions to growth curve analysis by Wishart (1938), Box (1950), Rao (1958), Potthoff and Roy (1964), and Grizzle and Allen (1969). These early contributions to growth curve modelling laid the foundation for the linear mixed-effects model. The idea of randomly varying regression coefficients was also a common theme in the so-called two-stage approach to analyzing longitudinal data. In the two-stage formulation, the repeated measurements on each individual are assumed to follow a regression model with distinct regression parameters for each individual. The distribution of these individual-specific regression parameters, or "random effects," is modelled in the second stage. A version of the two-stage formulation was popularized by bio-statisticians working at the U.S. National Institutes of Health (NIH). They proposed a method for analyzing repeated measures data, where, in the first stage, subject-specific regression coefficients are estimated using ordinary least-squares regression. In the second stage, the estimated regression coefficients are then analyzed

as summary measures using standard parametric (or non-parametric) methods. Interestingly, this method for analyzing repeated-measures data became known as the "NIH method." In the agricultural sciences, a similar approach was popularized in a highly cited paper by Rowell and Walters (1976). Rao (1965) put this two-stage approach on a more formal footing by specifying a parametric growth curve model that assumed normally distributed random growth curve parameters.

In the early 1980's, Laird and Ware (1982), drawing upon a general class of mixed models introduced earlier by Harville (1977), proposed a flexible class of linear mixed-effects models for longitudinal data. These models could handle the complications of mistimed and incomplete measurements in a very natural way. The linear mixed-effects model is given by

$$\boldsymbol{Y}_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{b}_i + e_i,$$

where $\boldsymbol{Z}_{ij}$ is a design vector for the random effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{G})$, and $e_i \sim N(\boldsymbol{0}, R_i)$. The linear mixed-effects model proposed by Laird and Ware (1982) included the univariate repeated-measures ANOVA and growth curve models for longitudinal data as special cases. In addition, the model of Laird and Ware had two desirable features: first, there were fewer restrictions on the design matrices for the fixed and random effects; second, the model parameters could be estimated efficiently *via* likelihood based methods. Previously, difficulties with estimation of mixed-effects models had held back their widespread application to longitudinal data. Laird and Ware (1982) showed how the expectation-maximization (EM) algorithm (Dempster *et al* (1977)) could be used to fit this general class of models for longitudinal data. Soon after, Jennrich and Schluchter (1986) proposed a variety of alternative algorithms, including

Fisher scoring and Newton-Raphson. Currently, maximum likelihood and restricted maximum likelihood estimation, the latter devised to diminish the small-sample bias of maximum likelihood, are the most frequently employed routes for estimation and inference (see Verbeke and Molenberghs (2000); Fitzmaurice *et al* (2004)).

Since the pioneering work of Laird and Ware (1982), statistical methods for the analysis of longitudinal data have advanced dramatically. They showed that generalized mixed-effects regression models could be used to perform a more complete analysis of all of the available longitudinal data under much more general assumptions regarding the missing data (i.e., missing at random; MAR). The net result was a more powerful set of statistical tools for analysis of longitudinal data that led to more powerful statistical hypothesis tests, more precise estimates of rates of change (and differential rates of change between experimental and control groups), and more general assumptions regarding missing data, for example because of study dropout. This early work has led to considerable related advances in statistical methodology for analysis of longitudinal data (for excellent reviews of this development, see Diggle *et al* (2003), Fitzmaurice *et al* (2004), Goldstein (1995), Hedeker and Gibbons (2006), Longford (1993), Raudenbush and Bryk (2002), Singer and Willett (2003), and Verbeke and Molenberghs (2000)). Notable among these advances and relevant to this review are generalizations of the original Laird-Ware type model to the nonlinear case (relevant for the analysis of binary, ordinal, nominal, count, and time-to-event outcomes), even more general forms of missing data (not missing at random; NMAR), higher levels of nesting such as three-level models (*e.g.*, repeated observations nested within subjects and subjects nested within hospitals or clinics), alternative distributional assumptions for residual errors of measurement, correlated residual errors of

measurement, multivariate mixed-effects regression models, and advances in sample size determination in the context of longitudinal studies. Computational advances in parameter estimation have also been seen, particularly in the area of nonlinear mixed effects regression models, where numerical evaluation of the likelihood function is more complex and requires high dimensional numerical integration (*e.g.*, adaptive quadrature or Monte Carlo-type integration for full Bayes estimation of model parameters).

So, by the mid-1980's, a very general class of linear models for longitudinal data had been proposed that could handle issues of unbalanced data, due to either mistimed measurement or missing data, could handle both time-varying and time-invariant covariates, and provided a flexible, yet parsimonious, model for the covariance. Moreover, these developments appeared at a time when there were great advances in computing power. It was not too long before these methods were available at the desktop and were being applied to longitudinal data in a wide variety of disciplines.

1.1.3 Models for non-Gaussian longitudinal data

The early advances in methods for longitudinal data analysis have been based on linear models for continuous responses that may be approximately normally distributed. The developments in methods for analyzing a continuous longitudinal response span more than a century, from the early work on simple random-effects models by the British astronomer Airy (1861) through the landmark paper on linear mixed-effects

models for longitudinal data by Laird and Ware (1982). In contrast, many of the advances in methods for discrete longitudinal data have been concentrated in the last four decades, harnessing the high-speed computing resources available at the desktop.

When the longitudinal response is discrete, linear models are no longer appropriate for relating changes in the mean response to covariates. Instead, statisticians have developed extensions of generalized linear models (Nelder and Wedderburn, 1972) for longitudinal data. Generalized linear models provide a unified class of models for regression analysis of independent observations of a discrete or continuous response. A characteristic feature of generalized linear models is that a suitable non-linear transformation of the mean response is assumed to be a linear function of the covariates. This non-linearity raises some additional issues concerning the interpretation of the regression coefficients in models for longitudinal data. Statisticians have extended generalized linear models to handle longitudinal observations in a number of different ways; here we consider three broad, but quite distinct, classes of regression models for longitudinal data: (i) marginal or population averaged models, (ii) random-effects or subject-specific models, and (iii) transition or response conditional models. These models differ not only in how the correlation among the repeated measures is accounted for, but also have regression parameters with discernibly different interpretations. These differences in interpretation reflect the different targets of inference of these models. Because binary data are so common, we focus much of our review on models for longitudinal binary data. Most of the developments apply to, say, categorical data and counts equally well.

(i) Marginal or population-averaged models

The extensions of generalized linear models from the univariate to the multivariate response setting have followed a number of different research threads. In this section we consider an approach for extending generalized linear models to longitudinal data that leads to a class of regression models known as marginal or population-averaged models. The term marginal in this context is used to emphasize that the model for the mean response at each occasion depends only on the covariates of interest, and not on any random effects or previous responses. This is in contrast to mixed-effects models, where the mean response depends not only on covariates but also on a vector of random effects, and to transition or generally conditional models (*e.g.*, Markov models), where the mean response depends also on previous responses.

Marginal models provide a straightforward way to extend generalized linear models to longitudinal data. They directly model the mean response at each occasion, using an appropriate link function. Because the focus is on the marginal mean and its dependence on the covariates, marginal models do not necessarily require full distributional assumptions for the vector of repeated responses, only a regression model for the mean response. This can be advantageous, as there are few tractable likelihoods for marginal models for discrete longitudinal data.

A marginal model for longitudinal data has the following three-part specification:

1. The mean of each response, $E(Y_{ij}|\boldsymbol{X}_{ij}) = \mu_{ij}$, is assumed to depend on the

covariates through a known link function

$$h^{-1}(\mu_{ij}) = \boldsymbol{X}'_{ij}\boldsymbol{\beta}.$$

2. The variance of each Y_{ij}, given the covariates, is assumed to depend on the mean according to

$$\text{Var}(Y_{ij}|\boldsymbol{X}_{ij}) = \phi\ \nu(\mu_{ij}),$$

where $\nu(\mu_{ij})$ is a known variance function and ϕ is a scale parameter that may be known or may need to be estimated.

3. The conditional within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of an additional set of association parameters, α (and may also depend upon the means, μ_{ij}).

The first of above is the key component of a marginal model and it specifies the model for the mean response at each occasion, $E(Y_{ij}|\boldsymbol{X}_{ij})$, and its dependence on the covariates. However, there is an implicit assumption in the first component that is often overlooked. Marginal models assume that the conditional mean of the j^{th} response, given $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in}$, depends only on $\boldsymbol{X}_{ij}$, that is,

$$E(Y_{ij}|\boldsymbol{X}_i) = E(Y_{ij}|\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in}) = E(Y_{ij}|\boldsymbol{X}_{ij}),$$

where obviously $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in})$; see Fitzmaurice *et al* (1993) and Pepe and Anderson (1994) for a discussion of the implications of this assumption. With time

invariant covariates, this assumption necessarily holds. Also, with time-varying covariates that are fixed by design of the study (*e.g.*, time since baseline, treatment group indicator in a crossover trial), the assumption also holds, as values of the covariates are determined a priori by study design and in a manner unrelated to the longitudinal response. However, when a time-varying covariate varies randomly over time, the assumption may no longer hold. As a result, somewhat greater care is required when fitting marginal models with time-varying covariates that are not fixed by design of the study. This problem has long been recognized by econometricians (see, for example, Engle *et al* 1983), and there is now an extensive statistical literature on this topic (see, for example, Robins *et al* 1999).

The second component specifies the marginal variance at each occasion, with the choice of variance function depending upon the type of response. For balanced longitudinal designs, a separate scale parameter, ϕ_j, can be specified at each occasion; alternatively, the scale parameter could depend on the times of measurement, with $\phi(t_{ij})$ being some parametric function of t_{ij}. Restriction to a single unknown parameter ϕ is especially limiting in the analysis of continuous responses where the variance of the repeated measurements is often not constant over the duration of the study.

The third component recognizes the characteristic of lack of independence among longitudinal data by modelling the within-subject association among the repeated responses from the same individual.

From a historical perspective, it is difficult to pinpoint the origins of marginal models. In the case of linear models, the earliest approaches based on the ANOVA

paradigm fit squarely within the framework of marginal models. In a certain sense, the necessity to distinguish marginal models from other classes of models becomes critical only for discrete responses. The development of marginal models for discrete longitudinal data has its origins in likelihood-based approaches, where the three-part specification given above is extended by making full distributional assumptions about the $n \times 1$ vector of responses. Next, we trace some of these early developments and highlight many of the issues that have complicated the application of marginal models to discrete data, leading to the widespread use of alternative, semi-parametric methods. At least three main research threads can be distinguished in the development of likelihood based marginal models for discrete longitudinal data. Because binary data are so common, we focus much of this review on models for longitudinal binary data. One of the earliest likelihood-based approaches was proposed by Gumbel (1961), who posited a latent-variable model for multivariate binary data. In this approach, there is a vector of unobserved latent variables, say $L_{i1}, \ldots, L_{in}$, and each of these is related to the observed binary responses *via*

$$Y_{ij} = \begin{cases} 1, & L_{ij} \leq \boldsymbol{X}'_{ij}\boldsymbol{\beta}, \\ 0, & L_{ij} > \boldsymbol{X}'_{ij}\boldsymbol{\beta}. \end{cases}$$

Assuming a multivariate joint distribution for $L_{i1}, \ldots, L_{in}$ identifies the joint distribution for $Y_{i1}, \ldots, Y_{in}$, with

$$\begin{aligned} \Pr(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{in} = 1) &= \Pr(L_{i1} \leq \boldsymbol{X}'_{i1}\boldsymbol{\beta}, L_{i2} \leq \boldsymbol{X}'_{i2}\boldsymbol{\beta}, \ldots, L_{in} \leq \boldsymbol{X}'_{in}\boldsymbol{\beta}) \\ &= F(\boldsymbol{X}'_{i1}\boldsymbol{\beta}, \boldsymbol{X}'_{i2}\boldsymbol{\beta}, \ldots, \boldsymbol{X}'_{in}\boldsymbol{\beta}), \end{aligned}$$

where $F(.)$ denotes the joint cumulative distribution function of the latent variables. Furthermore, any dependence among the L_{ij} induces dependence among the Y_{ij}. For example, a bivariate logistic distribution for any L_{ij} and L_{ik} induces marginally a logistic regression model for Y_{ij} and Y_{ik},

$$E(Y_{ij}|\boldsymbol{X}_{ij}) = \frac{\exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta})},$$

with positive correlation between Y_{ij} and Y_{ik}.

A closely related work, assuming a multivariate normal distribution for the latent variables, appeared in Ashford and Sowden (1970), Cox (1972), and Ochi and Prentice (1984). In the latter model, the Y_{ij} marginally follow a probit model,

$$E(Y_{ij}|\boldsymbol{X}_{ij}) = \phi(\boldsymbol{X}'_{ij}\boldsymbol{\beta}),$$

where $\phi(\cdot)$ denotes the normal CDF and the model allows both positive and negative correlation among the repeated binary responses, depending on the sign of the correlation among the underlying latent variables. This model is often referred to as the multivariate probit model. Interestingly, the multivariate probit model can also be motivated through the introduction of random effects.

At around the same time as Gumbel (1961) proposed his latent-variable formulation, a second approach to likelihood-based inferences was proposed by Bahadur (1961). Bahadur (1961) proposed an elegant expansion for an arbitrary probability mass function for a vector of responses $Y_{i1}, \ldots, Y_{in}$. The expansion for the repeated

binary responses is of the form

$$f(y_{i1},\ldots,y_{in}) = \left\{\prod_{j=1}^{n}(\pi_{ij})^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}\right\} \times\left\{1+\sum_{j<k}\rho_{ijk}z_{ij}z_{ik}+\sum_{j<k<l}\rho_{ijkl}z_{ij}z_{ik}z_{il}+\cdots+\rho_{i1\ldots n_i}z_{i1}\cdots z_{in}\right\},$$

where

$$Z_{ij}=\frac{Y_{ij}-\pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}},$$

$\pi_{ij}-E(Y_{ij})$, and $\rho_{ijk}=E(Z_{ij}Z_{ik}),\ldots,\rho_{i1\cdots n}=E(Z_{i1}\cdots Z_{in})$. Here, ρ_{ijk} is the pairwise or second-order correlation and the additional parameters relate to the third and higher order correlations among the responses.

The Bahadur expansion has a particularly appealing property, shared with the multivariate probit model and many other marginal models, of being "reproducible" or "upwardly compatible" in the sense that the same model holds for any subset of the vector of responses. In addition, the multinomial probabilities for the vector of binary responses are relatively straightforward to obtain given the model parameters. Kupper and Haseman (1978) and Altham (1978) discussed applications of this model, albeit with very simple pairwise correlation structure and assuming that the higher-order terms are zero. The chief drawback of the Bahadur expansion that has limited its application to longitudinal data is its parameterization of the higher-order associations in terms of correlation parameters. As noted earlier, for discrete data there are severe restrictions on the correlations and dependence of the correlations on the means. Thus, for discrete data, the Bahadur model requires a complicated set of

inequality constraints on the model parameters that make maximization of the likelihood very difficult. Except in very simple settings with a small number of repeated measures, the Bahadur model has not been widely applied to longitudinal data.

Because of the restrictions on the correlations, alternative multinomial models for the joint distribution of the vector of discrete responses have recently been proposed where the within-subject association is parameterized in terms of other metrics of association. For example, Dale (1984), McCullagh and Nelder (1989), Lipsitz *et al* (1990), Liang *et al* (1992), Becker and Balagtas (1993), Molenberghs and Lesaffre (1994), Lang and Agresti (1994), Glonek and McCullagh (1995), and others have proposed full likelihood approaches where the higher-order moments are parameterized in terms of marginal odds ratios. In closely related work, Ekholm (1991) parameterizes the association directly in terms of the higher-order marginal probabilities (see also Ekholm *et al* (1995)). An alternative approach is to parameterize the within-subject association in terms of conditional associations, leading to so-called "mixed-parameter" models (Fitzmaurice and Laird (1993); Glonek (1996); Molenberghs and Ritter (1996)). However, except in certain special cases (*e.g.*, Markov models), these conditional association parameters have somewhat less appealing interpretations in the longitudinal setting; moreover, their interpretation is straightforward only in balanced longitudinal designs.

In virtually all of these later advances, the application of the methodology has been hampered by at least three main factors. First, unlike in the Bahadur model, there are no simple expressions for the joint probabilities in terms of the model parameters. This makes maximization of the likelihood somewhat difficult. Second, even

with the current advances in computing, these models are difficult to fit except when the number of repeated measures is relatively small. Finally, many of these models are not robust to misspecification of the higher-order moments. That is, many of the likelihood-based methods require that the entire joint distribution be correctly specified. Thus, if the marginal model for the mean responses has been correctly specified but the model for any of the higher-order moments has not, then the maximum likelihood estimators of the marginal mean parameters will fail to converge in probability to the true mean parameters. The "mixed-parameter" models are exception to the rule; however, even these models lose this robustness property when there are missing data.

A third approach to likelihood-based marginal models is to specify the entire multinomial distribution of the vector of repeated categorical responses and estimate the multinomial probabilities non-parametrically. This was the approach first proposed in Grizzle *et al* (1969). Specifically, they proposed a weighted least-squares (WLS) method for fitting a general family of models for categorical data; in recognition of its developers, the method is often referred to as the "GSK method." Koch and Reinfurt (1971) and Koch *et al* (1977) later recognized how these models could be applied to discrete longitudinal data; Stanish *et al* (1978), Stanish and Koch (1984), and Woolson and Clarke (1984) further developed the methodology for longitudinal analysis.

The GSK method provides a very general family of models for repeated categorical data, allowing non-linear link functions to relate the marginal expectations to

covariates. The GSK method stratifies individuals according to values of the covariates and fully specifies the multinomial distribution of the vector of repeated categorical responses within each stratum. This method, for example, allows the fitting of logistic regression models to repeated binary data, albeit with the restrictions that the longitudinal study design be balanced on time, all covariates must be categorical, and there are sufficient numbers of individuals within covariate strata to estimate the multinomial probabilities non-parametrically as the sample proportions. The method requires the estimation of the covariance among the repeated responses, within strata defined by covariate values; the covariance follows directly from the properties of the multinomial distribution. Asymptotically, the GSK method is equivalent to maximum likelihood estimation; thus, this approach was appealing for analyzing discrete longitudinal data when all of the conditions required for its use were met.

Although the GSK method was a landmark technique for the analysis of repeated categorical data, it had many restrictions that limited its usefulness. Specifically, it required that all covariates be categorical and sample sizes be of sufficient size to allow for stratification and separate estimation of the multinomial covariance in each covariate stratum. However, as the number of categorical covariates in the model increases, sparse data problems quickly arise due to Bellman's (1961) "curse of dimensionality." Furthermore, missing data are not easily handled by the GSK method because they require additional stratification by patterns of missingness. Thus, the GSK method was restricted to balanced designs with categorical covariates and relatively large sample sizes.

In the mid-1980's, remarkable advances in methodology for analyzing discrete

longitudinal data were made when Liang and Zeger (1986) proposed the generalized estimating equations (GEE) approach. Because marginal models separately parameterize the model for the mean responses from the model for the within-subject association, Liang and Zeger (1986) recognized that it is possible to estimate the regression parameters in the former without making full distributional assumptions. The avoidance of distributional assumptions is potentially advantageous because, as we have discussed, there is no convenient and generally accepted specification of the joint multivariate distribution of $\boldsymbol{Y}_i$ for marginal models when the responses are discrete. The appeal of the GEE approach is that it only requires specification of that part of the probability mechanism that is of scientific interest, the marginal means. By avoiding full distributional assumptions for $\boldsymbol{Y}_i$, the GEE approach provided a remarkably convenient alternative to maximum likelihood estimation of multinomial models for repeated categorical data, without many of the inherent complications of the latter.

The GEE approach advocated in Liang and Zeger (1986) was a natural extension of the quasi-likelihood approach (Wedderburn, 1974) for generalized linear models to the multivariate response setting, where an additional set of nuisance parameters for the within-subject association must be incorporated. The foundation for the GEE approach relied on the theory of optimal estimating functions developed by Godambe (1960) and Durbin (1960). Liang and Zeger (1986) highlighted how the GEE provides a unified approach to the formulation and fitting of generalized linear models to longitudinal and clustered data. They demonstrated the versatility of the GEE method in handling unbalanced data, mixtures of discrete and continuous covariates, and arbitrary patterns of missingness. Until the publication of their landmark paper (Liang and Zeger, 1986), methods for the analysis of discrete longitudinal data had

lagged behind corresponding methods for continuous responses. Soon after, marginal models were being widely applied to address substantive questions about longitudinal change across a broad spectrum of disciplines. Their work also generated much additional theoretical and applied research on the use of this methodology for analyzing longitudinal data. For example, to improve upon efficiency, Prentice (1988) proposed joint estimating equations for both the main regression parameters, $\boldsymbol{\beta}$, and the nuisance association parameters, $\boldsymbol{\alpha}$.

The GEE approach has a number of appealing properties for estimation of the regression parameters in marginal models. First, in many longitudinal designs the GEE estimator of $\boldsymbol{\beta}$ is almost efficient when compared to the maximum likelihood estimator. For example, it can be shown that the GEE has a similar expression to the likelihood equations for $\boldsymbol{\beta}$ in a linear model for continuous responses that are assumed to have a multivariate normal distribution. The GEE also has an expression similar to the likelihood equations for $\boldsymbol{\beta}$ in certain models for discrete longitudinal data. As a result, for many longitudinal designs, there is relatively little loss of precision when the GEE approach is adopted as an alternative to maximum likelihood. Second, the GEE estimator has a very appealing robustness property, yielding a consistent estimator of $\boldsymbol{\beta}$ even if the within-subject associations among the repeated measures have been misspecified. It only requires that the model for the mean response be correct. This robustness property of GEE is important because the usual focus of a longitudinal study is on changes in the mean response. Although the GEE approach yields a consistent estimator of $\boldsymbol{\beta}$ under misspecification of the within-subject associations, the usual standard errors obtained under the misspecified model for the within-subject association are not valid. However, valid standard errors for the resulting estimator $\widehat{\boldsymbol{\beta}}$

can be obtained using the empirical or so-called sandwich estimator of Cov($\widehat{\boldsymbol{\beta}}$). The sandwich estimator is also robust in the sense that, with sufficiently large samples, it provides valid standard errors when the assumed model for the covariances among the repeated measures is not correct.

(ii) The Generalized Linear Models

In the previous section, we discussed how marginal models can be considered as an extension of generalized linear models that directly incorporate the within-subject association among the repeated measurements. In a certain sense, marginal models account for the consequences of the correlation among the repeated measures, but do not provide any explanation for its potential source. An alternative approach for accounting for the within-subject association, and one that provides a source for the within-subject association, is *via* the introduction of random effects in the model for the mean response. Following the same basic ideas as in linear mixed-effects models, generalized linear models can be extended to longitudinal data by allowing a subset of the regression coefficients to vary randomly from one individual to another. These models are known as generalized linear mixed (effects) models (GLMMs), and they extend in a natural way the conceptual approach represented by the linear mixed-effects models. In GLMMs the model for the mean response is conditional upon both measured covariates and unobserved random effects; it is the inclusion of the latter that induces correlation among the repeated responses marginally, when averaged over the distribution of the random effects. However, as we discuss later, with non-linear link functions, the introduction of random effects has important ramifications

for the interpretation of the "fixed-effects" regression parameters.

The generalized liner mixed model can be formulated using the following two-part specification:

1. Given a $q \times 1$ vector of random effects $\boldsymbol{b}_i$, the Y_{ij} are assumed to be conditionally independent and to have exponential family distributions with conditional mean depending upon both fixed and random effects,

$$h^{-1}\{E(Y_{ij}|b_i)\} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\boldsymbol{b}_i,$$

   for some known link function, $h^{-1}(\cdot)$. The conditional variance is assumed to depend on the conditional mean according to $\text{Var}(Y_{ij}|\boldsymbol{b}_i) = \phi\ \nu\{E(_{Y}ij|\boldsymbol{b}_i)\}$, where $\nu(\cdot)$ is a known variance function and ϕ is a scale parameter that may be known or may need to be estimated.

2. The random effects, $\boldsymbol{b}_i$, are assumed to be independent of the covariates, $\boldsymbol{X}_{ij}$, and to have a multivariate normal distribution, with zero mean and $q \times q$ covariance matrix G.

These two components completely specify a broad class of generalized linear mixed models. Any multivariate distribution can be assumed for the $\boldsymbol{b}_i$; in practice, however, it is common to assume that the $\boldsymbol{b}_i$ has a multivariate normal distribution.

Generalized linear mixed models have their foundation in simple random-effects models for binary and count data. The early literature on random-effects models for discrete data can be traced back to the development of random compounding models

that introduced random effects on the response scale. For example, Greenwood and Yule (1920) introduced the negative binomial distribution as a compound Poisson distribution for count data, while Skellam (1948) provided an early discussion of the beta-binomial distribution for binary data. The model has been used in a wide variety of different clustered data applications (*e.g.*, Chatfield and Goodhardt (1970); Griffiths (1973); Williams (1975); Kupper and Haseman (1978); Crowder (1978,1979); Otake and Prentice (1984); Aerts *et al* (2002)).

The main feature of the beta-binomial model that has limited its usefulness for analyzing longitudinal data is that it produces the same marginal distribution at each measurement occasion. While this may not be so problematic in certain clustered data settings (*e.g.*, in study designs where $\boldsymbol{X}_{i1} = \boldsymbol{X}_{i2} = \cdots = \boldsymbol{X}_{in}$), in a longitudinal study, where interest is primarily in changes in the marginal means over time, this restriction on the marginal distributions is very unappealing. Nonetheless, the beta-binomial and other random compounding models motivated the later development of more versatile random-effects models. Recall that in the beta-binomial model, it is assumed that success probabilities vary randomly about a mean and the latter can be related to covariates *via* an appropriate link function, such as a logit link function. In contrast to this formulation, Pierce and Sands (1975) proposed an alternative model where the logit of p_i is assumed to vary about an expectation given by $\boldsymbol{X}'_{i1}$,

$$\text{logit}\{E(Y_{ij}|b_i\} = \boldsymbol{X}'_{i1}\boldsymbol{\beta} + b_i,$$

where b_i has a normal distribution with zero mean and constant variance. The appealing feature of the model proposed by Pierce and Sands (1975) is that the fixed

and random effects are combined together on the same logistic scale. This model is often referred to as the simple logit-normal model and is very similar in spirit to the random intercept model for continuous outcomes. Although this model was remarkably simple, it proved to be difficult to fit at the time because maximum likelihood estimation required maximization of the marginal likelihood, averaged over the distribution of the random effect. This required integration, and no analytic solutions were available. The fact that the integral cannot be evaluated in a closed form limited the application of this model.

In closely related work, Ashford and Sowden (1970) proposed a very similar model, except with probit rather than logit link function. Interestingly, Ashford and Sowden's (1970) model with random intercept and probit link function was equivalent to the equicorrelated latent-variable model, leading to identical inferences provided the correlation is positive. Despite the fact that maximum likelihood estimation for even the simple logit-normal model was computationally demanding with the computer resources available at the time, Korn and Whittemore (1979) proposed a far more ambitious version of the model, where

$$\text{logit}\{E(Y_{ij}|\boldsymbol{b}_i\} = \boldsymbol{X}'_{i1}\boldsymbol{\beta} + \boldsymbol{Z}'_{i1}\boldsymbol{b}_i,$$

with $\boldsymbol{Z}_{ij} = \boldsymbol{X}_{ij}$. Although their model was very general and avoided some of the obvious drawbacks of the simple logit-normal model, it was difficult to fit and required a very long sequence of repeated measures on each subject.

From an historical perspective, the papers by Ashford and Sowden (1970), Pierce and Sands (1975), and Korn and Whittemore (1979) laid the conceptual foundations

for generalized linear mixed models; much of the work that followed focused on the thorny problem of estimation. In GLMMs the marginal likelihood is used as the basis for inferences for the fixed-effects parameters, complemented with empirical Bayes estimation for the random effects. In general, evaluation and maximization of the marginal likelihood for GLMMs requires integration over the distribution of the random effects. While this is, strictly speaking, true for the linear mixed-effects model as well, there the integration can be done analytically, so effectively a closed form for the marginal likelihood function arises, in which case the application of maximum or restricted maximum likelihood is straightforward. In the absence of an analytical solution, and because high-dimensional numerical integration can be very trying, a variety of approaches has been suggested for tackling this problem.

Because no simple analytic solutions were available, Stiratelli *et al* (1984) proposed an approximate method of estimation for the logit-normal model, based on empirical Bayes ideas, that circumvented the need for numerical integration. Specifically, they avoided the need for numerical integration by approximating the integrands with simple expansions, whose integrals have closed forms. The paper by Stiratelli *et al* (1984) led to the development of a general approach for fitting GLMMs, known as penalized quasilikelihood (PQL). Various authors (*e.g.*, Schall (1991); Breslow and Clayton (1993); Wolfinger (1993)) motivated PQL as a Laplace approximation to the marginal likelihood for GLMMs. Despite the generality of this method, and its implementation in a variety of commercially available software packages, the PQL method can often yield quite biased estimators of the variance components, which in turn leads to biased estimators of $\boldsymbol{\beta}$, especially for longitudinal binary data. This motivated research on bias corrections (*e.g.*, Breslow and Lin (1995)) and on more

accurate approximations based on higher-order Laplace approximations (*e.g.*, Raudenbush *et al* (2000)). In general, the inclusion of higher-order terms for PQL has been shown to improve estimation. Breslow and Clayton (1993) also considered an alternative approach, related to PQL, known as marginal quasi-likelihood (MQL). MQL differs from PQL by being based on an expansion around the current estimates of the fixed effects and around $b_i = 0$. In general, MQL yields severely biased estimators of the variance components, providing a good approximation only when the variance of the random effects is relatively small.

There has also been much recent research on alternative methods, including approaches based on numerical integration (*e.g.*, adaptive Gaussian quadrature) and Markov chain Monte Carlo algorithms. In particular, adaptive Gaussian quadrature, with the numerical integration centered around the empirical Bayes estimates of the random effects, permits maximization of the marginal likelihood with any desired degree of accuracy (*e.g.*, Anderson and Aitkin (1985); Hedeker and Gibbons (1994, 1996)). Adaptive Gaussian quadrature is especially appealing for longitudinal data where the dimension of the random effects is often relatively low. Monte Carlo approaches to integration, for example Monte Carlo EM (McCulloch (1997); Booth and Hobert (1999)) and Monte Carlo Newton-Raphson algorithms (Kuk and Cheng, 1997), have been proposed. The hierarchical formulation of GLMMs also makes Bayesian approaches quite appealing. For example, Zeger and Karim (1991) have proposed the use of Monte Carlo integration, *via* Gibbs sampling, to calculate the posterior distribution.

(iii). Conditional and transition models

There is a third way in which generalized linear models can be extended to handle longitudinal data. This is accomplished by modelling the mean and time dependence simultaneously *via* conditioning an outcome on other outcomes or on a subset of other outcomes (see, for example, Molenberghs and Verbeke (2005), Part III). A particular case is given by so-called transition, or Markov, models. Transition models are appealing due to the sequential nature of longitudinal data. In transition models, the conditional distribution of each response is expressed as an explicit function of the past responses and the covariates. Transition models can be considered conditional models in the sense of modelling the conditional distribution of the response at any occasion given the previous responses and the covariates. The dependence among the repeated measures is thought of as arising due to past values of the response influencing the present observation.

There is an extensive history to the use of Markov chains to model equally spaced discrete longitudinal data with a finite number of states or categories (*e.g.*, Anderson and Goodman (1957); Cox (1958); Billingsley (1961)). In the simplest of models for longitudinal data, a first-order Markov chain, the transition probabilities are assumed to be the same for each time interval. The resulting Markov chain can then be described in terms of the initial state and the set of transition probabilities. The transition probabilities are the conditional probabilities of going into each state, given the immediately preceding state. In a first-order Markov chain, there is dependence on the immediately preceding state but not on earlier outcomes. Among others, Cox (1972), Korn and Whittemore (1979), Zeger *et al* (1985), and Ware *et al* (1988)

discuss transition models applicable to longitudinal data.

Although Markov and autoregressive models have a long and extensive history of use for the analyses of time series data, their application to longitudinal data has been somewhat more limited. There are a number of features of transition models that limit their usefulness for the analysis of longitudinal data. In general, transition models have been developed for repeated measures that are equally separated in time; these models are more difficult to apply when there are missing data, mistimed measurements, and non-equidistant intervals between measurement occasions. In addition, estimation of the regression parameters $\boldsymbol{\beta}$ is very sensitive to assumptions concerning the time dependence; moreover, the interpretation of $\boldsymbol{\beta}$ changes with the order of the serial dependence. Finally, in many longitudinal studies $\boldsymbol{\beta}$ is not the usual target of inference because conditioning on the history of past responses may lead to attenuation of the effects of covariates of interest. That is, when a covariate is expected to influence the mean response at all occasions, its effect may be somewhat diminished if there is conditioning on the past history of the responses.

1.2 Terminology and Notations

The units being studied in a longitudinal analysis are referred to as individuals or subjects. The time points at which individuals are measured repeatedly are called measurement occasions or times. If the repeated measures are equally separated in time, the study is said to be balanced over time. If the repeated measures at all occasions are for each individual are not available, the data set is said to be

incomplete.

Y_{ij} : Response variable for the i^{th} individual ($i = 1, 2, \ldots, N$) at the j^{th} occasion (j= 1,2,. . . ,n_i). Realised values of Y_{ij} are denoted by y_{ij}.

$\boldsymbol{Y}_i : (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})', i = 1, 2, \ldots, N$. $\boldsymbol{Y}_i$ is a time ordered collection of n_i response variables for the i^{th} individual. The vector of responses $\boldsymbol{Y}_i$'s for N subjects are assumed to be independent of one another.

$\boldsymbol{\mu}_j$: $E(Y_{ij})$, the mean or expectation of Y_{ij}, where $E(\cdot)$ can be loosely thought of as denoting the long term average over a large population of subjects at the j^{th} occasion, or the weighted average of Y_{ij}'s with weight being probability of occurrence of each possible value.

μ_{ij} : $E(Y_{ij})$. The use of double letter subscripts is to allow the mean response to vary from individual to individual as a function of individual level covariates.

$\sigma_j^2 = E\{Y_{ij} - E(Y_{ij})\}^2 = E\{Y_{ij} - \mu_{ij}\}^2$, the variance of Y_{ij}

$\sigma_{jk} = E\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\}$, covariance between the responses at two different occasions.

$\rho_{jk} = \dfrac{E\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\}}{\sigma_j \sigma_k}$, correlation between Y_{ij} and Y_{ik}.

X_{ijp} : The p^{th} covariate associated with the i^{th} subject ($i = 1, 2, \ldots, N$) and j^{th} occasion ($j = 1, 2, \ldots, n_i$).

$\boldsymbol{X}_{ij} : (X_{ij1}, X_{ij2}, \ldots, X_{ijp})'$, the $p \times 1$ vector of covariates associated with response Y_{ij}, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, n_i$. The p rows of $\boldsymbol{X}_{ij}$ correspond to different covariates.

$$\boldsymbol{X}_i : (\boldsymbol{X}'_{i1}, \boldsymbol{X}'_{i2}, \ldots, \boldsymbol{X}'_{ip})'$$

The variance-covariance matrix of $\boldsymbol{Y}_i$

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \mathrm{Cov}\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n_i} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_i1} & \sigma_{n_i2} & \cdots & \sigma_{n_i}^2 \end{pmatrix}$$

$$\mathrm{Corr}(\boldsymbol{Y}_i) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n_i} \\ \rho_{21} & 1 & \cdots & \rho_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i1} & \rho_{n_i2} & \cdots & 1 \end{pmatrix}$$

$\boldsymbol{X}_{ij}$ may include two main types of covariates: covariates whose values do not change throughout the duration of study (*e.g.* gender, fixed experimental treatments *etc.*) and covariates whose values change over time (*e.g.* time since baseline, current smoking status, environmental exposures *etc.*). The inclusion of time-varying covariates can raise subtle issues concerning the interpretation and estimation of the resulting models.

1.3 An Overview of the Thesis

In the previous part of this chapter, we first considered the importance and the necessity of longitudinal data analysis as an enhancement of the classical theory of data analysis, explaining the manner in which it is different from the cross sectional and time series analysis and mentioning the advantages of longitudinal data analysis over the other methods. Then we had an exploration of the advances in the area of longitudinal data analysis, highlighting the seminal works in the area. We collocate the seminal works that turned out to be the landmarks in the headway of longitudinal data analysis. All the important works that serve the foundation upon which the entire superstructure of longitudinal data analysis have been included. It further gives a brief introduction of various tools and models used in longitudinal analysis.

In Chapter 2, we consider the models for continuous data. Modelling the expected values and covariance structure are considered. Inferential procedures using maximum likelihood and restricted maximum likelihood are considered in the context of modelling expected values. For modelling the covariance structure the unstructured model as well as a large number of structured covariance models, which includes a few that are used little in practical applications, despite their attractive performances, are considered. For model selection besides the classical methods of AIC and BIC, the AIC_C that outperforms popularly used methods for small samples is also considered. A comparative study of various models considered in this chapter is made in chapter 5. Further procedures for fixed effects as well as for covariance structure are discussed. Parametric mixed effect models, random effects models and AR models are discussed under the framework of general multivariate models.

Chapter 3 considers the models for discrete data. Starting with the plausibility of the assumptions underlying the classical regression models, here we explain the necessity of alternative models to be used when the assumptions are likely to be violated and introduce the generalised linear models in this context. The two most popularly used generalised linear models *viz*, logistic regression for binary responses and log linear model for counts are considered.

Chapter 4 introduces the nonparametric methods longitudinal data analysis as an alternative to the usually used parametric methods for cases where the restrictive assumptions of parametric methods are not realistic. In this chapter we review four of the most popular smoothers that include local polynomial smoothers, regression splines, smoothing splines, and P-splines, and briefly describe linear smoothers, which include the above four popular smoothers as special cases.

In Chapter 5, we consider illustrations of the discussions made in earlier chapters using real life data, data from internet sites as well as simulated data. SAS codes using which the computations are done are listed. Performances of large number of covariance models are compared. We show that the UN model need not always be better than other models. The larger number of covariance parameters involved does not always make the UN model superior to others. Further we show that the heterogeneous models often performs better than the corresponding homogeneous models, that assume homogeneity of variance over time.

The thesis ends with the concluding remarks and direction for further work followed by the bibliography.

Chapter 2

Models for Continuous Longitudinal Data

2.1 Introduction

The last thirty years have seen remarkable advances in methods for analysing longitudinal and clustered data. The problems that we encounter in longitudinal studies are similar to those we face under the cross-sectional counterparts and may be classified as analysis of variance (ANOVA) or more generally as regression problems. The basic difference between the two approaches is that in cross sectional studies we deal with independent observations, while in the longitudinal cases it is necessary to consider a possible statistical dependence among them. There now exists a broad and flexible class of models for correlated data based on a regression paradigm.

In longitudinal study the data associated with each sampling unit may usually be expressed in terms of a vector with the values of the response variable at different measurement occasions and a matrix with the corresponding values of the explanatory (independent) variables which may be classificatory (*e.g.* treatment, sex *etc.*) or not (*e.g.* time, temperature *etc.*).

Longitudinal data present us with two aspects of data that require modelling: the mean response over time and the covariance among the repeated measures on the same individuals. Mean response is modelled using two main approaches *viz* the analysis of response profiles and parametric or semi-parametric curves. Overall modelling strategy that takes account of interdependence between the models for mean and covariates are also developed.

2.2 Modelling The Expected Values

A variety of models for Gaussian data have been developed during the past decades. They have well known properties and are excellent tools for practical applications. The general form of the model is

$$\boldsymbol{Y}_i = g(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}_i) + \boldsymbol{e}_i,\ i = 1, 2, \ldots, N, \tag{2.1}$$

where g is a convenient vector valued function, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ a vector of population parameters, $\boldsymbol{\gamma}_i = (\gamma_1, \gamma_2, \ldots, \gamma_q)'$ is a vector of random subject specific effects with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Gamma}$ and $\boldsymbol{e}_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})'$ is a vector

of random errors with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Xi}_i$. The within sample units covariance matrices $\boldsymbol{\Sigma}_i$'s are obtained as functions of $\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\Gamma}$ and $\boldsymbol{\Xi}_i$. In many instances, specifically for linear models, a simplification of (2.1) given by

$$\boldsymbol{Y}_i = g(\boldsymbol{X}_i, \boldsymbol{\beta}) + \boldsymbol{e}_i,\ i = 1, 2, \ldots, N, \tag{2.2}$$

is more serviceable. Here $\boldsymbol{\Sigma}_i = \boldsymbol{\Xi}_i$, so that the within sample covariance matrices are modelled independently of the location parameters.

The present state of methodological development favours situations in which the linear version of (2.2), namely

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i,\ i = 1, 2, \ldots, N, \tag{2.3}$$

where $\boldsymbol{X}_i$ is the design matrix for the i^{th} individual and $\boldsymbol{e}_i$ is a vector of deviations with a multivariate normal (MVN) distribution with mean vector $\mathbf{0}$ and dispersion matrix $\boldsymbol{\Sigma}$. The design matrix can contain functions of time and both within and within subject covariates. The design is said to be balanced if each subject is observed at the same p time points. When $\boldsymbol{X}_i$ is independent of i, the design is termed completely balanced. The linear location model (2.3) in combination with the different models for covariance structure is sufficiently general to handle a large variety of practical situations.

In most of the problems, the development of a linear model for the mean-value function is the primary goal of the analysis. The general representation $\boldsymbol{X}_i\boldsymbol{\beta}$ is familiar from ordinary linear regression.

Rao (1959, 1965, 1975) considered the problem of polynomial growth curve analysis of serial measurements from a single group of subjects. If the design is balanced, Rao's model can be written as $\boldsymbol{E(Y}_i\boldsymbol{)=A\beta}$, where the columns of $\boldsymbol{A}$ are powers of t (time) or polynomials defined by the times of observation. The rank of $\boldsymbol{A}$ equals the degree of the polynomial growth curve plus one. In the same setting Grizzle and Allen (1969) introduced covariates by defining the expected value as

$$\boldsymbol{E(Y}_i\boldsymbol{)=A\beta x}_i, \tag{2.4}$$

where $\boldsymbol{x}_i$ is the vector of covariate values for the i^{th} individual. If $\boldsymbol{x}_i$ is $q \times 1$, then $\boldsymbol{\beta}$ is $r \times q$, where r is the number of columns in $\boldsymbol{A}$. Model (2.4) can be written in the general form of (2.3) by defining $\boldsymbol{X}_i=\boldsymbol{x}_i'\otimes$A where $\otimes$ denotes the tensor product, and defining $\boldsymbol{\beta}'$ as $1 \times rq$ vector produced by writing out the elements of the $r \times q$ matrix row by row. If we think of $\boldsymbol{A}$ as the matrix whose columns contain powers of t_i, this representation shows that the Grizzle and Allen model assumes that every coefficient in the polynomial model depends on each element of $\boldsymbol{x}_i$. The representation suggests that this requirement can be relaxed by deleting columns containing specified products of powers $\boldsymbol{x}_i$ and powers of t from the design matrix. In short, the mean-value model can be determined directly by defining the expected value of each element of $\boldsymbol{Y}_{ij}$ as the desired function of the time of observation and the covariates.

This direct approach to modelling the mean-value function has some important advantages not offered by the special structure usually assumed for the growth model. First, individuals need not be observed at the same times or on the same number of occasions. Second, time varying covariates can be included in the model, provided

that their contribution to the expected response can be written linearly. Third, covariates can modify either the expected value of $\boldsymbol{Y}_i$ or the rate of change in $\boldsymbol{E(Y}_i\boldsymbol{)}$; the latter arises from interaction terms involving the covariate and the appropriate power of t.

2.2.1 Inference by Maximum Likelihood/REML

We have the general model of the form

$$\boldsymbol{Y}_i \sim MVN(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i). \tag{2.5}$$

Writing, $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m)'$, the joint density of $\boldsymbol{Y}$ is,

$$f(\boldsymbol{y}) = \prod_{i=1}^{m}(2\pi)^{-n_i/2}|\boldsymbol{\Sigma}_i|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}.$$

Maximising this likelihood is equivalent to minimising the quadratic form

$$\sum_{i=1}^{m}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \tag{2.6}$$

The MLE, which is same as the GLS estimator, can now be obtained as

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} X_i\Sigma_i^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i\Sigma_i^{-1}Y_i\right) \tag{2.7}$$

$$= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} \tag{2.8}$$

To estimate $\boldsymbol{\Sigma}_i$ (or θ) we equate the first derivative of the log-likelihood function to zero and solve it for θ. In general, this equation is non-linear and it is not possible to write down simple, closed form expression for the MLE of θ. Instead the MLE must be found by solving this equation using iterative technique or using computer algorithms.

The sampling distribution of $\widehat{\boldsymbol{\beta}}$ can be straightforwardly derived as

$$\widehat{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}). \tag{2.9}$$

MLE's of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_i$ have desirable large sample properties. But the MLE of $\boldsymbol{\Sigma}_i$ is biased for finite samples. The diagonal elements of $\boldsymbol{\Sigma}_i$ are underestimated. The theory of REML was developed to address this problem. The method of REML estimation, was introduced by Patterson and Thompson (1971) as a way of estimating variance components in a general linear models. The method uses an adjustment similar to changing the divisor in the linear regression to correct the bias of the estimator. The adjustment involves replacing the usual likelihood by the modified likelihood given by

$$\prod_{i=1}^{m}(2\pi)^{-n_i/2}|\boldsymbol{\Sigma}_i|^{-1/2}|\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i|\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}. \tag{2.10}$$

The estimator of $\boldsymbol{\beta}$ obtained by maximising this modified likelihood has the same form as the one obtained in (2.7). The difference is that $\boldsymbol{\Sigma}$ is now estimated by maximising (2.10).

2.3 Modelling The Covariance Structure

The correlation among longitudinal data necessitates appropriate modelling of the covariance or time dependence among the repeated measures obtained on the same individuals. This enables getting correct standard errors and making valid inferences about the regression parameters. Accounting for the covariance among the repeated measures usually increases the precision with which the regression parameters can be estimated, that is, the positive correlation among the repeated measures reduces the variability of the estimate of the change over time within individuals. In addition, when there are missing data, correct modelling of the covariance is often a requirement for obtaining valid estimates of the regression parameters. In general, the failure to take into account of the covariance among the repeated measures will result in incorrect estimates of the sampling variability and can lead to quite misleading inferences. The correlation between observations taken at two instants are expected to decrease as the distance between them decreases. See Kenward (1987), Lindsey (1993) or Diggle *et al* (1994) for practical examples.

The choice of models for the mean response and the covariance are interdependent. This interdependence arises because the vector of residuals (observed responses minus fitted responses) depends upon the specification of the model for the mean. Stating more precisely, the covariance between the pair of residuals, say $\{Y_{ij} - \mu_{ij}(\boldsymbol{\beta})\}$ and $\{Y_{ik} - \mu_{ik}(\boldsymbol{\beta})\}$, depends on the model for mean (*i.e.* depends on $\boldsymbol{\beta}$). As a result of this interdependence between the models for mean and covariance, we will need to develop an overall modelling strategy that takes this interdependence into account.

When $\boldsymbol{\Sigma}_i = \text{Cov}(\boldsymbol{e_i})$ are known, Aitkin's generalized least square estimator of $\boldsymbol{\beta}$ can be written as

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Y}_i \right). \tag{2.11}$$

When the covariate structure is unknown, most estimation procedures lead to estimators of the form (2.11) with estimates substituted for the population covariance matrix $\boldsymbol{\Sigma}_i$. Thus the estimation problem reduces to the problem of modelling and estimating $\boldsymbol{\Sigma}_i$.

Three broad approaches to modelling the covariance among the repeated measures are distinguished: (1) unstructured covariance, (2) covariance pattern models, and (3) random effects covariance structures. The first is to allow any arbitrary pattern of covariance among the repeated measures. The second and third approaches place structure on the covariance matrix. The key issue is to choose a covariance model $\Sigma_i(\theta)$ that is correct, but parsimonious, meaning that (a) the true covariance matrix is in the family defined by the covariance model and that (b) vector parameter θ has as few elements as possible. In this section we present how different covariance models can be used to model the dependence among the repeated measures on same individuals over time. We study a large number of covariance models, and explore the comparative performance of these models.

2.3.1 The Unstructured Covariance Model

The unstructured covariance (UN) matrix Σ is the most general covariance matrix possible. Covariance matrix is said to follow unstructured model, if there is no apparent systematic pattern of variance and correlation. With n measurement occasions, the unstructured covariance matrix has $\frac{n(n+1)}{2}$ parameters: the n variances at each occasion and the $\frac{n(n-1)}{2}$ pairwise covariances,

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix}.$$

The number of parameters is large, and they can be hard to estimate when the number of observations is small relative to $\frac{n(n+1)}{2}$.

2.3.2 Covariance Pattern Models

It might seem desirable to use an unstructured covariance model to model the covariance matrix of longitudinal data as it would seem guaranteed to model the data correctly. Unfortunately, use of the unstructured covariance model is usually not feasible nor recommended. For random observation times there are too many variance and covariance parameters, and they cannot all be estimated. The variance of responses at time t_{ij} will not be estimable if there is zero or one observation at that time. When the data set is balanced with possibly missing data, when the sample

size n is large and the number of repeated measures J is small or modest compared with n, then we can use the unstructured covariance model. Most data sets do not meet these constraints, and we need to use a parameterized covariance model.

A parameterized covariance matrix is one where all variances and covariances are functions of a small to moderate number of covariance parameters θ. The covariance model $\Sigma(\theta)$ defines a family of possible covariance matrices with members of the family indexed by θ. The key issue is to choose a covariance model $\Sigma(\theta)$ that is correct, but parsimonious, meaning that (a) the true covariance matrix is in the family defined by the covariance model and that (b) vector parameter has as few elements as possible.

i. Compound Symmetry (Uniform) Models

With a compound symmetry or uniform covariance it is assumed that the variances of the response variable at all observation instants are equal, say σ^2, and so are the covariances between the response variables observed at any two instances. Consequently $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$ for all j and k. That is

$$\text{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix} = (1-\rho)I + \rho J,$$

with the constraint that $\rho \geq 0$, where J is an $n \times n$ unit matrix. The compound symmetry covariance model has two parameters, σ^2 and ρ.

The key feature of the compound symmetry matrix is that for any time lag $t_{ij} - t_{il} \neq 0$ large or small, the correlation $\text{Corr}(Y_{ij}, Y_{il})$ is the same. This means that observations taken a few minutes apart and those taken a few years apart have the same correlation. This is unlikely for real data measured on human beings over long enough periods of time. In practice, measures that are very persistent over the data collection time frame may follow a compound symmetry covariance model. Weight might be an example; as adults, our weight often does not change over a period of years, other than changes due to typical daily energy consumption and expenditure.

ii. Toeplitz

The Toeplitz covariance pattern makes the assumption that any pair of responses that are equally separated have the same correlation and that the variance is constant across time. That is the lag i correlation is different from the lag j correlation, but all lag i correlations are the same. That is, $\text{Corr}(Y_{ij}, Y_{ij+k}) = \rho_k$ for all j and k and

$$\text{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}.$$

This structure is appropriate only when the measurements are made at equal or approximately equal intervals of time, since correlation among adjacent measurement occasions is a constant, ρ_1. The structure has n parameters (1 variance parameter and

$n-1$ correlation parameters). The first order autoregressive covariance is a special case of the Toeplitz covariance.

iii. Autoregressive

In the autoregressive model for the covariance, it is assumed that variance is a constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{ij+k}) = \rho^k$ for all j and k and $\rho \geq 0$. That is

$$\text{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}.$$

The autoregressive covariance has only two parameters, regardless of the number of measurement occasions. Because it has a Toeplitz form, it is appropriate only when the measurements are made at equal/approximately equal intervals of time. In the AR model, the correlation between two observations Y_{ij} and Y_{il} depends on the absolute value of the time between them:

$$\text{Corr}(Y_{ij}, Y_{il}) = \rho^{|t_{ij} - t_{il}|}.$$

The farther apart two observations are in time, the lower the correlation between them.

iv. Banded

Banding makes assumption about how quickly the correlation decreases to zero with increasing separation between the repeated measurements. The banded covariance pattern makes the assumption that the correlation is zero beyond some specified interval. For example, a banded covariance pattern with a band size of 3 assumes that $\text{Corr}(Y_{ij}, Y_{i,j+k}) = 0$ for $k \geq 3$. It is possible to apply a banded pattern to any of the covariance pattern models considered so far. Thus, a banded Toeplitz covariance pattern with a band size of 2 is given by

$$\text{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & 0 & \cdots & 0 \\ \rho_1 & 1 & \rho_1 & \cdots & 0 \\ 0 & \rho_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

v. Exponential

The formulation of autoregressive covariance model can be generalised to the case of unequally spaced measurement occasions as follows. Let $(t_{i1}, t_{i2}, \ldots, t_{in})$ denote the observation times for the i^{th} individual and assume that the variance is a constant across all measurement occasions, say σ^2, and

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|}, \text{ for } \rho \geq 0. \tag{2.12}$$

That is the correlation between any pair of repeated measures decreases exponentially with the time separation between them. This structure is known as an exponential covariance model because it can be re-expressed as (Jones (1993))

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \sigma^2 \rho^{|t_{ij} - t_{ik}|} && (2.13) \\ &= \sigma^2 \exp(-\theta\, |t_{ij} - t_{ik}|), && (2.14) \end{aligned}$$

where $\theta = -\log(\rho)$ or $\rho = \exp(-\theta)$ for $\theta \geq 0$. The Exponential model assumes that the correlation is one if measurements are made repeatedly at the same occasion and that the correlation decreases rapidly to zero as the time separation increases. The first aspect corresponds to an assumption that the responses are measured without error; an unrealistic assumption in most longitudinal studies in health sciences. The latter also is rarely observed in longitudinal studies.

vi. Antedependence

The antedependence model is a generalization of the autoregressive model. It allows for non-constant lag 1 correlations over time. For balanced data with J time points, there are $J-1$ correlation parameters and a variance parameter σ^2. Each correlation parameter ρ_j is the lag 1 correlation between observations at times t_j and t_{j+1}. Higher lag correlations are the product of the intervening lag 1 correlations. For example, It has four correlation parameters $\rho = (\rho_1, \rho_2, \rho_3, \rho_4)$ for data with 5 repeated measures

and the entire covariance matrix is

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_1\rho_2 & \rho_1\rho_2\rho_3 & \rho_1\rho_2\rho_3\rho_4 \\ \rho_1 & 1 & \rho_2 & \rho_2\rho_3 & \rho_2\rho_3\rho_4 \\ \rho_1\rho_2 & \rho_2 & 1 & \rho_3 & \rho_3\rho_4 \\ \rho_1\rho_2\rho_3 & \rho_2\rho_3 & \rho_3 & 1 & \rho_4 \\ \rho_1\rho_2\rho_3\rho_4 & \rho_2\rho_3\rho 4 & \rho_3\rho_4 & \rho_4 & 1 \end{pmatrix}.$$

It should be clear that if the lag 1 correlations ρ_j's are the same, then we are right back at the AR(1) correlation model.

vii. Autoregressive Moving Average

The autoregressive moving average or ARMA(1,1) model is a generalization of both the AR(1) model and the CS model. The ARMA(1,1) model has three parameters, a variance parameter $\mathrm{Var}(Y_{ij}) = \sigma^2$ and two correlation parameters, γ and ρ. The first correlation parameter γ is the lag one correlation

$$\mathrm{Corr}(Y_{ij}, Y_{i(j-1)}) = \gamma,$$

while ρ is the additional decrease in correlation for each additional lag. The lag k correlation is

$$\mathrm{Corr}(Y_{ij}, Y_{i(j-k)}) = \gamma\rho^{k-1}.$$

The full covariance matrix is

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 & \gamma\rho^3 \\ \gamma & 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma\rho & \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^3 & \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{pmatrix}.$$

There are at least three interesting special cases of the ARMA(1,1) model.

1. $\rho = 1$ is the compound symmetry model,
2. $\gamma = \rho$ is the autoregressive model, and
3. $\rho = 0$ is the moving average or MA model.

viii. Factor Analytic

The covariance matrix of the q^{th} order factor analytic (FA) structure is of the form $\Lambda'\Lambda + \boldsymbol{D}$, where Λ is a $n \times q$ rectangular matrix and $\boldsymbol{D}$ is a $n \times n$ diagonal matrix with n different parameters. When $q > 1$, the elements of Λ in its upper right-hand corner (that is, the elements in the i^{th} row and j^{th} column for $j > i$) are set to zero to fix the rotation of the structure. The covariance matrix of the first order factor

analytic structure (FA(1)), is of the form

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \lambda_1^2 + d_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 \\ \lambda_2\lambda_1 & \lambda_2^2 + d_2 & \lambda_2\lambda_3 & \lambda_2\lambda_4 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2 + d_3 & \lambda_3\lambda_4 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2 + d_4 \end{pmatrix}$$

The FA structure has two other forms. The no diagonal FA (FA0(q)) structure is similar to FA(q) structure except that no diagonal matrix $\boldsymbol{D}$ is included. When $q < n$, that is, when the number of factors is less than the dimension of the matrix, this structure is nonnegative definite but not of full rank. The equal diagonal FA (FA1(q)) structure is similar to the FA(q) structure except that all of the elements in $\boldsymbol{D}$ are constrained to be equal. This offers a useful and more parsimonious alternative to the full factor-analytic structure.

ix. Huynh-Feldt

The Huynh-Feldt covariance structure is similar to the CSH structure in that it has the same number of parameters and heterogeneity along the main diagonal. However, it constructs the off-diagonal elements by taking arithmetic rather than geometric means.

$$\text{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma_1^2 & \frac{\sigma_1^2+\sigma_2^2}{2} - \lambda & \frac{\sigma_1^2+\sigma_3^2}{2} - \lambda \\ \frac{\sigma_2^2+\sigma_1^2}{2} - \lambda & \sigma_2^2 & \frac{\sigma_2^2+\sigma_3^2}{2} - \lambda \\ \frac{\sigma_3^2+\sigma_1^2}{2} - \lambda & \frac{\sigma_3^2+\sigma_2^2}{2} - \lambda & \sigma_3^2 \end{pmatrix}$$

Remark 2.3.1. *Many of the covariance pattern models make strong assumptions about homogeneity of variance over time. But this assumption is not always realistic as longitudinal data often exhibit heterogeneity or non-constant variance over time. Practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. Some of the models discussed here can be extended by relaxing this assumption, leading to corresponding heterogeneous models. It can be shown that generally heterogeneous models outperforms the corresponding homogeneous models.*

The Valid values for covariance-structure in SAS and their descriptions along with the number of covariance parameters involved and the expression for the $(i, j)^{th}$ element of the covariance matrix are provided in Table 2.1.

2.4 Implication of Correlation among Longitudinal Data

The positive correlation among repeated measures can be used to advantage in the study of change over time. That is, we can capitalise on the positive correlation among the longitudinal data when the main focus of the analysis is on change in the mean response. Consider a sample longitudinal study where the change in the health outcome before and after receiving a health intervention is of interest. With only two

Table 2.1: Covariance Structures

Structure	Description	No of Cov Parametrers	$(i,j)^{th}$ element
UN	Unstructured	n(n+1)/2	σ_{ij}
UN(q)	Banded	[q/2](2n-q+1)	$\sigma_{ij}I(\|i-j\|<q)$
CS	Compound Symmetry	2	$\sigma_1^2+\sigma^2 I(i=j)$
CSH	Heterogeneous CS	n+1	$\sigma_i\sigma_j[\rho I(i\neq j)+I(i=j)]$
AR(1)	Autoregressive(1)	2	$\sigma^2\rho^{\|i-j\|}$
ARH(1)	Heterogeneous AR(1)	n+1	$\sigma_i\sigma_j\rho^{\|i-j\|}$
TOEP	Toeplitz	n	$\sigma_{\|i-j\|+1}$
TOEP(q)	Banded Toeplitz	q	$\sigma_{\|i-j\|+1}I(\|i-j\|<q)$
TOEPH	Heterogeneous TOEP	2n-1	$\sigma_i\sigma_j\rho_{\|i-j\|}$
TOEPH(q)	Banded Hetero TOEP	n+q-1	$\sigma_i\sigma_j\rho_{\|i-j\|}I(\|i-j\|<q)$
ANTE(1)	Antedependence	2n-1	$\sigma_i\sigma_j\prod_{k=i}^{j-1}\rho_k$
ARMA(1,1)	ARMA(1,1)	3	$\sigma^2[\gamma^{\|i-j\|-1}I(i\neq j)+I(i=j)]$
VC	Variance Components	q	$\sigma_k^2 I(i=j)$
FA(q)	Factor Analytic	[q/2](2n-q+1)+n	$\sum_{i=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}+\sigma_i^2 I(i=j)$
FA0(q)	No diagonal FA	[q/2](2n-q+1)	$\sum_{i=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}$
FA1(q)	Equal diagonal FA	[q/2](2n-q+1)+1	$\sum_{i=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}+\sigma^2 I(i=j)$
HF	Huynh-Feldt	n+1	$(\sigma_i^2+\sigma_j^2)/2+\lambda I(i\neq j)$

repeated measures, the analysis will focus on the difference score, say $Y_{i2}-Y_{i1}$, whose variance is given by

$$\begin{aligned} V(Y_{i2}-Y_{i1}) &= V(Y_{i1})+V(Y_{i2})-2\ \mathrm{Cov}(Y_{i1},Y_{i2}) \\ &= \sigma_1^2+\sigma_2^2-2\ \rho_{_{12}}\sigma_1\sigma_2, \end{aligned} \tag{2.15}$$

where $\rho_{_{12}}$ is the correlation among the pair of responses, Y_{i1} and Y_{i2}.

On the other hand, if a cross sectional study is adopted, the study participants are assigned to two groups, a group that receives the intervention and a control group

that does not. Then the variance of the difference between the responses of any two individuals, when one individual is randomly selected from the intervention group and the other from the second group, is given by

$$\begin{aligned} V(Y_{i2} - Y_{i1}) &= V(Y_{i1}) + V(Y_{i2}) \\ &= \sigma_1^2 + \sigma_2^2. \end{aligned} \tag{2.16}$$

Thus the variability of the within individual difference is always substantially smaller than the variability of the between individual differences, provided the correlation is relatively large and positive. It is in this sense that a longitudinal study can provide a more precise (*i.e.,* less variable) estimate of change in the mean response than a cross-sectional study with the same number and pattern of observations.

Further, failure to adequately account for correlation among repeated measures can result in misleading inferences, as it renders the standard error incorrect. With incorrect standard errors, test statistics and p-values will also be incorrect and they may lead to incorrect inferences.

2.5 Model Selection

Model selection involves the choice of an appropriate model from among a set of candidate models. Model selection is used when there is no particular clear choice among many different models. Model selection tools are a useful set of techniques for screening through the many different covariance models. The choice among models

can be made by comparing the maximum likelihoods for each of the covariance pattern models. Fixed effects models and covariance models are rather different; they are usually treated differently in modelling, and we discuss model selection for the covariance parameters and for the fixed effects separately. The covariance models for our data are often of secondary interest; we do want to pick the best covariance model for our data but if we make a modest error in covariance model specification it is not as costly as an error in the fixed effects specification.

2.5.1 Model Selection for The Covariance Model

When faced with a new data set, we want to determine a best covariance model for use in fitting the fixed effects, and we try out a large number of covariance models. Our goal is typically to pick a single best or most useable model for use in further analyses of fixed effects. While we do our covariance model selection, we will pick a single set of covariates and try out many different covariance models. We compare the performance of various covariance structures, by comparing the maximum likelihoods, AIC, AIC_C, and BIC for the covariance pattern models. We compare nested covariance models to each other, using likelihood ratio tests. To compare non-nested models, the most popular covariance model selection criteria AIC (Akaike information criterion), AIC_C (Corrected Akaike information criterion) and BIC(Bayes information criterion) are used.

In practice, a number of models may be considered for fitting the same longitudinal data. Hence, Jones (1993, Section 2.8), Singer and Willett (2003, Section 4.6),

and Fitzmaurice *et al* (2004, Section 7.5) recommended adopting the Akaike information criterion (AIC) (Akaike, 1973) and the Bayesian information criterion (BIC) (Schwarz, 1978) to select the best model from all possible candidate models. It is known that AIC is an efficient criterion, while BIC is a consistent criterion (McQuarrie and Tsai, 1998; Burnham and Anderson, 2002). There is no general agreement on which of these two categories of criteria is preferable. However, the performance of both AIC and BIC is often unsatisfactory when the sample size is small. In classical linear regression models, Hurvich and Tsai (1989) proposed an efficient criterion, the corrected Akaike information criterion (AIC_C), by directly minimising the expected Kullback-Leibler discrepancy. They showed that AIC_C outperforms AIC in small samples, while performing comparably to AIC in large samples. The criterion is asymptotically efficient if the true model is infinite dimensional. Furthermore, when the true model is of finite dimension, AIC_C is found to provide better model order choices than any other asymptotically efficient method. More recently, Shi and Tsai (2002) proposed a consistent model selection criterion, the residual information criterion (RIC), based on the expected Kullback-Leibler discrepancy of the residual log-likelihood function. They demonstrated that in linear regression models with the first order autoregressive process, RIC is superior to BIC when the signal-to-noise (SNR) ratio is not weak.

We apply the likelihood and the residual likelihood approaches to derive model selection criteria, AIC_C and RIC, respectively, for linear regression models with longitudinal data.

One approach to model selection is to use likelihood ratio tests (section 6.1

Weiss (2005)). Using likelihood ratio tests, we can compare nested covariance models to each other.

Two covariance pattern models are said to be nested when one (the reduced) model is a special case of the other (full) model. For pair of nested models, a likelihood ratio test statistic can be constructed that compares the "full" and "reduced" models. The likelihood ratio test is obtained by taking twice the difference in the maximised REML log-likelihood as

$$G^2 = 2(\widehat{l}_{full} - \widehat{l}_{red}).$$

This statistic is compared to the percentiles of a chi-square distribution with degrees of freedom equal to the difference between the number of covariance parameters in the full and reduced models.

However, a number of covariance models are not nested within each other and have the same or similar numbers of parameters. A number of criterion-based approaches to model selection have been developed. Criterion-based model selection approaches compare adjusted log likelihoods penalized for the number of parameters in the covariance model. The penalty for model m increases with the number of covariance parameters q_m. Models with more covariance parameters should fit better, meaning they should naturally have a higher log likelihood, than models with fewer parameters. The penalty function levels the playing field compared to what would happen if we compared models using raw log likelihood. The model with the best score on the criterion is selected as "best." To compare non-nested models, two most popular covariance model selection criteria are AIC (*Akaike information criterion*)

and BIC(*Bayes information criterion*). AIC for a given model m is defined as

$$AIC(m) = -2 \text{ log likelihood}(m) + 2q_m \tag{2.17}$$

and BIC is nearly identical, except that instead of the 2 multiplying the number of covariance parameters q_m, the penalty is log(N)

$$BIC(m) = -2 \text{ log likelihood}(m) + \log(N)q_m. \tag{2.18}$$

The model with the smallest value of AIC or BIC is selected as best. Decision theory suggests that, as sample size increases, we should use a decreasing type-one error rate. BIC does this for us automatically, by penalizing the model according to both the number of parameters q_m and the number of observations N. The log function does grow slowly in the number of observations, and so the penalty does not grow quickly in N. Because of its larger penalty on the number of parameters, BIC tends to pick smaller models than AIC does and so BIC more than AIC tends to pick the null model when the null model is in fact correct.

Conversely, for smaller data sets, one should make additional assumptions when modelling data; as the sample size increases, one should relax the assumptions and allow the data to determine modelling assumptions. In specifying a covariance model, additional assumptions means choosing a more parsimonious covariance model, one with fewer parameters. As the sample size n increases, one should allow for more complex covariance models with more unknown parameters. BIC goes against this advice, at least in comparison to AIC. Simulation studies have suggested that BIC

tends to select a model with too few parameters. On the other hand, having too many extra parameters in the covariance model can interfere with inferences.

There are 4 equivalent variants of *AIC* and *BIC*. We consider the versions of *AIC* and *BIC* in the smaller is the better form. Some programs multiply these definitions by -1 and have a "larger is better" version, and some programs may multiply *AIC* and *BIC* by $\pm$ 0.5 in the definition.

As r, the dimension (number of parameters) of the candidate model, increases in comparison to n, the sample size, AIC becomes a strongly negatively biased estimate of the information. A bias corrected to the Akaike information criterion, AIC, is derived by Hurvich and Tsai (1989). The correction is of particular use when the sample size is small, or when the number of fitted parameters is a moderate to large fraction of the sample size. For linear regression, the corrected method, called AIC_C, is exactly unbiased. In all cases, the reduction in bias is achieved without any increase in variance, since AIC_C may be written as the sum of AIC and a nonstochastic term. Among the efficient methods studied, AIC_C is found to perform best. For small samples, AIC_C is able to out-perform even the consistent methods. The bias corrected AIC, termed as AIC_C, is defined as

$$AIC_C = AIC + \frac{2(r+1)(r+2)}{n-r-2} \tag{2.19}$$

Thus AIC_C is the sum of AIC and an additional non-stochastic penalty term.

Model selection for covariance parameters is typically done with REML log likelihoods; as such, the log likelihoods in the formulas for AIC, AIC_C and BIC should be

REML likelihoods, q_m is the number of covariance parameters, and the fixed effects model must be the same in all models.

In covariance model selection, it is important that all models have the same fixed effects. Otherwise model testing or comparison procedures would be comparing both different fixed effects and different covariance parameters at the same time.

2.5.2 Model Selection Using Bayesian Probability

Bayesian testing can be used to compare nested or non-nested models. This method can compare two or more models at the same time. A Bayesian test of two models 1 against 2 is a probability statement that model 1 is the correct model given that either 1 or 2 is correct. BIC can also be used to approximate the probability that a given covariance model among a set of covariance models is correct. Let BIC(1) and BIC(2) be the BIC's for two models. The probability that model 1 is the correct model is

$$\begin{aligned} P(1|Y) &\approx \frac{\exp[-.5BIC(1)]}{\exp[-.5BIC(1)] + \exp[-.5BIC(2)]} \\ &= \frac{1}{1 + \exp\{-.5[BIC(2) - BIC(1)\}]}, \end{aligned}$$

and the probability that model 2 is correct is $P(2|Y) = 1 - P(1|Y)$. What is important in this calculation is the difference of BIC's. For example, if the difference, BIC(2)$-$ BIC(1) = 6, then $P(1|Y) = 0.9526$ and if the difference is greater than 12, $P(1|Y) > 0.9975$. If BIC(2)$-$BIC(1) =0, the two models are equally likely and $P(1|Y) =$

$P(2|Y) = 0.5$. Given K models, the probability that model k is the correct model is

$$P(k|Y) \approx \frac{\exp[-.5BIC(k)]}{\sum_{j=1}^{K} \exp[-.5BIC(j)]}.$$

2.5.3 Model Selection for Fixed Effects

The huge variety of different circumstances involving fixed effects makes, giving advice for model selection for fixed effects, more complex than for covariance models. The first step in choosing the fixed effects is to determine if there is a strong time trend. If there is, we specify fixed effects to adequately describe the time trend. This time trend must be included in all further analyses including covariance model specification.

Now suppose we wish to wade through a large number of covariates and to include those that are predictive of the response. The traditional model selection methods of forward selection and backward elimination can be used with longitudinal models. Suppose we are contemplating a finite set of potential additional covariates. Forward selection works by first specifying a base model. Each potential covariate is added to the base model in turn. The covariate with the most significant t or F statistic is added to the base model and this new base plus covariate model becomes the new base model. The remaining covariates are added in to this model one at a time and the most significant is included in the model. The process continues until no remaining covariate is significant enough to be included in the model. A significance level is set as a minimum level required to allow a new covariate into the model.

Backward selection works similarly, but starts with a model with all of the candidate covariates as predictors. The least significant covariate is dropped and the model is refit. The least significant covariate in the new model is dropped, unless its significance level is above some minimum level. Many variations of forward selection and backward elimination have been suggested, included algorithms that switch between forward and backward steps in various orders.

2.6 General Multivariate Models

When the design is balanced and there is no theoretical or empirical basis to assume special covariance structure, one need assume only that $\mathrm{Cov}(\boldsymbol{e}_i) = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is an arbitrary positive definite covariance matrix. Kleinbaum (1973) investigated multivariate methods for estimating the mean and covariance matrix in unbalanced data sets. This approach breaks down when the set of observation times become large relative to the number of individuals: estimation of the many parameters in the covariance matrix becomes computationally burdensome and the resulting estimators of location parameters are inefficient when simpler covariance structures apply. Thus when the data set is highly unbalanced or incomplete or when p is large relative to n, more parsimonious models for the covariance structure must be considered. Two natural candidates for the purpose are parametric mixed-effects models which includes random effects as a special case and autoregressive (AR) models.

2.6.1 Parametric Mixed-Effects Models

Parametric mixed-effects models or random-effects models are powerful tools for longitudinal data analysis. Linear and nonlinear mixed-effects models (including generalized linear and nonlinear mixed-effects models) have been widely used in many longitudinal studies. Good surveys on these approaches can be found in the books by Searle *et al* (1992), Davidian and Giltinan (1995), Vonesh and Chinchilli (1996), Verbeke and Molenberghs (2000), Pinheiro and Bates (2000), Diggle *et al* (2003), and Demidenko (2004), among others. In this section, we shall review various parametric mixed-effects models and emphasize the methods that we will use in later chapters.

2.6.2 Linear Mixed-Effects Model

In previous sections, we have introduced models for longitudinal data where changes in the mean response, and their relation to covariates, can be expressed as

$$E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \tag{2.20}$$

and where the primary goal is to make inferences about the population regression parameters, $\boldsymbol{\beta}$. In this section we consider an alternative, but closely related, approach for analyzing longitudinal data using linear mixed effects models. The underlying premise of linear mixed effects models is that some subset of the regression parameters vary randomly from one individual to another, thereby accounting for sources of natural heterogeneity in the population. That is, individuals in the population are

assumed to have their own subject-specific mean response trajectories over time and a subset of the regression parameters are now regarded as being random. The distinctive feature of linear mixed effects models is that the mean response is modelled as a combination of population characteristics, $\boldsymbol{\beta}$, that are assumed to be shared by all individuals, and subject-specific effects that are unique to a particular individual. The former are referred to as fixed effects, while the latter are referred to as random effects. The term mixed is used in this context to denote that the model contains both fixed and random effects.

Because linear mixed effects models explicitly distinguish between fixed and random effects, they allow the analysis of between-subject and within-subject sources of variation in the longitudinal responses. In addition, it is not only possible to estimate parameters that describe how the mean response changes in the population of interest, but it is also possible to predict how individual response trajectories change over time. For example, linear mixed effects models can be used to obtain predictions of individual growth trajectories over time. The latter will be of interest when the focus of inference is on the individual rather than the population of individuals. For example, in the physician-patient context, these predictions can be used to identify those patients who do not respond well to their assigned treatment in a clinical trial.

Model Specification

Harville (1976, 1977) and Laird and Ware (1982) first proposed the following general linear mixed-effects (LME) model:

$$
\begin{aligned}
y_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i + \epsilon_{ij}, \\
\boldsymbol{b}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i), \\
j &= 1, 2, \ldots, n_i; \quad i = 1, 2, \ldots, n,
\end{aligned}
\tag{2.21}
$$

where $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, \ldots, \epsilon_{in_i}]'$, y_{ij} and ϵ_{ij} denote the response and the measurement error of the j^{th} measurement of the i^{th} subject, the unknown parameters $\boldsymbol{\beta} : p \times 1$ and $\boldsymbol{b}_i : q \times 1$ are usually called the fixed-effects vector and random-effects vectors, respectively (for simplicity, they are often referred to fixed-effects and random-effects parameters of the LME model), and $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are the associated fixed-effects and random-effects covariate vectors. In the above expression, $\boldsymbol{D}$ and $\boldsymbol{R}_i, i = 1, 2, \ldots, n$ are known as the variance components of the LME model. In the above LME model, for simplicity, we assume that $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are independent with normal distributions, and the between-subject measurements are independent.

The LME model (2.21) is often written in the following form:

$$
\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \\
\boldsymbol{b}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i), \\
i &= 1, 2, \ldots, n,
\end{aligned}
\tag{2.22}
$$

where $\boldsymbol{y}_i = [y_{il}, \ldots, y_{in}]'$, $\boldsymbol{X}_i = [\boldsymbol{x}_{il}, \ldots, \boldsymbol{x}_{in_i}]'$, and $\boldsymbol{Z}_i = [\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i}]'$.

The above LME model includes linear random coefficient models (Longford 1993) and models for repeated measurements as special cases. For example, a two-stage linear random-coefficient model for growth curves (Longford 1993) can be written as

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i), \\ & i = 1, 2, \ldots, n, \end{aligned} \tag{2.23}$$

where $\boldsymbol{y}_i, \boldsymbol{Z}_i, \boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are similarly defined as in (2.22), $\boldsymbol{\beta}_i$ is a $q \times 1$ vector of random coefficients of the i^{th} subject, and $\boldsymbol{A}_i$ is a $q \times p$ design matrix containing between-subject covariates. It is easy to see that the linear random-coefficient model (2.23) can be written into the form of the general LME model (2.22) once we set $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{A}_i, i = 1, 2, \ldots, n$.

In fact, we can write a general two-stage linear random coefficient model into the form of the general LME model (2.22). A general two-stage random coefficient model can be written as (Davidian and Giltinan 1995, Vonesh and Chinchilli 1996)

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i), \\ & \qquad i = 1, 2, \ldots, n, \end{aligned} \tag{2.24}$$

where $\boldsymbol{B}_i$ is a $q \times k$ design matrix with elements of 0's and 1's arranged to determine the components of $\boldsymbol{\beta}_i$, that are random, and $\boldsymbol{b}_i$ is the associated k-dimensional random effects vector. This general two-stage random-coefficient model can be written into the form of the general LME model (2.22): $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i^*\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$ once we set $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{A}_i$ and $\boldsymbol{Z}_i^* = \boldsymbol{Z}_i\boldsymbol{B}_i,\ i = 1, 2, \ldots, n$. In fact, we can easily show that the general two-stage random coefficient model (2.24) is equivalent to the general LME model (2.22). In particular, when $\boldsymbol{B}_i = \boldsymbol{I}_q$, the general two-stage random coefficient

model (2.24) reduces to the random coefficient model (2.23) for growth curves. Notice that the general two-stage random coefficient model (2.24) is also known as a two-stage mixed-effects model and the general LME model (2.22) is also called a hierarchical linear model.

In matrix notation, the general LME model (2.22) can be further written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \\ \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{D}), &\quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{R}), \end{aligned} \tag{2.25}$$

where

$$\begin{aligned} \boldsymbol{y} &= [\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n']', & \boldsymbol{b} &= [\boldsymbol{b}_1', \ldots, \boldsymbol{b}_n']', \\ \boldsymbol{\epsilon} &= [\boldsymbol{\epsilon}_1', \ldots, \boldsymbol{\epsilon}_n']', & \boldsymbol{X} &= [\boldsymbol{X}_1', \ldots, \boldsymbol{X}_n']', \\ \boldsymbol{Z} &= \mathrm{diag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n), & \widetilde{\boldsymbol{D}} &= \mathrm{diag}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n), \\ \boldsymbol{R} &= \mathrm{diag}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n). \end{aligned} \tag{2.26}$$

It is usually assumed that the repeated measurements from different subjects are independent and they are correlated only when they come from the same subject. Based on the general LME model (2.25), we have $\mathrm{Cov}(\boldsymbol{y}) = \mathrm{diag}(\mathrm{Cov}(\boldsymbol{y}_1), \ldots, \mathrm{Cov}(\boldsymbol{y}_n))$ where the covariance matrix of repeated measurement vector $\boldsymbol{y}_i$ for the i^{th} subject is $\mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i$. We can see that the correlation among the repeated measurements can be induced either through the between-subject variation term $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'$ or through the within-subject covariance matrix $\boldsymbol{R}_i$. Thus, even if the intra-subject measurement errors $(\epsilon_i, i = 1, 2, \ldots, n)$ are independent, the repeated measurements $\boldsymbol{y}_i$ may be still correlated due to the between-subject variation. In

some problems, the correlation may come from both sources. However, for simplicity, we may assume that the correlation is induced solely *via* the between-subject variation or assume that $\boldsymbol{R}_i$ is diagonal in the development of methodologies.

Estimation of Fixed and Random-Effects

The inferences for $\boldsymbol{\beta}$ and $\boldsymbol{b}_i, i = 1, 2, \ldots, n$ for the general LME model (2.22) can be based on the likelihood method or generalized least squares method. For known $\boldsymbol{D}$ and $\boldsymbol{R}_i, i = 1, 2, \ldots, n$, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i, i = 1, 2, \ldots, n$ may be obtained by minimising the following twice negative logarithm of the joint density function of $\boldsymbol{y}_i,\ i = 1, 2, \ldots, n$ and $\boldsymbol{b}_i,\ i = 1, 2, \ldots, n$ (upto a constant):

$$\begin{aligned} GLL(\boldsymbol{\beta}, \boldsymbol{b}_i|\boldsymbol{y}) = \sum_{i=1}^{n} \ & \big\{[\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i]'\boldsymbol{R}_i^{-1}[\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i] \\ & + \boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i + \log|\boldsymbol{D}| + \log|\boldsymbol{R}_i|\big\} \end{aligned} \tag{2.27}$$

Since $\boldsymbol{b}_i, i = 1, 2, \ldots, n$ are random-effects parameter vectors, the expression (2.27) is not a conventional log-likelihood. For convenience, from now on and throughout this book, we call (2.27) a generalized log-likelihood (GLL) of the mixed-effects parameters $(\boldsymbol{\beta}, \boldsymbol{b}_i, i = 1, 2, \ldots, n)$. Note that the first term of the right-hand side of (2.27) is a weighted residual taking the within-subject variation into account, and the term $\boldsymbol{b}_i'\boldsymbol{D}^{-1}\boldsymbol{b}_i$ is a penalty due to random-effects $\boldsymbol{b}_i$ taking the between-subject variation into account.

For given $\boldsymbol{D}$ and $\boldsymbol{R}_i, i = 1, 2, \ldots, n$, minimising the GLL criterion (2.27) is equivalent to solving the so-called mixed model equations (Harville 1976, Robinson 1991):

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & Z'\boldsymbol{R}^{-1}\boldsymbol{Z}+\widetilde{\boldsymbol{D}}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix},$$

where $\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{X}, \boldsymbol{Z}, \widetilde{\boldsymbol{D}}$ and $\boldsymbol{R}$ are defined in (2.26). Using matrix algebra, the mixed model equations yield

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}, \tag{2.28}$$

$$\hat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}), i = 1, 2, \ldots, n, \tag{2.29}$$

where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i, i = 1, 2, \ldots, n$ and $\boldsymbol{V} = \text{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$. The covariance matrices of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}_i$ are:

$$\text{Cov}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}, \tag{2.30}$$

$$\begin{aligned} \text{Cov}(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i) &= \boldsymbol{D} - \boldsymbol{D}(\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i)\boldsymbol{D} + \boldsymbol{D}(\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i) \\ &\quad \times \left(\sum_{j=1}^{n} \boldsymbol{X}_j'\boldsymbol{V}_j^{-1}\boldsymbol{X}_j\right)^{-1} (\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i)\boldsymbol{D}, i = 1, 2, \ldots, n. \end{aligned} \tag{2.31}$$

Bayesian Interpretation

It is well known that the general LME model (2.22) has a close connection with a Bayesian model in the sense that the solutions (2.28) and (2.29) are the posterior expectations of the parameters of a Bayesian model under non-informative priors.

Before we go further, we state the following two useful lemmas whose proofs can

be found in some standard multivariate textbooks, *e.g.*, Anderson (1984).

Lemma 2.6.1. *Let* $\boldsymbol{A}$, $\boldsymbol{B}$ *and* $\boldsymbol{X}$ *be* $p \times p$, $q \times q$ *and* $p \times q$ *matrices so that* $\boldsymbol{A}$ *and* $\boldsymbol{A} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}'$ *are invertible. Then*

$$(\boldsymbol{A} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}')^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{B}(\boldsymbol{B} + \boldsymbol{B}\boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{B})^{-1}\boldsymbol{B}\boldsymbol{X}'\boldsymbol{A}^{-1}. \tag{2.32}$$

In particular; when $q = 1$, $\boldsymbol{B} = 1$ *and* $\boldsymbol{X} = \boldsymbol{x}$ *where* $\boldsymbol{x}$ *is a* $p \times 1$ *vector; we have*

$$(\boldsymbol{A} + \boldsymbol{x}\boldsymbol{x}')^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{x}'\boldsymbol{A}^{-1}/(1 + \boldsymbol{x}'\boldsymbol{A}^{-1}\boldsymbol{x}). \tag{2.33}$$

Lemma 2.6.2. *Let*

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right],$$

where $\boldsymbol{\Sigma}_{22}$ *is invertible. Then*

$$\boldsymbol{X}_1|\boldsymbol{X}_2 \sim N\left[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right].$$

We now define the following Bayesian problem:

$$\boldsymbol{y}|\boldsymbol{\beta}, \quad \boldsymbol{b} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{R}), \tag{2.34}$$

with prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{b}$:

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{H}), \quad \boldsymbol{b} \sim N(\boldsymbol{0}, \widetilde{\boldsymbol{D}}), \tag{2.35}$$

where $\boldsymbol{\beta}, \boldsymbol{b}$ and $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}$ are all independent of each other, and $\widetilde{\boldsymbol{D}}$ is defined in (2.26).

Notice that specification of $\boldsymbol{H}$ is flexible. For example, we may let $\boldsymbol{H} = \boldsymbol{\lambda}\boldsymbol{I}_p$. This indicates that the components of $\boldsymbol{\beta}$ are independent of each other. Moreover, when $\boldsymbol{\lambda} \to \infty$, we have $\boldsymbol{H}^{-1} = \boldsymbol{\lambda}^{-1}\boldsymbol{I}_p \to \boldsymbol{0}$. This indicates that the limit of the prior on $\boldsymbol{\beta}$ is non-informative.

Theorem 2.1. *The Best Linear Unbiased Predictors (2.28) and (2.29) that minimise the GLL criterion (2.27) are same as the limit posterior expectations of the Bayesian problem defined in (2.34) and (2.35) as* $\boldsymbol{H}^{-1} \to 0$. *That is,*

$$\widehat{\boldsymbol{\beta}} = \lim_{\boldsymbol{H}^{-1}\to 0} E(\boldsymbol{\beta}|\boldsymbol{y}), \quad \hat{\boldsymbol{b}} = \lim_{\boldsymbol{H}^{-1}\to 0} E(\boldsymbol{b}|\boldsymbol{y}). \tag{2.36}$$

Moreover, as $\boldsymbol{H}^{-1} \to 0$ *we have the following posterior distributions:*

$$\begin{aligned} \boldsymbol{\beta}|\boldsymbol{y} &\to N(\widehat{\boldsymbol{\beta}}, (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}), \\ \boldsymbol{b}|\boldsymbol{y} &\to N(\widehat{\boldsymbol{b}}\widetilde{\boldsymbol{D}} - \widetilde{\boldsymbol{D}}\boldsymbol{Z}'\boldsymbol{P}_{\boldsymbol{V}}\boldsymbol{Z}\widetilde{\boldsymbol{D}}), \\ \boldsymbol{\epsilon}|\boldsymbol{y} &\to N(\widehat{\boldsymbol{\epsilon}}, \boldsymbol{R} - \boldsymbol{R}\boldsymbol{P}_{\boldsymbol{v}}\boldsymbol{R}), \end{aligned} \tag{2.37}$$

where $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{b}}$ *and*

$$\boldsymbol{P}_{\boldsymbol{V}} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}. \tag{2.38}$$

Proof. By (2.34) and (2.35), we have

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \\ \boldsymbol{\epsilon} \\ \boldsymbol{y} \end{pmatrix} \sim N \left[\mathbf{0}, \begin{pmatrix} \boldsymbol{H} & \mathbf{0} & \mathbf{0} & \boldsymbol{H}\boldsymbol{X}' \\ \mathbf{0} & \widetilde{\boldsymbol{D}} & \mathbf{0} & \widetilde{\boldsymbol{D}}\boldsymbol{Z}' \\ \mathbf{0} & \mathbf{0} & \boldsymbol{R} & \boldsymbol{R} \\ \boldsymbol{X}\boldsymbol{H} & \boldsymbol{Z}\widetilde{\boldsymbol{D}} & \boldsymbol{R} & \boldsymbol{\Omega} \end{pmatrix} \right],$$

where $\boldsymbol{\Omega} = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}'$ with $\boldsymbol{V} = \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}' + \boldsymbol{R}$ as defined before. Then applying Lemma 2.6.2, we have

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left[\boldsymbol{H}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}, \boldsymbol{H} - \boldsymbol{H}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\boldsymbol{H}\right]$$

Applying Lemma 2.1, we have

$$\begin{aligned} \boldsymbol{\Omega}^{-1} &= (\boldsymbol{V} + \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}')^{-1} \\ &= \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{H}(\boldsymbol{H} + \boldsymbol{H}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{H})^{-1}\boldsymbol{H}\boldsymbol{X}'\boldsymbol{V}^{-1} \\ &= \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{H}^{-1} + \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}. \end{aligned}$$

It follows that

$$\begin{aligned} \boldsymbol{H}\boldsymbol{X}'\boldsymbol{\Omega}^{-1} &= \boldsymbol{H}\boldsymbol{X}'\boldsymbol{V}^{-1} - \boldsymbol{H}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{H}^{-1} + \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})\boldsymbol{X}'\boldsymbol{V}^{-1} \\ &= \boldsymbol{H}\boldsymbol{H}^{-1}(\boldsymbol{H}^{-1} + \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1} \\ &= (\boldsymbol{H}^{-1} + \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}. \end{aligned}$$

Moreover,

$$\begin{aligned} \boldsymbol{H} - \boldsymbol{H}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\boldsymbol{H} &= \boldsymbol{H} - (\boldsymbol{H}^{-1} + \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{H} \\ &= (\boldsymbol{H}^{-1} + \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}. \end{aligned}$$

It follows that as $\boldsymbol{H} \to \boldsymbol{0}$, we have the first expression in (2.37) and the first expression in (2.36) due to (2.28). Similarly, we can show the other expressions in (2.37) and (2.36). The theorem is then proved. □

Notice that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}$ involve the unknown parameters $\widetilde{\boldsymbol{D}}$ and $\boldsymbol{R}$. If we substitute point estimates of $\widetilde{\boldsymbol{D}}$ and $\boldsymbol{R}$ (We shall discuss how to estimate them in the next subsections), the Bayesian estimates, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}$ are usually referred to as empirical Bayes estimates, although empirical Bayes estimation is conventionally applied only to the random-effects $\boldsymbol{b}_i, i = 1, 2, \ldots, n$.

Theorem 2.1 gives the limit posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{b}$ band $\boldsymbol{\epsilon}$ under the Bayesian framework (2.34) and (2.35) when $\boldsymbol{H}^{-1} \to \boldsymbol{0}$ or when the prior on $\boldsymbol{\beta}$ is non-informative. Sometimes, it is of interest to know the posterior distributions of $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ when $\boldsymbol{\beta}$ is given, *e.g.*, when $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. Actually, this knowledge is a basis for the maximum likelihood-based EM-algorithm that we shall review in the next subsection. The following theorem gives the related results.

Theorem 2.2. *Under the Bayesian framework (2.34) and (2.35), we have*

$$\begin{aligned} \boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta} &\sim N\left[\widetilde{\boldsymbol{D}}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \widetilde{\boldsymbol{D}} - \widetilde{\boldsymbol{D}}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\widetilde{\boldsymbol{D}}\right], \\ \boldsymbol{\epsilon}|\boldsymbol{y}, \boldsymbol{\beta} &\sim N\left[\boldsymbol{R}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{R} - \boldsymbol{R}\boldsymbol{V}^{-1}\boldsymbol{R}\right]. \end{aligned} \tag{2.39}$$

Proof. Set $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. Under the Bayesian framework (2.34) and (2.35), we have

$$\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \\ \widetilde{\boldsymbol{y}} \end{pmatrix} \sim N \left[\boldsymbol{0}, \begin{pmatrix} \widetilde{\boldsymbol{D}} & \boldsymbol{0} & \widetilde{\boldsymbol{D}}\boldsymbol{Z}' \\ \boldsymbol{0} & \boldsymbol{R} & \boldsymbol{R} \\ \boldsymbol{Z}\widetilde{\boldsymbol{D}} & \boldsymbol{R} & \boldsymbol{V} \end{pmatrix} \right].$$

Applying Lemma 2.6.2 directly, we have

$$\boldsymbol{b}|\boldsymbol{y} \sim N\left[\widetilde{\boldsymbol{D}}\boldsymbol{Z}'\boldsymbol{V}^{-1}\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{D}} - \tilde{\boldsymbol{D}}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\widetilde{\boldsymbol{D}}\right],$$

$$\boldsymbol{b}|\widetilde{\boldsymbol{y}} \sim N\left[\boldsymbol{R}\boldsymbol{V}^{-1}\widetilde{\boldsymbol{y}}, \boldsymbol{R} - \boldsymbol{R}\boldsymbol{V}^{-1}\boldsymbol{R}\right].$$

Then (2.39) follows by replacing $\widetilde{\boldsymbol{y}}$ by $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ in the above expressions. The theorem is proved. □

It is worthwhile to notice that by Theorem 2.2, we have $E(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{b}}$ and $E(\boldsymbol{\epsilon}|\boldsymbol{y}, \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\epsilon}}$.

Estimation of Variance Components

If the covariance matrices, $\boldsymbol{D}$ and $\boldsymbol{R}_i$, are unknown, but their point estimates, say, $\widehat{\boldsymbol{D}}$ and $\widehat{\boldsymbol{R}_i}$, are available, then we can have $\widehat{\boldsymbol{V}}_i = \boldsymbol{Z}_i\widehat{\boldsymbol{D}}\boldsymbol{Z}_i' + \widehat{\boldsymbol{R}}_i$. The estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ thus can be obtained by substitution of $\widehat{\boldsymbol{V}}_i$ and $\widehat{\boldsymbol{D}}$ in (2.28) and (2.29). Their corresponding standard errors are given by (2.30) and (2.32) after replacing $\boldsymbol{V}_i$ and $\boldsymbol{D}$ by their estimates. However, these standard errors are underestimated since the estimation errors of $\widehat{\boldsymbol{V}}_i$ and $\widehat{\boldsymbol{D}}$ are not accounted for.

Under the normality assumption, the maximum likelihood (ML) method and the restricted maximum likelihood (REML) method are two popular techniques to estimate the unknown components of $\boldsymbol{D}$ and $\boldsymbol{R}_i$, although this may not be appropriate if the normality assumption is questionable.

Under the following normality assumptions,

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{R}), \quad \boldsymbol{b} \sim N(\boldsymbol{0}, \widetilde{\boldsymbol{D}}),$$

the generalized likelihood function can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{b}, \widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) = \; & (2\pi)^{-N/2}|\boldsymbol{R}|^{-1/2} \exp\left\{-\frac{1}{2}[\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}]'\boldsymbol{R}^{-1} \times [\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}]\right\} \\ & \times (2\pi)^{-qn/2}|\widetilde{\boldsymbol{D}}|^{-1/2} \exp\left(-\tfrac{1}{2}\boldsymbol{b}'\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{b}\right), \end{aligned}$$

where qn is the dimension of $\boldsymbol{b}$ and $N = \sum_{i=1}^{n} n_i$. If the random-effects vector $\boldsymbol{b}$ is integrated out, we can obtain the following conventional likelihood function:

$$\begin{aligned} \boldsymbol{L}(\boldsymbol{\beta}, \widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) &= \int L(\boldsymbol{\beta}, \boldsymbol{b}, \widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) d\boldsymbol{b} \\ &= (2\pi)^{-N/2}|\boldsymbol{V}|^{-1/2} \exp\left\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\} \end{aligned}$$

The ML method for estimation of variance components is to maximize the following log-likelihood function:

$$\log \boldsymbol{L}(\boldsymbol{\beta}, \widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) = -\frac{1}{2}N\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \quad (2.40)$$

with respect to the variance components for a given $\boldsymbol{\beta}$. However, joint maximization with respect to the variance components $\widetilde{\boldsymbol{D}}, \boldsymbol{R}$ and the fixed-effects parameter vector

$\boldsymbol{\beta}$ also results in the estimate of $\boldsymbol{\beta}$ in (2.28).

The REML method is used to integrate out both $\boldsymbol{b}$ and $\boldsymbol{\beta}$ from $L(\boldsymbol{\beta}, \boldsymbol{b}, \widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y})$ in order to adjust for loss of degrees of freedom due to estimating $\boldsymbol{\beta}$ from the ML method, i.e., to maximize

$$\boldsymbol{L}(\widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) = \int\int L(\boldsymbol{\beta}, \boldsymbol{b}, \widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y})\ d\boldsymbol{b}\ d\boldsymbol{\beta}.$$

It can be shown that

$$\boldsymbol{L}(\widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) = \frac{(2\pi)^{p/2}|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|^{-1/2}}{(2\pi)^{N/2}|\boldsymbol{V}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}'\boldsymbol{P}_{\boldsymbol{V}}\boldsymbol{y}\right),$$

where $\boldsymbol{P}_{\boldsymbol{V}} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$ as defined in (2.38). Thus, we have

$$\boldsymbol{L}(\widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) = (2\pi)^{p/2}|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|^{-1/2} L(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}).$$

The REML estimates of variance components can be obtained *via* maximizing

$$\boldsymbol{L}(\widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) = \log \boldsymbol{L}(\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{D}}, \boldsymbol{R}|\boldsymbol{y}) + \frac{1}{2}p\log(2\pi) - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|. \tag{2.41}$$

More detailed derivations for these results can be found in Davidian and Giltinan (1995).

The EM-algorithms

The implementation of the ML and REML methods is not trivial. To overcome this implementation difficulty, the EM-algorithm and Newton-Raphson methods have been proposed (Laird and Ware 1982, Dempster *et al* 1981, Laird, Lange and Stram 1987, Jennrich and Schluchter 1986, Lindstrom and Bates 1990). The books by Searle *et al* (1992), Davidian and Giltinan (1995), Vonesh and Chinchilli (1996) and Pinheiro and Bates (2000) also provide a good review of these implementation methods. The standard statistical software packages such as SAS and S-PLUS now offer convenient functions to implement these methods (*e.g.*, the S-PLUS function lme and the SAS procedure PROC MIXED). We shall briefly review the EM-algorithm here.

Recall that we generally assume that $\boldsymbol{R}_i$ has the following simple form:

$$\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{n_i}, \quad i = 1, 2, \ldots, n. \tag{2.42}$$

When $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ were known, under the normality assumption, the natural ML estimates of σ^2 and $\boldsymbol{D}$ would be

$$\widehat{\sigma}^2 = N^{-1} \sum_{i=1}^{n} \epsilon_i' \epsilon_i, \quad \widehat{\boldsymbol{D}} = \sum_{i=1}^{n} \boldsymbol{b}_i \boldsymbol{b}_i'. \tag{2.43}$$

This is the M-step of the EM-algorithm. Because $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$ are unknown, the above estimates are not computable. There are two ways for overcoming this difficulty, associated, respectively, with the ML or REML-based EM-algorithm.

Notice that the ML estimates of $\boldsymbol{D}$ and σ^2 are obtained *via* maximizing the

loglikelihood function (2.40) with the fixed-effects parameter vector $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ given. Therefore, the key for the ML-based EM-algorithm is to replace the $\widehat{\sigma}^2$ and $\widehat{\boldsymbol{D}}$ in (2.43) with

$$E(\widehat{\sigma}^2|\boldsymbol{y},\ \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}) \quad \text{and} \quad E(\hat{\boldsymbol{D}}|\boldsymbol{y},\ \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}), \tag{2.44}$$

respectively. The underlying rationale is that the variance components $\boldsymbol{D}$ and σ^2 are estimated based on the residuals after the estimated fixed-effects component $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is removed from the raw data, and the estimation will not take the variation of $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ into account. This is the E-step of the ML-based EM-algorithm. Using Theorem 2.2, we can show the following theorem.

Theorem 2.3. *Assume the Bayesian model defined in (2.34) and (2.35) holds, and assume* $\boldsymbol{R}_i, i = 1, 2, \ldots, n$ *satisfy (2.42). Then we have*

$$\begin{aligned} E(\widehat{\sigma}^2|\boldsymbol{y},\ \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}) &= N^{-1}\textstyle\sum_{i=1}^{n}\left\{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} + \sigma^2[n_i - \sigma^2 tr(\boldsymbol{V}_i^{-1})]\right\}, \\ E(\hat{\boldsymbol{D}}|\boldsymbol{y},\ \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}) &= n^{-1}\textstyle\sum_{i=1}^{n}\left\{\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i' + [\boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\boldsymbol{D}]\right\} \end{aligned} \tag{2.45}$$

In the right-hand sides of the expressions (2.45), the variance components σ^2 and $\boldsymbol{D}$ are still unknown. However, when they are replaced by the current available values, the updated values for $\widehat{\sigma}^2$ and $\widehat{\boldsymbol{D}}$ can be obtained. In other words, provided some initial values for $\boldsymbol{D}$ and σ^2, we can update $\widehat{\sigma}^2$ and $\widehat{\boldsymbol{D}}$ using (2.45) until convergence. This is the main idea of the EM-algorithm. For simplicity, the initial values can be taken as $\widehat{\boldsymbol{D}} = \boldsymbol{I}_q$, and $\widehat{\sigma}^2 = 1$. The major cycle for the ML-based EM-algorithm is as follows:

1. Given $\widehat{\boldsymbol{D}}$ and $\widehat{\sigma}^2$, compute $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}_i$ using (2.28) and (2.29).

2. Given $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}_i$, update $\widehat{\boldsymbol{D}}$ and $\widehat{\sigma}^2$ using (2.45).

3. Iterate between (1) and (2) until convergence.

Let $r = 0, 1, 2, \ldots,$ index the sequence of the iterations, and $\widehat{\boldsymbol{\beta}}_{(r)}$ and $\widehat{\boldsymbol{b}}_{i(r)}$ the estimated values for, $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ at iteration r. Other notations such as $\boldsymbol{V}_{i(r)}, \sigma^2_{(r)}$ are similarly defined. Then more formally, the ML-based EM-algorithm may be written as follows:

ML Based EM-algorithm

Step 0. Set $r = 0$. Let $\sigma^2_{(r)} = 1$, and $\widehat{\boldsymbol{D}}_{(r)} = \boldsymbol{I}_q$.

Step 1. Set $r = r + 1$. Update $\widehat{\boldsymbol{\beta}}_{(r)}$ and $\widehat{\boldsymbol{b}}_{i(r)}$ using

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{(r)} &= \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \widehat{\boldsymbol{V}}^{-1}_{i(r-1)} \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \widehat{\boldsymbol{V}}^{-1}_{i(r-1)} \boldsymbol{y}_i\right), \\
\widehat{\boldsymbol{b}}_{i(r)} &= \boldsymbol{D}_{(r-1)} \boldsymbol{Z}_i' \widehat{\boldsymbol{V}}^{-1}_{i(r-1)} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_{(r)}\right), i = 1, 2, \ldots, n,
\end{aligned}$$

where

$$\widehat{\boldsymbol{V}}_{i(r-1)} = \boldsymbol{Z}_i \widehat{\boldsymbol{D}}_{(r-1)} \boldsymbol{Z}_i' + \widehat{\sigma}^2_{(r-1)} \boldsymbol{I}_{n_i}, i = 1, 2, \ldots, n.$$

Step 2. Update $\sigma^2_{(r)}$, and $\widehat{\boldsymbol{D}}_{(r)}$ using

$$\sigma^2_{(r)} = N^{-1}\sum_{i=1}^{n}\left\{\hat{\epsilon}'_{i(r)}\hat{\epsilon}_{i(r)} + \widehat{\sigma}^2_{(r-1)}[n_i - \widehat{\sigma}^2_{(r-1)}\mathrm{tr}(\widehat{\boldsymbol{V}}^{-1}_{i(r-1)})]\right\},$$

$$\widehat{\boldsymbol{D}}_{(r)} = n^{-1}\sum_{i=1}^{n}\left\{\hat{\boldsymbol{b}}_{i(r)}\hat{\boldsymbol{b}}'_{i(r)} + [\widehat{\boldsymbol{D}}_{(r-1)} - \widehat{\boldsymbol{D}}_{(r-1)}\boldsymbol{Z}'_i\widehat{\boldsymbol{V}}^{-1}_{(r-1)}\boldsymbol{Z}_i\widehat{\boldsymbol{D}}_{(r-1)}]\right\},$$

where $\hat{\epsilon}_{i(r)} = \boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_{(r)} - \boldsymbol{Z}_i\widehat{\boldsymbol{b}}_{i(r)}$.

Step 3. Repeat Steps 1 and 2 until convergence.

The REML-based EM-algorithm can be similarly described. The main differences include

(a) The REML-based EM-algorithm is developed to find the REML estimates of σ^2 and $\boldsymbol{D}$ that maximize (2.41).

(b) The key for the REML-based EM-algorithm is to replace $\widehat{\sigma}^2$ and $\widehat{\boldsymbol{D}}$ in (2.43) by $E(\widehat{\sigma}^2|\boldsymbol{y})$ and $E(\widehat{\boldsymbol{D}}|\boldsymbol{y})$ instead of their expectations conditional on $\boldsymbol{y}$ and $\widehat{\boldsymbol{\beta}}$ as given in (2.44). These conditional expectations can be easily obtained using Theorem 2.1 and we shall present them in Theorem 2.4 below for easy reference.

(c) The REML-based EM-algorithm can be simply obtained *via* replacing all the $\widehat{\boldsymbol{V}}^{-1}_{i(r-1)}$ the Step 2 of the ML-based EM-algorithm above with $\boldsymbol{P}_{\widehat{\boldsymbol{V}}_{i(r-1)}}$, where

$$\boldsymbol{P}_{\widehat{\boldsymbol{V}}_{i(r-1)}} = \widehat{\boldsymbol{V}}^{-1}_{i(r-1)} - \widehat{\boldsymbol{V}}^{-1}_{i(r-1)}\boldsymbol{X}_i\left(\sum_{j=1}^{n}\boldsymbol{X}'_j\widehat{\boldsymbol{V}}^{-1}_{j(r-1)}\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}'_i\widehat{\boldsymbol{V}}_{i(r-1)}$$

Theorem 2.2 below is similar to Theorem 2.3 but it is based on Theorem 2.1.

Theorem 2.4. *Assume the Bayesian model defined in (2.34) and (2.35) holds, and assume* $\boldsymbol{R}_i, i = 1, 2, \ldots, n$ *satisfying (2.42). Then as* $\boldsymbol{H}^{-1} \to \boldsymbol{0}$,

$$\begin{aligned} E(\widehat{\sigma}^2|\boldsymbol{y}) &\to N^{-1}\sum_{i=1}^{n}\left\{\hat{\boldsymbol{\epsilon}}_i'\hat{\boldsymbol{\epsilon}}_i + \sigma^2[n_i - \sigma^2 tr(\boldsymbol{P}_{\boldsymbol{V}_i})]\right\}, \\ E(\widehat{\boldsymbol{D}}|\boldsymbol{y}) &\to n^{-1}\sum_{i=1}^{n}\left\{\widehat{\boldsymbol{b}}_i\widehat{\boldsymbol{b}}_i' + [\boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{P}_{\boldsymbol{V}_i}\boldsymbol{Z}_i\boldsymbol{D}]\right\}, \end{aligned} \tag{2.46}$$

where $\boldsymbol{P}_{\boldsymbol{V}_i} = \boldsymbol{V}_i^{-1} - \boldsymbol{V}_i^{-1}\boldsymbol{X}_i\left(\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\boldsymbol{X}_i'\boldsymbol{V}_i$.

2.6.3 Random Effects Models

An alternative perspective on explicit modelling of longitudinal response is to think directly of the fact that each unit appears to have its own trajectory or inherent trend with its own peculiar features. The resulting statistical model, called a random coefficient model. Random effects models for serial measurements have a long history(Wishart 1938). Rao(1965, 1975) described a family of two-stage random effects models for serial measurements and developed estimation and testing procedures for data sets balanced on time and with no between subject covariates. In Rao's formulation, the first stage consists of a linear model conditioned on the individual growth curve parameters, $\boldsymbol{\beta_i}$. At the second stage, the growth curve parameters are assumed to depend linearly on fixed covariates.

Stage 1. Individual Model.

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta_i} + \boldsymbol{\varepsilon_i}, \tag{2.47}$$

where, $\boldsymbol{\varepsilon_i} \sim\ MVN(\boldsymbol{0}, \boldsymbol{R}_i)$. This is like a regression model for the i^{th} unit, with design matrix $\boldsymbol{Z}_i$ and $(k \times 1)$ regression parameter $\boldsymbol{\beta_i}$.

Stage 2. Population Model.

Stage 1 tells only part of the story; it describes what happens at the level of a unit, and includes explicit mention of within unit variation. To accommodate the among unit variation, we write $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{b}_i$, where $\boldsymbol{\beta}$ is the mean vector of the population of all $\boldsymbol{\beta}_i$. Here, $\boldsymbol{b}_i$ is a vector of random effects with $\boldsymbol{0}$ mean, describing how the intercept and slope for the i^{th} unit deviates from the mean value and dispersion matrix $\boldsymbol{\Lambda}$. $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon_i}$ are assumed to be independent.

$$\boldsymbol{\beta_i} \sim MVN(\boldsymbol{\beta}, \boldsymbol{\Lambda}). \tag{2.48}$$

If there are two groups in the population, like males females, the groups may be allowed to differ in their responses by modifying the relation as $\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i$. If we combine the two parts of the model, we get

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i) + \boldsymbol{\varepsilon_i} = \boldsymbol{Z}_i\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon_i}. \tag{2.49}$$

If we assume that there is only one group, so that $\boldsymbol{A}_i = \boldsymbol{I}$,

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon_i}. \tag{2.50}$$

Using the independence of $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon_i}$, this implies that

$$\text{var}(\boldsymbol{\varepsilon_i}) = \boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{\Lambda}\boldsymbol{Z}_i' + \boldsymbol{R}_i. \tag{2.51}$$

Thus the model expresses the covariance vector of the data vector as the sum of two pieces representing the within unit and among unit variations. Assuming that both $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon_i}$ are both well-represented by multivariate normal distributions and are independent, we may conclude that

$$\boldsymbol{Y}_i \sim MVN(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i), \tag{2.52}$$

where $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{A}_i$, and $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{\Lambda}\boldsymbol{Z}_i' + \boldsymbol{R}_i$.

Inference on Regression and Covariance Parameters

As this model resembles with the classical model, the methods of maximum likelihood and restricted maximum likelihood may be used to estimate the parameters $\boldsymbol{\beta}$, $\boldsymbol{D}$ and the parameters that make up $\boldsymbol{R}_i$. The generalized least squares estimator for $\boldsymbol{\beta}$ and its large sample approximate sampling distribution will have the same form, with $\boldsymbol{X}_i$ and $\boldsymbol{\Sigma}_i$ defined as in (2.52).

2.6.4 AR Models

The AR models (Box and Jenkins 1970) offer a natural alternative to the random effects models. First and second order AR models have been found specially attractive for serial measurements.

The general approach to parameter estimation for AR models can be illustrated by a discussion of the first order AR model. If

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i, i = 1, \ldots, n, \tag{2.53}$$

and $\boldsymbol{e}_i$ is AR(1), then the likelihood can be written in terms of Y_{i1} and $Y_{ij}, j = 2, \ldots, n$ by using the representations

$$Y_{i1} = \boldsymbol{X}'_{i1}\boldsymbol{\beta} + e_{i1}, \tag{2.54}$$

and

$$Y_{ij} - \rho Y_{i,j-1} = (\boldsymbol{X}_{ij} - \rho\boldsymbol{X}_{i,j-1})\boldsymbol{\beta} + \nu_{ij}, \tag{2.55}$$

where $\boldsymbol{X}'_{ij}$ is the j^{th} row of $\boldsymbol{X}_i$, ρ is the correlation between successive observations, and $\nu_{ij} = e_{ij} - \rho e_{i,j-1}$ is distributed as $N(0, \sigma^2(1-\rho^2))$

MLE of the parameters of this model can be obtained by a modified Gauss-Seidel procedure (Louis and Spiro, 1984). One part of the procedure consists of weighted least squares estimation of $\boldsymbol{\beta}$ from (2.54) and (2.55), with ρ set to its current estimate and treated as known. The other part consists of maximisation of the likelihood residuals $\boldsymbol{Y}_{ij} - \boldsymbol{X}'_{ij}\widehat{\boldsymbol{\beta}}$ with respect to ρ, with $\widehat{\boldsymbol{\beta}}$ treated as known.

This analysis assumes an AR structure for the e_{ij}, not the Y_{ij}. Some investigators have studied models of the form

$$\boldsymbol{Y}_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \rho\,\boldsymbol{Y}_{i,j-1} + \nu_{ij}. \tag{2.56}$$

The above three models for covariance of serial measurements- multivariate, random effects and AR- offer a rich set of alternatives from which to choose a covariance model for a particular application.

Chapter 3

Models for Discrete Responses.

3.1 Generalized Linear Models

In our discussion hitherto, we have focused on situations where

1. The response is continuous and reasonably assumed to be normally distributed.
2. The variance of the response is constant regardless of the setting of $\boldsymbol{X}_i$.
3. The model relating mean response to time and possibly other covariates is linear in parameters that characterize the relationship.

For example, regardless of how we modelled covariance (by direct modelling or by introducing random effects), we had models for the mean response of a data vector

of the form

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta}. \tag{3.1}$$

Under these conditions, we were led to methods that were based on the assumption that

$$\boldsymbol{Y}_i \sim MVN(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i), \tag{3.2}$$

the form of the matrix $\boldsymbol{\Sigma}_i$ is dictated by what one assumes about the nature of variation. To fit the model, we used the methods of maximum likelihood and restricted maximum likelihood under the assumption that the data vectors are distributed as multivariate normal. Thus, the fitting method was based on the normality assumption.

The assumption of normality is not always relevant for some data. This issue is not confined to longitudinal data analysis. It is an issue even in ordinary regression modelling. If the response is in the form of small counts, or is in fact binary (yes/no), it is clear that the assumption of normality would be quite unreasonable. Thus, the modelling and methods we have discussed so far, including the classical techniques, would be inappropriate for these situations.

Many bio-medical applications involves discrete (*e.g.* binary or count) longitudinal responses, for example, the presence or absence of respiratory illness, or counts of the number of elliptic seizures in an interval. When the response is discrete, linear models are no longer appropriate for relating changes in the mean response to covariates. Instead we consider extensions of GLMs for longitudinal data. Classical linear regression methods were extended to allow the assumption that these distributions,

rather than the normal distribution, are sensible probability models for the data. The term generalized linear models is used to refer to the models and techniques used. We will focus on three models in particular; a more extensive catalogue of models may be found in McCullagh and Nelder (1989):

1. The Bernoulli probability distribution as a model for binary data (discrete) (this may be extended to model data in the form of proportions
2. The Poisson probability distribution as a model for count data (discrete)
3. The gamma probability distribution as a model for continuous but non-normal data with constant coefficient of variation.

We will see that all of these probability models are members of a special class of probability models. This class also includes the normal distribution with constant variance (the basis for classical linear regression methods for normal data); thus, generalized linear models will be seen to be an extension of ordinary linear regression models.

Generalised linear models provide a unified class of models for regression analysis of independent observations of a discrete or continuous response. The correlation among observations obtained from the same individual makes a straightforward application of generalised linear models to longitudinal data inappropriate. Hence we consider extensions of this broad class of models,

We present a unified methodology for analysing diverse types of longitudinal responses, that avoids making assumptions about the distribution of the vector of

responses. The method relies solely on the assumption about the mean response.

Generalised linear models provide a unified method for analysing diverse types of discrete as well as continuous univariate responses. They include as special cases, the standard linear regression and analysis of variance (ANOVA) models for a normally distributed continuous response, logistic regression for a binary or dichotomous response and log-linear or Poisson regression models for counts.

Generalised linear models were introduced in a seminal paper by Nelder and Wedderburn (1972) as a generalisation of the method described by Finney (1952). In these papers the authors show that maximum likelihood estimates for a large class of commonly used regression models can be obtained by the method of iteratively weighted least squares, in which the weights and the response are adjusted from one iteration to the next. The proposed algorithm, known as 'Fisher's scoring' is an extension of Fisher's (1935) method for computing maximum likelihood estimates in linear probit models. The same result was obtained independently by Bradley (1973) and Jenrich and Moore (1975), though not exploited to its full extent.

The logistic regression model for binary responses and Poisson regression model for count are two popular special cases. The logistic regression model is a generalization of the normal linear model. It is used to relate cross-sectional binary data $\boldsymbol{y}_i$ to a vector of covariates $\boldsymbol{x}_i$. Similarly, the Poisson regression model allows us to relate a single count $\boldsymbol{y}_i$ to a vector of covariates $\boldsymbol{x}_i$. The logistic and Poisson regression models can be generalized to handle longitudinal binary and count data by introducing random effects into the model. These models are called the binomial or Bernoulli or logistic random effects model and the Poisson random effects model.

McCullagh and Nelder (1989) provides a comprehensive description of the theory and application of generalised linear models. A general overview of logistic regression, Poisson regression and generalised linear models can be found in Neter *et al* (1996). Dobson (1990), Firth (1991) and Gill (2000) provide excellent introductions to generalised linear models. Hosmer and Lemeshow (2000) provides an accessible and comprehensive description of logistic regression models for binary data.

3.1.1 Distributional Assumptions

GLMs assume that response variable has a probability distribution belonging to the exponential family of distributions, that includes many distributions like normal, Bernoulli, binomial and Poisson distributions. GLM extends the standard linear regression by taking a suitable transformation of the mean response and relating the transformed mean response to covariates. This is achieved by the introduction of link function. The link function applies a transformation to the mean and then links the covariates, *via* the linear predictor, to the transformed mean of the distribution of the responses,

$$g(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} = \sum_{k=1}^{p} \beta_k X_{ik} = \boldsymbol{X}_i' \boldsymbol{\beta}, \tag{3.3}$$

where the link function $g(\cdot)$ is some known function, for example, $\log(\mu_i)$. This implies that it is the transformed mean response that changes linearly with changes in the values of the covariates. Thus, while in a standard linear regression model the mean response is related directly to the linear combination of covariates, in GLMs, it is

some appropriate transformation of the mean response that is related to the linear combination of covariates.

The term linear in GLM means that η_i must be linear in regression parameters and it allows η_i to be non-linear in X_i, if the non-linearity can be accommodated by appropriate transformation of covariates (*e.g.*. $\log(X)$) and/or by including a polynomial in X.

When viewed as a GLM, the standard linear regression model adopts the identity link function $g(\mu_i) = \mu_i$. which gives the standard linear regression model,

$$\mu_i = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \tag{3.4}$$

The primary motivation for considering link functions other than the identity is to ensure that the linear predictor produces predictions of the mean response that are within the allowed range. For example, when analysing a binary response, μ_i has interpretation in terms of the probability of success and we must have $0 < \mu_i < 1$. Hence the identity link is not appealing since, when the mean response (here the probability of success) is directly related to the linear combination of covariates, the model can yield predicted probabilities outside the allowable range from 0 to 1 for sufficiently large or small values of the covariates. The use of certain non-linear link function ensures that this cannot happen. It is preferable to use a link function that maps μ_i from the range [0,1] onto the unrestricted range $(\infty, -\infty)$.

Every distribution belonging to the exponential family has a special link function called canonical link function g$(\cdot)$ defined such that $g(\mu_i) = \theta_i$, where θ_i is the canonical location parameter. The canonical link function for normal distribution is the identity link function $g(\mu_i) = \mu_i$, which gives the standard linear regression model,

$$\mu_i = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \tag{3.5}$$

For Poisson distribution the canonical link function is the log link function, $g(\mu_i) = \log \mu_i$, which gives the log-linear regression model

$$\log \mu_i = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \tag{3.6}$$

For Bernoulli distribution, where $0 < \mu_i < 1$, the canonical link function is the logistic or logit link function $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$, which gives the logistic regression model

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \tag{3.7}$$

However one can choose other link functions when they seem appropriate to the application in hand. For example, when Y_i is Bernoulli, one may use the log-log link function $g(\mu_i) = \log(-\log(1-\mu_i))$ and the probit link function $g(\mu_i) = \phi^{-1}(\mu_i)$, where $\phi(\cdot)$ is the standard normal cumulative distribution function.

3.2 Logistic Regression for Binary Responses

Logistic regression is widely used to describe the relationship between a binary response variable and a set of covariates. In this model, the dependent variable is a logit (log-odds). Let Y_i denote a binary response variable, assuming values 0 and 1. The probability distribution of Y_i is Bernoulli, with $P(Y_i = 1) = \mu_i$, and $\mathrm{P}(Y_i = 0) = 1-\mu_i$. For ease of exposition, first we will assume that there is a single covariate X_i. The analytical goal is to investigate the relationship between μ_i and X_i. The linear regression

$$E(Y_i/X_i) = \beta_1 + \beta_2 X_i, \tag{3.8}$$

will yield predicted probabilities outside the range from 0 to 1 for sufficiently large X_i. Further, we cannot always expect a linear relationship between μ_i and X_i. Also the usual assumption of homogeneity of variance would be violated, since the variance of a binary variable depends on the mean, with

$$V(Y_i) = \mu_i(1 - \mu_i). \tag{3.9}$$

If the logit or logistic function, $\log\left(\dfrac{\mu_i}{1-\mu_i}\right)$ is adopted, the resulting model,

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathrm{logit}(\mu_i) = \beta_1 + \beta_2 X_i, \tag{3.10}$$

is known as logistic regression model.

If the predictor variable X_i is dichotomous, taking values 0 and 1,

$$\text{logit}(\mu_i/X_i = 1) - \text{logit}(\mu_i/X_i = 0) = (\beta_1 + \beta_2) - \beta_1 = \beta_2. \tag{3.11}$$

Thus $\exp(\beta_2)$ has interpretation as the odds ratio of the response for the possible values of the covariates. β_2 is the change in log odds for a unit change in X_i. Equivalently, a unit change in X_i changes the odds of success multiplicatively by $\exp(\beta_2)$.

The logistic regression can also be expressed as

$$\mu_i = \frac{\exp(\beta_1 + \beta_2 X_i)}{1 + \exp(\beta_1 + \beta_2 X_i)} \ . \tag{3.12}$$

If $\boldsymbol{X}_i$ is a $p \times 1$ vector of covariates, the logistic regression model becomes,

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \tag{3.13}$$

where $X_{i1} = 1, \ \forall \ i = 1, 2, \ldots, N$. Here β_k represents change in log odds for a unit change in X_{ik}, given that all other predictor variables remain constant. Equivalently, a unit change in X_{ik} changes the odds of success multiplicatively by a factor $\exp(\beta_k)$.

Illustration

As an illustrative example, we consider a study conducted by Van Marter *et al.* (1990) concerning low birth weight infants in a neonatal care unit. The interest was in the development of brichopulmonary dysplasia (DPD), a chronic lung disease, in a sample of 223 infants weighing less than 1750 gms. To examine whether there is association

between the risk of BPD and birth weight (in gms $\times 10^{-2}$) a logistic regression model was considered. Letting $Y_i = 1$ if the i^{th} infant develops BPD by day 28 of life and $Y_i = 0$ otherwise, the logistic regression model

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2\ \texttt{Weight}_i, \tag{3.14}$$

where $\mu_i = E(Y_i) = P(Y_i = 1)$, is estimated as

$$\log\left(\frac{\widehat{\mu}_i}{1-\widehat{\mu}_i}\right) = 4.0343 - 0.4229\ \texttt{Weight}_i, \tag{3.15}$$

with $\text{SE}(\beta_2) = 0.0641$ and β_2 is significantly different from zero ($Z = -6.599$). This shows that the BPD decreases with increasing birth weights. Specifically a 100 gms increases in birth weight causes the log odds of BPD to decrease by 0.42. Thus the predicted probability corresponding to any specific birth weights can be computed using the relation

$$\widehat{\mu}_i = \frac{\exp(\widehat{\beta} + \widehat{\beta}_2 X_i)}{1 + \exp(\widehat{\beta} + \widehat{\beta}_2 X_i)}\ .$$

For an infant weighing 1200 gms, this relation yields, $\widehat{\mu}_i \approx 0.27$.

Next we include two additional covariates, gestational age (in weeks) and the presence of toxemia (with 1 denoting the presence of toxemia and 0 denoting its absence and then

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2\ \texttt{Weight}_i + \beta_3\ \texttt{Age}_i + \beta_4\ \texttt{Toxemia}_i, \tag{3.16}$$

The estimated regression line is

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = 13.9361 - 0.2644\ \texttt{Weight}_i - 0.3885\ \texttt{Age}_i - 1.3437\ \texttt{Toxemia}_i, \quad (3.17)$$

with standard errors SE(β_2) = 0.0812, SE(β_2)=0.1149 and SE(β_2)=0.6075. The estimated coefficient for birth weight has now decreased. Comparing mothers who were diagnosed with toxemia to mothers who were not, the estimated coefficient of toxemia has now the following interpretation. The adjusted odds ratio (adjusting the effects of birth weight and gestational age) is 0.26 (=exp(−1.34)) and this indicates that infants of mothers diagnosed with toxemia has approximately a quarter the risk of developing BPD.

3.3 Log-linear(Poisson) Regression for Counts

Log-linear regression, often referred to as Poisson regression, is used widely for the analysis of counts of number of times of occurrence of some events. The logarithm of the mean of the response variable is related to the explanatory variables in this model. The objective of the log-linear regression is to relate the mean or expected count to the set of covariates.

If the occurrences of some event are counted within an interval of time, then the rate at which the event occurs is of more interest than the count, since rates are more meaningfully compared if the duration of time during which occurrences are observed are different.

When the response is a count, it is often reasonable to assume that Y_i has a Poisson distribution with

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},\ y_i = 0, 1, 2, \cdots \tag{3.18}$$

Note that μ_i is the expected count. The expected rate is $\frac{\mu_i}{T_i}$ where T_i is the interval of time, Since rate cannot be negative, standard linear regression model relating $\frac{\mu_i}{T_i}$ directly to X_i will be unappealing. Instead, we can relate a transformation of the rate directly to X_i. The log-linear regression model,

$$\log\left(\frac{\mu_i}{T_i}\right) = \beta_1 + \beta_2 X_i, \tag{3.19}$$

is one such, where a logarithmic transformation is adopted. It may be also expressed as

$$\log\,\mu_i = \log\,T_i + \beta_1 + \beta_2 X_i. \tag{3.20}$$

When viewed as a GLM, log-linear model is simply the special case where the distribution of Y_i is assumed to be Poisson.

To interpret β_2, for the special case where the predictor X_i is dichotomous, consider

$$\log(\mu_i/X_i = 1) - \log(\mu_i/X_i = 0) = \{\log T_i + \beta_1 + \beta_2\} - \{\log T_i + \beta_1\} = \beta_2. \tag{3.21}$$

Thus $\exp(\beta_2)$ has interpretation as the rate ratio $\frac{(\mu_i/X_i = 1)}{(\mu_i/X_i = 0)} \cdot \beta_2$ has the interpretation as the change in the log expected rate for a single unit change in X_i. Equivalently,

a unit change in X_i changes the rate of occurrence of the event multiplicatively by $\exp(\beta_2)$.

If $\boldsymbol{X}_i$ is a $p \times 1$ vector of covariates, the log-linear regression model becomes

$$\log(\mu_i) = \log(T_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \tag{3.22}$$

where $X_{i1} = 1, \ \forall \ i = 1, 2, \ldots, N$. Just as in the case when X_i is univariate, here, β_k can be interpreted in terms of influence of the k^{th} component of $\boldsymbol{X}_i$ on the log expected rate, given that all other predicator variables except X_{ik} remain constant.

Illustration

The data for this illustration arise from a prospective study of potential risk factors for coronary heart disease (CHD) (Rosenman *et al.* 1975). The study observed 3154 men aged 40-50 for an average of 8 years and recorded the incidence of cases of CHD. The potential risk factors included smoking, blood pressure, and personality/behaviour type. Let Y_i denote the count of number of cases of CHD and T_i denote the duration of follow up. To examine whether the rates of CHD are related to the smoking exposure, we consider the log-linear model,

$$\log\left(\frac{\mu_i}{T_i}\right) = \beta_1 + \beta_2 \ \texttt{Smoke}_i, \tag{3.23}$$

where $\texttt{Smoke}_i$ is the measure of smoke exposure (0 for non smoker, 10 for 1-10 cigarettes per day, 20 for 11-20 cigarettes per day, 30 for 21-30 cigarettes per day).

The MLE of β_2 is 0.0318 with standard error 0.0056 and β_2 is significantly different from zero. The expected rate of CHD for individuals with smoking habit is $e^{0.0318\ \texttt{Smoke}_i}$ times as high as the rate of CHD for non-smokers.

Including blood pressure and behaviour pattern in the set of covariates, the log-linear regression model becomes

$$\log\left(\frac{\mu_i}{T_i}\right) = \beta_1 + \beta_2\ \texttt{Smoke}_i + \beta_3\ \texttt{BP}_i + \beta_4\ \texttt{Type}_i, \tag{3.24}$$

where $\texttt{BP}_i = 1$ if BP$\geq$140, and 0 otherwise, $\texttt{Type}_i = 1$ if the person is impatient, aggressive, tense or competitive and 0 if not. The estimated regression is

$$\log\left(\frac{\mu_i}{T_i}\right) = -5.4202 + 0.0273\ \texttt{Smoke}_i + 0.7526\ \texttt{BP}_i + 0.7534\ \texttt{Type}_i, \tag{3.25}$$

with SE(β_2) = 0.0056, SE(β_3) =0.1362 and SE(β_4) = .1292. All the regression coefficients are significant.

3.4 Estimation

In standard linear regression models, we estimate the regression coefficients using the method of least squares and the estimators are same as the MLE's when Y_i's are normally distributed with constant variance. The parameters of a GLM are estimated using a more general method of ML estimation. In a GLM, the response is assumed to have a distribution belonging to the exponential family of distributions, with density

in the form

$$f(y_i, \theta_i, \phi) = \exp\left[\{y_i\theta_i - a(\theta_i)\}/\phi + b(y_i, \phi)\right]. \tag{3.26}$$

Thus the likelihood function can be expressed as

$$L = \prod_{i=1}^{n} \exp[\{y_i\theta_i - a(\theta_i)\}/\phi + b(y_i, \phi)]. \tag{3.27}$$

Note that L is a function of β, since θ_i is a known function of the mean μ_i, and

$$\mu_i = g^{-1}\left(\sum_{i=1}^{p} \beta_k X_{ik}\right), \tag{3.28}$$

where $g^{-1}(\cdot)$ denotes the inverse of the link function. MLE's of β's are obtained by substituting this expression of μ_i into the likelihood function and finding those values of β that produces the largest value for the likelihood function.

Instead of maximising the likelihood, it is usually more convenient to maximise the log-likelihood

$$l = \log L = \sum_{i=1}^{N} \left[\{y_i\theta_i - a(\theta_i)\}/\phi + b(y_i, \phi)\right]. \tag{3.29}$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{N} \left(\frac{\partial \theta_i}{\partial \beta}\right)(y_i - \mu_i)/\phi. \tag{3.30}$$

When a canonical link function $g(\mu_i) = \theta_i = \eta_i$ has been assumed,

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{N} X_i(y_i - \mu_i)/\phi. \tag{3.31}$$

Solving this set of equations

$$\sum_{i=1}^{N} X_i(y_i - \mu_i) = 0, \tag{3.32}$$

yields the MLE's of β. Generally this requires iterative procedures.

Chapter 4

Nonparametric Regression Methods for Longitudinal Data Analysis.

4.1 Introduction.

Nonparametric regression methods for longitudinal data analysis have been a popular statistical research topic since the late 1990s. The needs of longitudinal data analysis from biomedical research and other scientific areas along with the recognition of the limitation of parametric models in practical data analysis have driven the development of more innovative nonparametric regression methods. Because of the flexibility in the form of regression models, nonparametric modelling approaches can play an important role in exploring longitudinal data, just as they have done for independent cross-sectional data analysis.

Parametric mixed-effects models are a powerful tool for modelling the relationship between a response variable and covariates in longitudinal studies. However, for many applications, although parametric mean models enjoy simplicity, they may be too restrictive, inflexible or limited, and sometimes unavailable for preliminary data analyses as well as for modelling complicated relationships between the response and covariates in various longitudinal studies. One of the basic assumptions for the parametric models is that the response variable (or *via* a known link function) is a known parametric function of fixed-effects and random-effects. That is, for each individual, the underlying relationship between the response and the mixed-effects covariates is parametric. However, this assumption is not always satisfied in practical applications. To overcome this difficulty, nonparametric regression techniques have been developed for longitudinal data analysis in recent years.

If an inappropriate parametric model is used, it is possible to produce misleading conclusions from the regression analysis. To overcome the difficulty caused by the restrictive assumption of a parametric form of the regression function, one may remove the restriction that the regression function belongs to a parametric family. This approach leads to so-called nonparametric regression. Practical applications have placed a strong demand on developing non-parametric and semiparametric regression methods for longitudinal data, where flexible functional forms can be estimated from the data to capture possibly complicated relationships between longitudinal outcomes and covariates.

The importance of nonparametric modelling methods has been recognized in longitudinal data analysis and for practical applications, since nonparametric methods

are flexible and robust against parametric assumptions. Such flexibility is useful for exploration and analysis of longitudinal data, when appropriate parametric models are unavailable.

Non-parametric regression methods can be broadly classified into kernel methods, which are often based on local likelihoods, and splines, which include smoothing splines, penalized splines, and regression splines.

Let the data be denoted as

$$(t_i, y_i), i = 1, 2, \ldots, n, \tag{4.1}$$

where $t_i, i = 1, 2, \ldots, n$ are known as design time points, and $y_i, i = 1, 2, \ldots, n$ are the responses at the design time points. The design time points may be equally spaced in an interval of interest, or be regarded as a random sample from a continuous design density, namely, $\pi(t)$. For simplicity, let us denote the interval of interest, or the support of $\pi(t)$ as $\mathcal{T}$, which can be a finite interval, *e.g.*, $[a, b]$ or the whole real line $(-\infty, \infty)$.

A simple nonparametric regression model is usually written as

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.2}$$

where $f(t)$ models the underlying regression function that we want to estimate, but cannot be approximated using a parametric model adequately, and $\epsilon_i = 1, 2 \ldots, n$ denote the measurement errors that cannot be explained by the regression function

$f(t)$. Mathematically, $f(t)$ is the conditional expectation of y_i, given $t_i = t, i.e.,$

$$f(t) = E(y_i|\ t_i = t), \quad i = 1, 2, \ldots, n.$$

There are many existing smoothers, with different strengths in one aspect or another, that can be used to estimate $f(t)$ in (4.2). For example, smoothing splines may be good for handling sparse data, while local polynomial smoothers may be computationally advantageous for handling dense designs. We shall review four of the most popular smoothers that include local polynomial smoothers (Wand and Jones 1995, Fan and Gijbels 1996), regression splines (Eubank 1988,1999, Stone *et al* 1997), smoothing splines (Wahba 1990, Green and Silverman 1994), and P-splines (Eilers and Marx, 1996, Ruppert *et al* 2003), and briefly describe linear smoothers, which include the above four popular smoothers as special cases.

4.2 Local Polynomial Kernel Smoother

4.2.1 General Degree LPK Smoother

Local polynomial kernel (LPK) smoothing considers locally approximating the f in (4.2) by a polynomial of appropriate degree, using Taylor expansion, which states that any smooth function can be locally approximated by a polynomial of some degree.

Let t_0 be an arbitrary fixed time point where the function f in (4.2) will be

estimated. Assume $f(t)$ has a $(p+1)$'st continuous derivative for some integer $p \geq 0$ at t_0. By Taylor expansion, $f(t)$ can be locally approximated as,

$$f(t) \approx f(t_0) + (t - t_0)f^{(1)}(t_0) + \cdots + (t - t_0)^p f^{(p)}(t_0)/p\ !\ \ ,$$

in a neighbourhood of t_0 that allows the above expansion where $f^{(r)}(t_0)$ denotes the r^{th} derivative of $f(t)$ at t_0.

Set $\beta_r = f^{(r)}(t_0)/r!, \quad r = 0, \ldots, p$. Let $\widehat{\beta},\ r = 0, 1, 2, \ldots, p$ be the minimisers of the following weighted least squares (WLS) criterion:

$$\sum_{i=1}^{n} \{y_i - [\beta_0 + (t_i - t_0)\beta_1 + \cdots + (t_i - t_0)^p \beta_p]\}^2 K_h(t_i - t_0), \tag{4.3}$$

where $K_h(\cdot) = K(\cdot|h)/h$, which is obtained *via* re-scaling a kernel function $K(\cdot)$ with a constant $h > 0$, called the bandwidth or smoothing parameter. The bandwidth h determines the size of the local neighbourhood, namely,

$$I_h(t_0) = [t_0 - h, t_0 + h], \tag{4.4}$$

where local fitting is conducted. The kernel function, $K(\cdot)$, determines how observations within $I_h(t_0)$ contribute to the fit at t_0. Let $\hat{f}_h^{(r)}(t_0)$ denote the estimate of the r^{th} derivative $f^{(r)}(t_0)$. Then

$$\hat{f}_h^{(r)}(t_0) = r!\hat{\beta}_r, \quad r = 0, 1, \ldots, p.$$

An explicit expression for $\hat{f}_h^{(r)}(t_0)$ is useful and can be made *via* matrix notation. Let

$$\boldsymbol{X} = \begin{pmatrix} 1 & (t_1 - t_0) & \cdots & (t_1 - t_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (t_n - t_0) & \cdots & (t_n - t_0)^p \end{pmatrix},$$

and

$$\boldsymbol{W} = \text{diag}(K_h(t_1 - t_0), \ldots, K_h(t_n - t_0)),$$

be the design matrix and the weight matrix for the LPK fit around t_0. Then the WLS criterion (4.3) can be rewritten as

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{4.5}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$. It follows that

$$\hat{f}_h^{(r)}(t_0) = r!\boldsymbol{e}'_{r+1}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}, \quad r = 0, 1, \ldots, p,$$

where $\boldsymbol{e}_{r+1}$ denotes a $(p+1)$ - dimensional unit vector whose $(r+1)^{st}$ entry is 1 and the other entries are 0. When t_0 runs over the whole support $\mathcal{T}$ of the design time points, a whole range estimation of $f^{(r)}(t)$ is obtained. The derivative estimator $\hat{f}_h^{(r)}(t), t \in \mathcal{T}$ is usually called the LPK smoother of the underlying derivative function $f^{(r)}(t)$. The derivative smoother$\hat{f}_h^{(r)}(t_0)$ is usually calculated on a grid of t's in $\mathcal{T}$.

We focus only on the curve smoother

$$\hat{f}_h(t_0) = \boldsymbol{e}'_1(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}, \tag{4.6}$$

unless we discuss derivative estimation. Set $\widehat{y}_i = \hat{f}_h(t_i)$ to be the fitted value of $f(t_i)$. By (4.6), it is seen that

$$\hat{f}_h(t_i) = \boldsymbol{a}(t_i)'\boldsymbol{y},$$

where $\boldsymbol{a}(t_i)$ is $\boldsymbol{e}_1'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}$ after replacing t_0 with t_i. Let $\widehat{\boldsymbol{y}}_h = [\widehat{y}_1, \ldots, \widehat{y}_n]'$ denote the fitted values at all the design time points. Then $\widehat{\boldsymbol{y}}_h$ can be expressed as

$$\widehat{\boldsymbol{y}}_h = \boldsymbol{A}_h\boldsymbol{y}, \tag{4.7}$$

where

$$\boldsymbol{A}_h = (\boldsymbol{a}(t_1), \ldots, \boldsymbol{a}(t_n))', \tag{4.8}$$

is known as the smoother matrix of the LPK smoother. Since $\boldsymbol{A}_h$ does not depend on the response vector $\boldsymbol{y}$, the LPK smoother $\hat{f}_h$ is known as a linear smoother.

4.2.2 Local Constant and Linear Smoothers

Local constant and linear smoothers are the two simplest and most useful LPK smoothers. The local constant smoother is known as the Nadaraya-Watson estimator (Nadaraya 1964, Watson 1964). This smoother results from the LPK smoother $\hat{f}_h(t_0)$ (4.6) by simply taking $p = 0$:

$$\hat{f}_h(t_0) = \frac{\sum_{i=1}^n K_h(t_i - t_0)y_i}{\sum_{i=1}^n K_h(t_i - t_0)}. \tag{4.9}$$

Within a local neighbourhood $I_h(t_0) = [t_0 - h, t_0 + h]$, it fits the data with a constant. That is, it is the minimiser β_0 of the following WLS criterion:

$$\sum_{i=1}^{n} (y_i - \beta_0)^2 K_h(t_i - t_0).$$

The Nadaraya-Watson estimator is simple to understand and easy to compute. Let $I_A(t)$ denote the indicator function of some set A. When the kernel function K is the Uniform kernel

$$K(t) = I_{[-1,1]}(t)/2 = \begin{cases} \frac{1}{2}, & t \in [-1, 1], \\ 0, & \text{otherwise}, \end{cases} \tag{4.10}$$

as depicted in the left panel of Figure 4.1, the Nadaraya-Watson estimator (4.9) is exactly the local average of y_i's that are within the local neighbourhood $I_h(t_0)$ (4.4):

$$\hat{f}_h(t_0) = \frac{\sum_{i=1}^{n} I_{[t_0-h,t_0+h]}(t_i) y_i}{\sum_{i=1}^{n} I_{[t_0-h,t_0+h]}(t_i)} = \left\{ \sum_{t_i \in I_h(t_0)} y_i \right\} / m_h(t_0),$$

where $m_h(t_0)$ denotes the number of the observations falling into the local neighbourhood $I_h(t_0)$. However, when t_0 is at the boundary of $\mathcal{T}$, fewer design points are within the neighbourhood $I_h(t_0)$ so that $\hat{f}_h(t_0)$ has a slower convergence rate than the case when t_0 is inside $\mathcal{T}$. For a detailed explanation of this boundary effect, one may refer to Fan and Gijbels (1996) and Cheng *et al* (1997).

The local linear smoother (Stone 1984, Fan 1992, 1993) is obtained *via* fitting a data set locally with a linear function. Let $\widehat{\beta}_0, \widehat{\beta}_1$ minimise the following WLS criterion

$$\sum_{i=1}^{n} [y_i - \beta_0 - (t_i - t_0)\beta_1]^2 K_h(t_i - t_0).$$

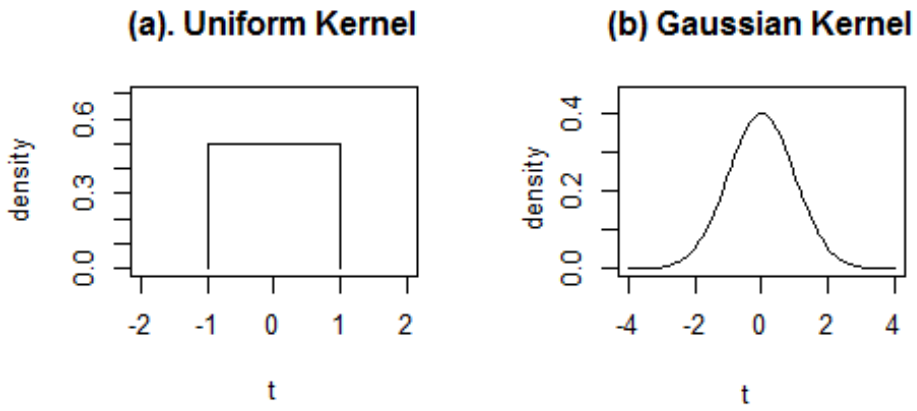


Figure 4.1: Two widely used kernel functions

Then the local linear smoother is $\hat{f}_h(t_0) = \widehat{\beta}_0$. It can be easily obtained from the LPK smoother $\hat{f}_h(t_0)$ (4.6) by simply taking $p = 1$. It is known as a smoother with a free boundary effect (Cheng *et al* 1997). That is, it has the same convergence rate at any point in $\mathcal{T}$. It also exhibits many good properties that the other linear smoothers may lack. Good discussions on these properties can be found in Fan (1992, 1993), Hastie and Loader (1993), and Fan and Gijbels (1996), among others. A local linear smoother can be simply expressed as

$$\hat{f}_h(t_0) = \frac{\sum_{i=1}^{n}[s_2(t_0) - s_1(t_0)(t_i - t_0)]K_h(t_i - t_0)y_i}{s_2(t_0)s_0(t_0) - s_1^2(t_0)}, \tag{4.11}$$

where

$$s_r(t_0) = \sum_{i=1}^{n} K_h(t_i - t_0)(t_i - t_0)^r, \quad r = 0, 1, 2.$$

Usually, the choice of the LPK fitting degree, p, is not as important as the choice

of the bandwidth, h. A local constant ($p = 0$) or a local linear ($p = 1$) smoother is often good enough for most application problems if the kernel function K and the bandwidth h are adequately determined. Fan and Gijbels (1996) pointed out that for curve estimation (not valid for derivative estimation) an odd p is preferable. This is true since a LPK fit with $p = 2q + 1$, introduces an extra parameter compared to a LPK fit with $p = 2q$, but does not increase the variance of the associated LPK estimator. However, the associated bias may be significantly reduced especially in the boundary regions (Fan 1992, 1993, Hastie and Loader 1993, Fan and Gijbels 1996, Cheng *et al* 1997). Thus, the local linear smoother is strongly recommended for most problems in practice.

4.2.3 Kernel Function

The kernel function $K(.)$ used in the LPK smoother (4.6) is usually a symmetric probability density function. While the bandwidth h specifies the size of the local neighbourhood $I_h(t_0)$, the kernel $K(.)$ specifies how the observations contribute to the LPK fit at t_0.

Figure 4.1 shows two widely-used kernel functions. The left panel shows the Uniform kernel (4.10) and the right panel shows the Gaussian kernel (standard normal probability density function). When the Uniform kernel is used, all the t_i's within the local neighbourhood $I_h(t_0)$ contribute equally (the weights are the same) in the LPK fit at t_0, while all the t_i's outside the neighbourhood contribute nothing. When the Gaussian kernel is used, however, the contribution of the t_i's is determined by

the distance of t_i from t_0, that is, the smaller the distance $(t - t_0)$, the larger the contribution. This is because the Gaussian kernel is bell-shaped and peaked at the origin. The Uniform kernel has a bounded support which allows the LPK fit to use the data only in the neighbourhood $I_h(t_0)$. This makes a fast implementation of the LPK fit possible, which is advantageous especially for large data sets. The use of the Gaussian kernel often results in good visual effects of the LPK smoothers, but pays a price of requiring more computation effort.

The choice of a kernel is usually not so crucial since it does not determine the convergence rate of the LPK smoother (4.6) to the underlying curve. However, it does determine the relative efficiency of the LPK smoother. For more discussion about kernel choice, see Gasser *et al* (1985), Fan and Gijbels (1996), Zhang and Fan (2000) and references therein.

4.2.4 Bandwidth Selection

A smoother is considered to be good if it produces a small prediction error, usually measured by the Mean Squared Error (MSE) or the Mean Integrated Squared Error (MISE) of the smoother. For the LPK smoother $\hat{f}_h(t_0)$, its MSE and MISE are defined as

$$\begin{aligned} \text{MSE}\left(\hat{f}_h(t_0)\right) &= E\left(\hat{f}_h(t_0) - f(t_0)\right)^2 \\ &= \text{Bias}^2\left(\hat{f}_h(t_0)\right) + \text{Var}\left(\hat{f}_h(t_0)\right), \\ MISE(\hat{f}_h) &= \int \text{MSE}\left(\hat{f}_h(t)\right) w(t)dt, \end{aligned} \tag{4.12}$$

where $w(t)$ is a weight function, often used to specify a particular range of interest.

Under some regularity conditions including that t_0 is an interior point, we can show that as $n \to \infty$,

$$\text{Bias}\left(\hat{f}_h(t_0)\right) = \begin{cases} O_P\left(h^{p+1}\right), \text{as } p \text{ odd}, \\ O_P\left(h^{p+2}\right), \text{as } p \text{ even}, \end{cases} \tag{4.13}$$

$$\text{Var}\left(\hat{f}_h(t_0)\right) = O_P\left((nh)\right)^{-1}, \tag{4.14}$$

where $X = O_P(Y)$ means X/Y is bounded in probability. See, for example, Fan and Gijbels (1996) for more details. From this, we can see that the bandwidth h controls the trade-off between the squared bias and the variance of the LPK smoother $\hat{f}_h(t_0)$. When h is small, the squared bias is small but the variance is large. On the other hand, when h is large, the squared bias is large while the variance is small. A good choice of h will generally trade-off these two terms so that the associated MSE or MISE is minimised.

The role played by the bandwidth h can also be seen intuitively. As mentioned previously, the bandwidth h specifies the size of the local neighbourhood $I_h(t_0) = [t_0 - h, t_0 + h]$. When h is small, $I_h(t_0)$ contains only a few observations so that $\hat{f}_h(t_0)$ can be well adjusted based on the WLS criterion (4.3) to closely approximate $f(t_0)$. This implies a small bias for $\hat{f}_h(t_0)$. However, since only a few observations are involved in the LPK fit, the variance of the estimator is large. With a similar reasoning, when h is large, $I_h(t_0)$ contains many observations so that $\hat{f}_h(t_0)$ has a large bias but a small variance.

It is then natural to select a global bandwidth h so that the MISE (MSE for a local bandwidth) of $\hat{f}_h(t_0)$ is minimised. Unfortunately, the MISE (4.12) is not computable since f is, after all, unknown and is the target to be estimated. This problem can be overcome by selecting h to minimise some estimator of the MISE. An estimator of the MISE may be obtained *via* estimating the unknown quantities in the asymptotic MISE expression using some higher degree LPK fit, resulting in the so-called plug-in bandwidth selectors (Fan and Gijbels 1992, Ruppert *et al* 1995).

4.3 Regression Splines

Polynomials are not flexible in their ability to model data across a large range of values. However, this is not the case when the range is small enough. In local polynomial kernel (LPK) smoothing introduced in Section 4.2, local neighbourhoods were specified by a bandwidth h and a fixed time point t_0. In regression spline smoothing that we shall discuss in this section, local neighbourhoods are specified by a group of locations, say,

$$\tau_0, \tau_1, \tau_2, \ldots, \tau_K, \tau_{K+1}, \tag{4.15}$$

in the range of interest, say, an interval $[a, b]$ where $a = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = b$. These locations are known as knots. These knots divide the interval of interest, $[a, b]$, into K subintervals (local neighbourhoods):

$$[\tau_r, \tau_{r+1}), \quad r = 0, 1, \ldots, K,$$

so that within any two neighbouring knots, a Taylor's expansion up to some degree is valid. In other words, a regression spline is a piecewise polynomial which is a polynomial of some degree within any two neighbouring knots τ_r and τ_{r+1} for $r = 0, 1, \ldots, K$ and is joined together at knots properly but allows discontinuous derivatives at the knots.

4.3.1 Truncated Power Basis

A regression spline can be constructed using the following so-called k^{th} degree truncated power basis with K knots $\tau_1, \tau_2 \ldots, \tau_k$:

$$1, t, \ldots, t^k, (t - \tau_1)_+^k, \ldots, (t - \tau_K)_+^k, \tag{4.16}$$

where $w_+^k = [w_+]^k$ denotes power k of the positive part of w with $w_+ = \max(0, w)$. Note that the first $(k + 1)$ basis functions of the truncated power basis (4.16) are polynomials of degree up to k, and the others are all the truncated power functions of degree k.

Using the above truncated power basis (4.16), a regression spline can be expressed as

$$f(t) = \sum_{s=0}^{k} \beta_s t^s + \sum_{r=1}^{K} \beta_{k+r} (t - \tau_r)_+^k, \tag{4.17}$$

where $\beta_0, \beta_1, \ldots, \beta_{k+K}$ are the associated coefficients. For convenience, it may be called a regression spline of degree k with knots $\tau_1, \tau_2 \ldots, \tau_K$. The regression splines (4.17) associated with k= 1,2 and 3 are usually called linear, quadratic, and cubic

regression splines, respectively.

We can see that within any subinterval or local neighbourhood $[\tau_r, \tau_{r+1})$, we have

$$f(t) = \sum_{s=0}^{k} \beta_s t^s + \sum_{l=1}^{r} \beta_{k+l}(t - \tau_l)_{,}^{k}$$

which is a k^{th} degree polynomial. However, for $r = 1, 2, \ldots, K$,

$$f^{(k)}(\tau_r -) = k! \left(\beta_k + \sum_{l=1}^{r-1} \beta_{k+l} \right),$$

$$f^{(k)}(\tau_r +) = k! \left(\beta_k + \sum_{l=1}^{r} \beta_{k+l} \right).$$

Therefore

$$f^{(k)}(\tau_r +) - f^{(k)}(\tau_r -) = k! \beta_{k+r}. \tag{4.18}$$

That is, $f^{(k)}(t)$ jumps at τ_r with amount $k!$ β_{k+r} for $r = 1, 2, \ldots, K$. In other words, a regression spline of degree k with knots $\tau_1, \tau_2, \ldots, \tau_K$, has continuous derivatives up to $k - 1$ times, and has a discontinuous k-times derivative; the coefficient β_{k+r} of the r^{th} truncated power basis function measures how large the jump is (up to a constant multiplicity of $k!$).

Figure 4.2(a) presents a cubic truncated power basis with knots .2, .4, .6, and .8. It includes the first four polynomials $1, t, t^2, t^3$, and the four truncated power functions with nonzero-values starting at knots .2, .4, .6 and .8, respectively. Figure 4.2(b) displays three cubic regression splines as simple linear combinations (4.17) of the truncated power basis using coefficients generated randomly. It can be seen that the

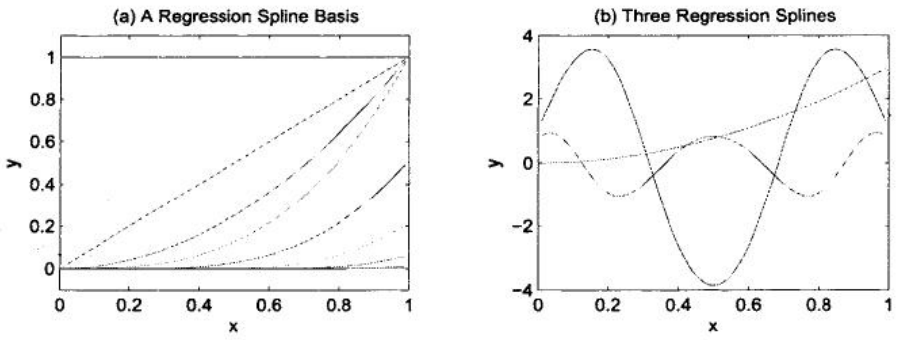


Figure 4.2: Example of Cubic regression spline basis, and three cubic regression splines

truncated power basis is flexible in describing functions from simple to complicated ones.

4.3.2 Regression Spline Smoother

For convenience, it is often useful to denote the truncated power basis (4.16) as

$$\mathbf{\Phi}_p(t) = (1, t, \dots t^k, (t-\tau_1)_+^k, \dots, (t-\tau_K)_+^k)', \tag{4.19}$$

where $p = K + k + 1$ denotes the number of the basis functions involved. Similarly, denote the associated coefficients as

$$\boldsymbol{\beta} = (\beta_0, \dots, \beta_k, \beta_{k+1}, \dots, \beta_{k+K})'.$$

Then the regression spline (4.17) can be re-expressed as

$$f(t) = \boldsymbol{\Phi}_p(t)'\boldsymbol{\beta}, \tag{4.20}$$

so that the model (4.2) can be re-written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.21}$$

where,

$$\begin{aligned} \boldsymbol{y} &= (y_1, \ldots, y_n)', \\ \boldsymbol{X} &= (\boldsymbol{\Phi}_p(t_1), \ldots, \boldsymbol{\Phi}_p(t_n))', \\ \boldsymbol{\epsilon} &= (\epsilon_1, \ldots, \epsilon_n)'. \end{aligned}$$

Since $\boldsymbol{\Phi}_p(t)$ is a basis, $\boldsymbol{X}$ is of full rank, and hence $\boldsymbol{X}'\boldsymbol{X}$ is invertible when $n \geq p$. A natural estimator for β, which solves the approximation linear model (4.21) by the ordinary least squares (OLS) method, is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{4.22}$$

It follows that the regression spline fit of the function $f(t)$ in (4.2) is

$$\hat{f}_p(t) = \boldsymbol{\Phi}_p(t)'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{4.23}$$

which is often called a regression spline smoother of f. In particular, the values of $\hat{f}_p(t)$ evaluated at the design time points $t_i, i = 1, 2, \ldots, n$ are collected in the

following fitted response vector:

$$\widehat{\boldsymbol{y}}_p = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{A}_p\boldsymbol{y}, \tag{4.24}$$

where, $\widehat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)'$ with $\hat{\boldsymbol{y}}_i = \hat{f}_p(t_i), i = 1, 2, \ldots, n$, and

$$\boldsymbol{A}_p = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \tag{4.25}$$

is called the regression spline smoother matrix. It is easy to notice that $\boldsymbol{A}_p$, is a projection matrix satisfying $\boldsymbol{A}_p' = \boldsymbol{A}_p, A_p^2 = \boldsymbol{A}_p$, and $\mathrm{tr}(\boldsymbol{A}_p) = p$. The trace of the smoother matrix $\boldsymbol{A}_p$, is often called the degrees of freedom of the regression spline smoother. It measures the complexity of the regression spline model used.

4.3.3 Selection of Number and Location of Knots

Good performance of the regression spline smoother (4.23) strongly depends on good knot locations, $\tau_1, \ldots, \tau_k$, and good choice of the number of knots, K. The degree of the regression spline basis (4.16), k, is usually less crucial, and it is often taken as 1, 2 or 3 for computational convenience. Three widely-used methods for locating the knots are listed as follows:

Equally Spaced Method. This method takes K equally spaced points in the range of interest, say, [a, b], as knots. That is, the K knots are defined as

$$\tau_r = a + (b - a)r/(K + 1), \quad r = 1, 2, \ldots, K. \tag{4.26}$$

This method of knots placing is independent of the design time points. It is usually employed when the design time points are believed to be uniformly scattered in the range of interest.

Equally Spaced Sample Quantiles as Knots Method. This method uses equally spaced sample quantiles of the design time points $t_i, i = 1, 2, \ldots, n$ as knots. Let $t_{(1)}, \ldots, t_{(n)}$ be the order statistics of the design time points. Then the K knots are defined as

$$\tau_r = t_{(1+[rn/(K+1)])}, \quad r = 1, 2, \ldots, K, \tag{4.27}$$

where $[a]$ denotes the integer part of a. This method of knots placing is design adaptive. It locates more knots where more design time points are scattered. When the design time points are uniformly scattered, it is approximately equivalent to the equally spaced method.

Model Selection Based Method. This method uses all the distinct design time points as knot candidates. Note that for the truncated power basis (4.16), deletion of a knot is equivalent to deletion of a truncated power basis function associated with the knot. This is equivalent to the deletion of a covariate in the approximation linear model (4.21). Therefore, knot selection can be done *via* model selection methods such as forward selection, backward elimination, or stepwise regression for introducing or deleting a knot from the knot candidates (Stone *et al* 1997).

For the last method, the selection of the number of knots is done at the same time as knot introduction or deletion. But for the first two methods, after a knot placing method is specified, the number of knots, K, is generally chosen based on a

smoothing parameter selection criterion such as those to be presented in Section 4.6 for linear smoothers. For more knot selection methods for regression splines, see, for example, Friedman and Silverman (1989), Friedman (1991), and Smith and Kohn (1996), among others.

4.4 Smoothing Splines

For regression splines, once the knot locating method is specified, the next task is to choose the number of knots, K, which in general is smaller than the sample size n. Since K must be an integer, the opportunity for adjusting K is limited, and the adjustment is usually rather crude. Alternatively, we can use all the distinct design time points as knots. But too many knots leads to undersmoothing and this causes curve to be quite rough, showing dramatic curve changes over a short interval. Smoothing splines overcome this difficulty by introducing a penalty for the roughness of the curve under consideration.

Without loss of generality, we may assume the range of interest of f in (4.2) to be a finite interval, say, $[a, b]$ for some finite numbers a and b. The roughness of f is usually defined as the integral of its squared k-times derivative

$$\int_a^b \left\{ f^{(k)}(u) \right\}^2 du, \tag{4.28}$$

for some $k \geq 1$. Then the smoothing spline smoother of the f in (4.2) is defined as

the minimizer $\hat{f}_\lambda(t)$ of the following penalized least squares (PLS) criterion:

$$\sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \lambda \int_a^b \left\{f^{(k)}(t)\right\}^2 \, dt, \tag{4.29}$$

over the k^{th} order Sobolev space $\mathcal{W}_2^k[a,b]$:

$$\left\{ f : f^{(r)} \text{ absolute continuous for } 0 \leq r \leq k-1, \int_a^b \left\{f^{(k)}(t)\right\}^2 \, dt < \infty \right\}, \tag{4.30}$$

where $\lambda > 0$ is a smoothing parameter controlling the size of the roughness penalty, and it is usually used to trade-off the goodness of fit, indicated by the first term in (4.29), and the roughness of the resulting curve. The $\hat{f}_\lambda(t)$ is known as a natural smoothing spline of degree $(2k-1)$. For example, when $k = 2$, the associated $\hat{f}_\lambda(t)$ is a natural cubic smoothing spline (NCSS). For a detailed description of smoothing splines, see for example, Eubank (1988,1999), Wahba (1990), Green and Silverman (1994) and Gu (2002), among others.

4.4.1 Cubic Smoothing Splines

To minimize the PLS criterion (4.29), we need to compute the integral that defines the roughness, and estimate parameters numbering up to the sample size n. This is a challenging aspect for computing a smoothing spline.

When $k = 2$, however, the associated cubic smoothing spline is less computationally challenging. Actually, there is a way to compute the roughness term quickly as stated in Green and Silverman (1994). That is one of the reasons why cubic

smoothing splines are popular in statistical applications.

Let $\tau_1, \tau_2, \ldots, \tau_K$ be all the distinct design time points and be sorted in an increasing order. They are all the knots of the cubic smoothing spline $\hat{f}_\lambda(t)$ that minimizes (4.29) when $k = 2$. Set

$$h_r = \tau_{r+1} - \tau_r, r = 1, 2, \ldots, K-1.$$

Define $\boldsymbol{A} = (a_{rs})$ as a $K \times (K-2)$ matrix with all the entries being 0 except for $r = 1, 2, \ldots, K-2$,

$$a_{r,r} = h_r^{-1}, \quad a_{r+1,r} = -(h_r^{-1} + h_{r+1}^{-1}), \quad a_{r+2,r} = -h_{r+1}^{-1}.$$

Define $\boldsymbol{B} = (b_{rs})$ as a $(K-2) \times (K-2)$ matrix with all the entries being 0 except

$$b_{11} = (h_1 + h_2)/3, \quad b_{21} = h_2/6,$$

and for $r = 1, 2, \ldots, K-4$,

$$b_{r,r+1} = h_{(r+1)}/6, \quad b_{r+1,r+1} = (h_{(r+1)} + h_{(r+2)})/3, \quad b_{r+2,r+1} = h_{(r+2)}/6,$$

and

$$b_{K-3,K-2} = h_{(K-2)}/6, \quad b_{K-2,K-2} = (h_{(K-2)} + h_{(K-1)})/3.$$

Finally define a $K \times K$ matrix $\boldsymbol{G}$:

$$\boldsymbol{G} = \boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}'. \tag{4.31}$$

Let $\boldsymbol{f} = (f_1, \ldots, f_K)'$ where $f_r = f(\tau_r), r = 1, 2, \ldots, K$. Then it is easy to show that the roughness, (4.28), of $f(t)$ for $k = 2$ can be expressed as

$$\int_a^b \{f'(t)\}^2 dt = \boldsymbol{f}'\boldsymbol{G}\boldsymbol{f}. \tag{4.32}$$

Therefore, we often refer to $\boldsymbol{G}$ as a roughness matrix. It follows that the PLS criterion (4.29) can be written as

$$\| \boldsymbol{y} - \boldsymbol{W}\boldsymbol{f} \|^2 + \lambda \boldsymbol{f}'\boldsymbol{G}\boldsymbol{f}, \tag{4.33}$$

where $\boldsymbol{W} = (w_{ir})$ is an $n \times K$ incidence matrix with $w_{ir} = 1$ if $t_i = \tau_r$ and 0 otherwise, and $\| \boldsymbol{a} \|^2 = \sum_{i=1}^n a_i^2$ denotes the usual L^2-norm of $\boldsymbol{a}$. Therefore an explicit expression for the cubic smoothing spline $\hat{\boldsymbol{f}}_\lambda$, evaluated at the knots τ_r, $r = 1, 2, \ldots, K$, is as follows:

$$\hat{\boldsymbol{f}}_\lambda = (\boldsymbol{W}'\boldsymbol{W} + \lambda \boldsymbol{G})^{-1}\boldsymbol{W}'\boldsymbol{y}. \tag{4.34}$$

The fitted response vector at the design time points is

$$\widehat{\boldsymbol{y}}_\lambda = \boldsymbol{A}_\lambda \boldsymbol{y}, \tag{4.35}$$

where

$$\boldsymbol{A}_\lambda = \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W} + \lambda \boldsymbol{G})^{-1}\boldsymbol{W}', \tag{4.36}$$

is known as the cubic smoothing spline smoother matrix. The expression (4.34) indicates that the cubic smoothing spline smoother is a linear smoother as defined later in Section 4.6. When all the design time points are distinct,

$$\widehat{\boldsymbol{y}}_\lambda = \hat{\boldsymbol{f}}\lambda = (\boldsymbol{I}_n + \lambda \boldsymbol{G})^{-1}\boldsymbol{y},$$

since $\boldsymbol{W} = \boldsymbol{I}_n$, an identity matrix of size n. For more details about cubic smoothing splines, see Green and Silverman (1994), among others.

4.4.2 General Degree Smoothing Splines

For $k \neq 2$, computation of the roughness term in (4.29) is quite challenging. However, it becomes easier if a good basis is provided for the underlying function $f(t)$ in (4.2). Let $\boldsymbol{\Phi}_p(t) = (\phi_1(t), \ldots, \phi_p(t))'$ denote a basis so that $\phi_r^{(k)}(t)$, $r = 1, 2, \ldots, p$ are squared integrable. Then we can express the $f(t)$ in (4.2) as $f(t) = \boldsymbol{\Phi}_p(t)'\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is a p dimensional coefficient vector, and the roughness

$$\int_a^b \{f^{(k)}(t)\}^2 dt = \boldsymbol{\alpha}'\boldsymbol{G}\boldsymbol{\alpha},$$

where obviously

$$\boldsymbol{G} = \int_a^b \boldsymbol{\Phi}_p^{(k)}(t)\boldsymbol{\Phi}_p^{(k)}(t)' dt. \tag{4.37}$$

It follows that the PLS criterion (4.29) can be expressed as

$$\| \boldsymbol{y} - \boldsymbol{W}\boldsymbol{\alpha} \|^2 + \lambda\boldsymbol{\alpha}'\boldsymbol{G}\boldsymbol{\alpha}, \tag{4.38}$$

where $\boldsymbol{W} = (\boldsymbol{\Phi}_p(t_1), \ldots, \boldsymbol{\Phi}_p(t_n))'$. Therefore, $\widehat{\boldsymbol{y}}_\lambda$ can be expressed as

$$\widehat{\boldsymbol{y}}_\lambda = \boldsymbol{A}_\lambda \boldsymbol{y}, \tag{4.39}$$

where

$$\boldsymbol{A}_\lambda = \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W} + \lambda\boldsymbol{G})^{-1}\boldsymbol{W}', \tag{4.40}$$

is the associated smoother matrix.

Note that $\mathbf{\Phi}_p(t)$ can be a truncated power basis (4.16) of degree $(2k-1)$ with knots at all the distinct design time points among $t_i, i = 1, 2, \ldots, n$, a B-spline basis (de Boor 1978) or a reproducing kernel Hilbert space basis (Wahba 1990) or any other basis.

4.4.3 Choice of Smoothing Parameters

The smoothing parameter λ plays an important role in the smoothing spline smoother (4.39). It trades-off the bias with the variance of the smoothing spline smoother $\hat{\boldsymbol{f}}_\lambda$ (4.39). Since $\hat{\boldsymbol{f}}_\lambda$ is a linear smoother, good choice of λ can be obtained by applying the smoothing parameter selectors such as cross validation (CV), generalised cross validation (GCV), AIC or BIC.

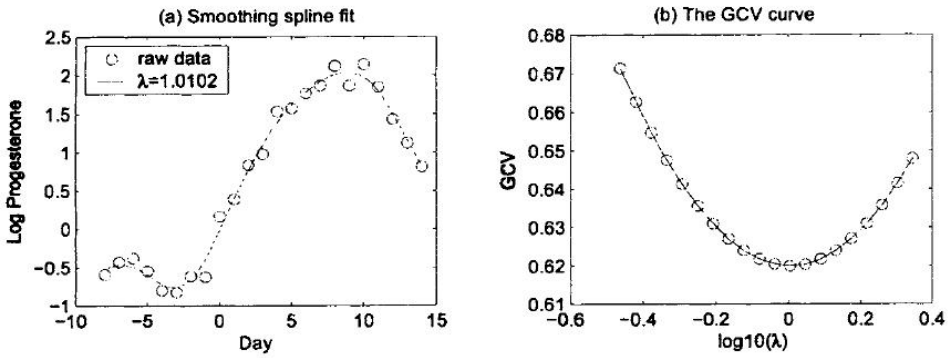


Figure 4.3: (a) A cubic smoothing spline fit to the progesterone data from a subject. (b) The associated GCV curve.

Figure 4.3(a) presents a NCSS fit (solid curve) of the progesterone data from a

subject. The data were collected in a study of early pregnancy loss conducted by the Institute for Toxicology and Environmental Health at the Reproductive Epidemiology Section of the California Department of Health Services, Berkeley, USA. The fit was obtained using the cubic smoothing spline smoother (4.35) with a smoothing parameter $\lambda = 1.0102$, selected by GCV. Figure 4.3(b) presents the GCV curve against λ in a $\log_{10}$-scale.

4.5 Penalized Splines

Let $\boldsymbol{\Phi}_p(t)$ be the truncated power basis (4.19) of degree k with K knots $\tau_1, \tau_2, \ldots, \tau_K$. Then $f(t)$ can be first written as a regression spline $\boldsymbol{\Phi}_p(t)'\boldsymbol{\alpha}$ where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{k+K})'$ is the associated coefficient vector. Let

$$\boldsymbol{G} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_K \end{pmatrix}, \tag{4.41}$$

be a $p \times p$ diagonal matrix. Then P-spline smoother of the function $f(t)$ in (4.2) is defined as $\hat{f}_\lambda(t) = \boldsymbol{\Phi}_p(t)'\widehat{\boldsymbol{\alpha}}$ where $\widehat{\boldsymbol{\alpha}}$ is the minimizer of the following PLS criterion:

$$\sum_{i=1}^{n} (y_i - \boldsymbol{\Phi}_p(t_i)'\boldsymbol{\alpha}^2 + \lambda \boldsymbol{\alpha}'\boldsymbol{G}\boldsymbol{\alpha}. \tag{4.42}$$

Note that

$$\boldsymbol{\alpha}'\boldsymbol{G}\boldsymbol{\alpha} = \sum_{i=1}^{K} \alpha_{k+r}^2,$$

and α_{k+r} is the coefficient of the r^{th} truncated power basis function $(t-\tau_r)_+^2$ in the basis $\boldsymbol{\Phi}_p(t)$. The coefficient α_{k+r}, measures the jump of the k-times derivative of $f(t)$ at the knot τ_r. It follows that the penalty in (4.42) is imposed just for the derivative jumps of the resulting P-spline at the knots. Therefore, the larger the coefficients $\alpha_{k+r},\ r=1,2,\ldots,K$, the rougher the resulting P-spline. Thus, the term $\boldsymbol{\alpha}'\boldsymbol{G}\boldsymbol{\alpha}$ is a measure of the roughness of the resulting P-spline. For convenience, we call $\boldsymbol{G}$ the P-spline roughness matrix.

Since there is essentially no difference between (4.42) and (4.38), it is expected that a P-spline has essentially the same formula as that for a smoothing spline (4.39). Actually, the P-spline smoother $\hat{f}_\lambda(t)$ can be expressed as

$$\hat{f}_\lambda(t) = \boldsymbol{\Phi}_p(t)'(\boldsymbol{W}'\boldsymbol{W} + \lambda\boldsymbol{G})^{-1}\boldsymbol{W}'\boldsymbol{y}, \tag{4.43}$$

where $\boldsymbol{W} = (\boldsymbol{\Phi}_p(t_1), \boldsymbol{\Phi}_p(t_2), \ldots, \boldsymbol{\Phi}_p(t_n))'$.

Based on (4.43), the fitted response vector at the design time points is

$$\widehat{\boldsymbol{y}}_\lambda = \boldsymbol{A}_\lambda \boldsymbol{y}, \tag{4.44}$$

where the smoother matrix is

$$\boldsymbol{A}_\lambda = \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W} + \lambda\boldsymbol{G})^{-1}\boldsymbol{W}'. \tag{4.45}$$

Therefore, a P-spline is also a linear smoother as defined later in Section 4.6.

4.5.1 Choice of the Knots and Smoothing Parameter Selection

Choice of the knots for P-splines is not so crucial as long as the number of knots K is large enough so that when $\lambda = 0$, the P-spline smoother (4.43) is under smoothing. One may use the equally spaced method or the equally spaced sample quantiles as knots method to locate the knots as described in Section 4.3. The number of knots K may be pre-specified subjectively, for example, taken as a third or a quarter of the total number of distinct design time points.

4.6 Linear Smoother

Linear smoothers express the fitted response vector $\widehat{\boldsymbol{y}}$ at the design time points as linear combinations of the response vector $\boldsymbol{y}$ (Buja *et al* 1989). Specifically, a smoother $\hat{f}_p(t)$ is called a linear smoother if it makes the fitted response vector $\widehat{\boldsymbol{y}}$ to be connected with $\boldsymbol{y}$ *via* the following simple formula:

$$\widehat{\boldsymbol{y}} = \boldsymbol{A}_\rho \boldsymbol{y}, \tag{4.46}$$

where $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)'$ with $\widehat{y}_i = \hat{f}_\rho(t_i)$, $\boldsymbol{A}_\rho$, is the so-called $n \times n$ smoother matrix, determined by the smoother with a smoothing parameter ρ. The smoother matrix $\boldsymbol{A}_\rho$, usually depends on the design time points but is required to be independent of

the response vector $\boldsymbol{y}$. It follows that

$$\begin{aligned} E(\widehat{\boldsymbol{y}}|t_1,\ldots,t_n) &= \boldsymbol{A}_\rho E(\boldsymbol{y}|t_1,\ldots,t_n), \\ \mathrm{Cov}(\widehat{\boldsymbol{y}}|t_1,\ldots,t_n) &= \boldsymbol{A}_\rho \mathrm{Cov}(\boldsymbol{y}|t_1,\ldots,t_n)\boldsymbol{A}_\rho'. \end{aligned}$$

These two formulas can be used to derive the biases and variances of the fitted response vector, and hence allow construction of pointwise standard deviation bands for the underlying function $f(t)$ based on $\widehat{\boldsymbol{y}}$. Moreover, in next section, we shall see that we can develop unified smoothing parameter selection methods for linear smoothers.

From the previous discussions, it is clear that the local polynomial, regression spline, smoothing spline and P-spline smoothers are all linear smoothers.

Chapter 5

Some Comparative & Case Studies

In this chapter we make some comparative studies and a case study that lead us to some rewarding findings. In the first comparative study we compare the performance of the UN model with that of several structured ones and refute the well accepted belief that the covariance structure with larger number of covariance parameters will give better fit to data than those with fewer number of parameters. We establish that there do exist situations in which the structured models with much lesser number of covariance parameters perform better than the UN model. In the second study we make a comparison between the performances of various heterogeneous covariance models with the corresponding homogeneous ones, and show that the heterogeneous models often outperform the homogeneous models by way of giving better values for fit statistics. We corroborate that the heterogeneous covariance models should be preferred if the economy of number of covariance parameters in the homogeneous models does not outweigh the gain in efficiency. In the case study we explore the

impact of a drug in health of rats and the redemption capacity of an extract on these ill effects.

We use some real data, data downloaded from internet and simulated data for illustration of the theoretical procedures and to draw conclusions. SAS codes are used for computation. Performance of various covariance models are compared.

5.1 Structured versus Unstructured Covariance Patterns

In this comparative study we show that the belief that the comparatively larger number of covariance parameters onvolved in the UN makes the model superior to the others, is not always true. We establish that, there can be data for which some other model gives much better results. But in practice, the structured models are being made use of few often compared to the UN model disregarding their savingness of degrees of freedom and competency. We suggest that all covariance models should be tried and the best be picked rather than taking the UN model for granted. We establish evidences in favour of this by using real data as well as simulated data sets.

5.1.1 Illustration 1 (TLC Data)

The Treatment of Lead-Exposed Children (TLC) example considers a placebo controlled, randomized trial of succimer conducted in 100 children. The TLC data is

downloaded from the web site www.biostat.harvard.edu/∼fitzmaur/ala. Children received up to three 26-day courses of succimer or placebo and were followed for 3 years. The blood lead levels at baseline, week 1, week 4, and week 6 are measured for each child.

Table 5.1 presents data on blood lead levels at baseline, week 1, week 4 and week 6 for 10 randomly selected children from the study.

Table 5.1: Blood Lead Levels

ID	Group	Baseline	Week 1	Week 4	Week 6
79	P	30.8	26.9	25.8	23.8
8	S	26.5	14.8	19.5	21
44	S	25.8	23	19.1	23.2
11	P	24.7	24.5	22	22.5
69	S	20.4	2.8	3.2	9.4
29	S	20.4	5.4	4.5	11.9
46	P	28.6	20.8	19.2	18.4
13	P	33.7	31.6	28.5	25.1
74	P	19.7	14.9	15.3	14.7
53	P	31.1	31.2	29.2	30.1

The mean blood levels at each measurement occasion for random subset of 100 children, broken down by treatment group are presented in Table 5.2.

Table 5.2: Mean Blood Lead Levels.

Group	Baseline	Week 1	Week 4	Week 6
Succimer	26.5	13.5	15.5	20.8
	(5.0)	(7.7)	(7.8)	(9.2)
Placebo	26.3	24.7	24.1	23.6
	(5.0)	(5.5)	(5.8)	(5.6)

The mean response profiles for the two groups randomised to succimer and

placebo are presented in the Figure 5.1.

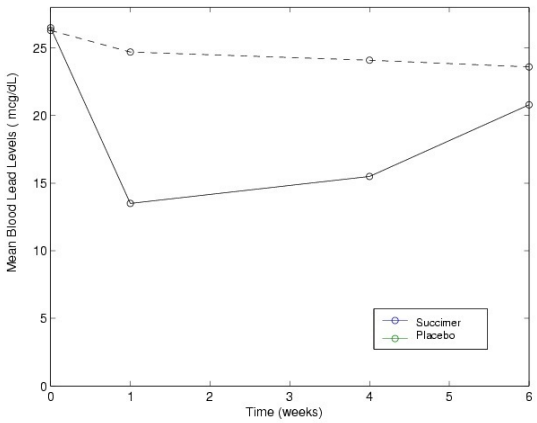


Figure 5.1: Mean Response Profiles

Due to randomisation, the mean response at baseline is similar in the two treatment groups. However, there are discernible differences in the patterns of change in the mean response over time. At week 1 there is a dramatic drop in blood lead levels among the children treated with succimer followed by a rebound in blood lead level as lead stored in the bones and tissues is mobilized and new equilibrium is achieved.

To compare their relative performances, fitting of various covariances models is done using the SAS codes. The SAS code for getting information like covariance matrix, simple statistics, and correlation matrix and for fitting the unstructured covariance model is given below.

```
data TLC;
infile ‘D:TLC.txt’ DSD delimiter=‘09’x firstobs=2;
input id  Group $  Week0 Week1 Week4 Week6;
```

```
y = Week0; time=0; OUTPUT;
y = Week1; time=1; OUTPUT;
y = Week4; time=4; OUTPUT;
y = Week6; time=6; OUTPUT;
run;
-------------------------------------------------
ods rtf file="tlc.rtf";
-------------------------------------------------
PROC SORT DATA=TLC;
by group;
run;
-------------------------------------------------
PROC MEANS DATA=TLC MEAN STD;
by group;
var Week0 Week1 Week4 Week6 ;
run;
-------------------------------------------------
PROC CORR DATA=TLC COV;
var Week0 Week1 Week4 Week6 ;
run;
-------------------------------------------------
PROC CORR DATA=TLC COV;
by group; var Week0 Week1 Week4 Week6 ;
run;
```

```
-------------------------------------------------
PROC MIXED DATA=TLC COVTEST;
CLASS id group time;
MODEL y=group time group*time/ S CHISQ;
REPEATED time/TYPE=UN SUBJECT=id R RCORR;
run;
-------------------------------------------------
ods rtf close;
-------------------------------------------------
```

Remark 5.1.1. *The MIXED procedure in the above SAS code considers the unstructured covariance structure. Output corresponding to other covariance structures can be obtained by replacing the argument UN, in TYPE=UN with their SAS names CS, AR, TOEP, ANTE* etc. *These SAS names are shown in column 2 of Table 5.3.*

SAS output and interpretation

Selected output in the rtf (Rich Text Format) of SAS and their interpretation are given below. We have deleted some part of the output of PROC MIXED that is irrelevant to our purposes here to shorten the presentation.

The procedure MEANS with the argument by groups, computes means and standard deviations for each of the groups.

The CORR procedure is used to compute the covariance matrix, simple statistics and Pearson correlation coefficient with their tests of significance for the data.

The MEANS Procedure

Group=A

Variable	Mean	Std Dev
Week0	26.5400000	4.9829458
Week1	13.5220000	7.6144346
Week4	15.5140000	7.7927942
Week6	20.7620000	9.1763710

Group=P

Variable	Mean	Std Dev
Week0	26.2720000	4.9860928
Week1	24.6600000	5.4198593
Week4	24.0700000	5.7095970
Week6	23.6460000	5.5971353

Addition of the argument by group computes the covariance matrices for each of the groups separately. Output of the CORR procedure for the overall data as well as for each group are given below. The table shows that all the correlation coefficients are significant.

4 Variables:	Week0	Week1	Week4	Week6

Covariance Matrix, DF = 399				
	Week0	Week1	Week4	Week6
Week0	24.80116692	18.02411429	18.77919599	21.61842206
Week1	18.02411429	74.65926717	58.79561704	37.20504862
Week4	18.77919599	58.79561704	64.89377043	36.26760100
Week6	21.61842206	37.20504862	36.26760100	59.70665063

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Week0	400	26.40600	4.98008	10562	19.70000	41.10000
Week1	400	19.09100	8.64056	7636	2.80000	40.80000
Week4	400	19.79200	8.05567	7917	3.00000	40.40000
Week6	400	22.20400	7.72701	8882	4.10000	63.90000

Pearson Correlation Coefficients, N = 400 Prob > |r| under H0: Rho=0				
	Week0	Week1	Week4	Week6
Week0	1.00000	0.41887 <.0001	0.46810 <.0001	0.56179 <.0001
Week1	0.41887 <.0001	1.00000	0.84470 <.0001	0.55725 <.0001
Week4	0.46810 <.0001	0.84470 <.0001	1.00000	0.58265 <.0001

The CORR Procedure

Group=A

4 Variables:	Week0	Week1	Week4	Week6

Covariance Matrix, DF = 199				
	Week0	**Week1**	**Week4**	**Week6**
Week0	24.82974874	15.23228141	14.90978894	22.63891457
Week1	15.23228141	57.97961407	43.36531859	35.42375477
Week4	14.90978894	43.36531859	60.72764221	32.52415276
Week6	22.63891457	35.42375477	32.52415276	84.20578492

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
Week0	200	26.54000	4.98295	5308	19.70000	41.10000
Week1	200	13.52200	7.61443	2704	2.80000	39.00000
Week4	200	15.51400	7.79279	3103	3.00000	40.40000
Week6	200	20.76200	9.17637	4152	4.10000	63.90000

Pearson Correlation Coefficients, N = 200 Prob > |r| under H0: Rho=0				
	Week0	**Week1**	**Week4**	**Week6**
Week0	1.00000	0.40146 <.0001	0.38397 <.0001	0.49511 <.0001
Week1	0.40146 <.0001	1.00000	0.73082 <.0001	0.50697 <.0001
Week4	0.38397 <.0001	0.73082 <.0001	1.00000	0.45482 <.0001
Week6	0.49511 <.0001	0.50697 <.0001	0.45482 <.0001	1.00000

The CORR Procedure

Group=P

4 Variables:	Week0	Week1	Week4	Week6

Covariance Matrix, DF = 199				
	Week0	**Week1**	**Week4**	**Week6**
Week0	24.86112161	22.40651256	23.89523618	21.09496281
Week1	22.40651256	29.37487437	26.63356784	23.03159799
Week4	23.89523618	26.63356784	32.59949749	27.79354774
Week6	21.09496281	23.03159799	27.79354774	31.32792362

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
Week0	200	26.27200	4.98609	5254	19.70000	38.10000
Week1	200	24.66000	5.41986	4932	14.90000	40.80000
Week4	200	24.07000	5.70960	4814	15.30000	38.60000
Week6	200	23.64600	5.59714	4729	13.50000	43.30000

Pearson Correlation Coefficients, N = 200 Prob > |r| under H0: Rho=0				
	Week0	**Week1**	**Week4**	**Week6**
Week0	1.00000	0.82914 <.0001	0.83935 <.0001	0.75588 <.0001
Week1	0.82914 <.0001	1.00000	0.86067 <.0001	0.75922 <.0001
Week4	0.83935 <.0001	0.86067 <.0001	1.00000	0.86971 <.0001
Week6	0.75588 <.0001	0.75922 <.0001	0.86971 <.0001	1.00000

The covariance matrices show that the variances are not constant over time. This suggests that the assumption of homogeneity of variances across time is not valid in this case.

The MIXED procedure is used to fit various covariance structures. The COVTEST option requests asymptotic tests of all of the covariance parameters. By default SAS uses the REML method to estimate the unknown covariance parameters. The additional argument METHOD=ML can be added to use the maximum likelihood method.

The Mixed Procedure

Model Information	
Data Set	WORK.TLC
Dependent Variable	y
Covariance Structure	Unstructured
Subject Effect	id
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

The following tables show that 10 covariance parameters result from the 4×4 unstructured blocks of R. Two Newton-Raphson iterations are required to find the REML estimates.

Dimensions	
Covariance Parameters	10
Columns in X	15
Columns in Z	0
Subjects	100
Max Obs Per Subject	4
Observations Used	400
Observations Not Used	0
Total Observations	400

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	2626.25517748	
1	1	2416.07594087	0.00000000

Convergence criteria met.

Estimated R Matrix for id 1				
Row	Col1	Col2	Col3	Col4
1	25.2257	19.1074	19.6995	22.2016
2	19.1074	44.3458	35.5351	29.6750
3	19.6995	35.5351	47.3778	30.6205
4	22.2016	29.6750	30.6205	58.6510

The preceding 4×4 matrix is the estimated unstructured covariance matrix.

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	id	25.2257	3.6037	7.00	<.0001
UN(2,1)	id	19.1074	3.8911	4.91	<.0001
UN(2,2)	id	44.3458	6.3351	7.00	<.0001
UN(3,1)	id	19.6995	4.0194	4.90	<.0001
UN(3,2)	id	35.5351	5.8587	6.07	<.0001
UN(3,3)	id	47.3778	6.7683	7.00	<.0001
UN(4,1)	id	22.2016	4.4863	4.95	<.0001
UN(4,2)	id	29.6750	5.9604	4.98	<.0001
UN(4,3)	id	30.6205	6.1581	4.97	<.0001
UN(4,4)	id	58.6510	8.3787	7.00	<.0001

The preceding table lists the 10 estimated covariance parameters in order; note their correspondence to the first block of R displayed previously. The parameter estimates are labelled according to their location in the block in the Cov Parm column, and all of these estimates are associated with id as the subject effect. The Std Error column lists approximate standard errors of the covariance parameters obtained from the inverse Hessian matrix. These standard errors lead to approximate Wald Z-statistics, which are compared with the standard normal distribution. The results of these tests indicate that all the parameters are significantly different from 0.

The succeeding tables show that the null model likelihood ratio test (LRT) is highly significant for this model, indicating that the unstructured covariance matrix is preferred to the diagonal one of the ordinary least-squares null model. The degrees

Fit Statistics	
-2 Res Log Likelihood	2416.1
AIC (smaller is better)	2436.1
AICC (smaller is better)	2436.7
BIC (smaller is better)	2462.1

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
9	210.18	<.0001

of freedom for this test is 9, which is the difference between 10 and the 1 parameter for the null model's diagonal matrix

Type 3 Tests of Fixed Effects						
Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
Group	1	98	25.43	25.43	<.0001	<.0001
time	3	98	184.48	61.49	<.0001	<.0001
Group*time	3	98	107.79	35.93	<.0001	<.0001

The "Type 3 Tests of Fixed Effects" table displays Type III tests for all of the fixed effects. The table shows that all fixed effects are highly significant. It is usually best to consider higher-order terms first, and in this case the Group*time test reveals a difference between the slopes that is statistically significant. The time test is one

for an overall growth curve accounting for possible heterogeneous slopes, and it is also highly significant. Finally, the Group row tests the null hypothesis of a common intercept, and this hypothesis cannot be rejected from these data.

Table 5.3, provides summary of the fit statistics of a large number of covariance structures.

Observe that in Table 5.3 some of the models have fit statistics that are smaller than those for the UN model suggesting that these models give better fit to data than the UN model does. This clearly shows that there are models that perform better than the UN model and that the larger number of covariance parameters involved does not always make the UN model superior to others. In terms of the BIC, the FA1(2), CSH, FA1(3), FA(1) and HF models perform either better than or at least as good as the UN model. The comparison using the Bayesian test, that follows, shows that the performance of the FA1(2) model is significantly better than that of the UN model and it gives the best fit to the data. This is quite contrary to the existing belief that the UN model usually outperforms all other models, which has lead to parsimonious use of the model disregarding the fact the model involves comparatively greater number of parameters. That is, though the performance of models improves with increase in number of parameters in general, there are exceptions.

In Table 5.3, the UN model has BIC = 2462.1 and the FA1(2) model has BIC = 2454.1. Hence, if we restrict attention to just these two models, then

$$P(\text{FA1(2)}/Y) \approx \frac{1}{1+\exp[-0.5\times(2462.1-2454.1)} = 0.9820138.$$

Table 5.3: Fit Statistics for different covariance models

Covariance Pattern	SAS Option	# of cov pars	-2 Res logL	AIC	AIC_C	BIC
Unstructured	UN	10	2416.1	2436.1	2436.7	2462.1
Banded1	UN(1)	4	2608.8	2616.8	2616.9	2627.2
Banded2	UN(2)	7	2497.5	2511.5	2511.8	2529.7
Banded3	UN(3)	9	2455.8	2473.8	2474.3	2497.2
Compound Symmetry	CS	2	2460.6	2464.6	2464.7	2469.8
Heterogeneous CS	CSH	5	2434.0	2444.0	2444.1	2457.0
Autoregressive	AR(1)	2	2472.6	2476.6	2476.7	2481.8
Heterogeneous AR	ARH(1)	5	2451.6	2461.6	2461.8	2474.7
Toeplitz	TOEP	4	2457.2	2465.2	2465.3	2475.6
Banded Toeplitz	TOEP(1)	1	2626.3	2628.3	2628.3	2630.9
	TOEP(2)	2	2518.7	2522.7	2522.7	2527.9
	TOEP(3)	3	2491.6	2497.6	2497.6	2505.4
Heterogeneous TOEP	TOEPH	7	2431.1	2445.1	2445.4	2463.4
Banded Hetero TOEP	TOEPH(1)	4	2608.8	2616.8	2616.9	2627.2
	TOEPH(2)	5	2498.4	2508.4	2508.5	2521.4
	TOEPH(3)	6	2471.9	2483.9	2484.1	2499.5
Antedependence	ANTE(1)	7	2439.7	2453.7	2454.0	2471.9
AR Moving Average	ARMA(1,1)	3	2458.8	2464.8	2464.8	2472.6
Factor Analytic	FA(1)	8	2421.7	2437.7	2438.1	2458.5
	FA(2)	11	2416.1	2438.1	2438.8	2466.7
	FA(3)	13	2416.1	2442.1	2443.0	2475.9
	FA(4)	14	2416.1	2444.1	2445.2	2480.5
No Diagonal FA	FA0(4)	10	2416.1	2436.1	2436.7	2462.1
Equal Diagonal Factor Analytic	FA1(1)	5	2440.0	2450.0	2450.2	2463.0
	FA1(2)	8	2417.3	2433.3	2433.7	2454.1
	FA1(3)	10	2416.1	2434.1	2434.5	2457.5
	FA1(4)	11	2416.1	2436.1	2436.7	2462.1
Huynh-Feldt	HF	5	2436.7	2446.7	2446.8	2459.7

Hence there is overwhelming evidence in favour of the FA1(2) model. The Bayesian posterior probability for the other models mentioned above, in comparison to the UN model are

$$P(CSH/Y) \approx \frac{1}{1+\exp[-0.5 \times (2462.1 - 2457)} = 0.927573515.$$
$$P(FA1(3)/Y) \approx \frac{1}{1+\exp[-0.5 \times (2462.1 - 2457.5)} = 0.908877039.$$
$$P(FA(1)/Y) \approx \frac{1}{1+\exp[-0.5 \times (2462.1 - 2458.5)} = 0.858148935.$$
$$P(HF/Y) \approx \frac{1}{1+\exp[-0.5 \times (2462.1 - 2459.7)} = 0.768524783.$$

From the above, it is clear that the FA1(2), CSH, FA1(3), FA(1) and HF models have Bayesian posterior proababilities significantly higher than 0.5. Thus all these models give better fit to the data, with FA1(2) in the top position.

5.1.2 Illustration 2 (Simulation Study)

In the previous illustration we have seen that at times there are models that perform better than the UN model. The improvement is made out using the log likelihood, AIC, BIC and AIC_C. We have also quantified the probability of models using the Bayesian probability that a model is correct. The FA1(2) model had the least BIC which is smaller than that of the UN model by 8, putting the Bayesian probability of the FA1(2) model as high as 0.98201379. Now we show with the help of a simulated model that the difference of the BIC's can be more striking and there can be very

strong or undeniable evidence against the widely used UN model, at times. We keep the values of blood lead level at Week0 as such. Assuming Markovian dependence for Week1, Week4 and Week 6 values, we simulate values at Week1, Week4 and Week6 so that these values hold the same linear dependence with the just preceding values as in the original data. That is we maintain the relation $Y_t = a + \rho Y_{t-1}$ where a and ρ are estimated from the original data. The estimated relationships that responses at Week1, Week4 and Week6 hold with those at their immediate preceding measurement occasions were obtained using R as follows.

$$\begin{aligned} Week1 &= -0.09942 + 0.72674 * Week0 \\ Week4 &= 4.7575 + 0.7875 * Week1 \\ Week6 &= 11.1427 + 0.5589 * Week4 \end{aligned} \tag{5.1}$$

The result is quite promising in that the difference between the BIC's rises as high as 15.2 for the ARH(1) model, making it far better than the UN model. It may be noticed that the number of covariance parameters for the ARH(1) model is only 5 where as that for the UN model is 10. The BIC's of the best three models and that of the UN model along with their Bayesian probabilities in comparison to the UN model are summarised in Table 5.4.

In the simulation study considered we have maintained the same correlation structure as in the original data. Next we show that the situation can be still attractive if the correlation increases. In the second simulation we keep the values at Week0 as such. The values at Week1, Week4 and Week6 are simulated so as to have

Table 5.4:

Covariance Pattern	# of cov pars	BIC	difference of BIC with that of UN	Posterior probability
UN	10	1495.8	—	—
TOEP	7	1487.6	8.2	0.983697501
ANTE(1)	7	1485	10.8	0.995503727
ARH(1)	5	1480.6	15.2	0.999499799

Table 5.5:

Covariance Pattern	# of cov pars	BIC	difference of BIC with that of UN	Posterior probability
UN	10	2849.7	—	—
FA1(3)	10	2845.1	4.6	0.908877039
HF	5	2842.6	7.1	0.972077426
ARH(1)	15	2840.8	8.9	0.988456248
ANTE(1)	14	2838.6	11.1	0.996127597
FA1	6	2823.2	26.5	0.99999824

correlations with the values at the preceding measurement occasion larger than that in the original data. The values at Week1, Week4 and Week6 are simulated to have correlation of 0.9, 0.8 and 0.7 with the values at the immediately preceding occasion. This causes larger difference of BIC's for some models than that in the first simulation, leading us to the conclusion that the appropriateness of the UN model worsens for larger correlation. Simulations with still higher correlations ratify this argument. The BIC's of the best five models and that of the UN model along with their Bayesian probabilities in comparison to the UN model for the second simulation are summarised in Table 5.5. The FA(1) model is found to be overwhelmingly better than the UN model in this case. The difference between the BIC's has explosively increased to 26.5 in this case. The ANTE(1) model also has a very attractive

performance. When simulation is done a large number of times, it was found that the ANTE model and the FA(1) model outperforms the UN model in a significantly better way if the correlation is high.

5.2 Heterogeneous versus Homogeneous Covariance Patterns

In this section we compare the performances of various heterogeneous covariance structures with the corresponding homogeneous covariance structures. We examine the appropriateness of the covariance structures that assume homogeneity of variances over time, for varied structures of correlations among repeated-measures. We elucidate that the relaxation of the strong assumption of homoscedasticity, improves the covariance models by way of giving smaller AIC, AIC_C and BIC. Quite often the heteroscedastic model outperforms the corresponding homoscedastic model. But heeding little attention to this, homoscedastic models are being made of use of parsimoniously.

5.2.1 Illustration 3

To illustrate the above arguments we use the TLC data cited earlier. The part of the computations and tables used for this comparison, that has already appeared in section 5.1 are not reproduced here.

The estimated variances at Week0, Week1, Week4, and Week6 for the entire group are 24.8012, 74.6593, 64.8938, and 59.7067 respectively, which shows that the assumption of homogeneity of variances is quite invalid in this case. In Table 5.3, it may be observed that the fit statistics for the heterogeneous models are smaller than that for the corresponding homogeneous models. That is, disputing to the usual practise of adopting the prototypical homogeneous models, table 5.3 show that it is the heterogeneous covariance model that perform better in practical situations. This shows that the heteroscedastic models are better performers than the corresponding homoscedastic models.

In terms of BIC, the CSH, ARH(1), TOEPH, TOEPH(1), TOEPH(2) and TOEPH(3) models perform better than the corresponding homogeneous models. The comparison using the Bayesian posterior probability, that follows, shows that the performances of the homogeneous and heterogeneous versions of the same structure are significantly different. This leads to the observation that the existing practise of using homogeneous parsimoniously is debatable. For example the CS model has BIC= 2469.8 whereas the corresponding heterogeneous model, viz, the CSH model has BIC=2457.0, which is smaller than that of the former by 12.8, which is a highly significant difference. If we restrict attention to just these two models, then

$$P(\text{CSH}/Y) \approx \frac{1}{1+\exp[-0.5\times(2469.8-2457.0)} = 0.998341199.$$

Hence there is overwhelming evidence in favour of the CSH model in comparison to the CS model. Table 5.6 gives Bayesian posterior probabilities for comparing the relative performances in different cases, which further vindicate this argument.

Table 5.6:

Homogeneous stucture	No. of Pars	BIC	Hetereogeneous structure	No. of Pars	BIC	Posterior Probability
CS	2	2469.8	CSH	5	2457	0.998341199
AR(1)	2	2481.8	ARH(1)	5	2474.75	0.972077426
TOEP	4	2475.6	TOEPH	7	2463.4	0.997762151
TOEP(1)	1	2630.5	TOEPH(1)	4	2627.2	0.83889105
TOEP(2)	2	2527.9	TOEPH(2)	5	2521.4	0.947846437
TOEP(3)	3	2505.4	TOEPH(3)	6	2499.5	0.950263488

5.2.2 Illustration 4 (Simulation Study).

In the previous illustration we have seen that the heterogeneous models perform better than the corresponding homogeneous models. The betterment is made out using the log likelihood, AIC, BIC and AIC_C. We have also quantified the probability of models using the Bayesian probability that a model is correct. Now we show with the help of a simulated model that the difference of the BIC's can be alarmingly high or there can be undeniable evidence against the parsimonious use of the homogeneous models if the heterogeneity among variances is large. In the simulated data we maintain the dependence among the repeated measures same as that in the original data. But the heterogeneity among variances are increased. As in illustratin 2, we keep the values of blood lead level at Week0 as such. Assuming Markovian dependence for Week1, Week4 and Week6 values, we simulate values at Week1, Week4 and week6, so that these values hold the same linear dependence with the just preceding values as in the original data. The estimated equations showing these relationship are given in (5.1).

In the simulation, we added a mean zero normal error term with variance 25 to

the estimated relation. The result is quite alarming in that the difference between the BIC's rises above 200 in all the cases, raising the Bayesian posterior probability approximately 1 for the heterogeneous models throwing the homogeneous models rubbish. Thus the heterogeneous models are found to be overwhelmingly better than the than the homogeneous models. It may be noticed that the number of covariance parameters is slightly larger for the heterogeneous models. But this is immaterial for such a huge hike in the Bayesian posterior probability. The results are summarised for six models and their heterogeneous versions in Table 5.7.

Table 5.7:

Homogeneous structure	No. of Pars	BIC	Hetereogeneous structure	No. of Pars	BIC	Posterior Probability
CS	2	3617.4	CSH	5	3399.0	1
AR(1)	2	3597.4	ARH(1)	5	3388.3	1
TOEP	4	3605.1	TOEPH	7	3396.6	1
TOEP(1)	1	3670.	TOEPH(1)	4	3445.1	1
TOEP(2)	2	3603.1	TOEPH(2)	5	3399.5	1
TOEP(3)	3	3601.1	TOEPH(3)	6	3393.4	1

5.3 Longitudinal Analysis of effect of a drug in rats - A Case Study

In this study we explore the effect of drug 'vincristine' in rats and study the curing ability of extract of the plant Sida Cordifolia on the ill effects of the drug. This case study is based on an experiment conducted by the Department of Life Sciences, University of Calicut, and the data generated there from. In the study, paw flick

responses and tail flick responses to water at 50°C of 19 rats were observed at four occasions - the 0^{th}, 28^{th}, 42^{nd}, 56^{th} and 77^{th} days. The rats were divided into 3 groups - 7 test rats, 6 control and 6 normal rats. The test rats were injected a drug 'vincristine' and plant extract was fed to the animals. To the control rats the same drug was fed, but no plant extract was given. The normal rats were given no drug or plant extract. At each measurement occasion responses on paw flick response latency and tail flick response latency were made. The drug is expected to induce neuropathy which causes damage to natural body responses to temperature differences. Objectives were to learn whether the the drug causes any change paw flick response latency and tail flick response latency and to see if the plant extract helps in redemption of the drug effect, that is to see if there is significant difference among these responses in the three groups. The analysis is done using SAS codes.

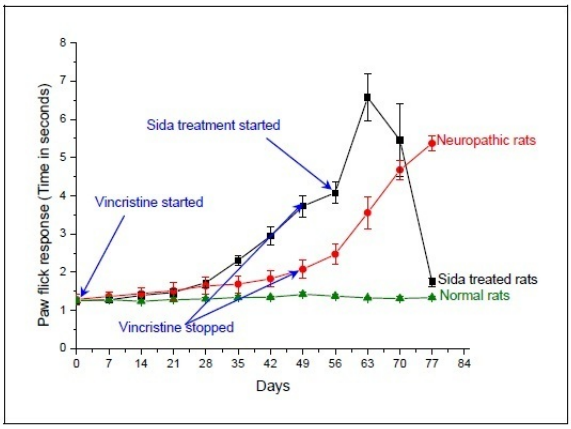
5.3.1 Effect on Paw Flick Responses

The succeeding figure shows the graph of paw flick responses over different measurement occasions. The figure shows that there is difference among the latency periods for the three groups and that the extract brings down the latency period almost to the natural level. This finding is formally established by appropriate test procedures later.

The tables of the MEANS procedure further establishes these findings. The latency period is comparatively high for the group injected with the drug (Group 1) than the normal group (Group 3). However the mean latency is brought down very

Figure 5.2:

Graph showing the comparison of paw flick response of normal rats, neuropathic rats (Control) and *sida* treated neuropathic rats (Test)



close to the normal level in the case of test rats at the end of the study, showing the redemption ability of the extract.

In comparison to the control group, the mean latency period is slightly greater for the test group. But this difference is shown to be insignificant by the comparison of means as shown in the table of type 3 tests of fixed effects that follows.

Group=1

Variable	Mean	Std Dev
TIME0	1.2500000	0.1668920
TIME28	1.7242857	0.3332927
TIME42	2.9485714	0.5778364
TIME56	4.0800000	0.7168354
TIME77	1.7428571	0.2766243

Group=2

Variable	Mean	Std Dev
TIME0	1.2800000	0.2944311
TIME28	1.6383333	0.5198613
TIME42	1.8283333	0.4809623
TIME56	2.4700000	0.6109631
TIME77	5.3716667	0.4366572

Group=3

Variable	Mean	Std Dev
TIME0	1.2500000	0.1466876
TIME28	1.2950000	0.0852481
TIME42	1.3400000	0.0869007
TIME56	1.3750000	0.1195898
TIME77	1.3316667	0.0657625

Type 3 Tests of Fixed Effects						
Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
Group	1	11	0.77	0.77	0.3813	0.4000
time	4	11	1394.08	348.52	<.0001	<.0001
Group*time	4	11	656.77	164.19	<.0001	<.0001

The succeeding table gives the estimates of covariances over the measurement occasions. The matrix shows that the strong assumption of homogeneity over time is not valid for this data, as we have remarked in section 2 of this chapter.

Estimated R Matrix for ID 1					
Row	**Col1**	**Col2**	**Col3**	**Col4**	**Col5**
1	0.05106	0.05501	0.01725	0.02914	0.03726
2	0.05501	0.1478	0.1198	0.1249	0.01818
3	0.01725	0.1198	0.2285	0.2246	0.01010
4	0.02914	0.1249	0.2246	0.3589	-0.02254
5	0.03726	0.01818	0.01010	-0.02254	0.1032

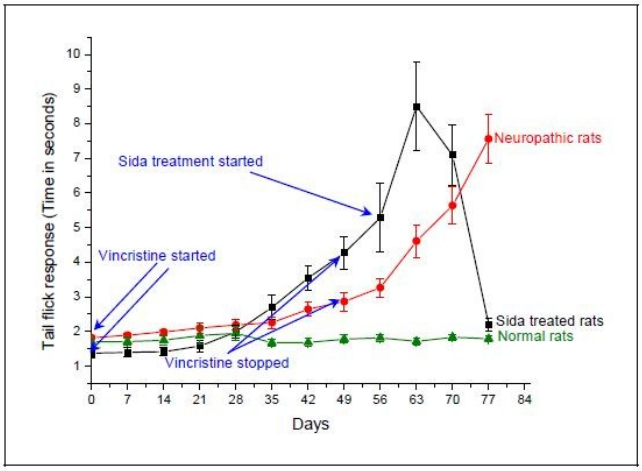
The results corresponding to the type 3 tests, shown below, points to formal justification of the findings arrived at roughly in the earlier discussion. All the three p values are significantly smaller than 0.05. The test corresponding to the group effect rejects homogeneity of mean responses for the groups. This shows that the means of paw flick response times for the groups are significantly different. That is, the drug induces neuropathic problems to the rats by way of prolonging paw flick responses to temperature. The time test testifies the violation of the homogeneity assumption. That is the means are not constant over time. The test concerning group*time interaction shows that the patterns of changes in the response over time are not same across the groups. This formally establishes the rough conclusion arrived at earlier that the extract helps in redemption of the damage to paw flick responses.

Type 3 Tests of Fixed Effects						
Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
Group	2	16	60.67	30.34	<.0001	<.0001
time	4	16	1201.10	300.28	<.0001	<.0001
Group*time	8	16	1323.89	165.49	<.0001	<.0001

5.3.2 Effect on Tail Flick Responses

The following figure show the graph of tail flick responses over different measurement occasions. There is obvious difference among the latency periods for the three groups and that the extract brings down the latency period almost to the natural level. As in the earlier case, this finding is formally established by appropriate test procedures later.

Graph showing the comparison of tail flick response of normal rats, neuropathic rats (Control) and *sida* treated neuropathic rats (Test)



The output of the MEANS procedure also supports the findings suggested by the profile plot. The latency period is comparatively high for the group injected with the drug and in the case of test rats the mean latency is brought down very close to the normal level towards the end of the study, testifying the redemption ability of the extract.

Group=1

Variable	Mean	Std Dev
TIME0	1.3700000	0.3328575
TIME28	1.9842857	0.4783506
TIME42	3.5428571	0.9086628
TIME56	5.2800000	2.4631579
TIME77	2.2057143	0.4642469

Group=2

Variable	Mean	Std Dev
TIME0	1.8166667	0.1160658
TIME28	2.1883333	0.3659855
TIME42	2.6383333	0.4563882
TIME56	3.2633333	0.6088618
TIME77	7.5583333	1.6034598

In the case of tail flick response data as well, the covariance matrix shows that

Group=3

Variable	Mean	Std Dev
TIME0	1.7016667	0.2093023
TIME28	1.9483333	0.4689466
TIME42	1.6783333	0.2656428
TIME56	1.8150000	0.2538497
TIME77	1.7883333	0.2237931

the strong assumption homogeneity over time is not valid.

Estimated R Matrix for ID 1					
Row	Col1	Col2	Col3	Col4	Col5
1	0.06785	0.06726	0.02686	-0.05156	-0.04298
2	0.06726	0.2255	0.1892	0.1874	0.08175
3	0.02686	0.1892	0.4520	0.9049	0.08456
4	-0.05156	0.1874	0.9049	2.7363	0.2462
5	-0.04298	0.08175	0.08456	0.2462	1.0418

The results corresponding to the type 3 tests points to formal justification of the findings arrived earlier. The test corresponding to the group effect shows that mean responses for the groups are significantly different. This shows that the means of tail flick response times for the groups are significantly different. That is the drug prolongs the tail flick responses to temperature. The time test shows that the

means are not constant over time. The test concerning group*time interaction shows that the patterns of changes are different for different groups in the responses. This establishes that the extract helps in redemption of the damage to paw flick responses.

Type 3 Tests of Fixed Effects						
Effect	**Num DF**	**Den DF**	**Chi-Square**	**F Value**	**Pr > ChiSq**	**Pr > F**
Group	2	16	60.67	30.34	<.0001	<.0001
time	4	16	1201.10	300.28	<.0001	<.0001
Group*time	8	16	1323.89	165.49	<.0001	<.0001

Chapter 6

Summary and Concluding Remarks

Correct modelling of the covariance is often a requirement for obtaining valid estimates of the regression parameters. In general, the failure to take into account of the covariance among the repeated measures will result in incorrect estimates of the sampling variability and can lead to quite misleading scientific inferences.

If the pattern of covariance does show a systematic structure, then not acknowledging this by maintaining the unstructured assumption involves estimation of many more parameters than might otherwise be necessary, thus making inefficient use of the available data. Models with more covariance parameters are intuitively expected to give better fit to data, meaning they should naturally have a higher log likelihood, than models with fewer parameters. We consider models that represent the correlation structure in terms of fewer parameters and illustrate that the larger number of covariance parameters involved does not always make the UN model superior to others. We show that there are models that perform better than the UN model. With

regard to the data in illustration 1, we show that in terms of the BIC, the FA1(2), CSH, FA1(3), FA(1) and HF models perform either better than or at least as good as the UN model. The comparison using the Bayesian test shows that at least the performance of the FA1(2) model is significantly better than that of the UN model. This is quite contrary to the existing belief that the UN model usually outperforms all other models, which has lead to parsimonious use of the model disregarding the fact the model involves comparatively greater number of parameters. The situation is more glaring in the simulation studies. In the simulation study we have shown that when the correlation between values at successive measurement occasions increase, the performance of the UN model becomes poorer and there can be situations when the use of UN model should definitely be replaced by models like FA, ANTE *etc.* Thus we substantiate that choice among the prominent covariance models should not merely be aimed at the efficient use of available data alone, and should also aim at finding model that perform the best by in terms of some selection criteria, (*e.g.* by yielding the least BIC).

We further vindicate the avoidance of the use of covariance pattern models that make the strong assumption that the variances are constant over time. As mentioned earlier, our practical experience with many longitudinal studies has led to the empirical observation that the variances are rarely constant over time. We show that when the variances are heterogeneous over time, the homogeneous models are not suitable. Therefore, because the assumption of constant variance is the one that is not valid in many setting, we recommend that covariance pattern models with heterogeneous variances, allowing the variances to depend arbitrarily on time, should generally be adopted. Empirical support to this argument has been constructed in

our illustrations.

We further reiterate that, that there has been a lag between the recent developments and their widespread application to substantive problems, need special attention while one goes for longitudinal data analysis. This lag as well as improper model selection has parented analysis that leads to inefficient conclusion at least in some cases.

Further, quite often little attention is paid to the plausibility of the assumptions using which the theory of longitudinal analysis is built. Data in fields like bio-medical and health sciences are generally heavy tailed and does not conform to the assumption of normality. Investigation regarding enhancement of the existing tools to accommodate nonnormality is therefore necessary.

In ayurveda, Sida Cordifolia is usually used to treat rheumatic complaints and digestive fire. In the case study we establish that the extract of this plant is effective in curing neuropathic disorders in rats. This gives scope for studies in human beings in this direction, which is presumed to be a quite rewarding and pioneering work.

It is worthwhile to study the regression models for longitudinal data, which suffer from attrition: units drop out of the study before its completion and thus present incomplete data records. The problem of some missing measurements from some units is noteworthy in longitudinal studies, where the units leaving the study may contribute to bias the entire design. The problem of greater interest is, how non-ignorable attrition can be treated in this framework. In these cases the properties of estimates like consistency is to be confirmed by choosing the appropriate drop out

mechanism.

Another problem of interest will be the joint modelling of longitudinal and survival data. Joint modelling may be accomplished using latent variables that link the longitudinal models and the survival models together. The joint models may result in unbiased and more efficient estimates.

Details of Research Papers/Presentations

Papers presented

1. Presented the paper entitled "Statistical Modeling Approach to Longitudinal Data" in the INTERNATIONAL CONFERENCE ON ACTUARIAL STATISTICS, BIO-STATISTICS AND STOCHASTIC MODELING (INCABS 11) organized by the Department of Statistical Sciences, Kannur University during 10-14 January, 2011.

2. Presented the paper entitled "On Selection of Covariance Structure in Modelling of Longitudinal Data" in the NATIONAL CONFERENCE ON RECENT DEVELOPMENTS IN THE APPLICATIONS OF RELIABILITY THEORY AND SURVIVAL ANALYSIS (NCRSA - 2012) organized by the Department of Statistics, Pondicherry University during 02-03 January, 2012.

3. Presented the paper entitled "Essential Statistical Tools as Applied in Empirical Research" in the DCE sponsored NATIONAL SEMINAR ON METHODOLOGY OF EMPIRICAL RESEARCH IN ECONOMICS: APPLICATION OF STATISTICAL TOOLS & SOFTWARE jointly organized by the Departments of Statistics and Economics, Govt. Brennen College, Thalassery during 5th, 6th and 7th March 2012.

Papers Comunicated

1. Paper entitled "Structured versus Unstructured Covariance patterns in Modelling Longitudinal data" Communicated to Journal of Probability and Statistical Science (JPSS).

Papers published

1. "On Selection of Covariance Structure in Modelling of Longitudinal Data." (jointly with M Manoharan) to appear in the proceedings of NCRSA - 2012

2. "Heterogeneous versus Homogeneous Covariance patterns in Modelling Longitudinal data" (jointly with M Manoharan) to appear in the Journal of Modern Mathematics Frontier.

Bibliography

[1] Aerts M.and Geys H.and Molenberghs G. and Ryan L. (2002). *Topics in Modelling of Clustered Data*. Boca Raton, FL: Chapman & Hall/CRC.

[2] Airy G. B. (1861). *On the Algebraical and Numerical Theory of Errors of Observation and the Combination of Observations*. Macmillan, London.

[3] Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. *In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory*, pages 261–281.

[4] Altham P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statistics*, **27**(162-167).

[5] Anderson T. W. and Goodman L. A. (1957). Statistical inference about markov chains. *Annals of Mathematical Statistics*, **28**:89–110.

[6] Anderson T.W. (1984). *An introduction to Multivariate Statistical Analysis*. Wiley.

[7] Anderson D. A. and Aitkin M. (1985). Variance components models with binary response: Interviewer variability. *Journal of the Royal Statistical Society*, Series B **47**:203–210.

[8] Ashford J. R. and Sowden R. R. (1970). Multivariate probit analysis. *Biometrics*, **26**:535–546.

[9] Bahadur R. R. (1961). *A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (ed.), Studies in Item Analysis and Prediction, pp. 158-168.* Stanford University Press, Palo Alto,CA.

[10] Baksalary J.K., Corston C.A., and Kala R. (1978). Reconciliation of two different views on estimation of growth curve parameters. *Biometrika*, **65**:662–665.

[11] Baltagi B.H. (1995). *Economic Analysis of Panel Data.* John Wiley & Sons.

[12] Becker M. P. and Balagtas C. C. (1993). Marginal modeling of binary cross-over data. *Biometrics*, **49**:997–1009.

[13] Bellman R. E. (1961). *Adaptive Control Processes.* Princeton University Press, Princeton NJ.

[14] Billingsley P. (1961). Statistical methods in markov chains. *Annals of Mathematical Statistics*, **32**:12–40.

[15] Booth J. G. and Hobert J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society*, Series B**61**:265–285.

[16] Box G.E.P. (1950). Problems in analysis of growth and wear data. *Biometrics*, **6**:362–389.

[17] Box G.E.P. and Jenkins G.M. (1970). *Time Series Analysis: Forecasting and Control*. Holden day, San Fransisco, California, revised edition.

[18] Bradely E.L. (1973). The equivalence of maximum likelihood and weighted least squares in exponential family. *J. Statist. Assoc*, **68**:199–200.

[19] Breslow N. E. and Clayton D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**:9–25.

[20] Breslow N. E. and Lin X. (1995). Bias correction in generalized linear models with a single component of dispersion. *Biometrika*, **82**:81–91.

[21] Buja A., Hastie T., and Tibshirani R. (1989). Linear smoothers and additive models. *Annals of Statistics*, **17**:453–510.

[22] Burnham K.P. and Anderson D.R. (2002). *Model Selection and Multimodel InferenceA Practical Information- Theoretical Approach*. Wiley, New York, second ed edition.

[23] Calinski T. and Caussinus H. (1989). A note on the analysis of covariance: Efficiency of concomitant variables. *J. Statist. Palnn. Infer.*, **21**:315–326.

[24] Chatfield C. and Goodhardt G. J. (1970). The beta-binomial model for consumer purchasing behaviour. *Applied Statistics*, **19**:240–250.

[25] Cheng M. Y., Fan J., and Marron J. S . (1997). On automatic boundary corrections. *Annals of Statistics*, **25**:1691 – 1708.

[26] Cole J.W. L. and Grizzle J. E. (1966). Applications of multivariate analysis of variance to repeated measurements experiments. *Biometrics*, **22**:810–828.

[27] Cox D. R. (1958). The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society*, Series B **20**:215–242.

[28] Cox D. R. (1972). The analysis of multivariate binary data. *Applied Statistics*, **21**:113–120.

[29] Crowder M. J. (1978). Beta-binomial anova for proportions. *Applied Statistics*, **27**:34–37.

[30] Crowder M. J. (1979). Inference about the intra-class correlation coefficient in the beta-binomial anova for proportions. *Journal of the Royal Statistical Society*, Series B **41**:230–234.

[31] Crowder M.J. and Hand J. (1990). *Analysis of Repeated Measures*. Chapman and Hall.

[32] Dale J. R. (1984). Local versus global association for bivariate ordered responses. *Biometrika*, **71**:507–514.

[33] Danford M. B., Hughes H. M., and McNee R. C. (1960). On the analysis of repeated measurements experiments. *Biometrics*, **16**:547–565.

[34] Davidian M. and Giltinan D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, London.

[35] de Boor C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

[36] Demidenko E. (2004). *Mixed Models: Theory and Applications*. Wiley, New York.

[37] Dempster A. P., Laird N. M., and Rubin D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**:1–38.

[38] Dempster A.P., Rubin D.B., and Tsutakawa R.K. (1981). Estimation in covariance components models. *Journal of American Statistical Association*, **76**:341–353.

[39] Diggle P.J. and Kenward M.G. (1994). Informative dropout in longitudinal data analysis (withdiscussion). *Applied Statistics*, **43**:49–93.

[40] Diggle P.J., Heagerty P., Liang K.Y., and Zeger S.L. (2003). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK, 2 edition.

[41] Dobson A.J. (1990). *An Introduction to Generalised Linear Models*. Chapman and Hall.

[42] Duncan G.J. and Kalton G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, **55**:97–117.

[43] Durbin J. (1960). Estimation of parameters in time-series regression models. *Biometrika*, **47**:139–153.

[44] Eilers P. H. C. and Man B. D. (1996). Flexible smoothing with b-splines and penal ties. *Statistics Science*, **11**:89–102.

[45] Ekholm A.(1991). Fitting regression models to a multivariate binary response. In G. Rosenqvist K. Juselius K. Nordstrom and J. Palmgren (eds.). *A Spectrum of Statistical Thought: Essays in Statistical Theory, Economics, and Population Genetics in Honour of Johan Fellman.* pp.19-32 Swedish School of Economics and Business Administration, Helsingfors.

[46] Ekholm A.and Smith P. W. F. and McDonald J. W. (1995). Marginal regression analysis of a multivariate binary response. *Biometrika*, **82**(847-854).

[47] Elston R.C. and Grizzle J.E. (1962). Time response curves and their confidence bands. *Biometrics*, **18**:148–159.

[48] Engle R. F. and Hendry D. F.and Richard J. F. (1983). Exogeneity. *Econometrica*, **51**:277–304.

[49] Eubank R.L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

[50] Eubank R.L. (1999). *Nonparametric Regression and Spline Smoothing.* Marcel Dekker, New York.

[51] Fan J. (1992). Design-adaptive nonparametric regression. *Journal of American Statistical Association*, **87**:998–1004.

[52] Fan J. (1993). Local linear regression smoothers and their minimax efficiency. *Annals of Statistics*, **21**:196–216.

[53] Fan J. and Gijbels I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, **20**:2008–36.

[54] Fan J. and Gijbels I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.

[55] Finney D.J. (1952). *Probit Analysis*. Cambridge University Press.

[56] Firth D. (1991). *Generalised linear models. In:* Statistical Theory and Modelling. In Honour of Sir David Cox. Chapman and Hall, Eds. Hinkley, D.V., Reid, N. and Snell, E.J.

[57] Fisher R.A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, **52**:399–433.

[58] Fisher R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**:3–32.

[59] Fisher R.A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

[60] Fisher R.A. (1935). The case of zero survivors [appendix to bliss (1935)]. *Annals of Applied Biology*, **22**:164–165.

[61] Fitzmaurice G. M. and Laird N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**:141–151.

[62] Fitzmaurice G. M., Laird N. M., and Rotnitzky A. G. (1993). Regression models for discrete longitudinal responses (with discussion). *Statistical Science*, **8**:248–309.

[63] Fitzmaurice G.M., Laird N.M., and Ware J.H. (2004). *Applied Longitudinal Analysis.* Wiley, New York.

[64] Friedman J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, **19**:1–68.

[65] Friedman J. H. and Silverman B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics*, **31**:3–39.

[66] Gasser T.and Miiller H. G. and Mammitzsch V. (1985). Kernels for nonparametric curve estimation. *Journal of Royal Statistical Society, Series B*, **86**:665–672.

[67] Geisser S. and Greenhouse S.W. (1958). An extension of box's results on the use of f-distribution in multivariate analysis. *Annals of Mathematical Statistics*, **29**:885–891.

[68] Geisser S. (1963). Multivariate analysis of variance for a special covariance case. *Journal of the American Statistical Association*, **58**:660–669.

[69] Geisser S. (1970). A bayesian analysis of growth curves. *Sankhya*, A **32**:53–64.

[70] Geisser S. (1980). Growth curve analysis. *Handbook of Statistics (Ed. Krishnaiah, P.R.) North Holland*, **1**:89–115.

[71] Gill J. (2000). *Generalised Linear Models: A Unified Approach.* Sage Publications.

[72] Glonek G. F. V. (1996). A class of regression models for multivariate categorical responses. *Biometrika*, **83**:15–28.

[73] Glonek G. F. V. and McCullagh P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society*, Series B**57**:533–546.

[74] Godambe V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**:1208–1212.

[75] Goldstein H. (1979). *The design and Analysis of Longitudinal Studies*. North Holland Mathematical Library, Academic Press, New York.

[76] Goldstein H. (1995). *Multilevel Statistical Models*. Halstead Press., New York, 2^{nd} edition.

[77] Graybill F.A. (1976). *Theory and Applications of the Linear Model*. North Holland Mathematical Library, Duxbury Press.

[78] Green P. and Silverman B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

[79] Greenhouse S.W. and Geisser S. (1959). On methods in the analysis of profile data. *Psychometrika*, **32**:95–112.

[80] Greenwood M. and Yule G. U. (1920). An enquiry into the nature of frequency distributions representative of multiple happenings with particular reference of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, **83**:255–279.

[81] Griffiths D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, **29**:37–48.

[82] Grizzle J. and Allen D. (1969). Analysis of growth and dose response curves. *Biometrics*, **25**:357–381.

[83] Grizzle J. E., Starmer C. F., and Koch G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, **15**:489–504.

[84] Gu C. (2002). *Smoothing Spline ANOVA Models.* Springer -Verlag, New York.

[85] Gumbel E. J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association*, **56**:335–349.

[86] Harville D.A. (1976). Extension of the gauss-markov theorem to include the estimation of random effects. *Annals of Statistics*, **4**:384–395.

[87] Harville D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**:320–338.

[88] Hastie T.J. and Loader C. (1993). Local regression: automatic kernel carpentry (with discussion). *Statistics Science*, **8**:120–143.

[89] Hedeker D. and Gibbons R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**:933–944.

[90] Hedeker D. and Gibbons R. D. (1996). Mixor: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, **49**:157–176.

[91] Hedeker D and Gibbons R.D. (2006). *Longitudinal Data Analysis.* Wiley, New York.

[92] Henderson C. R. (1963). *Selection index and expected genetic advance.In W. D. Hanson and H. F. Robinson (eds.), Statistical Genetics and Plant Breeding.* Washington, D.C, National Academy of Sciences-National Research Council.

[93] Henderson R., Diggle P., and Dobson A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**:465–480.

[94] Hosmer D.W.Jr. and Lemeshow S. (2000). *Applied Logistic Regression.* John Wiley & Sons, 2^{nd} edition.

[95] Hurvich C.M. and Tsai C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**:297–307.

[96] Jennrich R.I. and Moore R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Amer. Statist. Assoc. Proc. Statist. Computing Section*, :57–65.

[97] Jennrich R.I. and Schluchter M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**:805–820.

[98] Jones R.H. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach.* Chapman and Hall.

[99] Kass R.E. and Steffey D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, **84**:717–726.

[100] Kenward M.G. (1985). The use of fitted higher order polynomial coefficients as covariates in the analysis of growth curves. *Biometrics*, **41**:19–28.

[101] Kenward M.G.(1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, **36**:296–308.

[102] Kleinbaum D.G. (1973). A generalisation of the growth curve model which allows missing data. *Journal of Multivariate Analysis*, **3**:117–124.

[103] Koch G. G. and Reinfurt D. W. (1971). The analysis of categorical data from mixed models. *Biometrics*, **27**:157–173.

[104] Koch G. G., Landis J. R., Freeman J. L., Freeman D. H., and Lehnen R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, 33:133–158.

[105] Korn E. L. and Whittemore A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, **35**:795802.

[106] Kshirsagar A.M. and Smith W.B. (1995). *Growth Curves*. Marcel Dekker.

[107] Kuk A. Y. C. and Cheng Y. W. (1997). The monte carlo newton-raphson algorithm. *Journal of Statistical Computation and Simulation*, **59**:233–250.

[108] Kupper L. L. and Haseman J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics*, **34**:69–76.

[109] Laird N.M. and Ware J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**:963–974.

[110] Laird N. M., Lange N., and Stram D. (1987). Maximum likelihood computations with repeated measures: Application of the em algorithm. *Journal of American Statistical Association*, **82**:97–105.

[111] Lang J. B. and Agresti A. (1994). Simultaneous modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, **89**:625–632.

[112] Lee J.C. (1988). Prediction and estimation of growth curves with special covariance structures. *Journal of the American Statistical Association*, **83**:432–440.

[113] Lee J.C. and Geisser S. (1972). Growth curve prediction. *Sankhya*, A **34**:393–412.

[114] Li L., Shao J., and Palta M. (2005). A longitudinal measurement error model with a semicontinuous covariate. *Biometrics*, **61**:824–830.

[115] Liang K.Y. and Zeger S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**:13–22.

[116] Liang K.Y., Zeger S. L., and Qaqish B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society*, Series B **54**:2–24.

[117] Lindsey J. K. (1993). *Models for Repeated Measurements*. Oxford University Press, 1993.

[118] Lindstrom M. J. and Bates D. M. (1990). Nonlinear mixed-effects models for repeated measures. *Biometrics*, **46**:673–687.

[119] Lipsitz S. R., Laird N. M., and Harrington D. P. (1990). Maximum likelihood regression methods for paired binary data. *Statistics in Medicine*, **9**:1417–1425.

[120] Longford N.T. (1993). *Random Coefficient Models*. Oxford Univ. Press, New York.

[121] Louis T.A. and Spiro. A. (1984). Fitting first order autoregressive models with covariates. *Biometrics*, **38**:963–974.

[122] Marron J. S. and Nolan D (1988). Canonical kernels for density estimation. *Statistics and Probability Letters*, **7**:195–9.

[123] Mason W.B. and Fienberg S.E. (Eds.) (1985). *Cohort Analysis in Social Research: Beyond the Identification Problem*. Springer Verlag.

[124] McCullagh P. and Nelder J. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2^{nd} edition.

[125] McCulloch C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**:162–170.

[126] McQuarrie A.D.R. and Tsai C.L. (1998). *Regression and Time Series Model Selection.* World Scientific., Singapore.

[127] Molenberghs G. and Lesaffre E. (1994). Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association*, **89**:633–644.

[128] Molenberghs G. and Ritter L. (1996). Methods for analyzing multivariate binary data, with the association between outcomes of interest. *Biometrics*, **52**:1121–1133.

[129] Molenberghs G. and Verbeke G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.

[130] Nadaraya E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**:141–42.

[131] Naveen V P (2011). *Role of Sida Cordifolia in the Management of Peripheral Neuropathy in Wistar Albino Rats*. M.Sc (Human Phisiology) Dissertation Work (Unpublished).

[132] Nelder J. A. and Wedderburn R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Series A **135**:370–384.

[133] Neter J, Kutner M.H., Nachtsheim C.J., and Wasserman W. (1996). *Applied Linear Regression Models,*. Richard D. Irvin, 3^{rd} edition.

[134] Ochi Y. and Prentice R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, **71**:531–543.

[135] Otake M. and Prentice R. L. (1984). The analysis of chromosomally aberrant cells based on beta-binomial distribution. *Radiation Research*, **98**:456–470.

[136] Patterson H.D. and Thompson R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**:545–554.

[137] Pepe M. S. and Anderson G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics Simulation and Computation*, **23**:939–951.

[138] Pierce D. A. and Sands B. R. (1975). Extra-bernoulli variation in binary data. *Technical Report*, **46**.

[139] Pinheiro J. and Bates D. (2000). *Mixed-effectsModels insands-plus*. Springer-Verlag, New York.

[140] Potthoff R.F. and Roy S.W. (1964). A generalised multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**:313–326.

[141] Prentice R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**:1033–1068.

[142] Rao C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, **14**:1–17.

[143] Rao C. R. (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, **46**:49–58.

[144] Rao C. R. (1965). The theory of least squares when the parameters are stochastic and its applications to the analysis if growth curves. *Biometrika*, **52**:447–458.

[145] Rao C. R. (1975). Simultaneous estimations of parameters in different linear models and applications to biometric problems. *Biometrika*, **31**:545–554.

[146] Rao C. R. (1987). Prediction of future observations in growth curve models. *Statistical Science*, **4**:434–471.

[147] Raudenbush S. W., Yang H.-L., and Yosef M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational and Graphical Statistics*, **9**:141–157.

[148] Raudenbush S W and Bryk A S (2002). *Hierarchical Linear Models*. Sage, Thousand Oaks,CA, 2^{nd} edition.

[149] Robins J. M., Greenland S., and Hu F.C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association*, **94**:687–712.

[150] Robinson G. K. (1991). That blup, is a good thing: the estimation of random effects (with discussions). *Statistics Science*, **6**:15–32.

[151] Rogan W.J., Dietrich K.N., Ware J.H., Dockery D.W., Salganik M., Radcliffe J., Jones R.L., Ragan N.B., Chisolm J.J., and Rhoads G.G. (2001). The effect of chelation therapy with succimer on neuropsychological development in children exposed to lead. *New England Journal of Medicine*, **344**:1421–1426.

[152] Rosenman R.H., Brand R.J., Jenkins C.D., Friedman M., Straus R., and Wurm M. (1975). Coronary heart disease in the western collaborative study: Final follow-up experience of $8\frac{1}{2}$ years. *Journal of the American Medical Association*, **233**:872–877.

[153] Rowell J. G. and Walters D. E. (1976). Analysing data with repeated observations on each experimental unit. *Journal of Agricultural Science*, **87**:423–432.

[154] Ruppert D., Sheather S.J., and Wand M.P. (1995). An effective bandwidth selector for local least squares regression. *Journal of American Statistical Association*, **90**:1257–1270.

[155] Ruppert D., Wand M. P., and Carroll R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

[156] Schall R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**:719–727.

[157] Scheffe H. (1956). Alternative models for the analysis of variance. *Annals of Mathematical Statistics*, **27**:251–271.

[158] Schefe H. (1959). *The Analysis of Variance*. John Wiley & Sons.

[159] Schwarz G. (1978). Estimating the dimension of a model. *Ann. Statistics.*, **6**:461–464.

[160] Searle S.R, Casella G, , and McCulloch C.E. (1992). *Variance Components*. Wiley, New York.

[161] Shi P. and Tsai C.L. (2002). Regression model selectiona residual likelihood approach. *J. Roy. Statist. Soc. Ser. B*, **64**:237–252.

[162] Singer J.D. and Willett J.B. (2003). *Applied Longitudinal Data Analysis.* Oxford University Press, New York.

[163] Skellam J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society*, Series B **10**(257-261).

[164] Smith M. and Kohn R. (1996). Nonparametric regression via bayesian variable selection. *Journal of Econometrics*, **75**:317–344.

[165] Stanish W. M., Gillings D. B., and Koch G. G. (1978). An application of multivariate ratio methods for the analysis of a longitudinal clinical trial with missing data. *Biometrics*, **34**:3005–3117.

[166] Stanish W. M. and Koch G. G. (1984). The use of catmod for repeated measurement analysis of categorical data. *Proceedings of the Ninth Annual SAS Users Group International Conference*, **9**:761–770.

[167] Stiratelli R., Laird N. M., and Ware J. H. (1984). Random effects models for serial observations with binary response. *Biometrics*, **40**:961–971.

[168] Stone C. J. (1984). An asymptotically optimal window selection rule for kernel density estimation. *Annals of Statistics*, **12**:1285–97.

[169] Stone C. J., Hansen M.H., Kooperberg C., and Truong Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics*, **25**:1371–1425.

[170] Van Marter L.J., Leviton A., Kuban K.C.K., Pagano M., and Allred E.N. (1990). Maternal glucocorticoid therapy and reduced risk of bronchopulmonary dysplasia. *Pediatrics*, **86**:331–336.

[171] Verbeke G. and Molenberghs G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York.

[172] Vonesh E.F. and Chinchilli V.M. (1996). *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* Marcel Dekker, New York.

[173] Vonesh E.F. and Chinchilli V.M. (1997). Efficient inference for random-coefficients growth curve models with unbalanced data. *Biometrics*, **42**:601–610.

[174] Wahba G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. **59**.

[175] Wand M.P. and Jones M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

[176] Ware J. H.and Lipsitz S. R. and Speizer F. E. (1988). Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine*, **7**:95–107.

[177] Ware J.H. and Liang K.Y. (1996). *The design and analysis of longitudinal studies: A historical perspective. In:* Advances in Biometry *(Eds. Armitage, P. and David, H.A)*. John Wiley & Sons, 2 edition.

[178] Watson G. S . (1964). Smooth regression analysis. *Sankhya*, **26**:101–116.

[179] Wedderburn R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss- newton method. *Biometrika*, **61**:439–447.

[180] Weiss R E (2005). *Modelling Longitudinal Data*. Springer, New York.

[181] Williams D. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31**:949–952.

[182] Winer B.J. (1971). *Statistical Principles in Experimental Design*. McGraw Hill, 2 edition.

[183] Wishart J. (1938). Growth rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrics*, **30**:16–28.

[184] Wolfinger R. (1993). Laplaces approximation for nonlinear mixed models. *Biometrika*, **80**:791–795.

[185] Woolson R. F. and Clarke W. R. (1984). Analysis of categorical incomplete longitudinal data. *Journal of the Royal Statistical Society*, Series A **147**:87–99.

[186] Yates F. (1935). Complex experiments (with discussion). *Supplement to the Journal of the Royal Statistical Society. Biometrics*, **2**:181–247.

[187] Zhang J.T. and Fan J. (2000). Minimax kernels for nonparametric curve estimation. *Journal of Nonparametric Statistics*, **12**:417–445.

[188] Zeger S. L. and Karim M. R. (1991). Generalized linear models with random effects: A gibbs sampling approach. *Journal of the American Statistical Association*, **86**:79–86.

[189] Zeger S. L.and Liang K.Y. and Self S. G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, **72**:31–38.