University of Cape Town
Department of Statistical Sciences

A variance shift model for outlier detection and estimation in linear and linear mixed models

by

Freedom Nkhululeko Gumedze

Thesis Presented for the Degree of
Doctor of Philosophy in Statistical Sciences
in the Department of Statistical Sciences
University of Cape Town

2008

To my pillars of strength:
To my wife Lulu, daughter Siviwe and son Saneliso whose endurance and patience have been tested through the years of my research. This dedication will never compensate for the sacrifices you have made to support me during this research.

Abstract of the Dissertation

A variance shift model for outlier detection and estimation in linear and linear mixed models

by

Freedom Nkhululeko Gumedze

Outliers are data observations that fall outside the usual conditional ranges of the response data.They are common in experimental research data, for example, due to transcription errors or faulty experimental equipment. Often outliers are quickly identified and addressed, that is, corrected, removed from the data, or retained for subsequent analysis. However, in many cases they are completely anomalous and it is unclear how to treat them.

Case deletion techniques are established methods in detecting outliers in linear fixed effects analysis. The extension of these methods to detecting outliers in linear mixed models has not been entirely successful, in the literature. This thesis focuses on a variance shift outlier model as an approach to detecting and assessing outliers in both linear fixed effects and linear mixed effects analysis. A variance shift outlier model assumes a variance shift parameter, ω_i, for the ith observation, where ω_i is unknown and estimated from the data. Estimated values of ω_i indicate observations with possibly inflated variances relative to the remainder of the observations in the data set and hence outliers. When outliers lurk within anomalous elements in the data set, a variance shift outlier model offers an opportunity to include anomalies in the analysis, but down-weighted using the variance shift estimate $\hat{\omega}_i$. This down-weighting might be considered preferable to omitting data points (as in case-deletion methods). For very large values of ω_i a variance shift outlier model is approximately equivalent to the case deletion approach.

We commence with a detailed review of parameter estimation and inferential procedures for the linear mixed model. The review is necessary for the development of the variance shift outlier model as a method for detecting outliers in linear fixed and linear mixed models. This review is followed by a discussion of the status of current research into linear mixed model diagnostics. Different types of residuals in the linear mixed model are defined. A decomposition of the leverage matrix for the linear mixed model leads to interpretable leverage measures.

A detailed review of a variance shift outlier model in linear fixed effects analysis is given. The purpose of this review is firstly, to gain insight into the general case (the linear mixed model) and secondly, to develop the model further in linear fixed effects analysis. A variance shift outlier model can be formulated as a linear mixed model so that the calculations required to estimate the parameters of the model are those associated with fitting a linear mixed model, and hence the model can be fitted using standard software packages.

Likelihood ratio and score test statistics are developed as objective measures for the variance shift estimates. The proposed test statistics initially assume balanced longitudinal data with a Gaussian distributed response variable. The dependence of the proposed test statistics on the second derivatives of the log-likelihood function is also examined. For the single-case outlier in linear fixed effects analysis, analytical expressions for the proposed test statistics are obtained. A resampling algorithm is proposed for assessing the significance of the proposed test statistics and for handling the problem of multiple testing. A variance shift outlier model is then adapted to detect a group of outliers in a fixed effects model. Properties and performance of the likelihood ratio and score test statistics are also investigated.

A variance shift outlier model for detecting single-case outliers is also extended to linear mixed effects analysis under Gaussian assumptions for the random effects and the random errors. The variance parameters are estimated using the residual maximum likelihood method. Likelihood ratio and score tests are also constructed for this extended model. Two distinct computing algorithms which constrain the variance parameter estimates to be positive, are given. Properties of the resulting variance parameter estimates from each computing algorithm are also investigated.

A variance shift outlier model for detecting single-case outliers in linear mixed effects analysis is extended to detect groups of outliers or subjects having outlying profiles with random intercepts and random slopes that are inconsistent with the corresponding model elements for the remaining subjects in the data set. The issue of influence on the fixed effects under a variance shift outlier model is also discussed.

Acknowledgments

I record my deepest appreciation to Prof. Tim Dunne for his expert guidance and encouragement during this research. It has been a challenge to both teach and to conduct this research at the same time. Prof. Dunne understood my constraints and helped me patiently.

I also record my debt and my appreciation to Dr Sue Welham and Prof. Robin Thompson who provided the research topic and co-supervised this research during my sabbatical leave at Rothamsted Research. This thesis would not have been completed without their extensive support. I especially thank Dr Welham for checking my proofs and for her help with the programming in GenStat.

My sincere thanks to Dr Brian Cullis for introducing me to Dr Welham and Prof. Thompson, and for suggesting the use of the score tests discussed in this research.

I gratefully acknowledge Dr Beverley Gogel and Prof. Ari Verbyla for providing the framework for this research (the use of a variance shift outlier model to detect outliers).

To Assoc. Prof. Christien Thiart for critical reading of the thesis, I offer my thanks and gratitude.

Prof. Theo Stewart introduced me to LaTeX and gave me many LaTeX tips and tricks. Any inelegances in text are my own oversight.

I am very grateful to Prof. June Juritz for the introduction to REML and the area of linear and generalized linear mixed models, which gave me the courage to pursue this research.

My sincere thanks and gratitude to the academic and non-academic staff, and post-graduate students of the Department of Statistical Sciences for the friendly environment to conduct this research.

This research has been made possible by the financial assistance of the National Research Foundation, RSA, Rothamsted Research, UK and Commonwealth Scholarship Commission.

I am grateful to the University of Cape Town and the Faculty of Science for granting me the sabbatical leave in 2005 to focus on this research.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Problem statement

The linear mixed model (LMM) is used for analyzing correlated Gaussian data and appears in the literature under various names. These names include: covariance components model (Hocking, 1985), hierarchical linear model (Bryk and Raudenbush, 1992), linear mixed model (Searle, 1971, 1982a; Lindstrom and Bates, 1988; Searle et al., 1992; Brown and Prescott, 1999; Verbeke and Molenberghs, 2000; Demidenko, 2004), linear mixed-effects model (Hartley and Rao, 1967; Harville, 1977; Pinheiro and Bates, 2000), longitudinal data model (Hand and Crowder, 1996; Diggle et al., 2002; Fitzmaurice et al., 2004), multilevel linear model (Goldstein, 1995), random coefficient model (Swamy, 1971; Judge et al., 1980; Longford, 1993), random effects model (Laird and Ware, 1982; Stiratelli et al., 1984; Vonesh and Carter, 1987; Robinson, 1991) and repeated measures model (Laird et al., 1987; Crowder and Hand, 1990; Vonesh and Chinchilli, 1997). The linear mixed model is also related but not equivalent to the following models: empirical Bayes model (Lindley and Smith, 1972), growth curve model (Potthoff and Roy, 1964; Grizzle and Allen, 1969; von Rosen, 1991; Pan and Fang, 2002), and state-space model (Jones, 1993). Types of data that are usually analyzed using linear mixed models include: longitudinal data, repeated measures data, growth curve data, pharmacokinetic data, clustered (nested) data, multivariate data and panel data. These types of data occur in many disciplines including finance, economics, medicine, agriculture, biology, quality control and education. In most of these application areas the focus is on fixed effects but in the context of taking into account the dependence between observations, i.e. the random effects are treated as a nuisance source of variation. In some situations the random effects may be the focus of statistical inference. The estimation of random effects has been widely used in animal breeding applications and other areas, such as quantitative genetics, Kriging in geology, ecology and forestry, credibility theory in actuarial science and Kalman filtering in physics and finance. Robinson (1991) gives a detailed discussion of applications in which the random effects are of primary interest.

In the linear model diagnostic literature an outlier is defined as an observation which appears inconsistent with the remainder of the data (Beckman and Cook, 1983; Pan and Fang, 2002) or as an observation that does not follow the distributional pattern of the majority of the data (Rousseeuw and Leroy, 1987). The latter definition implies that the observation is a contaminant in the data set, i.e. it arises from a different distribution from the distribution from which the rest of the data come. Anscombe (1960) refers to such observations as spurious and states that they give rise to outliers; while Barnett and Lewis (1995) use the phrase 'discordant outlier' to refer to outlying but genuine observations. In this thesis we consider outliers to be data observations that fall outside the presumed or typical range of the response data. These observations are different from the the so-called influential observations (Belsley et al., 1980) which are typically located at outlying positions in the design space (Huber, 1981). Influential observations are observations which result in an unusually large influence on the fitted model. However, Velleman, commenting on Chatterjee and Hadi (1986) (see Chatterjee and Hadi, 1986, pp. 412-413), argues that this definition of an influential point (a point that alters the model estimates) is misleading since it ignores extreme points. He further argues that influential points may result in a high correlation because they lie on the line already described by the remaining data points. Hence all extreme points that affect the model estimates or their test statistics and conclusions of the study should be subject to scrutiny.

While outliers in linear fixed effects models have been extensively researched (Belsley et al., 1980; Barnett and Lewis, 1995; Beckman and Cook, 1983; Chatterjee and Hadi, 1986), research into outliers in linear mixed models is more recent (Tan et al., 2001; Haslett and Dillane, 2004; Zewotir and Galpin, 2005; Dillane, 2006; Haslett and Haslett, 2007; Zewotir and Galpin, 2007). We give a formal review of some these studies in Chapter 3. Linear mixed models extend fixed linear models to include more than one source of random variation and hence multiple variance parameters. In linear mixed models the variance parameters are not known but are estimated from the data. In the linear mixed model diagnostics literature very few authors (Langford and Lewis, 1998; Christensen et al., 1992a) explicitly define the term outlier because (i) the term 'outlier' now encompasses outliers for both the random terms in the model and for the variance parameter estimates, i.e. there are different types of outliers and (ii) the type of outlier(s) of particular interest is context-driven. We consider three broad definitions of outliers:

(i) influential for the assumed model (I),

(ii) high leverage (HL),

(iii) a combination of (i) and (ii) (HL & I).

An illustration of these different types of the outliers is given in Figure 1.1. Outliers of types (i) and (iii) are usually of most concern to us since they alter the model estimates and inference. In particular, type (i) outliers are the most problematic since they are camouflaged among the ordinary (D) observations and hence may be undetected in regression plots. Type (ii) outliers (high leverage) are usually of less concern as they do not necessarily alter the model estimates. Andrews and Pregibon (1978) call this type of outliers, 'outliers that do not matter'.

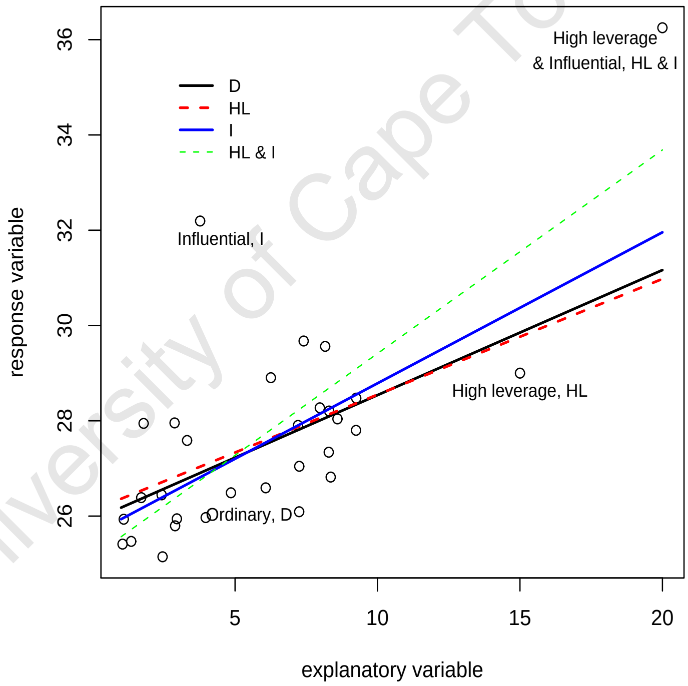

Figure 1.1: *Types of outliers in a typical data set with fitted lines.*

Once outliers are found, their treatment depends on the specific context. Nevertheless the following remedial action may be taken. If it is verified that the outliers are due to gross measurement errors, data entry errors or improper experimental conditions, then they should either be deleted or corrected, if possible. However, in many cases outliers are completely anomalous and it is unclear how to treat them. In these cases methods that retain the outliers in the analysis, but also minimize their influence on the model estimates, may be employed. These methods for accommodating the outliers are the so-called robust estimation methods (Huber, 1981; Rousseeuw, 1984; Fellner, 1986; Rousseeuw and Leroy, 1987). An outlier model (for example, Box and Tiao, 1968; Sharples, 1990a; Barnett and Lewis, 1995) and the use of heavy-tailed error distributions such as the t-distribution for outlying observations (for example, Lange et al., 1989; Pinheiro et al., 2001) are robust modelling techniques and are in the same spirit as robust estimation methods (accommodation of outliers).

Outliers are not necessarily bad data points as they may comprise the most interesting information about the phenomenon under study. In geology, for example searching for minerals (Wang, 1982); outliers may be observations which contain mineral deposit.

The most common method for the study of model diagnostics in linear models is case-deletion. This approach quantifies the effect of the ith observation on the fit when that observation is deleted. This quantification is achieved by computing the change in some aspect of the fit that occurs on deleting the ith observation. Many numerical measures summarizing the effects of deleting the ith observation are suggested in the literature on linear models (Belsley et al., 1980; Chatterjee and Hadi, 1986). Several researchers have extended the case-deletion measures for linear models to linear mixed models (see Chapter 3 for a detailed discussion). A major shortcoming of case-deletion measures for the linear mixed model is that the error variance-covariance matrix is not re-estimated with each successive deletion, which is a necessary limitation for correlated responses. Some attempts at solving this problem have been made by several authors (for example, Haslett, 1999; Haslett and Dillane, 2004). Moreover case-deletion measures for linear mixed models remain largely inaccessible to researchers. Oman (1995) noted that, for linear mixed models, most statistical packages concentrate on estimation and hypothesis testing regarding the model parameters and pay less attention to model diagnostics.

As part of the emerging research into linear mixed model diagnostics, this thesis adopts the variance shift model (Cook and Weisberg, 1982) as a method for outlier detection and accommodation in the linear model and in the linear mixed model. This model was also considered by Cook et al. (1982) and Thompson (1985). A variance shift model assumes that the variance of all the observations, except for one unknown case, is σ^2. The variance of the unknown case, say the ith unit is assumed to be $\alpha_i\sigma^2$ ($\alpha_i \geq 1$). The unknown parameter α_i is estimated from the data and acts as a variance inflator in the variance for the ith unit. In this thesis we will call this model a variance shift outlier model (VSOM) and parameterize the variance of the ith observation as $\sigma^2(\omega_i + 1)$, $\omega_i \geq 0$. The unknown parameter ω_i acts as variance shift parameter. A VSOM views observations with inflated variances as possible outliers. This is a different emphasis to case-deletion methods, despite the direct link when the variance shift parameter for an observation is infinite.

A related model is an outlier model of Box and Tiao (1968) which assumes that the data come from a Gaussian distribution with probability $1 - \pi$, and from a contaminating Gaussian distribution with probability π so that the error term e_i has a Gaussian mixture distribution, $(1-\pi)N(0, \sigma^2) + (\pi)N(0, k^2\sigma^2)$. A VSOM differs from the Box-Tiao outlier model which assumes that the proportion π of outliers in the data and the scale parameter k are known and fixed. Box and Tiao (1968) also envisage more than one outlier. Moreover, the purpose of their outlier model is primarily to accommodate the outlier(s) by assuming that they have a different distribution from the majority of the data, rather than to detect them.

The purpose of a VSOM is two fold. First, it allows us to (iteratively) highlight observations with inflated variances relative to the remainder of the observations and hence are outliers. Secondly, when it can be determined that the variance shift is large enough (using objective measures, such as likelihood ratio tests and score tests), we can either (a) remove the observation from the analysis or (b) include it in the analysis with the variance shift used as an input to weighting for the particular observation in the estimation process, generally by the inverse to the root of the estimated variance factor. A similar weighting is employed in weighted least squares (WLS) and generalized least squares (GLS) for linear models with a general covariance structure. Weighted least squares corrects for the heterogeneity in the errors by assigning a weight to each observation which is proportional to the inverse of its standard deviation. While GLS corrects for both variance heterogeneity and dependence among observations, by rescaling the data in a similar manner to weighted least squares, it also assumes a model for the correlation among observations e.g. a first-order autoregressive model. For parameter estimation, ordinary least squares is then applied to the weighted observations. Therefore, a VSOM approach differs from both WLS and GLS in that the weighting is applied to particular observation(s) and not *en masse*. An appealing feature of a VSOM is that the calculations required to estimate the parameters of the model are those associated with fitting a linear mixed model and so can be easily produced in standard software packages such as ASReml (Gilmour et al., 2002), GenStat

(Welham and Thompson, 2000), R (R Development Core Team, 2005), SAS (Littell et al., 1996), S-PLUS (Insightful Corp., 2001), SPSS (SPSS Inc., 1999) and Stata (StataCorp., 2005). Zewotir (2007) considered a model similar to a VSOM in the context of local influence analysis for the linear mixed model. He examined the effect of using known perturbations (weights) in the error variance of a single observation on changes in estimates of fixed and random effects, using a Cook's distance measure (Cook, 1977). Our approach differs in that the weight is estimated by REML and can only increase the error variance, so that any down-weighting is objectively determined. In this thesis we also provide objective measures for determining the amount or level of weighting required and for deciding for which particular observations the weights apply.

The main aim of this study is to develop statistical tests for detecting outliers or a group of outliers in the linear fixed effects analysis under a variance shift outlier model (VSOM) and to extend the results to linear mixed effects analysis. In the linear mixed models we assume the errors and the random effects are independent and have Gaussian distributions.

1.2 Motivation

The study is motivated by the desire to investigate

(a) aspects of a VSOM in linear fixed effects analysis that have not yet been studied. These aspects include statistical tests for a VSOM, dependence of the tests on the second derivatives of the REML log-likelihood function and distributional properties of these tests, and a derivation of a VSOM for detecting a group of outliers, and

(b) the extension a VSOM results in linear fixed effects analysis to linear mixed effects analysis.

1.3 Objective of the thesis

The objectives of this study are:

1. to develop statistical tests for detecting one or more outliers or a group of outliers in linear fixed effects analysis under a variance shift outlier model (VSOM),

2. to examine the dependence of the tests on the second derivatives of the REML log-likelihood function;
3. to assess the asymptotic, approximate, empirical and analytical (exact) distributions of these tests;
4. to extend a VSOM to the linear mixed model;
5. to extend the statistical tests developed for a VSOM in linear fixed effects analysis (VSOM in linear regression) to the standard linear mixed model and some of its variations;
6. to explore the use of the tests to detect one or more outliers or a group of outliers under both a VSOM in linear regression and a VSOM in linear mixed effects analysis (linear mixed VSOM).

The theoretical contributions of this thesis are as follows:

- exact forms of the likelihood ratio test (LRT) and score tests for a VSOM in linear regression,
- properties of one-step updates of the variance parameters under a VSOM,
- distribution of the score test statistic based on the expected information matrix,
- evaluation of the procedure for multiple testing and parametric bootstrap and,
- extension of a VSOM in linear regression to the matter of groups of outliers.
- general forms of the LRT, one-step LRTs and score tests for a VSOM in linear mixed effects analysis,
- approximations of the LRT, one-step LRTs and score tests for a VSOM in linear mixed effects analysis and,
- properties of one-step updates of the variance parameters under a VSOM in linear mixed effects analysis.
- extensions of a VSOM in linear mixed effects analysis to the matter of groups of outliers.

We use simulated data sets and a real data set from the literature to assess the properties and usefulness of the proposed test statistics. The annotated GenStat programs used in the analyses, are appended to the thesis.

1.4 Outline of the thesis

The purpose of this section is to provide an outline of the thesis. In the first chapter we have given the introduction, motivation and objectives for the study.

Chapter 2 provides a review the linear mixed model with a focus on parameter estimation and inference and introduces the central example in the thesis which sets the scene for considering outliers in the linear mixed model. The understanding of the form of the linear mixed model and parameter estimation in the model is important for the methods we develop in Chapters 4-7.

In Chapter 3 we review the literature in linear mixed model diagnostics.

Residuals and leverages play a key role in model diagnostics. While these quantities are well-defined and understood in linear models, this clarity has not been fully attained for linear mixed models. Haslett and Haslett (2007) give three basic types of residuals for the linear model with a general covariance structure, including the linear mixed model. With regard to leverages, the linear mixed model has fixed and random components and hence should have corresponding leverages or joint leverages. We focus our discussion in Chapter 4 on definitions of residuals and leverages for the linear mixed model. The chapter concludes with a demonstration of the usefulness of the proposed measures via a simulated data set and a real data set.

Chapter 5 reviews and extends a variance shift outlier model (VSOM) as a model for outliers in linear fixed effects analysis. The motivation for the review of a VSOM in linear regression is two-fold: (i) to gain insight into the general case (a linear mixed VSOM), our topic in Chapter 6 with some extensions in Chapter 7, and (ii) to develop the method further in this simple case. Likelihood ratio and score test statistics are developed as objective measures for testing for the variance shift estimate to determine whether individual observations are outliers. Below is an outline of Chapter 5.

(a) review of a variance shift outlier model (VSOM),

(b) construction of the likelihood ratio test (LRT),

(c) construction of likelihood ratio tests based on one-step estimates of the variance parameters (one-step LRTs) and score tests, and examination of their dependence on the one-step updates of variance and hence the second-order derivatives of the REML log-likelihood function, (i.e. dependence upon observed, expected and average information matrices),

(d) evaluation of the empirical and asymptotic distributions of the test statistics under the null hypothesis (no outliers),

(e) derivation of the exact distribution of the expected information score test giving insight into the distribution of the LRT,

(f) a resampling procedure to handle the problem of multiple (iterative) testing in using a VSOM approach to identify outliers successively,

(g) illustration of a VSOM approach to outlier detection in linear regression using a simulated data set and a real data set,

(h) evaluation of the performance of the likelihood ratio and score tests in terms of computing time, and type I and type II errors using simulation and

(i) extension of a VSOM to the case of multiple outliers.

Chapter 6 extends a linear regression VSOM results to the linear mixed model. A linear mixed VSOM can be viewed either as an extension of the linear mixed model with one or more additional random shift covariate(s) for the suspect observation(s), or as an extension of a linear regression VSOM (discussed in Chapter 5) with additional random terms for the random effects to account for dependence structures within the data. In Chapter 6 we will focus on a VSOM for individual observations whereas we dedicate Chapter 7 to the discussion of VSOMs for groups of observations. The outline of Chapter 6 is as follows:

(a) formulation of a linear mixed VSOM for a single observation in a given data set,

(b) variance parameter estimation in a linear mixed VSOM,

(c) construction of the LRT, one-step LRTs and score tests, for testing the significance of the variance shift estimate,

(d) updating schemes for use when the variance parameter estimates in a VSOM are less than or equal to zero,

(e) a modified resampling procedure to handle the problem of multiple (iterative) testing under a linear mixed VSOM,

(f) illustrations of a linear mixed VSOM using simulated data sets and two real data sets and

(g) evaluation of the performance of the likelihood ratio and score tests in terms of computing time, and type I and type II errors using simulation.

Chapter 7 gives possible extensions of a linear mixed VSOM discussed in Chapter 6. These extensions are illustrated using two simulated data sets and a real data set.

Chapter 8 presents conclusions of the thesis and possible future research directions.

In summary, Chapters 1 to 4 are essentially the broad scene-setting with highlighted elements that emerge in Chapters 5 to 8 as the thesis and its contributions.

CHAPTER 2

Review of the linear mixed model

The purpose of this chapter is two-fold. Firstly, it is to review the linear mixed model with a focus on parameter estimation and inference. The understanding of the form of the linear mixed model and parameter estimation in the model is important for the methods we develop in Chapters 4-7. Secondly, it is to introduce the central example in the thesis which sets the scene for considering outliers in the linear mixed model.

The chapter is structured as follows. First, we introduce the standard linear mixed model and its assumptions. This introduction is followed by a discussion of parameter estimation and inferential procedures for the various components of the model; the fixed effects parameters, random effects, variance parameters (or ratios). We also discuss inferential procedures for the estimated fixed effects and the variance parameter estimates. We then describe a specific data set which will be used as a typical example in the thesis.

2.1 The model

The linear mixed model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ known design matrix for the fixed effects, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector of fixed effects, $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_b]$, where $\boldsymbol{Z}_i$ is an $n \times q_i$ design matrix for the ith random effects factor, $\boldsymbol{u} = [\boldsymbol{u}_1', \ldots, \boldsymbol{u}_b']'$ is a $q \times 1$ vector of random effects where $\boldsymbol{u}_i$ is a $q_i \times 1$ vector such that $q = \sum_{i=1}^{b} q_i$, and $\boldsymbol{e}$ is an $n \times 1$ vector of random errors, with $\mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$. In addition it is assumed that $\boldsymbol{u}$ and $\boldsymbol{e}$ follow independent and multivariate Gaussian distributions such that

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{G}(\boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}(\boldsymbol{\rho}) \end{bmatrix} \right), \tag{2.2}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$ are $r \times 1$ and $s \times 1$ (with $s \leq n(n+1)/2$) vectors of unknown variance parameters corresponding to $\boldsymbol{u}$ and $\boldsymbol{e}$ respectively. If the random terms are correlated then the dimension of $\boldsymbol{\gamma}$ may exceed q, i.e. $\boldsymbol{\gamma}$ may be of dimension $r \leq q(q+1)/2$. Following Patterson and Thompson (1971) we write the variance-covariance matrix of the data, $\boldsymbol{y}$, as

$$\text{var}(\boldsymbol{y}) = \sigma^2(\boldsymbol{ZGZ}' + \boldsymbol{R}) = \sigma^2\boldsymbol{H}, \tag{2.3}$$

where

$$\boldsymbol{H} = \boldsymbol{ZGZ}' + \boldsymbol{R}. \tag{2.4}$$

The appeal of the parametrization (2.3), i.e. the factoring of the residual variance $\boldsymbol{R}$ out of the variance matrix for the data, is that it reduces the t-dimensional REML log-likelihood maximization problem by unity (Callanan and Harville, 1991), where $t = (r + s + 1)$ is the number of variance parameters in model (2.1). This variance matrix parametrization can also often useful for establishing overall scaling. However, it may not be useful in multivariate analysis of variance (MANOVA) problems where σ^2 has no meaningful interpretation. An alternative parametrization is when the model (2.1) is parametrized in terms of the variance components. For instance assuming $\boldsymbol{R} = \boldsymbol{I}$, the variance matrix is written as

$$\text{var}(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{ZG}^v\boldsymbol{Z}' + \sigma^2\boldsymbol{I}, \tag{2.5}$$

where $\boldsymbol{G}^v$ contains the variance components for each random effect factor and σ^2 is the residual error variance.

The matrix $\boldsymbol{H}$ consists of two components that are used to model heteroscedasticity and correlation: a random effects component $\boldsymbol{ZGZ}'$ and a within-group component $\boldsymbol{R}$. In some applications, the within-group component $\boldsymbol{R}$ is used to directly model the variance-covariance matrix of the data without the need to incorporate random effects in the model to account for dependence among observations.

2.2 Joint estimation of fixed and random effects

Once the model has been formulated, methods are needed to estimate the model parameters. In this section we first deal with the joint estimation of the fixed effects ($\boldsymbol{\beta}$) and random effects ($\boldsymbol{u}$) and then with estimation of the variance parameters ($\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and σ^2). There are many methods for obtaining the estimates of the fixed and random effects simultaneously (Searle et al., 1992, § 7.4c; Robinson, 1991). These methods include Henderson's (1950) mixed model equations (Henderson, 1950), Goldberger's (1962) approach of predicting a future observation, techniques based on two-stage regression, linearity in $\boldsymbol{y}$, partitioning of $\boldsymbol{y}$ and Bayes estimation. In this section we describe estimation using Henderson's mixed model equations because it produces sampling variances for the estimators and because it has a connection with maximum likelihood estimation of the variance parameters.

Henderson (1950) (also see Henderson et al., 1959) assumed $\boldsymbol{u}$ and $\boldsymbol{y}$ to be jointly Gaussian distributed as

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{X}\beta \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{G} & \boldsymbol{G}\boldsymbol{Z}' \\ \boldsymbol{Z}\boldsymbol{G} & \boldsymbol{H} \end{bmatrix} \right) \tag{2.6}$$

Thus $\boldsymbol{y}$ has the marginal probability density function $N[\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H}]$, where $\boldsymbol{H}$ is as defined in (2.4) with $\boldsymbol{G}$ and $\boldsymbol{R}$ assumed known. Henderson (1950) maximized the log-joint distribution of $(\boldsymbol{y}, \boldsymbol{u})$ to obtain estimators of $\boldsymbol{\beta}$ and $\boldsymbol{u}$. However, this logarithmic function is not a log-likelihood function as $\boldsymbol{u}$ is not observed. The marginal distribution is $\boldsymbol{u}$ from (2.6) is

$$\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\boldsymbol{G})$$

and the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$ is

$$\boldsymbol{y}|\boldsymbol{u} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma^2\boldsymbol{R}).$$

Hence the log-joint distribution of $(\boldsymbol{y}, \boldsymbol{u})$ is given by

$$\log f(\boldsymbol{y}, \boldsymbol{u}) = \log f(\boldsymbol{y}|\boldsymbol{u}) + \log f(\boldsymbol{u})$$

$$
\begin{aligned}
&= -\frac{1}{2}\left\{n \log \sigma^2 + \log \boldsymbol{R} + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})/\sigma^2\right\} \\
&\quad - \frac{1}{2}\left\{q \log \sigma^2 + \log \boldsymbol{G} + \boldsymbol{u}'\boldsymbol{G}^{-1}\boldsymbol{u}/\sigma^2\right\} \\
&= -\frac{1}{2}\left\{(n+q) \log \sigma^2 + \log \boldsymbol{R} + \log \boldsymbol{G} + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\sigma^2\right\} \\
&\quad - \frac{1}{2\sigma^2}\left\{\boldsymbol{u}'(\boldsymbol{Z}\boldsymbol{R}^{-1}\boldsymbol{Z}' + \boldsymbol{G}^{-1})\boldsymbol{u} - 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{u}\right\}.
\end{aligned}
$$

This function coincides with the h-likelihood function of Lee and Nelder (1996) for correlated Gaussian data, with the random effects also having a Gaussian distribution (i.e. linear mixed model). However, Lee and Nelder's approach can also handle correlated non-Gaussian data with conjugate distributions assumed for the random effects.

Estimates for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are obtained by solving the score equations

$$
\begin{aligned}
\boldsymbol{X}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\tilde{\boldsymbol{u}} &= \boldsymbol{0} \\
\boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\tilde{\boldsymbol{u}} &= \boldsymbol{0}.
\end{aligned}
$$

These equations are called the mixed model equations (MMEs) as proposed by Henderson (1950) and Henderson et al. (1959). They wrote the equations compactly in matrix form as

$$
\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}. \tag{2.7}
$$

Gilmour et al. (1995) rewrote the mixed model equations (2.7) as

$$
\boldsymbol{C}\boldsymbol{\psi} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y}, \tag{2.8}
$$

where $\boldsymbol{W} = [\boldsymbol{X}\ \boldsymbol{Z}]$, $\boldsymbol{\psi} = (\boldsymbol{\beta}', \boldsymbol{u}')'$ and

$$
\boldsymbol{C} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^{*+}
$$

with

$$\boldsymbol{G}^* = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G} \end{bmatrix} \quad \text{and} \quad \boldsymbol{G}^{*+} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}^{-1} \end{bmatrix},$$

where the superscript '+' denotes the Moore-Penrose inverse.

For the model (2.1) we have $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X\beta}$ and $\mathrm{var}(\boldsymbol{y}) = \sigma^2\boldsymbol{H}$. Assuming $\boldsymbol{H}$ is known, the fixed effects parameters $\boldsymbol{\beta}$ can be estimated by GLS to obtain

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}, \tag{2.9}$$

which is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$. If $\boldsymbol{X}$ is not full rank, then any generalized inverse $(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-}$ is used instead of $(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}$ to obtain a solution for $\boldsymbol{\beta}$. The resulting solution for $\boldsymbol{\beta}$ is not unique and is no longer unbiased. However, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is unique and unbiased for $\boldsymbol{X\beta}$.

The computational challenge of using GLS to estimate $\boldsymbol{\beta}$ is that it requires the inverse of $\boldsymbol{H}$ which is an $n \times n$ matrix. In contrast the joint estimators for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ can be obtained by solving either (2.7) or (2.8), i.e.

$$\tilde{\boldsymbol{\psi}} = \boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y}, \tag{2.10}$$

where $\tilde{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}', \tilde{\boldsymbol{u}}')'$ and $\boldsymbol{C}^{-1}$ is given in Lemma A.4 of Appendix A. It must be noted that (2.10) requires simply the inversion of $\boldsymbol{C}$, a $(p+q) \times (p+q)$ matrix, which is easier than finding the inverse of $\boldsymbol{H}$. We also note that although $\boldsymbol{R}^{-1}$ in (2.10) is also an $n \times n$ matrix, it usually has a structure that can be exploited (for example independence between subjects) which makes its computation easier.

Lemma 2.1 *The solutions for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ from solving the MMEs, for $\boldsymbol{G}$ and $\boldsymbol{R}$ known, are given by*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y} \tag{2.11}$$

$$\tilde{\boldsymbol{u}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{2.12}$$

with corresponding variance matrices

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}) &= \sigma^2[(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}] \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} var(\tilde{\boldsymbol{u}}) &= \sigma^2\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{H}\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\ &= \sigma^2\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}, \end{aligned} \tag{2.14}$$

respectively, where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$.

We also have that

$$var(\tilde{\boldsymbol{u}} - \boldsymbol{u}) = \sigma^2\boldsymbol{G} - var(\tilde{\boldsymbol{u}}), \tag{2.15}$$

which unlike (2.14) *takes into account the variability of* $\boldsymbol{u}$ *and can therefore be useful for constructing confidence intervals for* $\boldsymbol{u}$.

Proof. The proof of the lemma follows from the MMEs (2.7) and the matrix results given in Appendix A (Lemma A.4). ■

The predictor $\tilde{\boldsymbol{u}}$ is known as the best linear unbiased predictor (BLUP). It can also be viewed as the estimator of the conditional mean of $\boldsymbol{u}$ given $\boldsymbol{y}$. Applying Result A.7 directly to (2.6) gives

$$\boldsymbol{u}|\boldsymbol{y} \sim N\left[\boldsymbol{0} + \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \sigma^2(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G})\right].$$

Thus

$$\mathrm{E}(\boldsymbol{u}|\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

and

$$\mathrm{var}(\boldsymbol{u}|\boldsymbol{y}) = \sigma^2[\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G}],$$

which can be rewritten as

$$\begin{aligned} \mathrm{var}(\boldsymbol{u}|\boldsymbol{y}) &= \sigma^2[\boldsymbol{G} - (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}] \\ &= \sigma^2[\boldsymbol{G} - (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} - \boldsymbol{G}^{-1})\boldsymbol{G}] \end{aligned}$$

$$= \sigma^2(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}.$$

Though $\boldsymbol{u}$ is unobserved, (2.15) implies the reduced variation associated with the recovery of some information about $\boldsymbol{u}$ in $\tilde{\boldsymbol{u}}$.

The estimator $\tilde{\boldsymbol{u}}$ is also referred to as the Empirical Bayes estimator for $\boldsymbol{u}$. This label is justified by recognizing the random effects $\boldsymbol{u}$ as random variables and therefore the likelihood function $l(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u})$ corresponds to a complete density function, so that $p(\boldsymbol{u})$ is interpretable as the prior distribution of $\boldsymbol{u}$ and hence under the Gaussian assumptions of Result A.7 the posterior distribution of $\boldsymbol{u}|\boldsymbol{y}$ is Gaussian with mean $\tilde{\boldsymbol{u}}$ and variance $\sigma^2(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}$ (McCulloch and Searle, 2001).

The expressions in Lemma 2.1 assume that the variance parameters are known, but if estimates of the variance parameters $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and σ^2 are not known, $\boldsymbol{G}$, $\boldsymbol{R}$ and σ^2 can be replaced by the estimates $\hat{\boldsymbol{G}}$, $\hat{\boldsymbol{R}}$ and $\hat{\sigma}^2$ to obtain the estimates of the fixed effects and random effects and their standard errors using the expressions in Lemma 2.1. However, such standard errors of the fixed effects and of the random effects do not take into account the variability introduced by estimating $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and σ^2, and so underestimate the variability of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$. In the following section we discuss methods for estimating the variance parameters $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and σ^2.

2.3 Variance parameter estimation

Several methods for variance parameter estimation in linear mixed models are discussed in Searle et al. (1992, Ch. 5 and 11). These methods include the ANOVA method for balanced data which uses the expected mean squares approach. However, this method is difficult to apply when the data are unbalanced or when we wish to model the variation in the data using a more complex variance structure. Searle (1971) and Searle (1995) give a general discussion of the problems associated with estimating variance parameters using ANOVA methods in unbalanced data.

For unbalanced data, Rao (1971) proposed the minimum norm quadratic estimation (MINQUE) method for estimating the variance parameters, so-named because it produces quadratic unbiased estimators which have the minimum norm (MINQUE) property, i.e. the resulting estimates are translation invariant under unbiased quadratic forms of the observations. Earlier Henderson (1953) had proposed three methods for estimating variance parameters known as Henderson's methods I, II and III. Method I uses quadratic forms which are analogous to the sums of squares of generally balanced designs; Method II is an adaptation of Method I and takes account of the fixed effects in the model; Method III (also called fitting constants (FITCON) method (see Searle, 1971)) uses sums of squares from fitting the full mixed models as though all terms were fixed effects. A detailed account of Henderson's methods is given by Searle et al. (1992, § 5.3). Lee and Nelder (1998) give another way of estimating variance using extended quasi-likelihood i.e. using gamma-log generalized linear models.

Maximum likelihood (ML) and Residual Maximum Likelihood (REML), also known as restricted maximum likelihood, are now standard methods for estimating variance parameters for both balanced and unbalanced data. The main attraction of these methods is that they can handle a much wider class of variance models than simple variance components. ML estimators of the variance parameters (ratios) are biased downwards, especially in small samples, because they do not take into account the degrees of freedom lost in the estimation of the fixed effects (Lin and McAllister, 1984; Swallow and Monahan, 1984). Hence REML estimation of the variance parameters (or ratios) is preferable to ML estimation. It is for this reason that we adopt REML for variance parameter estimation in this thesis. Searle et al. (1992, § 6.8) discuss the advantages and disadvantages of ML and REML estimators for variance parameters (or ratios). ML estimation of the variance parameters (ratios) have been discussed by several researchers (e.g. Hartley and Rao, 1967; Jennrich and Sampson, 1976; Lindstrom and Bates, 1988; Searle et al., 1992; Verbeke and Molenberghs, 2000). Below, we describe REML estimation for variance parameters (or ratios) in linear mixed models.

Residual maximum likelihood

The downward biasedness of ML estimators of the variance parameters (or ratios), hidden in $\boldsymbol{H}$, can be overcome by using residual maximum likelihood (REML) estimation (Anderson and Bancroft, 1952; Patterson and Thompson, 1971). REML maximizes the likelihood of those linearly independent error contrasts, i.e. independent contrasts of linear combinations of the data $\boldsymbol{y}$, orthogonal to the design matrix $\boldsymbol{X}$. The linear combinations are chosen as $\boldsymbol{K}'\boldsymbol{y}$ so that $\boldsymbol{K}'\boldsymbol{y}$ is of maximal rank but is free of the fixed effects $\boldsymbol{\beta}$. These linear combinations are the residuals obtained after fitting the fixed effects hence the name residual maximum likelihood. Therefore

$\mathrm{E}(\boldsymbol{K}'\boldsymbol{y}) = 0$ which is true if and only if $\boldsymbol{K}'\boldsymbol{X} = 0$. This device results in performing maximum likelihood on $\boldsymbol{K}'\boldsymbol{y}$ instead of $\boldsymbol{y}$. Verbeke and Molenberghs (2000, § 5.3.1, pp. 43) illustrate the use of REML to obtain the estimate of σ^2 for a single Gaussian distributed random sample of size n and show that this estimate is restricted to $n-1$ error contrasts instead of the n contrasts used to obtain the MLE of σ^2 hence the name restricted maximum likelihood. In the context of the linear mixed model the MLE estimate of σ^2 is RSS/n, where RSS denotes the residual sums of squares, while the REML estimate is RSS/$(n-p)$ (also see equation (2.21) below). From a Bayesian view point, Harville (1974) showed that using only error contrasts to make inferences on the variance parameters is equivalent to ignoring any prior information on the fixed effects parameters. Verbyla (1990) shows that REML log-likelihood may also be regarded as a marginal likelihood, while Barndoff-Nielsen (1983) takes it as a modified profile log-likelihood. Lee et al. (2006) view the REML log-likelihood function as a conditional likelihood by assuming asymptotic (multivariate) Gaussian distribution for the fixed effects estimates given fixed variance parameter values. The REML log-likelihood also coincides with the conditional profile likelihood of Cox and Reid (1987).

For $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H})$ and $\boldsymbol{K}'\boldsymbol{X} = 0$ we have

$$\boldsymbol{K}'\boldsymbol{y} \sim N(0, \sigma^2\boldsymbol{K}'\boldsymbol{H}\boldsymbol{K}) \tag{2.16}$$

and the residual (REML) log-likelihood function is

$$\begin{aligned} l_R(\boldsymbol{\phi}; \boldsymbol{K}'\boldsymbol{y}) = -\frac{1}{2}\Big\{ &(n-p)\log(2\pi) + (n-p)\log\sigma^2 + \log\left|\boldsymbol{K}'\boldsymbol{H}^{-1}\boldsymbol{K}\right| \\ &+ \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{K}(\boldsymbol{K}'\boldsymbol{H}^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{y}, \Big\} \end{aligned} \tag{2.17}$$

where $\boldsymbol{\phi} = (\boldsymbol{\kappa}', \sigma^2)'$, $\boldsymbol{\kappa} = (\boldsymbol{\gamma}\prime, \boldsymbol{\rho}')'$. Patterson and Thompson (1971) derived the probability distribution of $\boldsymbol{K}'\boldsymbol{y}$ by carefully choosing $\boldsymbol{K}'$ as an $n-p) \times n$ matrix whose rows are $n-p$ linearly independent rows of $\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Since $\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is symmetric, idempotent and has rank $n-p$, it can be expressed as $\boldsymbol{K}\boldsymbol{K}'$ such that $\boldsymbol{K}'\boldsymbol{K} = \boldsymbol{I}$. Patterson and Thompson (1971) argued that since $\mathrm{E}(\boldsymbol{K}'\boldsymbol{y}) = 0$, $\boldsymbol{K}'\boldsymbol{y}$ lies in the error space, and hence contains no information about the fixed effects ($\boldsymbol{\beta}$), but it does contain information about the variance parameters. Then the REML log-likelihood function (ignoring constants) for the model is

$$l_{\mathrm{R}}(\boldsymbol{\phi};\boldsymbol{y}) = -\frac{1}{2}\Bigg\{(n-p)\log\sigma^2 + \log|\boldsymbol{H}| + \log|\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}| + \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\sigma^2}\Bigg\} \tag{2.18}$$

$$= -\frac{1}{2}\Bigg\{(n-p)\log\sigma^2 + \log|\boldsymbol{H}| + \log|\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}| + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^2}\Bigg\} \tag{2.19}$$

where $\hat{\boldsymbol{\beta}}$, the GLS estimate of $\boldsymbol{\beta}$, and $\boldsymbol{P}$ are given in Lemma 2.1. Khatri (1966) and Searle et al. (1992, pp. 15-18) showed that if $\boldsymbol{K}'\boldsymbol{X} = \boldsymbol{0}$, where $\boldsymbol{K}'$ has maximum row rank, and $\boldsymbol{H}$ is positive definite then

$$\boldsymbol{K}(\boldsymbol{K}'\boldsymbol{H}^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}' = \boldsymbol{P}$$

so that (2.17) and (2.19) are equivalent.

The equivalence between (2.18) and (2.19) is based on the relation

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}\\ &= \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\right)\boldsymbol{y}\\ &= \boldsymbol{H}\boldsymbol{P}\boldsymbol{y},\end{aligned}$$

and hence by Lemma A.2 of Appendix A

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \boldsymbol{y}'\boldsymbol{P}\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{H}\boldsymbol{P}\boldsymbol{y}\\ &= \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}.\end{aligned}$$

Differentiating the REML log-likelihood function (2.19) with respect to σ^2 and κ_j, $j = 1,\ldots,r+s$ gives (Gilmour et al., 1995)

$$\frac{\partial l_{\mathrm{R}}(\boldsymbol{\phi};\boldsymbol{y})}{\partial\sigma^2} = -\frac{n-p}{2\sigma^2} + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^4} \tag{2.20a}$$

$$\frac{\partial l_{\mathrm{R}}(\boldsymbol{\phi};\boldsymbol{y})}{\partial\kappa_j} = -\frac{1}{2}\Bigg\{\mathrm{tr}\Big(\boldsymbol{P}\dot{\boldsymbol{H}}_j\Big) - \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}\Bigg\}. \tag{2.20b}$$

Setting the equation (2.20a) equal to zero and solving gives a REML estimator for the error variance as

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}'\hat{\boldsymbol{P}}\boldsymbol{y}}{n-p}, \tag{2.21}$$

which must be computed iteratively since it depends on $\hat{\boldsymbol{\kappa}}$ through $\hat{\boldsymbol{P}}$. The REML estimate for $\boldsymbol{\kappa}$ must also be found iteratively (see Johnson and Thompson, 1995; Gilmour et al., 1995). Searle et al. (1992, § 6.6, pp. 251-254) give an iterative scheme for obtaining the REML estimates based on the variance parameters rather the variance ratios.

Result 2.1 *The REML log-likelihood function* (2.19) *can be rewritten as (Gilmour et al., 1995)*

$$l_R(\boldsymbol{\phi};\boldsymbol{y}) = -\frac{1}{2}\left\{(n-p)\log\sigma^2 + \log|\boldsymbol{C}| + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^2}\right\} \tag{2.22}$$

where $\boldsymbol{C}$ *is coefficient matrix in the MMEs* (2.7).

Proof. The proof uses matrix results given in Appendix A, and is not shown here. It can also be shown that the log-likelihood functions (2.19) and (2.22) are equivalent. ■

2.3.1 Iterative schemes

Below we describe three related iterative procedures that are used for the calculation of ML or REML estimates of the variance parameters (or ratios), namely: Newton-Raphson (NR), Fisher Scoring (FS) and the Average Information (AI) algorithms. The FS and AI algorithms are variations of the NR algorithm. Some variants of these algorithms have been explored by several authors for estimation of variance parameters in linear mixed models, for example Hemmerle and Hartley (1973), Corbeil and Searle (1976a), Jennrich and Schluchter (1986), Lindstrom and Bates (1988) and Callanan and Harville (1991).

The Newton-Raphson (NR) algorithm (Thisted, 1988, § 4.2.2) uses the first-order expansion of the score function around the current estimate $\boldsymbol{\phi}_{(m)}$ to produce the next estimate $\boldsymbol{\phi}_{(m+1)}$. This algorithm assumes concavity of log-likelihood function to get the quadratic approximation to the function. Each NR iteration requires the calculation of the score function and its derivative. Briefly, the NR procedure can be described as follows. Consider the log-likelihood function $l(\boldsymbol{\phi})$ for which we want to find the maximum at $\boldsymbol{\phi}$ with

$$\frac{\partial l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \mathbf{0}.$$

By first-order expansion we have the vector equation

$$\frac{\partial l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = U(\boldsymbol{\phi}) \approx U(\boldsymbol{\phi}_{(0)}) + \frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}(\boldsymbol{\phi} - \boldsymbol{\phi}_{(0)}). \tag{2.23}$$

Equating (2.23) to zero, and solving we have

$$U(\boldsymbol{\phi}_{(0)}) + \frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}(\boldsymbol{\phi} - \boldsymbol{\phi}_{(0)}) = \mathbf{0},$$

which gives

$$\boldsymbol{\phi} = \boldsymbol{\phi}_{(0)} - \left[\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right]^{-1} U(\boldsymbol{\phi}_{(0)}).$$

This equation can be used iteratively to refine the estimate of the maximum on the $(m+1)$th iteration:

$$\begin{aligned} \boldsymbol{\phi}_{(m+1)} &= \boldsymbol{\phi}_{(m)} - \left[\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right]^{-1} U(\boldsymbol{\phi}_{(m)}) \\ &= \boldsymbol{\phi}_{(m)} + \left[\boldsymbol{\mathcal{I}}_{\mathcal{O}(m)}\right]^{-1} U(\boldsymbol{\phi}_{(m)}), \end{aligned}$$

starting from a pre-specified initial value $\boldsymbol{\phi}_{(0)}$. $\boldsymbol{\mathcal{I}}_{\mathcal{O}(m)}$ is the observed information matrix evaluated at $\boldsymbol{\phi}_{(m)}$.

Fisher Scoring algorithm

The Fisher Scoring (FS) algorithm replaces the observed information matrix by the expected information matrix, $\mathrm{E}\left[-\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right]$, in the NR algorithm.

Average Information algorithm

More recently, Gilmour et al. (1995) and Johnson and Thompson (1995) introduced the Average Information (AI) algorithm for the estimation of variance parameters in a linear mixed model. The AI algorithm can be regarded as a modified Fisher Scoring algorithm since it replaces the expected information matrix in the FS algorithm with an average of the observed and expected information matrices called the average information matrix. This information matrix avoids the evaluation of trace terms in the observed and expected information matrix by approximating the trace terms by sums of squares with correct expected values , i.e. the use of the average information matrix is motivated by computational efficiency because the sums of squares terms are easier to calculate than the trace terms. Similar to the NR and FS algorithms, the AI algorithm is based on finding an efficient solution of the mixed model equations. At each iteration the current values for $\boldsymbol{\phi}$ are used to solve mixed model equations (2.8). Gilmour et al. (1995) describes how this solution is achieved using sparse matrix methods and an absorption and backsubstitution procedure which maximizes computational efficiency by avoiding calculation of unnecessary terms in $\boldsymbol{C}$ (and $\boldsymbol{C}^{-1}$) which come from the absorption process.

In the following we present the score functions for the elements of $\boldsymbol{\phi}$ as well as the observed, expected, and (approximate) average information matrices for $\boldsymbol{\phi}$, with the respective proofs given in Appendix A. These results were originally given by Johnson and Thompson (1995) and later by Gilmour et al. (1995). These score statistics and information matrices are required for the implementation of the iterative schemes described above and also to estimate the variance-covariance matrix of the variance parameters. The score functions and information matrices for the variance parameters will play an important role in the likelihood and score test statistics we develop in Chapters 5-7 of this thesis. Traditionally, the observed information and expected information matrices are used to obtain the variance-covariance matrix of parameters of a model. Efron and Hinkley (1978) give a comparison of the two methods when the observations are independent and identically distributed. They showed that in these situations the observed information is better than the expected information (also see Pawitan, 2001, pp. 245-250). We also explore the use of the exact average information matrix which is an evenly-weighted average of the observed and expected information matrices, i.e. the exact information matrix is constructed as a simple average of the observed and expected information matrix elements which involves evaluation of trace terms in the observed and expected information matrices. Hence our exact average information matrix differs from the average information matrix of Gilmour et al. (1995). We expect the exact average information matrix to give similar variance estimates to the approximate average information matrix as an indication of whether the approximate average information matrix adequately approximates the average of the observed and expected information matrices, i.e. whether the approximate average matrix approximates the trace terms in the observed and expected matrices adequately.

Result 2.2 *The score function for* κ_j *is given by*

$$U(\kappa_j) = \frac{\partial l_R(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \kappa_j} = -\frac{1}{2}\left\{\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) - \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}\right\},$$

where $l_R(\boldsymbol{\phi}; \boldsymbol{y})$ *is the REML log-likelihood function* (2.19), $\boldsymbol{\phi} = (\boldsymbol{\kappa}', \sigma^2)'$ *and* $\dot{\boldsymbol{H}}_j = \partial \boldsymbol{H}/\partial \kappa_j$; *for* $j = 1, \ldots, r+s$, *where* $r+s$ *is the number of variance parameters in* $\boldsymbol{\kappa}$. *The number of variance parameters in the model including* σ^2, *i.e. the number of parameters in* $\boldsymbol{\phi}$, *is* $t = r+s+1$.

Result 2.3 *The score function for* σ^2 *is given by*

$$U(\sigma^2) = \frac{\partial l_R(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \sigma^2} = -\frac{1}{2}\left\{\frac{(n-p)}{\sigma^2} - \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^4}\right\}.$$

Result 2.4 *The elements of the observed information matrix for the variance parameters,* κ_j *and* σ^2 *are*

$$\begin{aligned}
\mathcal{I}_{\mathcal{O}}(\kappa_j, \kappa_k) &= \frac{1}{2}\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) - \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) + \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{y} \\
&\quad - \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P}\boldsymbol{y} \\
\mathcal{I}_{\mathcal{O}}(\sigma^2, \kappa_j) &= \frac{\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{2\sigma^4} \\
\mathcal{I}_{\mathcal{O}}(\sigma^2, \sigma^2) &= -\frac{(n-p)}{2\sigma^4} + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^6}.
\end{aligned}$$

where $\ddot{\boldsymbol{H}}_{jk} = \partial^2 \boldsymbol{H}/\partial \kappa_j \kappa_k$.

Result 2.5 *The elements of the expected information matrix for the variance parameters,* κ_j *and* σ^2 *are*

$$\begin{aligned}
\mathcal{I}_{\mathcal{E}}(\kappa_j, \kappa_k) &= \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) \\
\mathcal{I}_{\mathcal{E}}(\sigma^2, \kappa_j) &= \frac{1}{2\sigma^2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \\
\mathcal{I}_{\mathcal{E}}(\sigma^2, \sigma^2) &= \frac{(n-p)}{2\sigma^4}.
\end{aligned}$$

Result 2.6 *The elements of the approximate average information matrix for the variance parameters, κ_j and σ^2 are*

$$\mathcal{I}_{\mathcal{A}}(\kappa_j, \kappa_k) = \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{y}$$
$$\mathcal{I}_{\mathcal{A}}(\sigma^2, \kappa_j) = \frac{1}{2\sigma^4}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}$$
$$\mathcal{I}_{\mathcal{A}}(\sigma^2, \sigma^2) = \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^6}.$$

Result 2.7 *The elements of the exact average information matrix for variance parameters are obtained by taking equally-weighted averages within the three pairs of terms in Result 2.4 and 2.5 and are given by*

$$\mathcal{I}_{\mathcal{A}e}(\kappa_j, \kappa_k) = \frac{1}{4}\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) + \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{y} - \frac{1}{4\sigma^2}\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P}\boldsymbol{y}$$
$$\mathcal{I}_{\mathcal{A}e}(\sigma^2, \kappa_j) = \frac{\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{4\sigma^4} + \frac{1}{4\sigma^2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right)$$
$$\mathcal{I}_{\mathcal{A}e}(\sigma^2, \sigma^2) = \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^6}.$$

In this thesis, we will refer to the average information matrix of Gilmour et al. (1995) as the approximate average information matrix to reflect the nature of the weighting of the observed and expected information matrix terms. Throughout the thesis we will use the subscripts $\mathcal{O}$, $\mathcal{E}$, $\mathcal{A}$, and $\mathcal{A}_e$ to denote quantities relating to the observed, expected, approximate average and exact average information matrices respectively.

Starting values

Much of the difficulty in estimating variance parameters (or ratios), using the algorithms just described, is centred on obtaining good starting values. Derivative-based algorithms, such as the AI, EM, Fisher Scoring and Newton-Raphson algorithms can be unreliable when estimating variance parameters, especially for models with complex variance structures, unless good starting values are available. Poor starting values may result in divergence of the algorithm or slow convergence. Thisted (1988, § 4.2.5) provides a general discussion of guidelines for starting values and convergence criteria of algorithms based on iterative schemes.

Searle et al. (1992) suggest the use of ordinary least squares estimates for starting values of the fixed effects and ANOVA estimators the variance parameters as starting values. Another method of obtaining starting values of variance parameters is a variant of the MINQUE of Rao (1973), namely MIVQUE0 (Goodnight, 1978; Seely, 1971). Corbeil and Searle (1976b) used MIVQUE0 to obtain starting values for REML estimation of variance parameters. Jennrich and Schluchter (1986) used MIVQUE0 estimates as starting values for the NR and FS algorithms for computing maximum likelihood estimates of the variance parameters. Jennrich and Schluchter (1986) and Laird et al. (1987) give further suggestions for starting values for variance parameters.

Convergence Criteria

The most commonly used criteria of convergence are based on the relative change in either the variance parameter values between successive iterations or score functions or information matrices, or differences between successive log-likelihood functions. For instance, the AI algorithm (in GenStat and ASReml), uses the relative change in the deviance as a check for convergence, whereas the FS method checks for changes in the variance parameter values (in GenStat). For assessing changes in variance parameter values, a measure that involves a multiplier of 0.005 is used. So, for convergence, the change in every variance parameter must be less than 0.005. When assessing change in deviance, convergence is declared when the absolute change in the deviance is less than 10^{-3}. Bates and Watts (1988) argue that these criteria may indicate lack of progress rather than convergence. They suggest a convergence criterion based on the relative Hessian (second derivative of the REML log-likelihood) matrix. Their criterion is defined as

$$U(\boldsymbol{\phi}^{(m)})' \left[\frac{\partial^2 l(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right]^{-1} U(\boldsymbol{\phi}^{(m)})[l(\boldsymbol{\phi}^{(m)})]^{-1},$$

where $l(\boldsymbol{\phi}^{(m)})$ is the REML log-likelihood function at the mth iteration and $U(\boldsymbol{\phi}^{(m)})$ is the score function at the mth iteration. This criterion can be used for all three of the NR, FS and AI algorithms and has the advantage that it can be calculated from the information available at each iteration.

2.4 Statistical inference

This section discusses the theory of inferential procedures used for the estimated parameters in the linear mixed model.

Inference for fixed effects

To compare two nested models with different fixed effects structures, a likelihood ratio test based on REML cannot be used. This difficulty arises because when the variance parameters are estimated using REML, the two models being compared use different error contrasts $\boldsymbol{K}'\boldsymbol{y}$. Hence the corresponding REML log-likelihood functions are no longer comparable since they are based on different observations. Welham and Thompson (1997) proposed an adjusted likelihood ratio test statistic for the comparison of two models with nested fixed effects, fitted using REML. An alternative would be to use the Wald test statistic (Wald, 1943). Consider testing the hypothesis

$$H_0 : \boldsymbol{L}'\boldsymbol{\beta} = \boldsymbol{l} \qquad \text{vs} \qquad H_A : \boldsymbol{L}'\boldsymbol{\beta} \neq \boldsymbol{l},$$

where $\boldsymbol{L}'$ is an $c \times p$ matrix and $\boldsymbol{l}$ is an $c \times 1$ vector. Then the Wald test statistic is given by

$$\begin{aligned} W &= (\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l})'[\text{var}(\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l})]^{-1}(\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l}) \\ &\doteq \frac{(\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l})'[\boldsymbol{L}'(\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{L}]^{-1}(\boldsymbol{L}'\hat{\boldsymbol{\beta}} - \boldsymbol{l})}{\hat{\sigma}^2} \end{aligned} \tag{2.28}$$

where $\hat{\sigma}^2\boldsymbol{L}'(\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{L}$ is the approximate covariance matrix of $\boldsymbol{L}'\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2\hat{\boldsymbol{H}}$ is the REML estimate for $\sigma^2\boldsymbol{H}$. Under H_0, W has an approximate chi-squared distribution with ν degrees of freedom, where $\nu = r_{\boldsymbol{L}}$.

The asymptotic property of the Wald test statistic is based on the assumption that the variance $\sigma^2\boldsymbol{H}$ is known without error, but $\sigma^2\boldsymbol{H}$ is not known and is estimated from the data using REML. This estimation introduces additional variability in the fixed effect estimates. In this way the Wald test statistic underestimates the variability in $\boldsymbol{L}'\hat{\boldsymbol{\beta}}$, so that the test statistic tends to be anti-conservative in small samples, i.e. the test indicates that an effect may be important more often than expected under the null hypothesis of no effect. Lill et al. (1988) reported little effect on the nominal size of the Wald test after replacing the unknown variance parameters by their REML estimates. Kenward and Roger (1997) suggested a scaled Wald statistic which is based on an adjusted covariance estimate, to account for the extra variability introduced by estimating the variance parameters, $\boldsymbol{\phi}$, using REML. This scaled Wald statistic improves the small sample behaviour of the test. They showed that the finite sampling distribution of the scaled Wald statistic was approximately an F distribution with denominator degrees of freedom estimated by a Satterthwaite-approximation method (Satterthwaite, 1946).

Inference for variance parameters

Fixed effect parameters are usually the focus of scientific interest in the linear mixed model. However, it is important to correctly specify the covariance structure to obtain valid statistical inferences for the fixed effects. Altham (1984) noted that overparametrization of the covariance structure may lead to inefficient estimates and poor standard errors for the fixed effects whereas a too restrictive parametrization of the covariance structure renders the inferences about the fixed effect invalid. Verbeke and Molenberghs (2000, Ch. 9) and Wolfinger (1993) give strategies for model building and covariance structure selection in linear mixed models.

Since the REML estimators of the variance parameters are asymptotically Gaussian, we may use approximate Wald tests for testing for their statistical significance. An alternative measure for comparing nested models with different variance parameters but with the same fixed effects is the likelihood ratio test which we describe below.

Lemma 2.2 *The Residual Maximum Likelihood Ratio Test (REMLRT) statistic for comparing two nested models R_0 and R_1 where R_1 includes an extra k variance parameters is given by*

$$REMLRT = -2(l_{R_0} - l_{R_1}), \tag{2.29}$$

where l_{R_i} is the REML log-likelihood function for model i, for $i = 0, 1$.

The REMLRT statistic is asymptotically chi-squared distributed with k degrees of freedom. However, when the null hypothesis is on the boundary of the parameter space, for example testing $H_0 : \sigma_a^2 = 0$ against $H_A : \sigma_a^2 > 0$, where σ_a^2 is the random effects variance, the standard asymptotic theory no longer holds, as regularity conditions are not met. For significance testing the $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution of Self and Liang (1987) has been used. The distribution χ_0^2 represents a distribution with a point mass at 0. Morrell (1998) compared the REMLRT (2.2) with its ML version in terms of type I errors using the $0.5\chi^2_0 + 0.5\chi^2_1$ mixture distribution. He found that the REML test statistic performed better than the ML statistic, i.e. on average, the empirical type I errors were closer to the nominal levels for the REML statistic than for the ML statistic. He did not compare these statistics in terms of type II errors.

In Chapters 5 and 6 we give a detailed discussion of the null distribution of the likelihood ratio test in variance parameter testing, in the context of a variance shift outlier model (VSOM). A VSOM is adopted as an approach to outlier detection in this thesis.

The score test statistic (Cox and Hinkley, 1990, § 9.3) can also be used for testing the significance of variance parameters instead of the likelihood ratio test statistic. The score test only involves the score vector and information matrix under the null hypothesis, i.e. with covariance parameter estimates obtained under the model that is to be tested. Its main advantage over the likelihood ratio test statistic is that it does not require fitting the model specified under the alternate hypothesis; only the null model fit is required to obtain the quantities involved in its calculation.

Lemma 2.3 *The score test statistic for comparing the model* (2.1) *with the model in which some of the specific variance parameters are equal to zero, i.e.* $H_0 : \boldsymbol{\kappa}_0 = \mathbf{0}$ *against* $H_A : \boldsymbol{\kappa}_0 \neq \mathbf{0}$*, where* $\boldsymbol{\kappa}_0$ *is a* $k \times 1$ $(k < r + s)$ *vector of variance parameters of interest, is given by*

$$S(\boldsymbol{\kappa}_0) = \boldsymbol{U}(\boldsymbol{\kappa}_0)' \boldsymbol{\mathcal{I}}^{\boldsymbol{\kappa}_0 \boldsymbol{\kappa}_0} \boldsymbol{U}(\boldsymbol{\kappa}_0)|_{\boldsymbol{\kappa}_0 = \mathbf{0}}, \tag{2.30}$$

where $\boldsymbol{\mathcal{I}}^{\boldsymbol{\kappa}_0 \boldsymbol{\kappa}_0}$ *is the portion of the inverse of the expected information matrix associated with* $\boldsymbol{\kappa}_0$*,* $\boldsymbol{U}(\boldsymbol{\kappa}_0)$ *is the score vector for* $\boldsymbol{\kappa}_0$ *and* $r+s$ *is the number of variance parameters in* $\boldsymbol{G}$ *and* $\boldsymbol{R}$*. Note that all terms in* $S(\boldsymbol{\kappa}_0)$ *are evaluated at* $\boldsymbol{\kappa}_0 = \mathbf{0}$*.*

The score test statistic (2.30) also has an asymptotic chi-squared distribution under the null hypothesis with k degrees of freedom, in line with the likelihood ratio test. It has been used for variance parameter testing in linear mixed models (Jaffrézic et al., 2003; Molenberghs and Verbeke, 2007; Verbeke and Molenberghs, 2003). The score test suffers from the same boundary problem (when the null hypothesis hypothesis is on the boundary of the parameter space) as the likelihood ratio and so the $0.5\chi^2_0 + 0.5\chi^2_1$ mixture distribution is used to assess the significance of the test. In Chapters 5-7 we will formulate score test statistics for variance parameters in a variance shift outlier model and some of its extensions. In these chapters we will also elaborate on the statistical issues which arise in using the score tests for variance parameter testing in a variance shift outlier model when it is formulated as a linear mixed model.

In the next section we describe a specific data set which we use as a typical example to illustrate the methods we propose in the thesis.

2.5 Example: The orthodont data

The data are taken from Potthoff and Roy (1964). The data consist of measurements of the distance in millimetres from the centre of the pituitary to the pterygomaxillary fissure at ages 8, 10, 12 and 14 years on 16 boys and 11 girls. The purpose of the study is to model the relationship between distance and age, with investigation of gender differences.

Figure 2.1 comprises plots of the distances by age for the boys and girls separately. Generally, the change in distance is approximately linear over the range 8-14 years. The data for boys appear more variable than for girls. The response profiles vary considerably between subjects. We use the prefixes "M" and "F" to number the male and female subjects, respectively. Subjects number 9 and 13 among the boys (M9 and M13) seem to have possible outlying observations. Male subject 4 appears to have a reduced slope while male subject 10 seems to have a higher intercept. Subject number 10 among the girls (F10) appears to have a suppressed response profile compared to other females, while subject number 11 among the girls (F11) seems to have an elevated response profile compared to other subjects in the group. Therefore any statistical modelling of these data would need to take account of the subject variation and possibly account for the presence of outliers within subjects and outlying subjects.

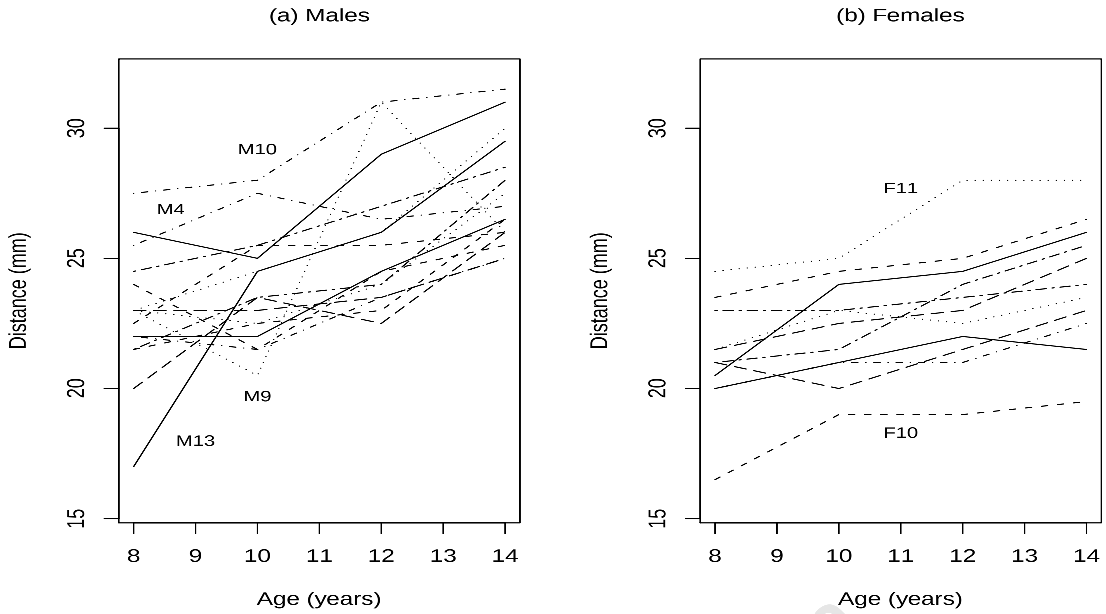

Figure 2.1: *Plots of distance against age for orthodont data.*

Following Pinheiro and Bates (2000) we fit the following linear mixed model to the data

$$\boldsymbol{y}_{jk} = (\mu + \beta_{0k} + u_{0jk})\mathbf{1}_4 + (\beta_1 + \beta_{1k} + u_{1jk})\boldsymbol{x} + \boldsymbol{e}_{jk}, \tag{2.31}$$

where, $\boldsymbol{y}_{jk}$ is the vector of distances for the jth subject of gender k, $j = 1, \ldots, 27$; $k = 1, 0$ with 1 for males and 0 for females, $\boldsymbol{x} = \{x_l - 11 : l = 1, \ldots, 4\}$, x_l is the age at measurement l, μ is the overall mean, β_{0k} is the intercept shift for gender k, β_1 is the overall slope, β_{1k} is the slope for gender k, u_{0jk} is the random additive effect of the jth subject of gender k and u_{1jk} is the random slope effect of the jth subject of gender k, and finally $\boldsymbol{e}_{jk}$ is the random error vector for subject j of gender k. The centering of the explanatory variable for age reduces the correlation between the slope and intercept. The random effects vector for the jth subject $\boldsymbol{u}'_{jk} = (u_{0jk}, u_{1jk})'$ is assumed to be Gaussian distributed with mean zero and variance matrix given by

$$\boldsymbol{G}_{\text{subject}} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}$$

where γ_{11} and γ_{22} are the variance ratios for the random intercepts and random slopes respectively, γ_{21} is the correlation between the variance ratios. and the corresponding error vector $\boldsymbol{e}_{jk}$ is assumed to have a Gaussian distribution with mean zero and variance matrix $\sigma^2 \boldsymbol{I}_4$, independently of $\boldsymbol{u}_{jk}$. The matrix $\boldsymbol{G}_{\text{subject}}$ specifies the subject variance structure and the identity matrix specifies random error structure. Then the matrices $\boldsymbol{G}$, $\boldsymbol{R}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ matrices, defined earlier, are given by

$$\boldsymbol{G} = \boldsymbol{I}_{27} \otimes \boldsymbol{G}_{\text{subject}},$$

$$\boldsymbol{R} = \boldsymbol{I}_{27} \otimes \sigma^2 \boldsymbol{I}_4 = \sigma^2 \boldsymbol{I}_{108},$$

$$\boldsymbol{X} = \mathbf{1}_{27} \otimes [\mathbf{1}_4 : \boldsymbol{x}]$$

and

$$\boldsymbol{Z} = \boldsymbol{I}_{27} \otimes [\mathbf{1}_4 : \boldsymbol{x}].$$

where $\otimes$ is the Kronecker product of rectangular matrices. The variance-covariance matrix for the data is of the form

$$\text{var}(\boldsymbol{y}) = \sigma^2(\boldsymbol{ZGZ}' + \boldsymbol{I}_{108}).$$

In the following we present results from the fitted model. To index the observations we use the notation $j.l$ to label the lth observation within the jth subject, $j = 1, \ldots, 27$: $j = 1, \ldots, 16$ for boys and $j = 17, \ldots, 27$ for girls. Figure 2.2 shows scatter plots of distance against age with fitted values superimposed for each subject. The scatter plots for boys are labelled as $M01, \ldots, M16$ (the first 64 observations) and the plots for girls are labelled as $F01, \ldots, F10$. Two points for male 9 and 2 observations for male 13 are not well fitted by the model (Figure 2.2). These four observations have large residuals relative to other observations in the data (Figure 2.3). Figure 2.3 suggests that the variability in the measurements may be greater among the boys than among the girls. We will investigate this feature of the data later in Chapter 7. Figure 2.4 is a scatter plot of the estimated random intercepts against the estimated random slopes and suggests that female 10 has the smallest random intercepts and male 13 has a large slope and may be quite different from other subjects.

Examining the results from the analysis reveals that there are suspect individual observations within subjects. There are also possibly outlying subject profiles indicated by the unusually large random intercept for F10 and the unusually large random slope for M13. If these observations or subjects have to be included in the analysis they have to be down-weighted. We may view the individually outlying observations as having inflated variances compared to the rest of the data points. Correspondingly we may consider the larger random intercept and slope as having increased variance. It is then necessary to quantify this increased variance and use that quantity in the down-weighting if the variance shift can be determined to be large enough. In this thesis we will adopt an approach which allows us to estimate the increased variance, develop statistical tests for it and use the estimate in the down-weighting of observations or groups of observations.

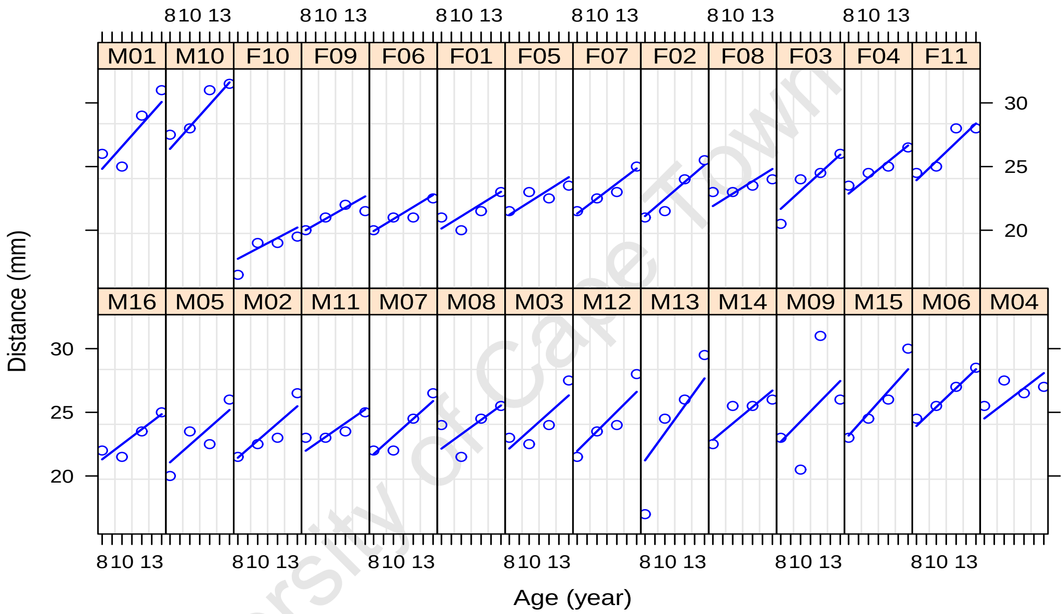

Figure 2.2: *Scatter plots of distance against age for each subject with fitted lines superimposed for orthodont data.*

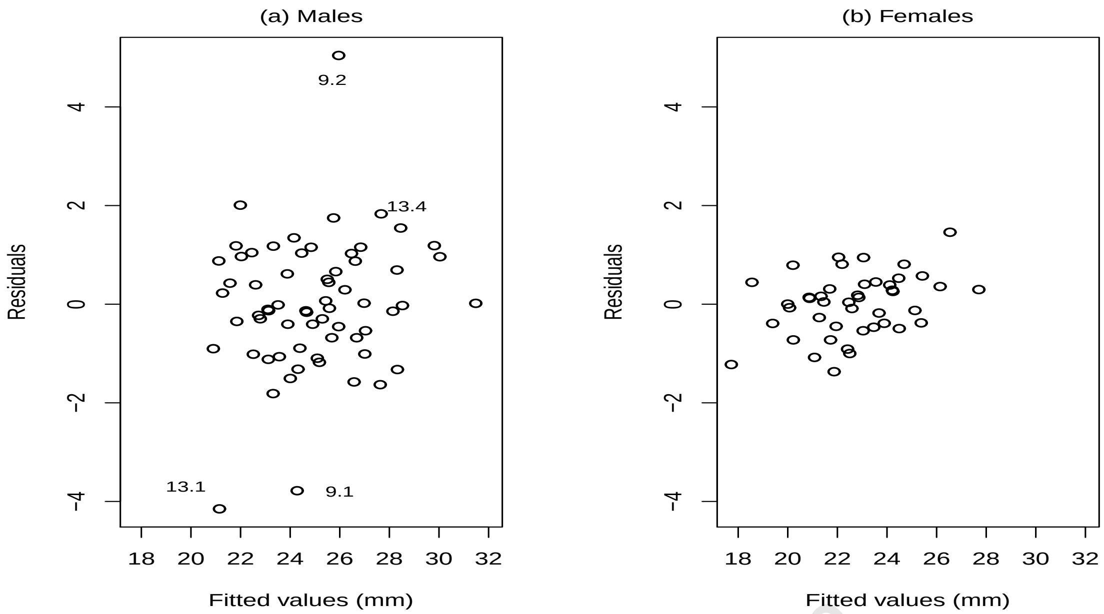

Figure 2.3: *Scatter plots of residuals against fitted values by gender for orthodont data.*

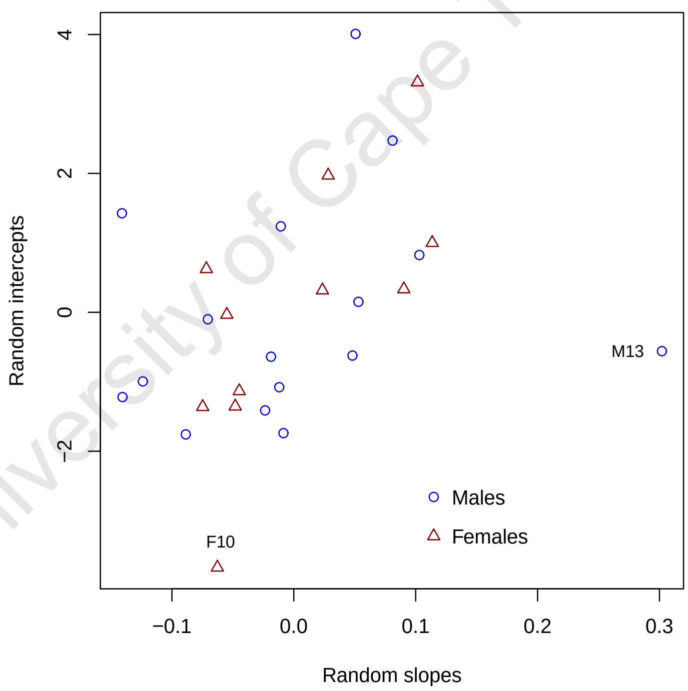

Figure 2.4: *Scatter plot of random intercepts against random slopes for orthodont data.*

2.6 Summary

In summary, we have reviewed parameter estimation and inference for the linear mixed model, for the variance parameters in particular. We will adopt REML estimation using the average information algorithm of Gilmour et al. (1995) in rest of the thesis since it gives unbiased variance parameter estimates. Likelihood ratio tests and score test statistics are usually used for testing for additional variance parameters in the model. The likelihood ratio test statistic (2.29) and the score test statistic (2.30) will play prominent roles in the methods we will develop in Chapters 5 to 7. In these chapters we will also examine the dependence on the (one-step) likelihood ratio tests and score test statistics on the four different information matrices: expected information, observed information, approximate average information and exact average information matrices. We also introduced a dataset which will be the central example of this thesis and have shown how the features of the data set stimulated our research.

CHAPTER 3

Review of linear mixed model diagnostics

3.1 Introduction

The purpose of this chapter is to review the literature on research into diagnostics for the linear mixed model and to critique the diagnostic techniques in current use. This review is also aimed at identifying gaps in the research into outliers in the linear mixed model.

Diagnostics in linear regression

Case-deletion

Outliers in linear fixed effects models have been widely researched. Cook (1977) developed case-deletion diagnostics for efficient identification of influential observations in linear regression analysis, and Cook and Weisberg (1982) and Belsley et al. (1980) give detailed discussion of the theory and application of these methods. There has also been widespread adoption of these ideas within other contexts, for example in survival analysis (Hall et al., 1982; Storer and Crowley, 1985) and in logistic regression analysis (Pregibon, 1981; Hosmer and Lemeshow, 1989).

Case-deletion diagnostics may suffer from masking and swamping effects. Masking occurs when an influential observation goes undetected due to the presence of another observation (Atkinson, 1985; Chatterjee and Hadi, 1988). Swamping occurs when an observation is incorrectly identified as influential due to the presence of another observation, or a subset of observations (Barnett and Lewis, 1995). Rousseeuw and Van Zomeren (1990) suggest using some robustified Mahalanobis distance for identifying multivariate outliers to deal with masking and swamping. Hoeting et al. (1996) discuss a Bayesian method for multiple outlier identification in linear regression which also detects masked outliers. Lawrence (1995) proposes joint influence and conditional influence measures for assessing masking effects in linear regression.

Local influence

Another approach to detecting influential observations is the local influence approach (Cook, 1986) in which the influence of observation(s) on parameter estimates is studied via minor perturbations of the model or of the data. In this approach the displacement of the log-likelihood function is used to evaluate the local influence of observation(s). The basic idea behind this approach may be described as follows. Let $l(\boldsymbol{\theta})$ denote the log-likelihood function of the observed data, where $\boldsymbol{\theta}$ is a $p \times 1$ vector of unknown model parameters (say the fixed effects, random effects and variance parameters in model (2.1)). Consider an $r-$dimensional perturbation vector $\boldsymbol{w} = (w_1, ..., w_r)^{'}$ varying in the open region $\Omega \subset R^r$. Let $l(\boldsymbol{\theta}|\boldsymbol{w})$ denote the log-likelihood function corresponding to the perturbed model for a given w in Ω. It is assumed that there exists a point w_0 in Ω such that $l(\boldsymbol{\theta}|w_0) = l(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. Finally, let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{w}}$ denote the maximum likelihood estimators under $l(\boldsymbol{\theta})$ and $l(\boldsymbol{\theta}|\boldsymbol{w})$, respectively. The local influence approach proceeds by comparing $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{w}}$ and interprets small differences in the estimates as indicating that the perturbations(s) have little effect on the parameter estimates. A measure of the distance between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{w}}$ is the likelihood displacement (Cook, 1986), defined by

$$\text{LD}(\boldsymbol{w}) = 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{\boldsymbol{w}})],$$

which takes into account the variability in $\hat{\boldsymbol{\theta}}$. The graph of $LD(\boldsymbol{w})$ versus $\boldsymbol{w}$ throughout Ω exhibits some essential information on the influence of the perturbation scheme. Large values of $\text{LD}(\boldsymbol{w})$ indicate the influential effect of the perturbation $\boldsymbol{w}$ on the estimation of $\boldsymbol{\theta}$.

Robust estimation and modelling

The preceding methods are concerned with detection of outliers and influential points. Youden (1975) presents a view towards robustness and comments as follows regarding the deletion of observations in interlaboratory studies: "If ... the results come from different laboratories, it hardly makes sense to discard a fair proportion of the population of laboratories. They are the only laboratories we have and, anyway, we have no power to make them vanish. Our task is that of presenting a realistic picture of the population of laboratories." This view is supported by Cook (1986) who contends that once influential observations and outliers have been identified, it may then be necessary to "accommodate" these observations in the analysis. Hence another action may involve refitting the model using a more robust method, yielding estimates which are robust against specified types of violations of model assumptions and are outlier-resistant.

There are three main classes of robust statistical procedures in linear regression: bounded influence procedures, such as generalized M-estimators discussed by Huber (1981), high breakdown methods, such as least median squares estimators of Rousseeuw (1984) or least trimmed sum of squares (LTS) estimators discussed in Venables and Ripley (1997) and the use of heavy-tailed distributions for the random errors in ML approaches (Lange et al., 1989). Both the bounded influence and high breakdown methods simultaneously identify outliers and estimate the parameters of the model. The use of heavy tail distributions for the random errors can be viewed as robust modelling rather than robust estimation. The difference between robust estimation and robust modelling is that in robust estimation one can assume a non-heavy tail distribution such as the normal distribution for the random errors and use a robust alternative to maximum likelihood to estimate the model parameters, whereas in robust modelling one assumes a flexible distribution for the random errors so that maximum likelihood, used in the estimation, is itself robust (Ruppert, 2004, pp. 217).

Other robust methods

Box and Tiao (1968) (also see Barnett and Lewis, 1995) introduced a model which assumes that the data come from a non-contaminating Gaussian distribution with probability $1 - \pi$, and from a contaminated Gaussian distribution with probability π so that the error term e_i has a Gaussian mixture distribution, $(1 - \pi)N(0, \sigma^2) + (\pi)N(0, k^2\sigma^2)$. This model is used in the robust modelling literature to accommodate outliers rather than to identify them. The usual assumptions are that the proportion π of outliers in the data and the scale parameter k are known and fixed. Aitkin and Wilson (1980) estimated the contaminating fraction π using maximum likelihood and Marks and Rao (1979) presented a similar example with π assumed known.

Cook et al. (1982) consider an alternative approach to case-deletion in which $(n-1)$ observations have the same variance σ^2 and one unknown observation has inflated variance but $\alpha_i\sigma^2$, $\alpha_i \geq 1$. Under Gaussian assumption, maximum likelihood is used to estimate $(\alpha_i, \boldsymbol{\beta}, \sigma^2)'$. They compared this alternative model with the case-deletion (mean shift outlier) model and found that maximum likelihood estimation does not result in the observation with the largest standardized residual being deemed the outlier. However, when the largest absolute standardized residual corresponds to the largest absolute residual, the position of the outlier will be same under both models. Cook and Weisberg (1982) discussed the same model as Cook et al. and called it the 'variance shift model' (Cook and Weisberg, 1982, § 2.2.2, pp. 82). Thompson (1985) also considered the same model as Cook et al. but used REML for estimating the model parameters and noted that 'the residual variance and outlier position are the same under both models'. The thesis adopts the same model but we call it a variance shift outlier model (VSOM) and parameterize the variance of the ith observation as $\sigma^2(\omega_i + 1)$, $\omega_i \geq 0$, with ω_i acting as variance shift parameter. We investigate a VSOM as an approach to detecting single-case and groups of outliers in Chapters 5-7. In these chapters we formulate a variance shift outlier model as a linear mixed model, and develop likelihood ratio and score test statistics for testing for the variance shift parameter.

3.2 Diagnostics in the linear mixed model

A great deal of effort in the literature for linear mixed models has been devoted to estimating the parameters of the model, and not on the ways its assumptions can be violated, including ways in which outlying observations may influence the model parameters. Ware (1985) suggested further research into methods for detecting influential and outlying observations in linear mixed models. In their review on small area estimation, Ghosh and Rao (1994) also wrote "...However, the literature on diagnostics for mixed linear models involving random effects is not extensive, unlike standard regression diagnostics. Only recently have some useful diagnostic tools been proposed...". Linear mixed models extend linear models to include more than one source of random variation and hence multiple variance parameters. In linear mixed models the variance parameters are not known but are estimated from the data. Subsequently, the term 'outlier' now encompasses outliers for both the random terms in the model and for the variance parameter estimates.

Nevertheless, we may consider the following estimated residuals for the fitted linear mixed model:

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Z}\tilde{\boldsymbol{u}} + \tilde{\boldsymbol{e}}$$

which comprise both the random and residual error component and

$$\tilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\tilde{\boldsymbol{u}}$$

to estimate residual errors.

In the context of longitudinal data models the residuals $\hat{\boldsymbol{e}}$ represent the deviation of a subject's specific profile from the overall population mean while the residuals $\tilde{\boldsymbol{e}}$ measure the difference between the observed values and the subject's own fitted regression line. The estimated random effect $\tilde{\boldsymbol{u}}$ (BLUP) may be viewed as a measure of how much the subject's specific profile deviates from the population average profile.

A challenge in outlier diagnostics for linear mixed models is that an entire vector of a subject's observations may be an outlier (e.g. an atypical individual) or a subset of the vector may be an outlier (an atypical period within a string of observations of an individual). It is also unclear how leverages should be defined since the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ usually have different dimensions. The issue of definitions of residuals and leverages has not received adequate attention in the literature on linear mixed model diagnostics. We address this topic in detail in Chapter 4.

Naes (1986) proposed an outlier detection method based on residuals from the GLS estimation of the fixed components in the mixed model. He also focused on the type of outliers that may be detected by this approach. Two types of outliers are considered: those observations that do not fit the corresponding fixed model, and those that fit this fixed model but have an 'abnormal' random component value. He focused on outlier detection for the fixed part of the model. Schall and Dunne (1988) discussed the detection of outliers and influential observations in the linear model with a general covariance structure. They derived F-type test statistics and adjusted parameter estimates associated with three different outlier models (distributional, additive-shift and transformational). They demonstrated that testing for their different types of outliers in the linear model with general variance covariance matrix, say $\sigma^2\boldsymbol{H}$ for $\boldsymbol{H}$ known, is equivalent to testing for outliers in a linear model with (i) the residuals from the null model treated as *data*, and (ii) the variance-covariance matrix being the variance-covariance matrix of these residuals. DeGruttola et al. (1987) describe measures of influence and leverage for a generalized three-step least squares estimator for the regression coefficients in a class of multivariate linear models for repeated measurements. However, their method does not apply to maximum likelihood estimation and it is also not clear how to extend their diagnostics to the case of unequal covariance matrices for individual subjects or groups of subjects.

Martin (1992) considered the roles of leverages and residuals in diagnostics in the linear model with correlated errors. Christensen et al. (1992b,a) demonstrated the role of prediction (deletion) residuals in diagnostics in a spatial context. More recently, Shi and Chen (2009) gave a summary of influence measures for general linear models with correlated errors and show that these measures can be written in terms of the generalized leverages and residuals. Haslett (1999) proposed the use of conditional residuals (known as prediction residuals) and marginal residuals for diagnostics in models for multivariate data.

Houseman et al. (2004) proposed the use of Cholesky residuals for checking the multivariate Gaussian assumption in a linear model with correlated responses. Their methodology involves multiplying the marginal residuals by a Cholesky decomposition of the estimated covariance matrix.

3.2.1 Case-deletion

Christensen et al. (1992a) develop case-deletion diagnostics for the linear mixed model when the variance parameters are estimated using REML. They consider deletion diagnostics for both fixed effects and for the variance parameters. For the fixed effects, they extended the Cook's (1977) distance notion to the linear mixed model analysis. They do not study the influence on the random effects. Haslett (1999) introduces a 'delete=replace' identity for linear models with general covariance structure. The focus is on deletion of arbitrary subsets of the data and the development is centred on the conditional residual. This approach is simpler than that of Christensen et al. (1992a). However, the error variance-covariance matrix is not re-estimated with each successive deletion, which is a limitation for correlated data (Haslett, 1999; Dillane, 2006). Haslett and Dillane (2004) present a more general discussion of the 'delete=replace' identity with a focus on deletion diagnostics for the REML variance parameter estimates. They simplify the one-step approximation of Christensen et al. (1992a) for the variance parameters. Their main contribution is an alternative approximation which is directly available as a by-product of the initial full fit of the model. Inference from the diagnostic is in terms of relative change in the variance parameter estimates when individual observations or groups of observations are removed from the data set.

Zewotir and Galpin (2005) also discussed influence diagnostics for the linear mixed model. They assessed influence on the fixed effects, linear functions of the fixed effects, variance parameters, random effects and predictions of the response variable. Zewotir (2008) extends the Zewotir and Galpin results to detection of multiple outliers. Banerjee and Frees (1997) considered subject deletion diagnostics for longitudinal data. They attempted to assess subject deletion results for fixed effect parameters and random effects. In their study all effects were treated as fixed and the ordinary least squares estimates are obtained to derive diagnostics analogous to Cook's distance (Cook, 1977). However, the ordinary least squares estimates obtained by treating random effects as fixed effects are different from the BLUP, which is more appropriate when the random effects are also of scientific interest.

Hurtardo (1993) provided another extension of linear model diagnostics to the linear mixed model. He used a modified Cook's distance to detect influential observations in the estimation of fixed effects and the prediction of the random effects. He also suggested an approach for identifying influential observations on the subject's error variance, say σ_a^2. The approach assumes that these observations do not belong to the population and follow the distribution $N(0, \sigma_a^2 \boldsymbol{I}_q)$. He proposed an F-type test statistic for testing the influential effect of the block/subject on σ_a^2. The proposed diagnostics were based on maximum likelihood estimation of the variance parameters and were evaluated on balanced data.

Tan et al. (2001) discuss diagnostics for detecting both influential subjects and observations. Haslett (1999) cautions that it is futile to compute diagnostics for all possible subsets and only natural subsets defined by context are useful. The motivation for deletion diagnostics which detect influential subjects is the fact that the subject is usually the experimental unit (Banerjee and Frees, 1997). This view is supported by Fung et al. (2002) who argue that observations from the same subject usually share the same values for the covariates and therefore it makes sense to identify influential subjects as opposed to influential observations. However, this focus on experimental subjects may not be sufficient for situations with time-varying covariates. Banerjee (1998) assessed the usefulness of the Cook's distance in fixed-effects models for longitudinal data with serial correlation and concluded that this statistic is not effective in detecting influential subjects since it is highly sensitive to nuisance parameters. Langford and Lewis (1998) discuss the detection of outliers in the context of multilevel models fitted using iterative GLS and, like Gray and Ling (1984), they also suggest cluster analysis for identifying a group of outliers. Longford (2001) proposed a simulation-based approach for flagging out single or multiple outliers in random coefficient models. This approach is essentially a parametric bootstrap which generates a sampling distribution of the largest residual or some chosen outlyingness statistic, for instance the likelihood ratio test statistic. This distribution is then used to obtain p-values for the residuals or test statistics. Gelman et al. (2003) and Marshall and Spiegelhalter (2007) use the posterior predictive checking approach of Rubin (1984) to detect outliers in hierarchical Bayesian models. Bayesian approaches to deletion diagnostics for linear mixed models were discussed by Sharples (1990b) and Hodges (1998).

The case-deletion approach is computationally demanding and does not ascribe influence to any of the subject's characteristics. It also does not provide diagnostics for the fixed effects and variance parameters simultaneously. Zewotir and Galpin (2006) showed that several case-deletion diagnostic measures for linear mixed models are susceptible to masking effects. They proposed a diagnostic strategy which first identifies high-leverage points and outliers separately then conducts influence analysis. Zhu et al. (2001) extend the joint influence and conditional influence measures of Lawrence (1995) for handling masking effects in linear mixed models.

Shi and Chen (2008) discussed the detection of single and multiple outliers in multilevel models. Multilevel models are special cases of linear mixed models, and are used for grouped or clustered data in which the pattern of clustering is known. The focus of their proposed diagnostics is the detection of outliers at different levels (hierarchies) in the multilevel data and not on influence on the model parameters. The proposed test are constructed under the mean-shift outlier model assuming a general covariance structure for the random errors.

3.2.2 Local influence

Beckman and Nachtsheim (1987) used the idea of local influence to develop methods for assessing the effect of perturbations from the usual assumptions in the mixed-models analysis of variance with uncorrelated random components. They investigated how the parameters change under small perturbations of the variance parameters and the response vector. Lesaffre and Verbeke (1998) (also see Verbeke and Molenberghs, 2000) proposed a case-weight perturbation scheme where they investigated how much the parameter estimates are affected by changes in the weights of the log-likelihood contributions of specific subjects in a repeated-measures setting. Here the perturbed log-likelihood is $\sum_{i=1}^{n} w_i l_i(\boldsymbol{\theta})$ where $l_i(\boldsymbol{\theta})$ is the contribution of the ith individual to the log-likelihood and w_i is the weight for the ith subject. This approach is the same as that of Beckman and Nachtsheim (1987) except that Lesaffre and Verbeke (1998) perturbed the log-likelihood whereas Beckman and Nachtsheim (1987) perturbed the parameter space. Verbeke and Molenberghs (2000, Chapter 11) conclude that case-weight perturbation approach is not useful in assessing the effect of an observation for global influence analysis. Case-deletion can be considered to be the limiting case of local influence. DeGruttola et al. (1987) found that in small samples case-deletion can detect influential observations where local influence cannot. Pan and Fang (2002) discuss deletion and local influence diagnostics for growth curve models, while Lee and Xu (2004) studied deletion and local influence diagnostics under the banner of nonlinear mixed models.

Local influence diagnostics are not affected by masking and swamping. However, some challenges still remain in using the local influence approach for assessing influence in a linear mixed model. Firstly, this approach has been used to detect influential subjects (clusters) and not observations within subjects. Secondly, the influence diagnostics are developed for the case where the covariance matrix for the random errors is of the simplest form $\boldsymbol{R} = \boldsymbol{I}$. Thirdly, distributional properties of local influence diagnostics for linear mixed model are not known. Fourthly, the diagnostics are derived for situations in which the random effects are not of scientific interest. It is not clear how the diagnostics can be extended to situations in which the random effects are not nuisance parameters. Furthermore, the diagnostics assume that the unknown variance parameters are estimated via maximum likelihood. There is a need for influence analysis based on REML variance estimates since these estimates are unbiased. Finally, the local influence measures are usually based on the Gaussian curvature of the log-likelihood, which is invariant under a change of scale (Poon and Poon, 1999). As a result there are no benchmarks for judging the largeness of Gaussian curvatures. Schall and Dunne (1992) established a close relationship between the concepts of parameter collinearity and local influence in regression diagnostics, and they introduced the scaled curvature, a modification of Cook's Gaussian curvature. This scaled curvature is invariant with respect to the perturbation scheme and has a upper bound of one. However, the two curvatures are generally not equivalent except in special circumstances. Poon and Poon (1999) proposed the conformal Gaussian curvature which is a one-to-one function of the curvature and assumes values in the interval $[0, 1]$. Since this curvature is invariant over reparametrization of the perturbation scheme, they suggest that it can be used to construct objective benchmarks for judging largeness. Zhu and Lee (2001) make an attempt at using conformal Gaussian curvature for local influence diagnostics in linear mixed models.

Demidenko (2004) introduces yet another approach to detecting influential observations, infinitesimal influence, which studies the influence of small changes in the individual observations or small perturbations of the model. He extended this approach to the linear mixed model. Following Demidenko (2004), Zewotir (2007) investigated changes in parameter estimates under infinitesimal change in the error variance. Their model perturbation approach assumed $e_i \sim N(0, \sigma^2/\omega)$, where e_i is the random error for the ith observation in the linear mixed model and $0 < \omega < \infty$ (ω being the perturbation). This approach was earlier considered by Hurtardo (1993) and is similar to the variance shift model of Cook and Weisberg (1982) except that here ω is a mathematical quantity and not a parameter to be estimated from the data. Hurtardo (1993) proposed an F-type test statistic for measuring the influence of the perturbation on the error variance estimate whereas Zewotir (2007) used both the likelihood ratio test and a modified Cook's distance as measures of influence on the fixed and random effects estimates, and variance ratios, for a given perturbation ω. For the influence on the variance ratios, he used one-step estimates to avoid the computation burden associated with obtaining REML or ML estimates of the variance ratios. He did not discuss the distributional properties of their proposed test statistics. He admits that the methodology does not lead to the conclusion that the ith observation has variance different from the remaining observations since the perturbation ω is not estimated but assumed.

3.2.3 Robust estimation and modelling

Fellner (1986) proposed a procedure for limiting the effect of outliers on the variance parameters estimates in the linear mixed model. He also obtained robust estimates of the variance parameters by modifying the equations for REML estimates as proposed by Huber (1981). He also proposed diagnostic displays for the identification of outliers. Rocke (1983, 1991) investigated robust ANOVA estimates of variance parameters in balanced linear mixed models. Following Fellner (1986), Burns (1992) proposed an algorithm for robust estimation of variance parameters under contaminated Gaussian distributions for the random effects and errors.

Richardson (1997) extends the bounded influence method that constrains the influence of any outlying observations on the parameter estimates in the linear model, to the linear mixed model. In principle the approach would indicate the degree of change in the estimator that can be wrought by an infinitesimal amount of contamination; and therefore the approach gives good stability and good efficiency against infinitesimal contamination. However, although the influence of outlying values of $\boldsymbol{y}$ and $\boldsymbol{X}$ on the parameter estimates is bounded, the influence of outlying values of $\boldsymbol{Z}$ is not, and there is no guideline on how to bound the influence of $\boldsymbol{Z}$. Welsh and Richardson (1997) discussed robust estimation in the linear mixed model by assuming a Student's t-distribution for the random effects (also see Pinheiro et al., 2001).

Gogel (1997) considered the extension of a VSOM to the spatial mixed model. However she only considered parameter estimation in the spatial mixed linear VSOM and did not give an evaluation of a VSOM.

3.3 Random effects distribution

When the random effects are of scientific interest, their distribution needs to be correctly specified in order to obtain reliable estimates. In linear mixed models, the Gaussian distribution is usually assumed for the random effects parameters. Hence diagnosing linear mixed models often includes checking the distributional assumptions for the random effects (Solomon, 1985). Dempster et al. (1984) proposed using half-normal plots as a graphical display for checking for evidence of non-normality of the random effects in linear mixed model with two variance parameters in which the variance parameters were estimated using REML. Lange and Ryan (1989) discuss quantile-quantile plots for checking the Gaussian assumption for the random effects in the linear mixed model. Jiang (2001) suggests a goodness of fit test for checking the distributional assumptions involved in the linear mixed model, including the Gaussian assumptions for the random effects.

Even when the random effects are not of scientific interest, their distribution needs to be correctly specified for reliable inference for the fixed effects. Verbeke and Lesaffre (1997) investigated the effect of miss-specifying the random effects distribution on the model parameters in the linear mixed model. They showed that maximum likelihood estimators of the fixed effects and variance parameters, obtained under the Gaussian assumption for the random effects, are consistent and asymptotically Gaussian distributed even when the random effects distribution is non-Gaussian. However, it was shown that the standard errors of all the parameters need correction in order to get valid inference. This correction involves replacing the 'naive' estimate for the asymptotic covariance matrix by a so-called 'information sandwich' estimate. Let $\boldsymbol{A}(\boldsymbol{\alpha})$ be minus the matrix of second-order derivatives of the log-likelihood function with respect to the elements of $\boldsymbol{\alpha}$, the vector of all parameters in model (2.1), and $\boldsymbol{B}(\boldsymbol{\alpha})$ be the matrix with cross-products of first-order derivatives of the log-likelihood function, also with respect to $\boldsymbol{\alpha}$. The estimates $\widehat{\boldsymbol{A}}(\boldsymbol{\alpha})$ and $\widehat{\boldsymbol{B}}(\boldsymbol{\alpha})$ are obtained by replacing $\boldsymbol{\alpha}$ by its ML or REML estimate. Then the corrected asymptotic covariance matrix (information sandwich) is $\widehat{\boldsymbol{A}}(\boldsymbol{\alpha})^{-1}\widehat{\boldsymbol{B}}(\boldsymbol{\alpha})\widehat{\boldsymbol{A}}(\boldsymbol{\alpha})^{-1}/n$. However, when the random effects are not Gaussian distributed, Verbeke and Lesaffre (1997) demonstrated, using simulation, that for the fixed effects, the 'naive' and sandwich standard errors are very similar. With regard to the standard errors of the variance parameters, they found that although the sandwich approach gives good estimates, they observed that it may yield incorrect confidence intervals for small samples. For large sample size the variance-covariance matrix for the variance parameters i.e. the inverse expected information matrix for the variance parameters yields valid standard errors.

Earlier, Butler and Louis (1992) had also presented simulation evidence that wrongly specifying the random effects distribution of univariate random effects has little effect on the fixed effects estimates or on estimates for the residual variance and variance of random effects. Zhang and Davidian (2001) considered the relaxation of the Gaussian assumption by modelling the random effects non-parametrically. Bayesian approaches to non-parametric estimation of the random effects are discussed by Bush and MacEachern (1996), Kleinman and Ibrahim (1998), and Van der Merwe and Pretorius (2002).

3.4 Covariance structure modelling

In addition to identifying outliers and influential observations, other assumptions of the linear mixed model that may be checked include the assumptions of independence of the random effects and homogeneity of their variances. Khuri (1989) proposed a test procedure for checking the validity of these assumptions. He assessed the consistency of the covariance structure of the random effects under the assumptions that the random effects and the errors have Gaussian distributions with zero means and variances $\boldsymbol{G}$ and $\boldsymbol{R}$, respectively. Along these lines Oman (1995) suggested a procedure for checking the covariance structure of the mixed model analysis of variance. The basic idea of the procedure is to express the model assumptions in terms of the underlying covariance structure and then transform the data into residuals, which are appropriate for checking this structure.

In the linear mixed model, a specific covariance structure is selected from a set of candidates. In most practical situations the dimension of this set is small with only a few structures available, for example, compound symmetry, first-order autoregressive (AR(1)), Toeplitz and unstructured covariance. Model selection with respect to the covariance structure in linear mixed models is discussed by several authors (Diggle, 1988; Wolfinger, 1993; Littell et al., 1996; Verbeke and Molenberghs, 2000). Lee (1991) and, Pan and Fang (2002) discuss covariance structure selection in growth curve settings. Although the fitted covariance structure may be "best", it may not be close to the true covariance structure. The misspecification of the covariance structure may affect the estimates of the fixed effects, random effects and variance parameters. In some circumstances, misspecification of the covariance structure may not have a large impact on the estimates of the fixed effects, but, rather, may bias their standard errors downwards. As a result, statistical inferences on the fixed effects may be misleading. Furthermore, misspecification of the covariance structure may lead to incorrect identification of influential subjects. Given the pivotal role of the covariance structure in model estimation and diagnostics. it is important to correctly specify the covariance structure. The role of the covariance structure in model diagnostics is not covered in this thesis.

In summary, research on diagnostics in the linear mixed model has focused on case-deletion, local influence and robust modelling approaches. Under case-deletion an observation deemed to be an outlier is deleted from the analysis, such a strategy may not be optimal if the context suggests that the observation should be included in the analysis. Furthermore, diagnostic measures for detecting multiple outliers, within which subjects form natural groups, in the linear mixed model are not well-developed. In this thesis we will adopt the approach of Thompson (1985) which we contend offers an objective compromise between case-deletion and robust estimation through the use formal tests of hypotheses. The estimation of the parameters in the model and the formulation of the statistics for the detection of outliers will be discussed in Chapters 5 to 7.

CHAPTER 4

Residuals and leverages in the linear mixed model

4.1 Introduction

In this chapter we review the different types of residuals and leverages which are used to identify outliers and influential points in the linear mixed model. In Section 4.2 we give definitions of both residuals and leverages in the linear mixed. A modified decomposition of the linear mixed model leverage matrix is presented. This description is followed by a discussion on the graphical displays for detecting high leverage observations, outliers or both (§ 4.2 to 4.5). Finally we explore the potential use of the different types of residuals and leverages using a simulated data set and two real data sets. In Chapter 8 we will highlight areas with regard to residuals and leverages in the linear mixed model requiring further research.

In fitting a linear mixed model to a given data set, we may wish (i) to model the mean population response $\boldsymbol{X\beta}$, and use the random effects as a mechanism to parameterize the variance-covariance matrix matrix of the data, $\sigma^2\boldsymbol{H}$, where $\boldsymbol{H}$ is as defined in (2.4) or (ii) to model population and individual behaviour and be interested in estimating the variance parameters (and random effects). We achieve the first objective by fitting only the fixed part of model (2.1) given in Chapter 2 and call the resulting deviations $\hat{\boldsymbol{e}}$ marginal residuals. The second objective is achieved by fitting the full linear mixed model (2.1) and obtain the conditional residuals, $\tilde{\boldsymbol{e}}$. The fitting process of the linear mixed model also yields the marginal residuals $\hat{\boldsymbol{e}}$ and the predictions of the random effects, $\boldsymbol{Z}\tilde{\boldsymbol{u}}$, as by-products. We also take the latter to be residuals, as they are deviations of the predicted responses for the ith subject from the population average.

Our use of the term marginal stems from the fact that $\widehat{\mathrm{E}}(\boldsymbol{y}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ while the use of the term term conditional stems from the fact that $\widehat{\mathrm{E}}(\boldsymbol{y}|\boldsymbol{u}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}$. Haslett and Hayes (1998) used the term conditional residuals in the context of the linear model with a general covariance matrix. They define the conditional residual as

$$\tilde{e}_i = y_i - \widehat{\mathrm{E}}(y_i|\boldsymbol{y}_{\setminus i})$$

where

$$\widehat{\mathrm{E}}(y_i|\boldsymbol{y}_{\setminus i}) = \tilde{y}_{(i)} = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{(i)} + (\boldsymbol{v}_i'\boldsymbol{V}^{-1})_{(i)}(\boldsymbol{y}_{\setminus i} - \boldsymbol{X}_i\hat{\beta}_{(i)})$$

where $\widehat{\mathrm{E}}(y_i|\boldsymbol{y}_{\setminus i})$, the BLUP of y_i, represents the estimated conditional expected value when the ith observation is omitted from the estimation set, $\boldsymbol{y}_{\setminus i}$ is a sub-vector of $\boldsymbol{y}$ with the ith observation omitted, the subscript (i) indicates computed with y_i omitted or computed from $\boldsymbol{y}_{\setminus i}$, and $\boldsymbol{V} = \sigma^2\boldsymbol{H}$. Haslett (1999) used the conditional residual to derive deletion measures for arbitrary subsets for the linear model with a general covariance structure. In a time series context, Fraccaro et al. (2000) introduced yet another definition of a conditional residual as the deviation of the observation at time t, y_t, from its expected value given the previous values of the observations, i.e. conditional upon the prior subset of the observations. Haslett and Haslett (2007) extend this idea of conditional residuals to a wide class of models including the linear mixed model.

In their paper on case-deletion measures for the linear mixed model Christensen et al. (1992a) did not discuss residuals and leverages and their roles in their suggested influence diagnostics. Zewotir and Galpin (2007) investigated residuals and leverages in the linear mixed model. They focused on the conditional residuals only and also defined joint leverages for the fixed and the random effects. Nobre and Singer (2007) discussed three types of residuals in the linear mixed: marginal, conditional and random effects residuals. They proposed a standardization of the conditional residuals, using their respective variances, for detecting outlying observations or subjects. They did not discuss standardization of the marginal residuals nor the leverages for the linear mixed model. The marginal residuals are used to check outliers with respect to the fixed effects and for checking the validity of the within-subjects covariance structure. The conditional residuals are used for checking the constancy of the error variance and the Gaussian assumption of the random errors. They are also used to detect outlying observations within subjects. Finally, the random effects residuals are used for checking the random effects covariance structure and the Gaussian assumption for the random effects. A summary of the use of each type of residual is given in Nobre and Singer (2007).

4.2 Residual diagnostics in linear mixed models

Consider the linear fixed effects model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{4.1}$$

for $\boldsymbol{y} = (y_i, \ldots, y_n)'$, $\boldsymbol{X}$ is an $n \times p$ design matrix for the fixed effects, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector of fixed effects and $\boldsymbol{e}$ is an $n \times 1$ vector of random errors. In the simple linear regression model the random errors, $\boldsymbol{e}$, are assumed be Gaussian distributed with zero mean and variance $\sigma^2\boldsymbol{I}$. Note that model (4.1) is a special case of model (2.1) with no random effects i.e. the model has $\boldsymbol{V} = \sigma^2\boldsymbol{I}$.

The parameter estimates for the fixed effects are $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ with variance-covariance matrix $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. The vector of fitted values is

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{M}\boldsymbol{y}\end{aligned}$$

with

$$\text{var}(\hat{\boldsymbol{y}}) = \sigma^2\boldsymbol{M},$$

where $\boldsymbol{M} = \partial\hat{\boldsymbol{y}}/\partial\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the hat matrix (Hoaglin and Welsch, 1978). This matrix has been studied by many authors in regression diagnostics and is used to detect high leverage observations. A common measure of leverage of the ith observation is the ith main diagonal element of the hat matrix, m_{ii}. An observation is considered to be a high leverage point (in this model) if $m_{ii} > 2p/n$.

The vector of residuals from the full data is

$$\begin{aligned}\hat{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{y}\end{aligned}$$

with

$$\text{var}(\hat{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{M}).$$

The internally (standardized) Studentized residuals are defined as

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-m_{ii}}}, \quad i = 1, \ldots, n,$$

where $\hat{\sigma}^2 = (\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}})/(n-p)$, m_{ii} is the ith diagonal element of $\boldsymbol{M}$. In contrast the externally Studentized residuals are given by

$$\begin{aligned} t_i^* &= \frac{\hat{e}_i}{\sigma_{(i)}\sqrt{(1-m_{ii})}} \\ &= t_i\sqrt{\left(\frac{n-p-1}{n-p-t_i^2}\right)} \end{aligned}$$

where

$$\begin{aligned} \sigma_{(i)}^2 &= \frac{(n-p)\hat{\sigma}^2}{n-p-1} - \frac{\hat{e}_i^2}{(n-p-1)(1-m_{ii})} \\ &= \hat{\sigma}^2\left(\frac{n-p-t_i^2}{n-p-1}\right) \end{aligned}$$

is the error variance estimate when the ith observation deleted.

Using the results of Ellenberg (1973), Cook and Weisberg (1982, pp. 19) showed that each $t_i^2/(n-p)$ is identically (but not independently) distributed as Beta$(1/2, (n-p)/2)$ with $1/n - p \leq t_i^2/(n-p) < 1$. In contrast Beckman and Trussell (1974) showed that t_i^{*2} has an F distribution with degrees of freedom 1 and $n-p$ (or $t_i^* \sim t_{n-p}$).

In the following we discuss marginal and conditional residuals and their related leverages in the linear mixed model.

Marginal residuals and leverages

In the linear mixed model (2.1), the deviations of the fitted values for the fixed effect part of the model, from the responses give the marginal residuals

$$\begin{aligned} \hat{\boldsymbol{e}}_{M} &= \boldsymbol{y} - \mathrm{E}(\boldsymbol{y}) \\ &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= (\boldsymbol{I} - \boldsymbol{M}_1)\boldsymbol{y}, \end{aligned}$$

where

$$\boldsymbol{M}_1 = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}, \tag{4.2}$$

where $\boldsymbol{H}$ is as defined in (2.4). It follows that

$$\mathrm{E}(\hat{\boldsymbol{e}}_M) = \boldsymbol{0}$$

and

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{e}}_M) &= \mathrm{var}[(\boldsymbol{I} - \boldsymbol{M}_1)\boldsymbol{y}] \\ &= \sigma^2(\boldsymbol{I} - \boldsymbol{M}_1)\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{M}_1)' \\ &= \sigma^2\boldsymbol{HPH} \\ &= \sigma^2[\boldsymbol{H} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'],\end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$, as defined in Lemma 2.1 of Chapter 2 (also see Lemma A.2 of Appendix A).

The matrix $\boldsymbol{M}_1$ has been discussed in the context of the linear model with correlated errors (DeGruttola et al., 1987; Schall and Dunne, 1988; Martin, 1992). Fung et al. (2002) considered $\boldsymbol{M}_1$ in the context of semi-parametric mixed models. Earlier Christensen et al. (1992a) discussed the use of $\boldsymbol{M}_1$ to detect high leverage observations, in a linear mixed model, with the covariance matrix matrix recomputed when an observation is deleted. Banerjee and Frees (1997) discussed $\boldsymbol{M}_1$ in the context of longitudinal data models. Christensen (2001) views $\boldsymbol{M}_1$ as a random projection operator since $\sigma^2\boldsymbol{H}$ is a function of $\boldsymbol{y}$ and hence random.

Although idempotent, with $\mathrm{tr}(\boldsymbol{M}_1) = p$, $\boldsymbol{M}_1$ it is not symmetric. Christensen (2002) uses the Cholesky decomposition, $\boldsymbol{LL}' = \sigma^{-2}\boldsymbol{H}^{-1}$ to obtain the projection matrix,

$$\boldsymbol{M}_1^* = \boldsymbol{X}^*(\boldsymbol{X}^{*'}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*'}, \tag{4.3}$$

where $\boldsymbol{X}^* = \boldsymbol{L}^{-1}\boldsymbol{X}$, which yields leverages associated with the independent data $\boldsymbol{L}^{-1}\boldsymbol{y}$. Schall and Dunne (1988) use the eigenvalue-eigenvector decomposition of $\sigma^2\boldsymbol{H}$ to obtain $\boldsymbol{L}$. Puterman (1988) takes $\boldsymbol{L}$ as the lower triangle of $\sigma^2\boldsymbol{H}$. Martin (1992) proposed $(\sigma^2\boldsymbol{H})^{-1}\boldsymbol{M}_1^*$ as the leverage matrix, whose elements depend on the scaling of the data.

DeGruttola et al. (1987) noted that the diagonal elements of the (non-symmetric) hat matrix $\boldsymbol{M}_1$ may be negative if observations are highly correlated, so that their lower and upper bounds differ from those of $\boldsymbol{M}$. Schabenberger (2004) supports this view and argues that the cause of this phenomenon is that surrounding observations receive a larger weight towards estimating the mean of $\boldsymbol{y}$, in such a way that the observation itself receives a negative weight.

Conditional residuals and leverages

Banerjee (1994) took $\boldsymbol{X}$ and $\boldsymbol{Z}$ as fixed in model (2.1) with $\boldsymbol{R} = \boldsymbol{I}$ and obtained ordinary least squares (OLS) estimators

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{y} \quad \text{and} \quad \hat{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{\Delta} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. The residuals from the extended linear model fit

$$\hat{\boldsymbol{e}} = [\boldsymbol{I} - \boldsymbol{\Delta}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Delta} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']\boldsymbol{y}$$

and hence the hat matrix is given by

$$\boldsymbol{M}_2 = \boldsymbol{\Delta}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Delta} + \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'. \tag{4.4}$$

Banerjee (1994) asserts that $\boldsymbol{\Delta}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Delta}$ and $\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ can be regarded as the leverages due to the variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ respectively. Hence, the amount of leverage for the i-th subject is represented by the whole i-th diagonal block of the hat matrix $\boldsymbol{M}_2$ which is an $n_i \times n_i$ symmetric and idempotent matrix.

Langford and Lewis (1998, pp. 127) give an obliquely similar formula for the hat matrix in multilevel models fitted using iterative GLS. They propose extracting separate projection matrices for the fixed and random parts of the model as

$$\boldsymbol{M}_X = \text{Diag}[\boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1/2}]$$

and

$$\boldsymbol{M}_Z = \text{Diag}[\boldsymbol{V}^{-1/2}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{V}^{-1/2}]$$

where $\boldsymbol{V} = \sigma^2\boldsymbol{H}$ and with $\boldsymbol{V}^{-1/2}$ having the property that $(\boldsymbol{V}^{-1/2})(\boldsymbol{V}^{-1/2})' = \boldsymbol{V}^{-1}$. An earlier discussion of the leverage matrix for the fixed effects $\boldsymbol{M}_X$ was given by Pregibon (1981). The relationship between $\boldsymbol{M}_X$ and $\boldsymbol{M}_Z$ can be investigated by a scatter plot of their respective diagonal elements. In modelling variance heterogeneity, Verbyla (1993) compared the REML hat matrix $\boldsymbol{M}_Z$ and the ML hat matrix

$$\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'.$$

He found that the leverage measures for variance estimation under ML and REML differed substantially leading to some points being flagged as influential under ML but not under REML.

In determining the degrees of freedom associated with a fitted linear mixed model Hodges and Sargent (2001) proposed the matrix (also see Vaida and Blancard, 2005)

$$\boldsymbol{M}_2 = [\boldsymbol{X}\ \boldsymbol{Z}]\boldsymbol{A}'\boldsymbol{A}[\boldsymbol{X}\ \boldsymbol{Z}]',$$

where $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{0} & -\Delta \end{bmatrix}$ and $\boldsymbol{G} = (\Delta'\Delta)^{-1}$. Hence, for $\boldsymbol{R} \neq \boldsymbol{I}$, $\boldsymbol{M}_2$ can be rewritten as

$$\boldsymbol{M}_2 = [\boldsymbol{X}\ \boldsymbol{Z}]\boldsymbol{C}^{-1}[\boldsymbol{X}\ \boldsymbol{Z}]'\boldsymbol{R}^{-1}, \tag{4.5}$$

where $\boldsymbol{C}$ is the coefficient matrix of the mixed model equations for the linear mixed model given in (2.7).

Even though, $\hat{\boldsymbol{y}} = \boldsymbol{M}_2\boldsymbol{y}$, Vaida and Blancard (2005) argued that $\boldsymbol{M}_2$ is not a projection matrix but the top-left square sub-matrix of the projection matrix $\boldsymbol{M} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$.

4.2.1 Hat matrix decompositions

The linear mixed model (2.1), with both fixed and random effects, has the fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}$. The deviations of these fitted values from the responses yield the conditional residuals

$$\tilde{\boldsymbol{e}}_C = \boldsymbol{y} - \widehat{\mathrm{E}}(\boldsymbol{y}|\boldsymbol{u})$$

$$
\begin{aligned}
&= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\tilde{\boldsymbol{u}} \\
&= [(\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}) - (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}]\boldsymbol{y} \\
&= (\boldsymbol{I} - \boldsymbol{M}_3)\boldsymbol{y},
\end{aligned}
$$

with expectation zero and variance

$$
\text{var}(\tilde{\boldsymbol{e}}_C) = \sigma^2(\boldsymbol{I} - \boldsymbol{M}_3)\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{M}_3)',
$$

where

$$
\boldsymbol{M}_3 = (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}. \tag{4.6}
$$

Alternatively, we can write the conditional residuals as

$$
\begin{aligned}
\tilde{\boldsymbol{e}}_C &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\tilde{\boldsymbol{u}} \\
&= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{y} \\
&= \boldsymbol{H}\boldsymbol{P}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{y} \\
&= \boldsymbol{R}\boldsymbol{P}\boldsymbol{y},
\end{aligned}
$$

so that the residuals have expectation zero and variance

$$
\text{var}(\tilde{\boldsymbol{e}}_C) = \sigma^2\boldsymbol{R}\boldsymbol{P}\boldsymbol{R} \qquad \text{using Lemma A.2 of Appendix A.}
$$

Hence the hat matrix is given by

$$
\boldsymbol{M}_3^* = \boldsymbol{I} - \boldsymbol{R}\boldsymbol{P}. \tag{4.7}
$$

An alternative form of $\boldsymbol{M}_3$ is given by Gilmour et al. (2002). They first define the conditional residuals as

$$
\begin{aligned}
\tilde{\boldsymbol{e}}_C &= \boldsymbol{y} - \boldsymbol{W}\tilde{\boldsymbol{\psi}}, \qquad \text{using (2.10)} \\
\tilde{\boldsymbol{e}}_C &= \boldsymbol{y} - \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} \\
&= \boldsymbol{R}(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1})\boldsymbol{y} \\
&= \boldsymbol{R}\boldsymbol{P}\boldsymbol{y} \qquad \text{using Lemma (A.5) of Appendix A,}
\end{aligned}
$$

with variance

$$\begin{aligned}\text{var}(\tilde{\boldsymbol{e}}_C) &= \sigma^2 \boldsymbol{R}\boldsymbol{P} \\ &= \sigma^2(\boldsymbol{R} - \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'),\end{aligned}$$

where $\boldsymbol{W} = [\boldsymbol{X}\ \boldsymbol{Z}]$.

Hence the hat matrix is given by

$$\boldsymbol{M}_4 = \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1} \tag{4.8}$$

Zewotir and Galpin (2005, 2007) assume $\boldsymbol{R} = \boldsymbol{I}$ and define the conditional residuals as

$$\tilde{\boldsymbol{e}}_C = \boldsymbol{P}\boldsymbol{y}$$

Since $\hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$ their hat matrix is given by

$$\boldsymbol{M}_5 = \boldsymbol{I} - \boldsymbol{P}, \tag{4.9}$$

so that for p_{ii} close to zero, $\hat{y}_i$ is determined by $(1 - p_{ii})y_i$, i.e. y_i is an influential observation.

The matrices $\boldsymbol{M}_2$, $\boldsymbol{M}_3$, $\boldsymbol{M}_3^*$ and $\boldsymbol{M}_4$ are essentially equivalent. Additionally, for $\boldsymbol{R} = \boldsymbol{I}$, they are all equal to $\boldsymbol{M}_5$.

It can be shown that $\boldsymbol{M}_3^* = \boldsymbol{M}_3$ and that $\boldsymbol{M}_4 = \boldsymbol{M}_3 = \boldsymbol{M}_2$ (see Appendix B).

4.3 Detection of high leverage observations

From now on we denote the matrix $\boldsymbol{M}_3$ (or $\boldsymbol{M}_2$ or $\boldsymbol{M}_3^*$ or $\boldsymbol{M}_4$) as $\boldsymbol{M}_{XZ}$. Demidenko and Stukel (2005) represent the generalized leverage matrix for the jth subject (cluster), for $j = 1, \ldots, g$, in the linear mixed model, as the sum of two matrices, i.e.

$$\begin{aligned}\boldsymbol{M}_{XZj} &= \boldsymbol{M}_{j1} + \boldsymbol{M}_{j2} \\ &= \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{H}_j^{-1}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{H}_j^{-1} + \boldsymbol{Z}_j\boldsymbol{G}_j\boldsymbol{Z}_j'\boldsymbol{P}_j \\ &= \boldsymbol{M}_{Zj} + (\boldsymbol{I} - \boldsymbol{M}_{Zj})\boldsymbol{M}_{Xj},\end{aligned} \tag{4.10}$$

where

$$\begin{aligned}
\boldsymbol{M}_{j1} &= \boldsymbol{M}_{Xj} = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{H}_j^{-1}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{H}_j^{-1}, \\
\boldsymbol{M}_{j2} &= \boldsymbol{Z}_j\boldsymbol{G}_j\boldsymbol{Z}_j'\boldsymbol{P}_j, \\
\boldsymbol{P}_j &= \boldsymbol{H}_j^{-1} - \boldsymbol{H}_j^{-1}\boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{H}_j^{-1}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{H}_j^{-1}
\end{aligned}$$

and

$$\boldsymbol{M}_{Z_j} = \boldsymbol{Z}_j\boldsymbol{G}_j\boldsymbol{Z}_j'\boldsymbol{H}_j^{-1}.$$

$\boldsymbol{M}_{j1}$ and $\boldsymbol{M}_{j2}$ are the leverages for the fixed effects and random effects respectively. Demidenko and Stukel (2005) propose $\text{tr}(\boldsymbol{M}_{XZj})$ as a leverage measure so that a large value of $\text{tr}(\boldsymbol{M}_{j1})$ indicates an outlier in the X-space, and correspondingly a large value of $\text{tr}(\boldsymbol{M}_{j2})$ points towards an outlier in the Z-space. But since $\boldsymbol{M}_{j2}$ involves the design matrix $\boldsymbol{X}$ of the fixed effects, their influence is not completely removed. Therefore a better alternative would be to use $\boldsymbol{M}_{Zj}$, or $\boldsymbol{M}_Z = \sum_{j=1}^{n}\boldsymbol{M}_{Zj}$ in our notation. Note that when $\boldsymbol{G} = \boldsymbol{0}$ and $\boldsymbol{R} = \boldsymbol{I}$, $\boldsymbol{M}_{XZj}$ is equivalent to the hat matrix $\boldsymbol{M}$ in linear regression.

From (4.10) we define the fixed effects leverage for the kth observation of the jth subject as the (k, k)th element of $\boldsymbol{M}_{Xj}$, i.e. $\boldsymbol{M}_{Xj(kk)}$. Likewise the random effects leverage for the kth observation of the jth subject is defined as the kth main diagonal element of $\boldsymbol{M}_{Z_j}$, i.e. $\boldsymbol{M}_{Zj(kk)}$. The leverage of both the fixed and random effects is defined as $\boldsymbol{M}_{XZj(kk)}$. The summary of the different leverage measures and their cut-off values is given in Table 4.1. The threshold values follow from suggestions in Hoaglin and Welsch (1978), and Demidenko and Stukel (2005).

Index plots of the leverage measures given in Table 4.1 (in conjunction with corresponding threshold values) are used as diagnostics for detecting high leverage observations with respect to the X-space and Z-space or both.

4.4 Detection of outliers

Marginal residuals

Lesaffre and Verbeke (1998) suggest using the marginal residuals to check the validity

Table 4.1: *Summary of leverage measures in the linear mixed model*

Influence on	Leverage quantity	Cut-off value
Fixed effects		
Observations	$\boldsymbol{M}_{Xj(kk)}$	$\boldsymbol{M}_{Xj(kk)} \geq 2p/n$
Subjects	$n_j^{-1}\text{tr}(\boldsymbol{M}_{Xj})$	$n_j^{-1}\text{tr}(\boldsymbol{M}_{Xj}) \geq 2p/n$
Random effects		
Observations	$\boldsymbol{M}_{Zi(kk)}$	$\boldsymbol{M}_{Zj(kk)} \geq 2n^{-1}\text{tr}(\boldsymbol{M}_Z)$
Subjects	$n_j^{-1}\text{tr}(\boldsymbol{M}_{Zj})$	$n_j^{-1}\text{tr}(\boldsymbol{M}_{Zj}) \geq 2n^{-1}\text{tr}(\boldsymbol{M}_Z)$
Random and fixed effects		
Observations	$\boldsymbol{M}_{XZj(kk)}$	$\boldsymbol{M}_{XZj(kk)} \geq 2n^{-1}\text{tr}(\boldsymbol{M}_{XZ})$
Subjects	$n_j^{-1}\text{tr}(\boldsymbol{M}_{XZj})$	$n_j^{-1}\text{tr}(\boldsymbol{M}_{XZj}) \geq 2n^{-1}\text{tr}(\boldsymbol{M}_{XZ})$

of the covariance structure, $\sigma^2\boldsymbol{H}$, for the case when $\boldsymbol{R} = \boldsymbol{I}$. They propose the statistic $||\boldsymbol{I}_{n_j} - \mathcal{R}_j\mathcal{R}_j'||$, where the $\mathcal{R}_j = (\sigma^2\boldsymbol{H}_j)^{-1/2}\boldsymbol{e}_{_{Mj}}$ are the internally standardized marginal residuals, $\boldsymbol{H}_j$ is the sub-matrix of $\boldsymbol{H}$ for the jth subject and $||\boldsymbol{A}|| = \sqrt{\text{tr}(\boldsymbol{A}\boldsymbol{A}')}$ is Euclidean norm of the matrix $\boldsymbol{A}$ (Golub and Van Loan, 1996, pp. 55).

Waternaux et al. (1989) proposed using the statistic

$$\begin{aligned} q_j &= \boldsymbol{e}_{_{Mj}}'(\sigma^2\boldsymbol{H}_j)^{-1}\boldsymbol{e}_{_{Mj}} \\ &= ||\mathcal{R}_j||^2, \qquad\qquad j = 1, \ldots, g \end{aligned}$$

for identifying outlying subjects/clusters in longitudinal data models, where $\boldsymbol{e}_{_{Mj}}$ is the vector of marginal residuals for subject j. Given the fitted model and the Gaussian assumption, the quadratic form of the standardized multivariate residual q_j has a chi-squared distribution with n_j degrees of freedom. The distribution of this statistic can be used to identify outliers in a linear mixed model. In data sets with a large number of subjects, the Bonferroni adjustment for multiple testing can be applied when using q_j to identify outlying units, but it can be conservative. Park and Lee (2004) suggested Q-Q plots of the q_j's as a diagnostic plot for goodness of fit and identification of outliers.

Since the marginal residuals are correlated, several authors (for instance Fraccaro et al., 2000) recommend orthogonalizing the marginal residuals to make them uncorrelated with the fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and hence orthogonal to $\hat{\boldsymbol{y}}$ and $\boldsymbol{X}$. This orthogonality to $\boldsymbol{X}$ is achieved by the transforming the marginal residuals to obtain the (orthogonal) residuals

$$\hat{\boldsymbol{e}}_O = \boldsymbol{H}^{-1}\hat{\boldsymbol{e}}_{_M}$$

$$= \boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}_{C},$$

with estimated covariance matrix

$$\begin{aligned} \text{var}(\hat{\boldsymbol{e}}_O) &= \text{var}(\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}_C) \\ &= \sigma^2\boldsymbol{P}. \end{aligned}$$

The distributional properties of these orthogonal residuals were not discussed in Fraccaro et al. (2000). Haslett and Haslett (2007, § 4.1) showed that these (orthogonal) residuals are equivalent to the conditional residuals.

Fitzmaurice et al. (2004, Ch. 9) also consider the use of the marginal residuals for 'checking departures from the mean response'. Similarly to Fraccaro et al. (2000), they make the residuals independent by taking the transformation $\boldsymbol{L}_i^{-1}\boldsymbol{e}_{Mi}$, where $\boldsymbol{L}_j$ arises from the Cholesky decomposition of var$(\hat{\boldsymbol{e}}_M)$.

Connection between conditional and marginal residuals

The conditional residuals can also be written as follows

$$\begin{aligned} \tilde{\boldsymbol{e}_C} &= \boldsymbol{RPy} \\ &= \boldsymbol{RH}^{-1}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{y} \\ &= \boldsymbol{RH}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{RH}^{-1}\hat{\boldsymbol{e}}_M. \end{aligned}$$

Thus, $\tilde{\boldsymbol{e}}_C = \boldsymbol{RH}^{-1}\hat{\boldsymbol{e}}_M$ or $\hat{\boldsymbol{e}}_M = \boldsymbol{HR}^{-1}\tilde{\boldsymbol{e}}_C$, which establishes the relationship between marginal residuals and conditional residuals. This connection between the marginal and conditional residuals assumes that the matrix $\boldsymbol{H}$ is positive definite (and so too $\boldsymbol{G}$), and both $\boldsymbol{X}$ and $\boldsymbol{Z}$ are of full-column rank (also see Haslett and Haslett, 2007, Appendix). Corresponding results for non-full rank matrices are given in Haslett and Haslett (2007, Appendix).

Conditional residuals

Pinheiro and Bates (2000, pp. 11) standardized the conditional residuals, using the error variance estimate rather than their estimated variances, as $\tilde{e}_i/\hat{\sigma}$, for $i = 1, \ldots, n$, where the $\tilde{e}_i$'s are elements of the conditional residual vector $\tilde{\boldsymbol{e}}_C$. They suggested index plots of these standardized residuals against the fitted values and Q-Q plots of these residuals for checking homoscedasticity and Gaussian assumption of the errors $\boldsymbol{e}$.

Both internally and externally Studentized conditional residuals, which use appropriate variance estimates for the conditional residuals, can also be used to detect outliers given a fitted linear mixed model. The internally and externally Studentized conditional residuals are, respectively, given by (also see Shi and Chen, 2009)

$$t_i = \frac{\tilde{e}_i}{\text{var}(e_i)} = \frac{\tilde{e}_i}{\hat{\sigma}\sqrt{(\boldsymbol{RPR})_{ii}}}$$

and

$$t_i^* = \frac{\tilde{e}_i}{\hat{\sigma}_{(i)}\sqrt{(\boldsymbol{RPR})_{ii}}},$$

where $(\boldsymbol{RPR})_{ii}$ is the diagonal element the matrix $\boldsymbol{RPR}$ with the denominator terms being $\hat{\sigma}\sqrt{\boldsymbol{P}_{ii}}$ when $\boldsymbol{R} = \boldsymbol{I}$.

Zewotir and Galpin (2007) used the internally Studentized residuals to identify outlying observations in the linear mixed model. While the properties of t_i and t_i^* have been investigated in linear regression, little is known about their properties in the linear mixed model. Hurtardo (1993) (also see Zewotir and Galpin, 2007) showed that $t_i^{*2} \sim \dfrac{n-1}{n-p-1} F_{1,n-p-1}$, where $F_{1,n-p-1}$ is an $F-$distribution with degrees of freedom 1 and $n-p-1$ or $t_i^{*2} \sim \chi_1^2$. This result was shown for a linear mixed model for balanced data; it is not known whether this finding would hold with an unbalanced design. Zewotir and Galpin (2007) suggested the statistic max$|t_i|$ for identifying the largest outlier. They derived the distribution of max$|t_i|$ using the connection between t_i and t_i^*.

Haslett and Haslett (2007) proposed the statistic

$$S = \hat{\mathbf{e}}'\boldsymbol{H}^{-1}\hat{\mathbf{e}}/\sigma^2 = \tilde{\mathbf{e}}'\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}\tilde{\mathbf{e}}/\sigma^2$$

as a goodness of fit measure for a linear mixed model. This goodness of fit statistic has a chi-squared distribution with degrees of freedom $n-p$ if the variance parameters in $\boldsymbol{H}$ are known. They suggested the decomposition of the statistic S to identify potential outliers i.e. observations with large contributions to the statistic. The above studies did not discuss the problems of masking and multiple testing.

Cressie (1991) proposed the variogram of conditional residuals for checking the covariance structure in spatial data. Gilmour et al. (1997) and Verbeke and Molenberghs (2000) also used the variogram of conditional residuals for identifying sources of variation in the analysis of field experiments and longitudinal data, respectively.

BLUPs as residuals

Plots of the estimated random effects, $\tilde{\boldsymbol{u}} = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_g)'$, where g is the number of subjects, are often used to identify outlying subjects (Fellner, 1986; Verbeke and Molenberghs, 2000; Pinheiro and Bates, 2000). The individual BLUPs, $\tilde{u}_j$, $j = 1, \ldots, g$ are usually scaled as follows

$$\zeta_j = \frac{\tilde{\boldsymbol{u}}_j}{\sqrt{\widehat{\text{var}}(\tilde{\boldsymbol{u}}_j)}},$$

where $\widehat{\text{var}}(\tilde{\boldsymbol{u}}_j) = \hat{\sigma}^2[\hat{\boldsymbol{G}} - \boldsymbol{G}\boldsymbol{Z}_j\hat{\boldsymbol{H}}_j^{-1}\boldsymbol{Z}_j\hat{\boldsymbol{G}}]$ and $\hat{\boldsymbol{H}}_j$ is the estimated covariance matrix portion for the jth subject.

If variance parameters are estimated using maximum likelihood, the individual statistics ζ_j, $j = 1, \ldots, g$ have a Student's t-distribution with degrees of freedom $n - r_{[\boldsymbol{X}\ \boldsymbol{Z}]}$ (Littell et al., 1996, pp. 502). This distribution of ζ_j is then used to detect outlying random effects. It must be noted that this test statistic is biased downwards since it depends on variance parameters estimated by maximum likelihood, and this criterion is known to produce downward biased variance estimates (Robin Thompson, 2006, in personal communication).

As an alternative the statistics $\tilde{\zeta}_j = \tilde{\boldsymbol{u}}_j{}'[\widehat{\text{var}}(\tilde{\boldsymbol{u}}_j - \boldsymbol{u}_j)]^{-1}\tilde{\boldsymbol{u}}_j$ (Waternaux et al., 1989) could be used to identify outlying subjects. We could also use the elements of the $\boldsymbol{Z}\tilde{\boldsymbol{u}}$. Hilden-Minton (1995) calls these quantities $\boldsymbol{Z}\tilde{\boldsymbol{u}}$ the random effects residuals, given by

$$\boldsymbol{Z}\tilde{\boldsymbol{u}} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{y}$$

with variance

$$\text{var}(\boldsymbol{Z}\tilde{\boldsymbol{u}}) = \sigma^2\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'.$$

Multiple outliers

In the linear mixed model context all observations from the same subject form a natural group of observations. As we have already discussed in this section the multivariate marginal residuals and BLUPs (the q_j and $\tilde{\zeta}_j$ statistics) can be used to identify outlying subjects (which form natural groups of observations). The statistics t_i^2 or t_i^{*2} can also be extended to detect outlying subjects using

$$t_I^2 = \frac{\tilde{\boldsymbol{e}}_I'(\boldsymbol{RPR})_I^{-1}\tilde{\boldsymbol{e}}_I}{\hat{\sigma}^2} \tag{4.11}$$

or

$$t_I^{*2} = \frac{\tilde{\boldsymbol{e}}_I'(\boldsymbol{RPR})_I^{-1}\tilde{\boldsymbol{e}}_I}{\hat{\sigma}_I^2}, \tag{4.12}$$

where $\tilde{\boldsymbol{e}}_I$ is the vector of conditional residuals for the jth subject, $(\boldsymbol{RPR})_I$ is the sub-matrix of $\boldsymbol{RPR}$ for the jth subject and $\hat{\sigma}_I^{*2}$ is the estimator of σ^2 when the all the observations for the jth subject are deleted. Hurtardo (1993) gave expressions for the fixed effects estimates and for predictions of the random effects, when an arbitrary subset of observations indexed by I, and for $\hat{\sigma}_I^2$. The statistics t_I^2 and t_I^{*2} are generalizations of the multiple-case deletion Studentized residuals in the linear regression model (Cook and Weisberg, 1982, pp. 28).

Haslett and Haslett (2007) give a multivariate extension of the conditional residuals i.e. conditional residuals associated with arbitrary subsets of the data $\boldsymbol{y}$. The residuals used in (4.11) and (4.12) are functions of the multivariate conditional residuals.

4.5 Detection of high leverage observations and outliers

Following Cook (1977), McCulloch and Meeter (1983) suggested a plot of raw (unstandardized) residuals against the leverages for a fitted linear regression model. Gray (1986) also proposed a modification of this plot, and called it a leverage-residual plot in which the leverage effect measured by m_{ii}, the ith diagonal element of $\boldsymbol{M}$ is plotted against the residual effect measured by the squared normalized residual $\hat{e}_i^2/\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$ (see Figure 4.1). This plot is then used to identify possible outliers and high leverage points. For instance, high leverage points are expected to fall in the upper left-hand corner and outliers are expected to fall in the lower right-hand corner (see Figure 4.1). Figure 4.1 is similar to Figure 1.1 given in Chapter 1 (pp. 1-3) with the positions of high leverage points interchanged, outliers correspond to influential observations and ordinary observations are represented by non-influential and non-outlier observations. In general, the usefulness of the plot in Figure 4.1 depends on the threshold values for the residuals and leverages.

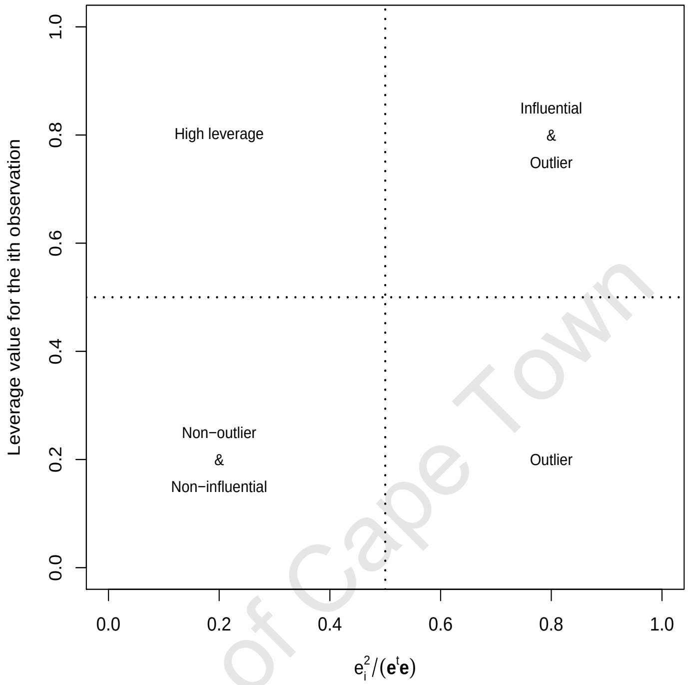

Figure 4.1: *Typical leverage and residual plot for a linear regression model.*

However, the leverage-residual plot is defined for marginal residuals (the only type of residuals in linear regression). Given that there are several types of residuals and leverages in the linear mixed model, it is not clear which leverage-residual combination should be used to construct the leverage-residual plot and how the plot should be interpreted. For illustrative purposes we will use the standardized conditional residuals and the joint leverages (i.e. the diagonal elements of $\boldsymbol{M}_{XZ}$) to construct the leverage-residual plots in the examples which follow. The standardized conditional residuals were also used by Nobre and Singer (2007) and are equivalent to the internally Studentized residuals of Zewotir and Galpin (2007). The advantage of the residual-leverage plot is that is uses information from the design space, from both matrices matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, and from the response $\boldsymbol{y}$, through the conditional residuals.

4.6 Example: Simulated data

To investigate the usefulness of the different types of residuals and leverages, we generate a data set from the following linear mixed model

$$y_{jk} = \mu + u_{0j} + \beta x_{jk} + u_{1j} z_k + e_{jk},$$

for $j = 1, \ldots, 10$, $k = 1, \ldots, 5$. The random effects u_{0j} and u_{1j} are independently generated from Gaussian distributions with zero means and variances σ_1^2 and σ_2^2, respectively. We set $\mu = 5$ and $\beta = 0.5$, and $\sigma_1^2 = \sigma_2^2 = 0.5$. The values of x_{jk} are drawn from the uniform distribution on $[0, 5]$ and the z_j's are also drawn from the uniform distribution on $[0, 5]$. The random errors e_{jk}'s are generated from a Gaussian distribution with mean zero and variance 1. From this stage on, the data set is fixed. Following Fellner's (1986, pp. 53) definition of outliers with respect to the rows of design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, we then introduce the following aberrant observations: $x_{1.5} = 8$, $x_{3.2} = 10$ and $z_{5.5} = 8$, $z_{8.1} = 10$, where the $j.l$ denotes the lth observation within the jth subject. Outliers in the Y-space are generated by setting $e_{9.1} = 5$ and $e_{10.3} = 10$, so that $e_{9.1}$ corresponds to a moderate outlier and $e_{10.3}$ to a severe outlier in the Y-space. Figure 4.2 shows plots of the responses (y) against the observation number (l) within each subject, and highlights outliers in the Y-space only (9.1 and 10.3). Figure 4.3 presents scatter plots of the responses against x for each subject and highlights outliers in both X-space and Y-space.

We then compute the different types of residuals and leverages and compare them mainly using index plots. The observations with outlying x_{jl}'s and z_{jl}'s are highlighted in Figure 4.4. The outlying observations (and subjects) with respect to fixed effects (1.5 and 3.2) are highlighted in plots (a) and (b) while the outlying observations with respect to random effects (5.5 and 8.1) are shown in plot (c). However, only subject 8 is identified as an outlier in plot (d) and not subject 5; this contrast implies that although there are outlying observations within subjects, the corresponding subject may not be considered influential. Only the observations with high leverage with respect to the random effects are highlighted in plots (e) and (f), suggesting that the joint generalized leverage for fixed and random effects may tend not to select influential observations with respect to the fixed effects.

The marginal residuals, conditional residuals and random effects residuals (and their standardized versions) do not require the fitting of a model and can be computed

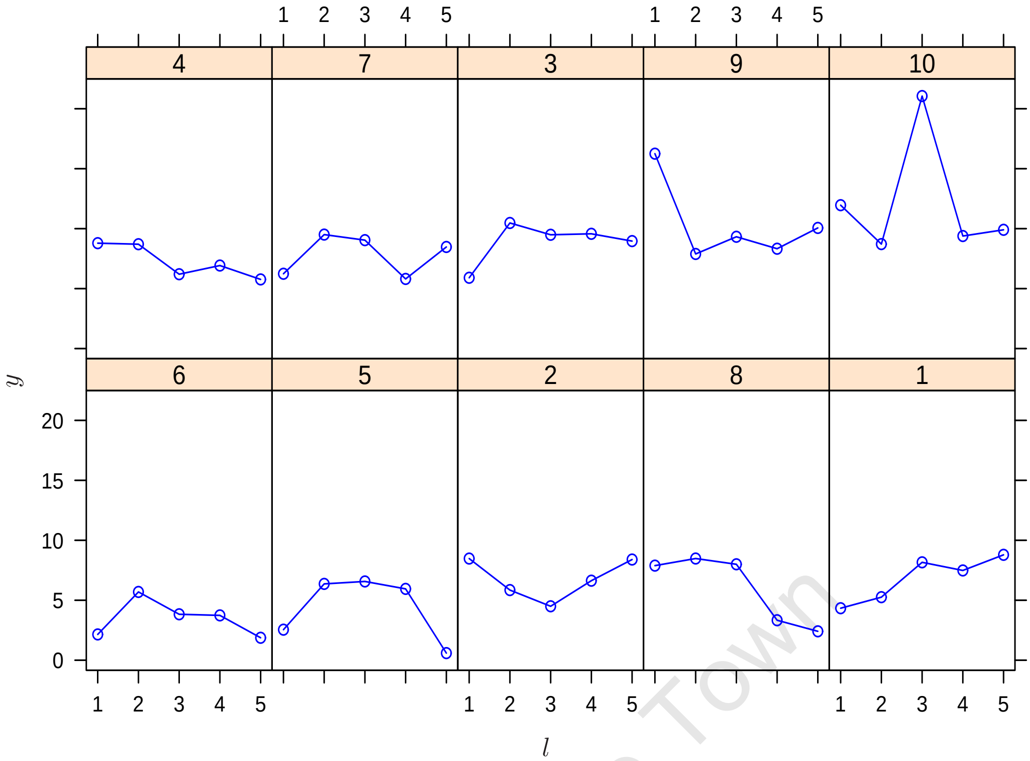

Figure 4.2: *Plots of responses (y) against observation number (l) within each subject for simulated data.*

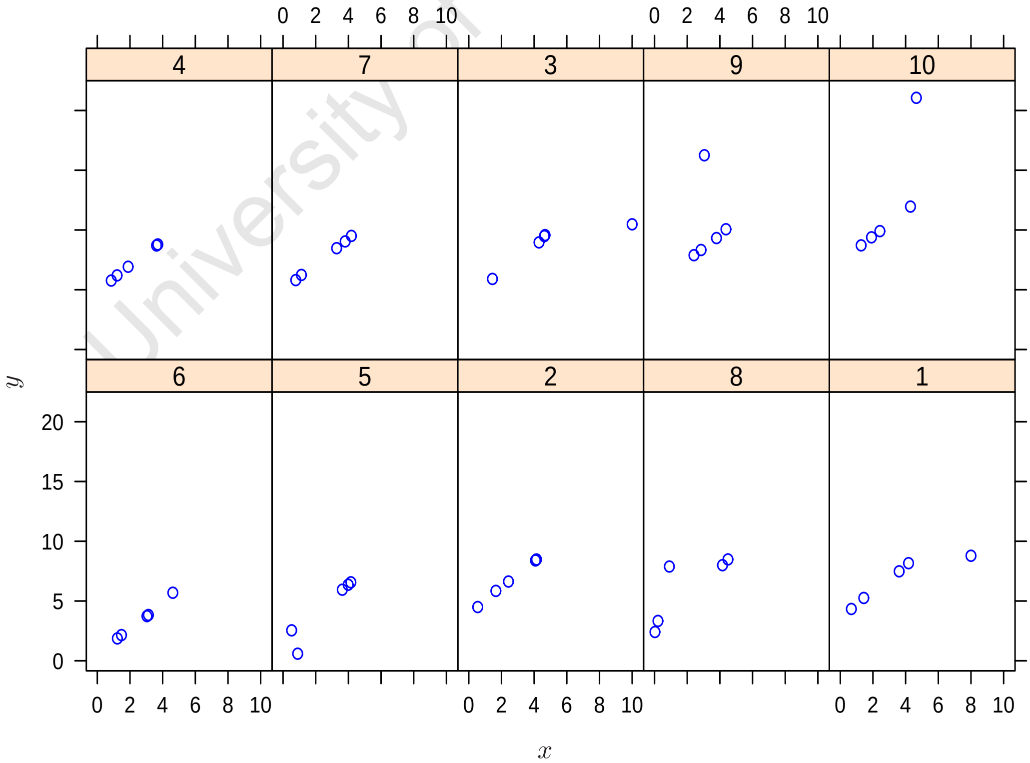

Figure 4.3: *Plots of responses (y) against x for simulated data.*

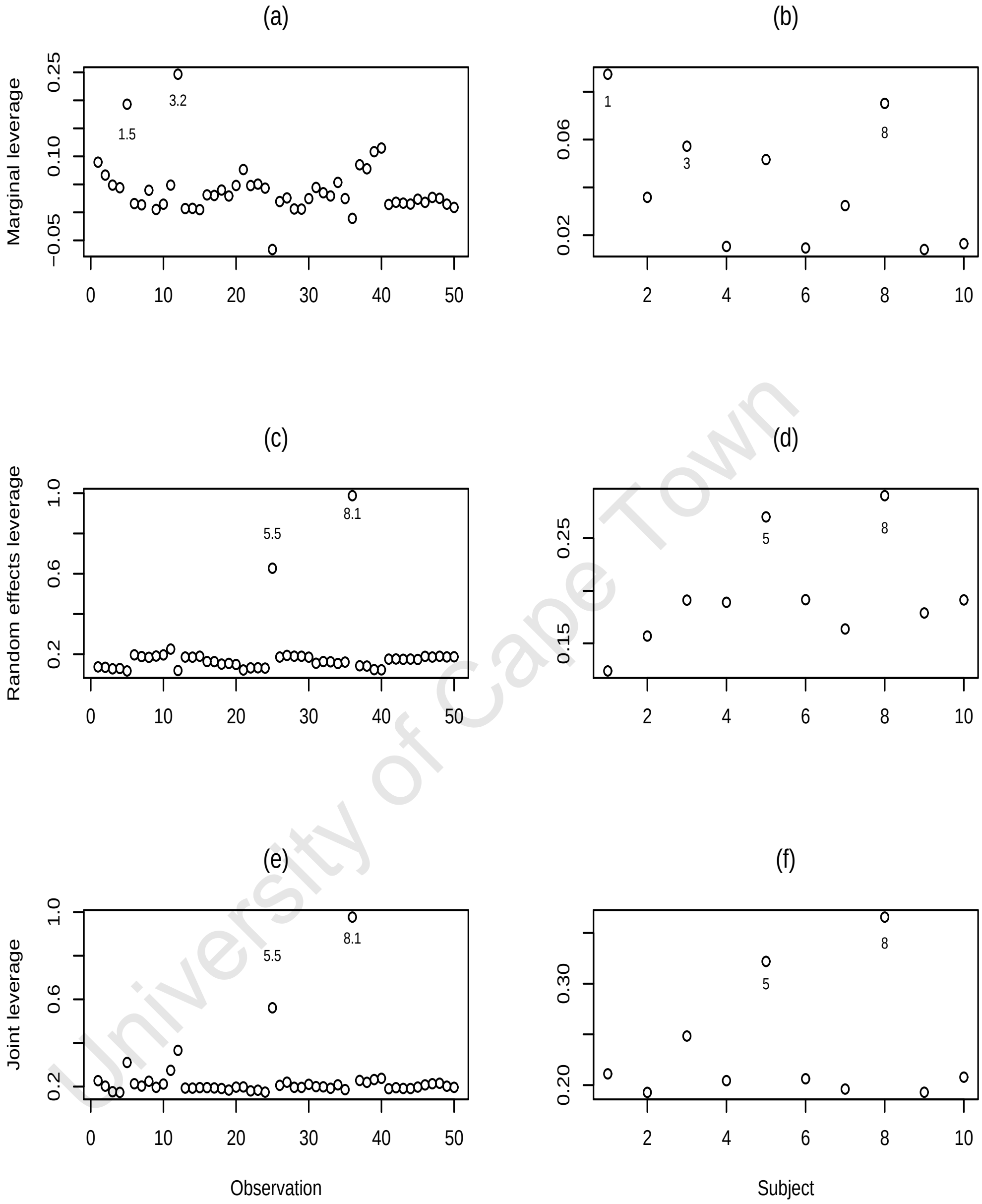

Figure 4.4: *Index plots of (a) marginal leverages for observations, (b) marginal leverages for subjects, (c) random effects leverages for observations, (d) random effects leverages for subjects, (e) joint leverages for observations and (f) joint leverages for subjects for simulated data.*

directly. Standardized marginal and conditional residuals are shown in Figures 4.5 and 4.6 respectively for both observations and subjects. The standardized marginal residuals picked out observations 1.1, 1.2 and 1.5 as outliers while only observations 3.2 is detected as an outlier by the standardized conditional residuals. The standardized subject marginal residuals are the q_j statistics of Waternaux et al. (1989) which individually have chi-squared distributions with degrees of freedom n_j ($n_j = 5$ for our example). Both the standardized subject marginal residuals and the standardized multivariate conditional identify subject 3 as an outlier. However, different individuals are identified as outliers according to the standardized subject marginal residuals (Figure 4.5 (b)) and the standardized random effects (Figure 4.6 (b)).

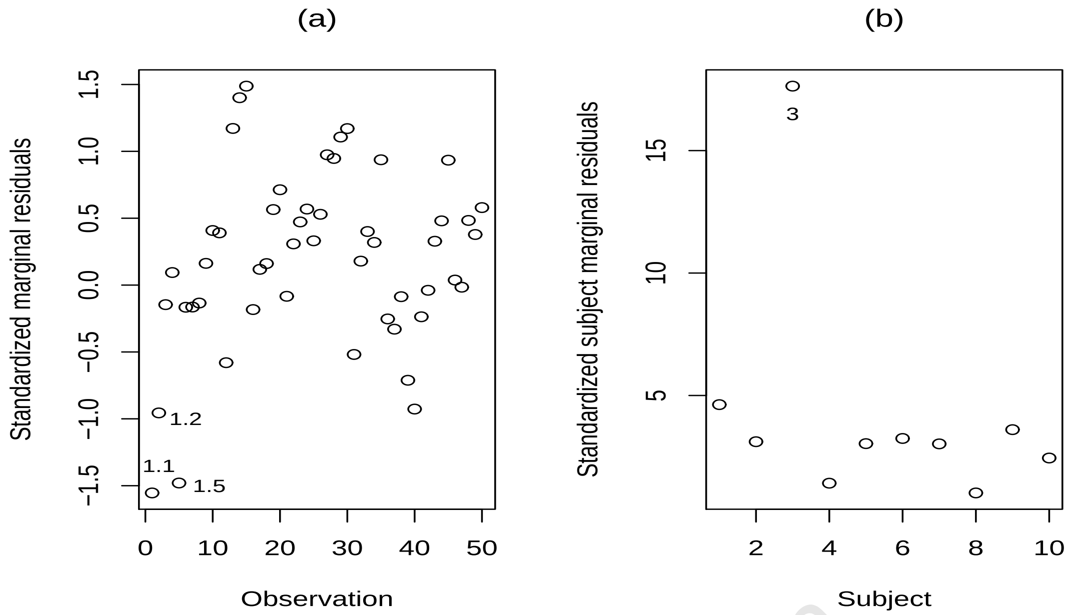

Figure 4.5: *Index plots of (a) Standardized marginal residuals and (b) Standardized subject marginal residuals for simulated data.*

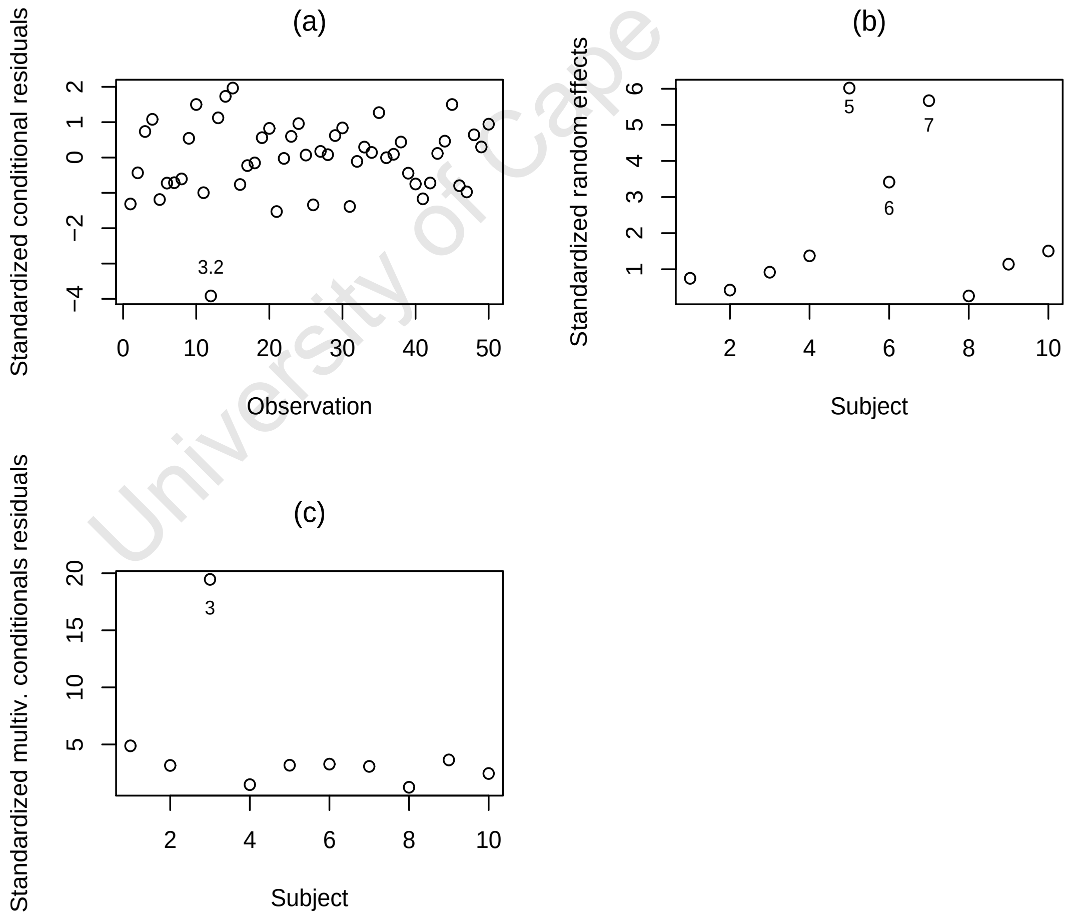

Figure 4.6: *Index plots of (a) Standardized conditional residuals, (b) Standardized random effects and (c) Standardized multivariate conditional residuals for simulated data.*

We use the leverage-conditional residual plot to identify high leverage observations and outliers. Observations 5.5 and 8.1 are highlighted as high leverage points in Figure 4.7 while observation 3.2 is identified as an outlier. Outliers with respect to the Y-space are not highlighted in the leverage-conditional residual plot.

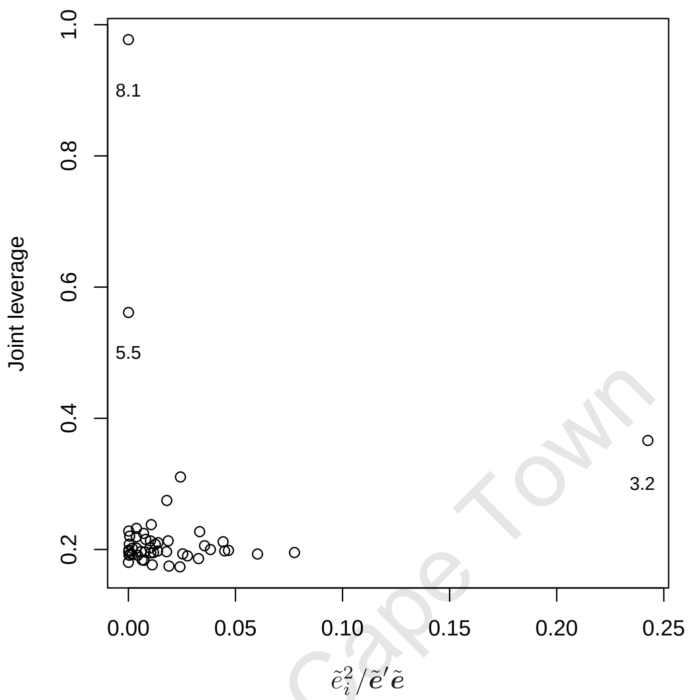

Figure 4.7: *Leverage-conditional residual plot for simulated data.*

4.7 Example: The orthodont data

We also reconsider the random coefficient model (2.31) fitted to the orthdont data in Chapter 2. We compute the various residuals and leverages for the fitted model and compare them.

Observations 13.2, 26.2 and 26.4 (observations for boy 13 and girl 10) (Figure 4.8, (a)) were identified as outlying according to standardized marginal residuals while only subjects 9 and 13 (both boys) were identified as outlying (Figure 4.8, (b)) using standardized subject marginal residuals. The standardized subject marginal residuals are the q_j's of Waternaux et al. (1989). The standardized conditional residuals identified the observations 9.2, 9.3 and 13.1 (observations for boys 9 and 13) as outliers (Figure 4.9, (a)). The profiles for these subject are highlighted in Figure 2.1 of Chapter 2, § 2.5. However, the standardized random effects identify subjects 1 and 10 (both boys) as outlying subjects but not subjects 9 and 13 (Figure 4.9, (b)).

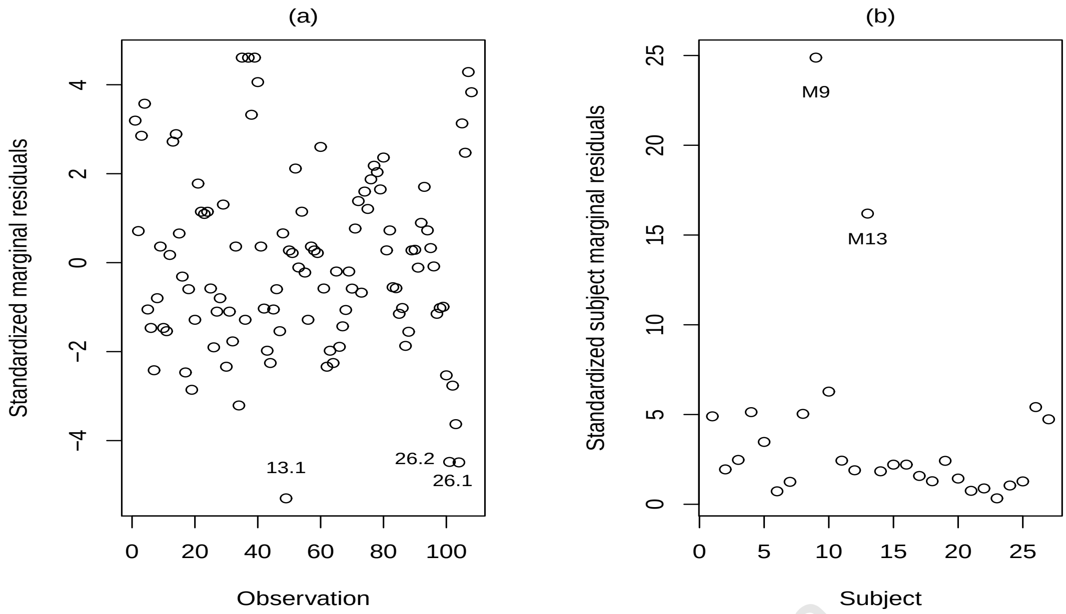

Figure 4.8: *Index plots of (a) Standardized marginal residuals and (b) Standardized subject marginal residuals, for orthodont data.*

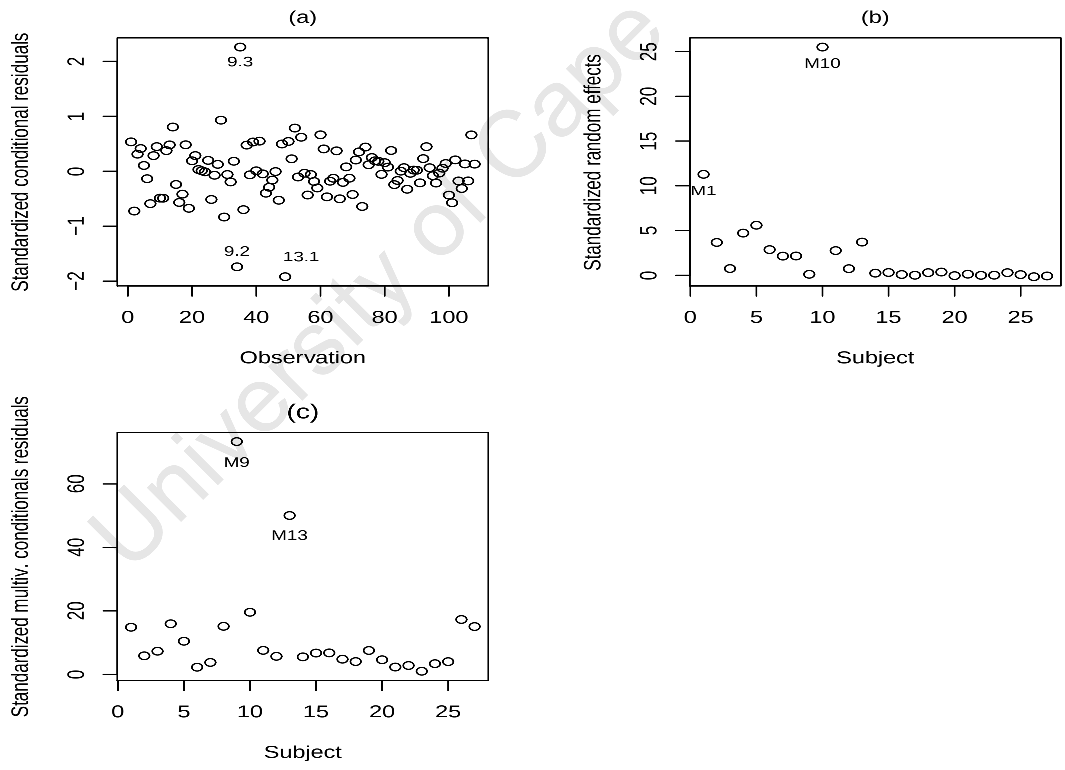

Figure 4.9: *Index plots of (a) Standardized conditional residuals (b) Standardized random effects and (c) Standardized multivariate conditional residuals for orthodont data.*

Figure 4.10 is a leverage-conditional residual plot for the orthodont data model and it identifies the observation labelled 13.1 as both influential and outlier, while the observations labelled 9.2 and 9.3 are identified as outliers. Only these influences seem to be supported on the basis of standardized conditional residuals. The joint leverages fall in 8 categories. These categories correspond to gender and age combinations. The leverages for boys are higher (in blue) than those for girls (in green) at all ages. The leverages are also higher for the ages 8 and 14 compared to the age categories 10 and 12.

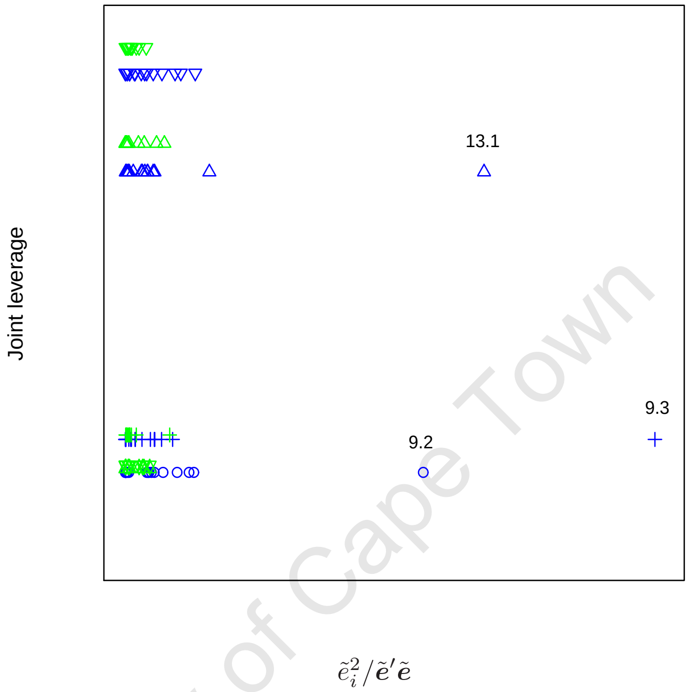

Figure 4.10: *Leverage-conditional residual plot for orthodont data:* $\circ$ = *age 10,* $+$ = *age 12,* $\triangle$= *age 8,* ∇ = *age 14. Points for boys are shown in blue and those for girls in green.*

4.8 Summary

We have presented a review of the different types of residuals and leverages in the linear mixed model and explore their uses. In the examples we have considered, the standardized marginal residuals and conditional residuals do not identify the same observations as outliers or high leverage points. The leverage matrix for the linear mixed model only involves the variance matrix $\boldsymbol{H}$ and the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, and not the observations $\boldsymbol{y}$, so that it is consequently a design and covariance issue. The decomposition of the leverage matrix for the linear mixed model leads to separate leverages for the fixed and random parts of the model. We noted that the joint leverages for the fixed and random effects only identify observations that are influential on the random effects. The fixed effects leverages, random effects leverages and joint leverages do not identify outliers in the Y-space. We also explored the use of the leverage-residual plot to identify possible high leverage observations or outliers or both. Possible areas of further research on residuals and leverages in the linear mixed model are given in Chapter 8.

In this chapter we have focused on the identification of outliers using standardized marginal and conditional residuals. In the next chapter we introduce an approach that can be used to detect outliers when the observations are independent and normally distributed. The approach also allows the inclusion of the outlier in the analysis. Formal statistical tests for outlier detection, under this approach, are also proposed. The approach we introduce in the next chapter will be extended to detect outliers in the linear mixed model in Chapters 6 and 7. In the linear mixed model with random slopes and intercepts, it may be of interest to identify subjects with unusual slopes or intercepts, such as in the orthodont data set.

CHAPTER 5

A variance shift outlier model for linear fixed effects analysis

The purpose of this chapter is to review and extend a variance shift outlier model (VSOM) as a model for outliers in linear fixed effects analysis. The motivation for the review of a VSOM in linear regression is two-fold: (i) to gain insight into the general case (a linear mixed VSOM), our topic in Chapter 6 with some extensions in Chapter 7 and (ii) to develop the method further in this simple case. Likelihood ratio and score test statistics are developed as objective measures for testing for the variance shift estimate to determine whether individual observations are outliers. Below we give an introduction of a VSOM and the outline of the chapter.

Cook et al. (1982) suggest an outlier detection method in which outliers are considered to be observations with inflated measurement error variances, as an alternative to additive effects under a common variance parameter. They used maximum likelihood to estimate the parameters in the fitted linear model. Cook and Weisberg (1982) discussed the same model and gave it the name 'variance shift model' (Cook and Weisberg, 1982, § 2.2.2, pp. 82). Thompson (1985) used REML for parameter estimation in the same model and found that the maximum of the REML log-likelihood function occurs when the inflated variance is associated with the unit having the largest standardized residual. The outcome is different when maximum likelihood is the principle of estimation. This property of REML estimation motivates our adoption of a VSOM approach to the detection of outliers in this thesis. Also, because a VSOM fits straight into the linear mixed model framework with REML in general.

A VSOM approach can be viewed as a compromise between robust estimation and case-deletion in the sense that it allows for the partial exclusion or inclusion ((down-weighting) of a unit in the estimation depending on the size of the variance shift estimate for the unit. The down-weighting of observations (or groups of observations) instead of deleting them (case-deletion) is a major advantage of a VSOM. A VSOM approach differs from weighted least squares or GLS, both of which weight all observations to achieve variance homogeneity (and to model the dependence between observations in case of GLS) in a linear model. A VSOM approach also differs from the outlier model of Box and Tiao (1968) and the local influence approach of Zewotir (2007) (see Chapter 3) in that the weight is estimated by REML (and can only increase the error variance) so that any down-weighting is objectively determined.

In this chapter we develop and examine likelihood ratio and score test statistics for determining whether individual observations have inflated variance or are possible outliers. The proposed likelihood ratio and score tests are standard tests in variance parameter testing in the linear mixed models but they serve a diagnostic purpose in our case i.e. the tests are used to evaluate evidence that the variance shift estimate(s) (one or more) is larger than zero, an indication that any corresponding observation is a possible outlier. Below is an outline of this chapter.

(a) review of a variance shift outlier model (VSOM),

(b) construction of the likelihood ratio test (LRT),

(c) construction of likelihood ratio tests based on one-step estimates of the variance parameters (one-step LRTs) and score tests, and examination of their dependence on the one-step updates of variance and hence the second-order derivatives of the REML log-likelihood function, (i.e. dependence upon observed, expected and average information matrices),

(d) evaluation of the empirical and asymptotic distributions of the test statistics under the null hypothesis (no outliers),

(e) derivation of the exact distribution of the expected information score test giving insight into the distribution of the LRT,

(f) a resampling procedure to handle the problem of multiple (iterative) testing in using a VSOM approach to identify outliers successively,

(g) illustration of a VSOM approach to outlier detection in linear regression using a simulated data set and a real data set.

(h) evaluation of the performance of the likelihood ratio and score tests in terms of computing time, and type I and type II errors using simulation and

(i) extension of a VSOM to the case of multiple outliers.

In reviewing a VSOM in linear regression and constructing likelihood ratio and score tests for the model, we first adopt a one-at-a-time approach similar to case-deletion and then extend the results to a VSOM for a group of outliers, i.e. detecting outliers more-than-one-at-a-time.

The new contributions introduced in this chapter are:

- exact forms of the LRT and score tests for a VSOM in linear regression,
- properties of one-step updates of the variance parameters under VSOM,
- distribution of the score test statistic based on the expected information matrix,
- evaluation of the procedure for multiple testing and parametric bootstrap and,
- extension of a VSOM in linear regression to the matter of groups of outliers.

5.1 Review of a variance shift outlier model

We reconsider the simple linear model (4.1), namely

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{5.1}$$

Under a variance shift outlier model (VSOM) for the ith observation, this observation has measurement error variance $\sigma^2(\omega_i + 1)$, $\omega_i \geq 0$, so that ω_i represents the factor increase in measurement error variance for the observation. We will refer to ω_i as the variance shift parameter. Our parametrization is different from that Cook and Weisberg (1982) (also see Cook et al., 1982; Thompson, 1985) who parameterized the variance of the ith observation as $\alpha_i\sigma^2$. In our parametrization $\alpha_i = (\omega_i + 1)$.

Using the parametrization $\theta_1 = \sigma^2$ and $\theta_{2i} = [1 + \omega_i(1 - v_i)]\sigma^2$, with $\sigma^2 > 0$ and $\omega_i \geq 0$ implies $\theta_1 > 0$ and $\theta_{2i} \geq \theta_1$, Thompson (1985) wrote the REML log-likelihood function for a VSOM as

$$l_i(\theta_1, \theta_{2i}; \boldsymbol{y}) = -\frac{1}{2}\left\{(n-p-1)\log\theta_1 + \log|\boldsymbol{X}'\boldsymbol{X}| + \log\theta_{2i} + \frac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y} - s_i^2}{\theta_1} + \frac{s_i^2}{\theta_{2i}}\right\} \tag{5.2}$$

where the subscript i in the log-likelihood $l_i(\theta_1, \theta_{2i}; \boldsymbol{y})$ stands for the log-likelihood under a VSOM for unit i, $\boldsymbol{P}_{X^\perp} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $s_i^2 = \hat{e}_i^2/(1 - v_i)$.

Under this parametrization, it is immediately obvious that

$$\hat{\theta}_1 = \frac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y} - s_i^2}{n-p-1} = \frac{(n-p-t_i^2)\hat{\sigma}_0^2}{(n-p-1)} \qquad \text{and} \qquad \hat{\theta}_{2i} = s_i^2$$

are the REML estimates of θ_1 and θ_{2i} from which estimates of ω_i and σ^2 can be derived as

$$\hat{\omega}_i = \begin{cases} \dfrac{(n-p)(t_i^2-1)}{(n-p-t_i^2)(1-v_i)} & t_i^2 > 1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

and

$$\hat{\sigma}^2 = \begin{cases} \dfrac{(n-p-t_i^2)\hat{\sigma}_0^2}{(n-p-1)} & t_i^2 > 1 \\ \hat{\sigma}_0^2 & \text{otherwise,} \end{cases} \tag{5.4}$$

where $t_i^2 = s_i^2/\hat{\sigma}_0^2$ is the squared standardized residual (the internally Studentized residual) for unit i under model (5.1) and $\hat{\sigma}_0^2$ is the REML estimate of the error variance under model (5.1). The variance estimates (5.3) and (5.4) are exactly the REML estimates from fitting the alternative model (5.5). Hence, given t_i^2, the variance estimates can be computed without the fitting of the alternative model.

Thompson (1985) noted that the REML estimator $\hat{\sigma}_0^2$ differs from the corresponding ML estimator used by Cook et al. (1982), being larger by a factor $n/(n-p)$. Atkinson (1985) showed that $t_i^2 \leq n-p$, leading to an infinite estimate of $\hat{\omega}_i$ in the case that $t_i^2 = n-p$. The latter corresponds to all the residual variation being with one point; it is difficult imagine an example in which this condition may occur. The fraction $(n-p-t_i^2)/(n-p-1)$ in (5.4) can be viewed as a downward adjustment to the estimator of $\hat{\sigma}_0^2$ when the ith observation is suspected to be an outlier and the complementary $(n-1)$ observations are not outliers.

We formulate a VSOM for unit i in the Gaussian linear fixed effects analysis as a linear mixed model, i.e.

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{d}_i\delta_i + \boldsymbol{e} \\
&\sim N\left(\boldsymbol{X\beta}, \sigma^2(\omega_i \boldsymbol{d}_i\boldsymbol{d}_i' + \boldsymbol{I})\right) \\
&\equiv N\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H}_i\right),
\end{aligned} \tag{5.5}
$$

where $\boldsymbol{d}_i$ is the ith unit vector of length n with a value of 1 in the ith position and 0 elsewhere, δ_i is an unknown random coefficient with zero mean and variance $\omega_i\sigma^2$ for $\omega_i \geq 0$, $\boldsymbol{H}_i = \omega_i\boldsymbol{d}_i\boldsymbol{d}_i' + \boldsymbol{I}$ is a diagonal matrix with $\omega_i + 1$ corresponding to y_i and 1 elsewhere and $\boldsymbol{I}$ is an identity matrix of order n. The variance-covariance matrix for the data under model (5.5) is $\sigma^2(\omega_i\boldsymbol{d}_i\boldsymbol{d}_i' + \boldsymbol{I})$ with the variance of the ith unit inflated from σ^2 to $\sigma^2(\omega_i + 1)$. The model requires here that we estimate $\boldsymbol{\beta}$, σ^2 and ω_i for $i = 1, \ldots, n$. It important to note that as the model (5.5) is essentially a linear mixed model with δ_i as a random effect with variance ω_i, and it can be fitted with any standard software which allows for fitting of a linear mixed model to appropriate data. In our case we use GenStat (Welham and Thompson, 2000). This form of the model allows us to easily extend (5.5) to the standard linear mixed model which may have extra random effect terms in addition to δ_i (see Chapters 6-7). An alternative parametrization of the variance of the data is in terms of the variance components in the form of (2.5). We have not explored this variance parametrization in this thesis because is it unclear how it can be extended to the linear mixed model in the context of a VSOM.

5.1.1 Joint estimation of fixed and random effects

The estimators for $\boldsymbol{\beta}$ and δ_i can be obtained using the MMEs for model (5.5). These MMEs are the general linear mixed model MMEs (2.7) with $\boldsymbol{Z}$ replaced by $\boldsymbol{d}_i$, $\boldsymbol{R}$ replaced by $\boldsymbol{I}$ and $\boldsymbol{G}$ replaced by ω_i , i.e.

$$
\begin{bmatrix} \boldsymbol{d}_i'\boldsymbol{d}_i + \frac{1}{\omega_i} & \boldsymbol{d}_i'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{d}_i & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\delta}_i \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{d}_i'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{y} \end{bmatrix}. \tag{5.6}
$$

The solutions for $\boldsymbol{\beta}$ and δ_i follow from Lemma 2.1 in Chapter 2 and are given by

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{H}_i^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}_i^{-1}\boldsymbol{y} \\
\tilde{\delta}_i &= \omega_i\boldsymbol{d}_i'\boldsymbol{H}_i^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})
\end{aligned} \tag{5.7}
$$

$$= \omega_i \boldsymbol{d}_i' \boldsymbol{P}_i \boldsymbol{y}, \tag{5.8}$$

where

$$\boldsymbol{P}_i = \boldsymbol{H}_i^{-1} - \boldsymbol{H}_i^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}_i^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}_i^{-1}$$

and

$$\boldsymbol{H}_i^{-1} = \boldsymbol{I} - \boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{d}_i + \omega_i^{-1})^{-1}\boldsymbol{d}_i' \qquad \text{using Result A.1.}$$

Since $\boldsymbol{d}_i'\boldsymbol{d}_i = 1$ then

$$\boldsymbol{H}_i^{-1} = \boldsymbol{I} - \frac{\omega_i}{1+\omega_i}\boldsymbol{d}_i\boldsymbol{d}_i'.$$

The unknown variance parameter ω_i in $\boldsymbol{H}_i$ replaced by its REML estimate. The solution (5.8) follows directly from Lemma 2.1 (equation 2.12) with $\boldsymbol{G}$ replaced by ω_i and $\boldsymbol{Z}'$ replaced by $\boldsymbol{d}_i'$.

Note that the expression for $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 - \frac{\omega_i \hat{e}_i}{[1 + \omega_i(1 - v_i)]}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d}_i, \tag{5.9}$$

where the subscript 0 indicates estimates under model (5.1). Thus $\hat{\boldsymbol{\beta}}_0$ is the estimate of $\boldsymbol{\beta}$ under model (5.1), i.e. the ordinary least squares estimate of $\boldsymbol{\beta}$, $\hat{e}_i = y_i - \boldsymbol{X}\hat{\boldsymbol{\beta}}_0$ is the residual for the ith observation and $v_i = \boldsymbol{d}_i'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d}_i$. The equation (5.9) relates the weighted estimate $\hat{\boldsymbol{\beta}}$ to the unweighted estimate $\hat{\boldsymbol{\beta}}_0$ with the last term $\omega_i\hat{e}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d}_i/[1 + \omega_i(1 - v_i)]$ as the adjustment to $\hat{\boldsymbol{\beta}}_0$ when $\omega_i > 0$.

5.1.2 Variance parameter estimation

Thompson (1985) parameterized a VSOM in terms of θ_1 and θ_{2i} and constructed the REML log-likelihood function (5.2) and maximized it to obtain $\hat{\theta}_1$ and $\hat{\theta}_{2i}$, and hence $\hat{\omega}_i$. This construction of the REML log-likelihood function is applicable to balanced data structures such as the linear regression model. In our case we view (5.5) as a linear mixed model with a single random effect, with the variance of the ith unit as $\sigma^2(\omega_i + 1)$. This simplicity of the model makes it easy to extend it to unbalanced data structures. We deal with models that can handle unbalanced data in Chapters 6-7.

In the following we give three related formulations of the REML log-likelihood function for model (5.5). These formulations are useful in the derivations of the likelihood ratio and score test statistics we use to evaluate a VSOM in this chapter. We note that for a VSOM in linear regression the Thompson's REML log-likelihood function (5.2) could also be used but it can not be used in the standard linear mixed model for unbalanced data structures, such as the models we consider in Chapters 6-7.

We denote the variance parameters under model (5.5) as $\boldsymbol{\phi} = (\omega_i, \sigma^2)'$. Note that this vector of variance parameters is analogous to $\boldsymbol{\phi}$ as defined in § 2.3, since (5.5) is in linear mixed model form, with replaced $\boldsymbol{\kappa}$ by ω_i. Then the REML log-likelihood function for estimating $\boldsymbol{\phi}$ under the model (5.5) is given by

$$\begin{aligned} l_i(\omega_i, \sigma^2; \boldsymbol{y}) &= -\frac{1}{2}\Big\{(n-p)\log\sigma^2 + \log|\boldsymbol{H}_i| + \log|\boldsymbol{X}'\boldsymbol{H}_i^{-1}\boldsymbol{X}| \\ &\qquad + \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{H}_i^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\sigma^2}\Big\} \\ &= -\frac{1}{2}\Big\{(n-p)\log\sigma^2 + \log|\boldsymbol{H}_i| + \log|\boldsymbol{X}'\boldsymbol{H}_i^{-1}\boldsymbol{X}| + \frac{\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{y}}{\sigma^2}\Big\}, \end{aligned} \tag{5.10}$$

Equation (5.10) follows directly from (2.19) with $\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{y} = (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{H}_i^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$.

After some algebra it can be shown that REML log-likelihood function (5.10) can be written as

$$\begin{aligned} l_i(\omega_i, \sigma^2; \boldsymbol{y}) &= -\frac{1}{2}\Big\{(n-p)\log\sigma^2 + \log(1+\omega_i) + \log|\boldsymbol{X}'\boldsymbol{X}| + \log(1+\omega_i-\omega_i v_i) \\ &\qquad - \log(1+\omega_i) + \frac{(\boldsymbol{y}-\hat{\boldsymbol{y}}_0)'(\boldsymbol{y}-\hat{\boldsymbol{y}}_0) - s_i^2}{\sigma^2} + \frac{s_i^2}{\sigma^2[1+\omega_i(1-v_i)]}\Big\} \end{aligned} \tag{5.11}$$

$$\begin{aligned} &= -\frac{1}{2}\Big\{(n-p-1)\log\sigma^2 + \log|\boldsymbol{X}'\boldsymbol{X}| + \log\sigma^2[1+\omega_i(1-v_i)] \\ &\qquad + \frac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y} - s_i^2}{\sigma^2} + \frac{s_i^2}{\sigma^2[1+\omega_i(1-v_i)]}\Big\}. \end{aligned} \tag{5.12}$$

Furthermore the REML log-likelihood function (5.12) can also be expressed as a function of the REML log-likelihood function for model (5.1) in the following form

$$l_i(\omega_i, \sigma^2; \boldsymbol{y}) = l_0(\sigma^2; \boldsymbol{y}) - \frac{1}{2}\Big\{\log[1+\omega_i(1-v_i)] - \frac{s_i^2\omega_i(1-v_i)}{\sigma^2[1+\omega_i(1-v_i)]}\Big\} \tag{5.13}$$

where

$$l_0(\sigma^2; \boldsymbol{y}) = -\frac{1}{2}\left\{(n-p)\log\sigma^2 + \log|\boldsymbol{X}'\boldsymbol{X}| + \frac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}}{\sigma^2}\right\} \tag{5.14}$$

is the log-likelihood function for the null model (5.1). The attraction of (5.13) is that only the second term involving ω_i has to be maximized to obtain estimates for ω_i.

5.1.3 Use of a variance shift outlier model

We contend that a VSOM may be applicable in the following two situations: (i) if an observation is suspicious prior to examining the residuals, for example the investigator identifies the observation as a possible outlier before the analysis; (ii) screening for outliers in a data set without prior information about which observations might be outlying. In the latter situation we wish to identify all potential anomalous observations that warrant inclusion in the modelling process. In the thesis we generally consider the second situation.

Thompson (1985) suggested an index plot of the estimates $\hat{\omega}_i$ as a graphical display for identifying observations (one-at-a-time) with inflated measurement error variances and hence potential outliers. Thompson (1985) also suggests the calculation of $l_i(\hat{\boldsymbol{\phi}}_i; \boldsymbol{y})$ for each unit i so that the unit with the largest log-likelihood value can be investigated as an outlier, and demonstrates that this value occurs for the unit with the largest squared standardized residual, t_i^2. Hence the following modelling approach is suggested in Thompson (1985)

(i) find the largest t_i^2, say t_k^2, $i = k$,

(ii) determine $\hat{\sigma}_k^2(\hat{\omega}_k + 1)$, the variance of the observation with inflated variance, with $\hat{\omega}_k$ and $\hat{\sigma}_k^2$ obtained using (5.3) and (5.4) respectively.

(iii) obtain weighted fixed effects estimates $\hat{\boldsymbol{\beta}}$ from (5.9).

However, this procedure does not provide information on their statistical significance, i.e. we can not objectively determine which observations have large enough inflated variances to be down-weighted in the analysis. We feel it is helpful to have objective measures as to whether potential outliers could reasonably have arisen from the underlying model. In the following we develop likelihood ratio and score test statistics for determining the significance of the variance shift estimates. Since the variance shift estimates ($\hat{\omega}_i$'s) are estimates of variance parameters, the test statistics (likelihood ratio and score test statistics) we will develop are special cases of the tests given in Lemmas 2.2 and 2.3. The likelihood ratio and score test statistics are frequently used in linear mixed models to test for variance parameters, i.e. to assess the need for random effects in the model. However, under a VSOM the conclusions from the tests relate to the outlyingness status of the unit(s) or groups of units, i.e. our likelihood ratio and score test statistics serve a diagnostic purpose. In general, if an observation is found to be inconsistent with the model we may (i) investigate the point whether it is error in the data, (ii) check whether the distributional assumptions are correct and (iii) if both the situations (i) and (ii) are not plausible, then we may down-weight the observation.

Since $t_i^2/(n-p)$ is distributed as Beta$(1/2,(n-p)/2)$ (Cook and Weisberg, 1982, pp. 19), we could use this distribution to flag observations with inflated error variance in the simple case of a single outlier model. However, in more complex models such as linear mixed models or a VSOM for groups of outliers we discuss in § 5.7, the distribution is not known so we have to consider test statistics such as the score, likelihood ratio and Wald test statistics, which are complex functions of t_i^2. In this thesis we use both the likelihood ratio and the score tests for comparing the null model (5.1), and a variance shift outlier model (5.5) in order to determine whether the ith observation has inflated variance and is therefore a possible outlier. These tests are asymptotically equivalent under the null hypothesis (Cox and Hinkley, 1990, § 9.3). The likelihood ratio test requires the fitting of both the null and alternative models, while the score test only involves the score vector and information matrix under the null hypothesis, and variance-covariance parameter estimates obtained with the null model, and hence requires fitting the null model only, leading to fewer computations. Another attractive feature of the score test is that it is invariant under transformation of parameters.

5.2 Likelihood ratio tests for the variance shift parameter

5.2.1 Likelihood ratio test

To evaluate the evidence that the variance shift parameter ω_i is larger than zero corresponds to testing the hypotheses

$$H_0 : \omega_i = 0 \qquad \text{versus} \qquad H_A : \omega_i > 0. \tag{5.15}$$

Note that H_A is a one-sided hypothesis as the variance shift parameter ω_i, defined as the variance of δ_i, must remain positive. Then the likelihood ratio test statistic for testing the hypotheses in (5.15) is

$$LRT_i = -2\Big\{l_0(\hat{\boldsymbol{\phi}}_0;\boldsymbol{y}) - l_i(\hat{\boldsymbol{\phi}}_i;\boldsymbol{y})\Big\}, \tag{5.16}$$

where $\hat{\boldsymbol{\phi}}_0 = (0, \hat{\sigma}_0^2)'$ and $\hat{\boldsymbol{\phi}}_i = (\hat{\omega}_i, \hat{\sigma}^2)'$ are the variance estimates for null model and the ith VSOM, respectively, and $l_0(\hat{\boldsymbol{\phi}}_0;\boldsymbol{y})$ and $l_i(\hat{\boldsymbol{\phi}}_i;\boldsymbol{y})$ are the REML log-likelihood functions for $\boldsymbol{y}$ evaluated at $\hat{\boldsymbol{\phi}}_0$ and $\hat{\boldsymbol{\phi}}_i$ respectively. The subscript i denotes a variance shift outlier model.

An explicit expression for (5.16) can be derived by first evaluating the REML log-likelihood functions under the null model (5.14) and alternative model (5.13) separately. A further simplification of the REML log-likelihood function under the alternative model (model 5.13) is obtained by using $\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}/\hat{\sigma}_0^2 = n - p$ in $l_0(\hat{\sigma}_0^2;\boldsymbol{y})$, to give

$$l_i(\hat{\boldsymbol{\phi}}_i;\boldsymbol{y}) = l_0(\hat{\sigma}_0^2;\boldsymbol{y}) - \frac{1}{2}\left\{(n-p-1)\log\frac{(n-p-t_i^2)}{(n-p-1)} + \log t_i^2\right\} \tag{5.17}$$

where

$$l_0(\hat{\sigma}_0^2;\boldsymbol{y}) = -\frac{1}{2}\left\{(n-p)\log\hat{\sigma}_0^2 + \log|\boldsymbol{X}'\boldsymbol{X}| + (n-p)\right\}. \tag{5.18}$$

The term $(n-p-1)\log(n-p-t_i^2) - \log t_i^2$ decreases as t_i^2 increases from 1 to $n-p$. The maximum value of l_i thus corresponds to the largest squared standardized residual, t_i^2, so that an index plot of t_i^2 (or $\hat{\omega}_i$) will also highlight observations with excess error variance and a large value of LRT_i.

The likelihood ratio test statistic is given by

$$LRT_i = \begin{cases} (n-p-1)\log\dfrac{(n-p-1)}{n-p-t_i^2} - \log t_i^2, & t_i^2 > 1 \\ 0 & \text{otherwise,} \end{cases} \tag{5.19}$$

which is a monotonically increasing function of t_i^2 (Figure 5.1). LRT_i can be calculated directly given the squared standardized residuals t_i^2.

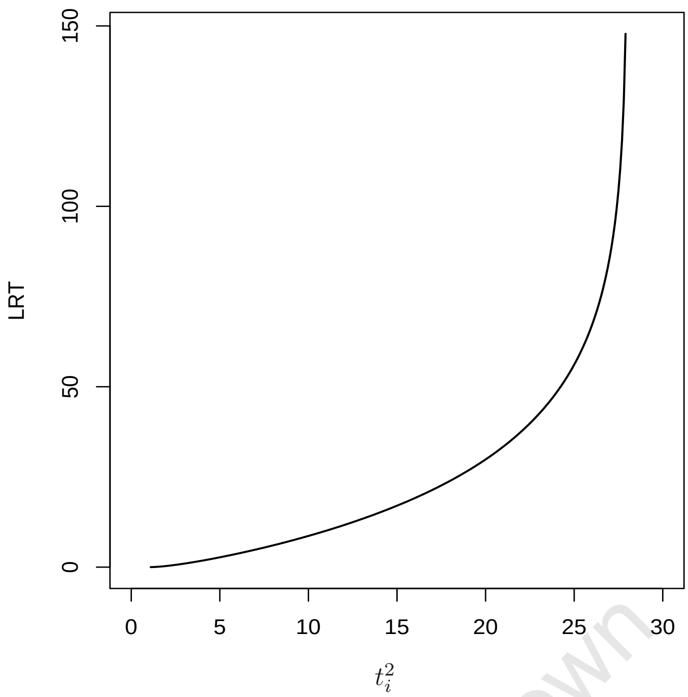

Figure 5.1: *Likelihood ratio test statistic as a function of* t_i^2 *for* $n = 30$ *and* $p = 2$ *assuming the linear model 5.1.*

We can use the relationship between t_i^2 and t_i^{*2}, introduced earlier in Chapter 4, namely,

$$t_i^{*2} = t_i^2 \left(\frac{n-p-1}{n-p-t_i^2} \right)$$

to express the LRT statistic, LRT_i, in terms of the (independent) externally Studentized residuals. However, this approach is not useful since as the LRT_i is again a complex function of t_i^{*2}. In § 5.4 we will investigate the distributional properties of LRT_i and propose a method for both calibrating this likelihood ratio test statistic and handling the problem of multiple testing.

5.2.2 One-step likelihood ratio tests

Christensen et al. (1992a) suggest the use of one-step updates for the variance parameter estimates in place of the REML estimates of the variance parameter in evaluating the likelihood function under the alternative model when using case-deletion to detect outliers in the linear mixed model; a strategy which reduces the amount of computation required. Since the linear regression VSOM is written as a linear mixed model in (2.1), the one-step LRT statistic may also be useful in testing for the significance of the variance shift estimate $\hat{\omega}_i$. In the case of the linear regression VSOM, the one-step LRT offers no computational advantage as we can compute the LRT directly using (5.19). However, the one-step LRT might be useful for more complex models where an analytical solution for the LRT statistic does not exist, for example, a VSOM for groups of outliers we discuss in § 5.7 and in extensions of a VSOM to linear mixed effects analysis in Chapters 6-7. Therefore we also investigate the one-step LRT statistic as an alternative to the LRT statistic in this simple case, in order to gain insight into more complex models.

The one-step likelihood ratio test statistic is usually constructed using one-step updates based on the expected information matrix, for instance in Christensen et al. (1992a). In our case we also consider three additional information matrices for obtaining one-step updates of the variance parameters: the observed, approximate average and exact (evenly-weighted) average information matrices. These matrices were introduced in Chapter 2 in the context of variance parameter estimation under the linear mixed model (Results 2.4 to 2.7). We consider the approximate average for computational efficiency. We expect the one-step LRT statistic based on the exact average information matrix to behave similarly to the one-step LRT statistic based on the approximate average information matrix if the approximate average information matrix gives a good approximation to the average of the observed and expected information matrices. Our motivation for also considering the observed information matrix is because of its involvement in the approximate average matrix, i.e. to assess whether the approximate average matrix approximates the trace terms in the observed information matrix adequately (see Result 2.4 in Chapter 2). Also some packages use updates of the variance parameters based on the observed information or have them as an option (for example SAS, Littell et al., 1996).

The likelihood ratio test statistic based on the one-step variance estimates of the ith VSOM is given by

$$LRT_{i(1)} = -2\Big\{ l_0(\hat{\boldsymbol{\phi}}_0; \boldsymbol{y}) - l_i(\hat{\boldsymbol{\phi}}_{i(1)}; \boldsymbol{y}) \Big\}, \tag{5.20}$$

where $l_i(\hat{\boldsymbol{\phi}}_{i(1)}; \boldsymbol{y})$ is the REML log-likelihood function for $\boldsymbol{y}$ evaluated at one-step variance estimates $\hat{\boldsymbol{\phi}}_{i(1)}$. We consider one-step estimates $\hat{\boldsymbol{\phi}}_{i(1)}$ calculated using a

Newton-Raphson type algorithm as

$$\hat{\boldsymbol{\phi}}_{i(1)} = \hat{\boldsymbol{\phi}}_0 + \boldsymbol{\mathcal{I}}_i(\hat{\boldsymbol{\phi}}_0)^{-1}\boldsymbol{U}_i(\hat{\boldsymbol{\phi}}_0) \tag{5.21}$$

where $\boldsymbol{U}_i(\hat{\boldsymbol{\phi}}_0)$ and $\boldsymbol{\mathcal{I}}_i(\hat{\boldsymbol{\phi}}_0)$ are the score vector and an information matrix respectively, evaluated at $\hat{\boldsymbol{\phi}}_0$. In computing (5.21), $\boldsymbol{\mathcal{I}}_i(\hat{\boldsymbol{\phi}}_0)$ is replaced by either the observed, expected, approximate average or exact average information matrix as defined in Results 5.1-5.4 below. In the following we describe the terms (that is, $\boldsymbol{U}_i$ and $\boldsymbol{\mathcal{I}}_i$) needed for the calculation of the one-step variance estimates (5.21) which are then used for the calculation of the likelihood ratio test statistic (5.20).

Score function for the variance shift parameter and information matrices

The score function for ω_i is obtained by differentiating (5.13) with respect to ω_i, i.e.

$$U_i(\omega_i) = \frac{1}{2}\left\{\frac{\hat{e}_i^2}{\sigma^2[(1-v_i)\omega_i+1]^2} - \frac{(1-v_i)}{[(1-v_i)\omega_i+1]}\right\}. \tag{5.22}$$

Thus evaluated at $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}_0$ where $\hat{\omega}_i = 0$ and $\sigma^2 = \hat{\sigma}_0^2$ the score function for ω_i is

$$U_i(\omega_i = 0) = \frac{(1-v_i)(t_i^2-1)}{2}, \tag{5.23}$$

where $t_i^2 = \hat{e}_i^2/\hat{\sigma}_0^2(1-v_i)$.

In the following we give the elements of the information matrices using results in Chapter 2 (Results 2.4 to 2.7).

Result 5.1 *The elements of the observed information matrix, $\boldsymbol{\mathcal{I}}_{\mathcal{O}i}$, for ω_i and σ^2, $i = 1, \ldots, n$, are*

$$\mathcal{I}_{\mathcal{O}}(\omega_i, \omega_i) = \frac{(1-v_i)^2}{2[1+\omega_i(1-v_i)]^2}\left\{\frac{2\hat{e}_i^2}{\sigma^2(1-v_i)} - 1\right\} \tag{5.24a}$$

$$\mathcal{I}_{\mathcal{O}}(\sigma^2, \omega_i) = \frac{\hat{e}_i^2}{2\sigma^4[1+\omega_i(1-v_i)]^2} \tag{5.24b}$$

$$\mathcal{I}_{\mathcal{O}}(\sigma^2, \sigma^2) = -\frac{(n-p)}{2\sigma^4} + \frac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}}{\sigma^6} - \frac{\hat{e}_i^2\omega_i}{\sigma^6[1+\omega_i(1-v_i)]}. \tag{5.24c}$$

Result 5.2 *The elements of the expected information matrix, $\boldsymbol{\mathcal{I}}_{\mathcal{E}i}$, for ω_i and σ^2, $i = 1, \ldots, n$, are*

$$\mathcal{I}_{\mathcal{E}}(\omega_i, \omega_i) = \frac{(1 - v_i)^2}{2[1 + \omega_i(1 - v_i)]} \tag{5.25a}$$

$$\mathcal{I}_{\mathcal{E}}(\sigma^2, \omega_i) = \frac{(1 - v_i)}{2\sigma^2[1 + \omega_i(1 - v_i)]} \tag{5.25b}$$

$$\mathcal{I}_{\mathcal{E}}(\sigma^2, \sigma^2) = \frac{(n - p)}{2\sigma^4}. \tag{5.25c}$$

The following two results are derived from Results 2.6 and 2.7 (in Chapter 2) respectively with the design matrix $\boldsymbol{Z}$ (in $\dot{\boldsymbol{H}}$) replaced by the covariate $\boldsymbol{d}_i$.

Result 5.3 *The elements of the approximate average information matrix, $\boldsymbol{\mathcal{I}}_{\mathcal{A}i}$, for ω_i and σ^2, $i = 1, \ldots, n$, are obtained by approximating the average of the observed and expected information matrix terms as in Gilmour et al. (1995):*

$$\mathcal{I}_{\mathcal{A}}(\omega_i, \omega_i) = \frac{\hat{e}_i^2(1 - v_i)}{2\sigma^2[1 + \omega_i(1 - v_i)]} \tag{5.26a}$$

$$\mathcal{I}_{\mathcal{A}}(\sigma^2, \omega_i) = \frac{\hat{e}_i^2}{2\sigma^4[1 + \omega_i(1 - v_i)]^2} \tag{5.26b}$$

$$\mathcal{I}_{\mathcal{A}}(\sigma^2, \sigma^2) = \frac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}}{2\sigma^6} - \frac{\hat{e}_i^2\omega_i}{2\sigma^6[1 + \omega_i(1 - v_i)]}. \tag{5.26c}$$

Result 5.4 *The elements of the exact average information matrix, $\boldsymbol{\mathcal{I}}_{\mathcal{A}ei}$, for ω_i and σ^2, $i = 1, \ldots, n$, (exact averages of the observed and expected information matrix terms) are*

$$\mathcal{I}_{\mathcal{A}e}(\omega_i, \omega_i) = \frac{\hat{e}_i^2(1 - v_i)}{2\sigma^2[1 + \omega_i(1 - v_i)]} \tag{5.27a}$$

$$\mathcal{I}_{\mathcal{A}e}(\sigma^2, \omega_i) = \frac{\hat{e}_i^2 + \sigma^2(1 - v_i)[1 + \omega_i(1 - v_i)]}{4\sigma^4[1 + \omega_i(1 - v_i)]^2} \tag{5.27b}$$

$$\mathcal{I}_{\mathcal{A}e}(\sigma^2, \sigma^2) = \frac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}}{2\sigma^6} - \frac{\hat{e}_i^2\omega_i}{2\sigma^6[1 + \omega_i(1 - v_i)]}. \tag{5.27c}$$

Gilmour et al. (1995) do not consider the exact average information matrix for reasons of computational convenience. However, we have found that the approximate average information matrix may lead to poor updates of the variance parameters and so investigate the exact average information matrix as well.

Denote the information matrix for the variance parameters ω_i and σ^2, with the notation for the type of information matrix suppressed, as

$$\boldsymbol{\mathcal{I}}_i = \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{bmatrix} \tag{5.28}$$

and its inverse as

$$\boldsymbol{\mathcal{I}}_i^{-1} = \begin{bmatrix} \mathcal{I}^{11} & \mathcal{I}^{12} \\ \mathcal{I}^{21} & \mathcal{I}^{22} \end{bmatrix}, \tag{5.29}$$

where $\mathcal{I}_{11} = \mathcal{I}(\omega_i, \omega_i)$ is the information matrix element for ω_i, $\mathcal{I}_{12} = \mathcal{I}(\omega_i, \sigma^2)$ corresponds to the information matrix element involving ω_i and σ^2, $\mathcal{I}_{21} = \mathcal{I}'_{12}$ and $\mathcal{I}_{22} = \mathcal{I}(\sigma^2, \sigma^2)$ is the information matrix term for σ^2. This conformable partitioning of $\boldsymbol{\mathcal{I}}_i$ will be generalized later in a VSOM for linear mixed effects analysis (see Chapter 6).

Under the null hypothesis, $H_0 : \omega_i = 0$, the observed, expected, approximate and exact average information matrices, under the null hypothesis, are respectively

$$\boldsymbol{\mathcal{I}}_{\mathcal{O}i}(\hat{\boldsymbol{\phi}}_0) = \begin{bmatrix} \dfrac{(1-v_i)^2}{2}\left\{\dfrac{2\hat{e}_i^2}{\sigma^2(1-v_i)} - 1\right\} & \dfrac{\hat{e}_i^2}{2\sigma^4} \\ \dfrac{\hat{e}_i^2}{2\sigma^4} & \dfrac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}}{\sigma^6} - \dfrac{(n-p)}{2\sigma^4} \end{bmatrix} \tag{5.30a}$$

$$\boldsymbol{\mathcal{I}}_{\mathcal{E}i}(\hat{\boldsymbol{\phi}}_0) = \begin{bmatrix} \dfrac{(1-v_i)^2}{2} & \dfrac{(1-v_i)}{2\sigma^2} \\ \dfrac{(1-v_i)}{2\sigma^2} & \dfrac{(n-p)}{2\sigma^4} \end{bmatrix} \tag{5.30b}$$

$$\boldsymbol{\mathcal{I}}_{\mathcal{A}i}(\hat{\boldsymbol{\phi}}_0) = \begin{bmatrix} \dfrac{\hat{e}_i^2(1-v_i)}{2\sigma^2} & \dfrac{\hat{e}_i^2}{2\sigma^4} \\ \dfrac{\hat{e}_i^2}{2\sigma^4} & \dfrac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}}{2\sigma^6} \end{bmatrix} \tag{5.30c}$$

$$\boldsymbol{\mathcal{I}}_{\mathcal{A}ei}(\hat{\boldsymbol{\phi}}_0) = \begin{bmatrix} \dfrac{\hat{e}_i^2(1-v_i)}{2\sigma^2} & \dfrac{[\sigma^2(1-v_i)+\hat{e}_i^2]}{4\sigma^4} \\ \dfrac{[\sigma^2(1-v_i)+\hat{e}_i^2]}{4\sigma^4} & \dfrac{\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}}{2\sigma^6} \end{bmatrix}. \tag{5.30d}$$

Note that the suffices (a)-(d) in the equation numbering distinguish the four different information matrices whereas in Results 5.1 to 5.4 the suffixes (a)-(c) distinguish the terms within each information matrix.

Evaluating the information structures under the null hypothesis at $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}_0$ where $\hat{\omega}_i = 0$ and $\sigma^2 = \hat{\sigma}_0^2$, the information matrices become

$$\boldsymbol{\mathcal{I}}_{\mathcal{O}i}(\hat{\boldsymbol{\phi}}_0) = \begin{bmatrix} \dfrac{(1-v_i)[2\hat{e}_i^2 - \hat{\sigma}_0^2(1-v_i)]}{2\hat{\sigma}_0^2} & \dfrac{\hat{e}_i^2}{2\hat{\sigma}_0^4} \\ \dfrac{\hat{e}_i^2}{2\hat{\sigma}_0^4} & \dfrac{(n-p)}{2\hat{\sigma}_0^4} \end{bmatrix} \tag{5.31a}$$

$$\boldsymbol{\mathcal{I}}_{\mathcal{E}i}(\hat{\boldsymbol{\phi}}_0) = \begin{bmatrix} \dfrac{(1-v_i)^2}{2} & \dfrac{(1-v_i)}{2\hat{\sigma}_0^2} \\ \dfrac{(1-v_i)}{2\hat{\sigma}_0^2} & \dfrac{(n-p)}{2\hat{\sigma}_0^4} \end{bmatrix} \tag{5.31b}$$

$$\boldsymbol{\mathcal{I}}_{\mathcal{A}i}(\hat{\boldsymbol{\phi}}_0) = \begin{bmatrix} \dfrac{\hat{e}_i^2(1-v_i)}{2\hat{\sigma}_0^2} & \dfrac{\hat{e}_i^2}{2\hat{\sigma}_0^4} \\ \dfrac{\hat{e}_i^2}{2\hat{\sigma}_0^4} & \dfrac{(n-p)}{2\hat{\sigma}_0^4} \end{bmatrix} \tag{5.31c}$$

$$\boldsymbol{\mathcal{I}}_{\mathcal{A}ei}(\hat{\boldsymbol{\phi}}_0) = \begin{bmatrix} \dfrac{\hat{e}_i^2(1-v_i)}{2\hat{\sigma}_0^2} & \dfrac{[\hat{\sigma}_0^2(1-v_i) + \hat{e}_i^2]}{4\hat{\sigma}_0^4} \\ \dfrac{[\hat{\sigma}_0^2(1-v_i) + \hat{e}_i^2]}{4\hat{\sigma}_0^4} & \dfrac{(n-p)}{2\hat{\sigma}_0^4} \end{bmatrix}. \tag{5.31d}$$

Updating schemes for one-step estimates of variance parameters

We propose two updating schemes for obtaining the one-step variance estimates required for evaluation of the one-step LRTs. These updating schemes must constrain the variance estimates to be positive, as this constraint defines the valid solution space.

Scheme A. Simultaneous updating of variance estimates

Step 1 Obtain one-step estimates of variance parameters using (5.21) with null model estimates $\hat{\boldsymbol{\phi}}_0 = (0, \hat{\sigma}_0^2)'$ as initial values. Label these updates as $\hat{\sigma}^2_{(1)}$ and $\hat{\omega}_{i(1)}$ respectively. If $\hat{\omega}_{i(1)} < 0$, then no updating is required and the one-step estimates of the variance parameters are replaced by their null model estimates, i.e. $\hat{\omega}_{i(1)} =$

0 and $\hat{\sigma}^2_{(1)} = \hat{\sigma}^2_0$.

Step 2 If $\hat{\sigma}^2_{(1)} < 0$, set $\hat{\sigma}^2_{(1)} = 10^{-4}$, and update ω_i given $\hat{\sigma}^2_{(1)} = 10^{-4}$. The value of $\hat{\sigma}^2_{(1)} = 10^{-4}$ is chosen to be very small (close to zero) but positive, and hence still within the appropriate parameter space of the error variance. This update computed as

$$\hat{\omega}_{i(1)} = 0 + \mathcal{I}(\omega_i, \omega_i)^{-1} U_i(\omega_i = 0),$$

with σ^2 replaced by 10^{-4} and $\mathcal{I}(\omega_i, \omega_i)$ is evaluated at $\omega_i = 0$.

Under the null hypothesis the score vector for all the variance parameters in the model is

$$\boldsymbol{U}_i\left(\hat{\boldsymbol{\phi}}_0\right) = \begin{pmatrix} \dfrac{(1-v_i)(t_i^2-1)}{2} \\ 0 \end{pmatrix}.$$

Note that the value for σ^2 is zero because of optimisation under the null model.

Then from step 1, the vectors of one-step estimates based on the observed, expected, approximate and exact average information matrices are respectively (if $t_i^2 > 1$)

$$\begin{aligned} \hat{\boldsymbol{\phi}}_{\mathcal{O}i(1)} &= \hat{\boldsymbol{\phi}}_0 + \boldsymbol{\mathcal{I}}_{\mathcal{O}_i}(\hat{\boldsymbol{\phi}}_0)^{-1} \boldsymbol{U}_i(\hat{\boldsymbol{\phi}}_0) \\ &= \begin{pmatrix} \dfrac{(n-p)(t_i^2-1)}{(1-v_i)[(n-p)(2t_i^2-1)-t_i^4]} \\ \dfrac{(n-p-t_i^2)(2t_i^2-1)\hat{\sigma}_0^2}{((n-p)(2t_i^2-1)-t_i^4)} \end{pmatrix}, \end{aligned} \tag{5.32a}$$

$$\begin{aligned} \hat{\boldsymbol{\phi}}_{\mathcal{E}i(1)} &= \hat{\boldsymbol{\phi}}_0 + \boldsymbol{\mathcal{I}}_{\mathcal{E}_i}(\hat{\boldsymbol{\phi}}_0)^{-1} \boldsymbol{U}_i(\hat{\boldsymbol{\phi}}_0) \\ &= \begin{pmatrix} \dfrac{(n-p)(t_i^2-1)}{(1-v_i)(n-p-1)} \\ \dfrac{(n-p-t_i^2)\hat{\sigma}_0^2}{(n-p-1)} \end{pmatrix}, \end{aligned} \tag{5.32b}$$

$$\hat{\boldsymbol{\phi}}_{\mathcal{A}i(1)} = \hat{\boldsymbol{\phi}}_0 + \boldsymbol{\mathcal{I}}_{\mathcal{A}_i}(\hat{\boldsymbol{\phi}}_0)^{-1}\boldsymbol{U}_i(\hat{\boldsymbol{\phi}}_0)$$

$$= \begin{pmatrix} \dfrac{(n-p)(t_i^2-1)}{t_i^2(1-v_i)(n-p-t_i^2)} \\ \dfrac{(n-p-2t_i^2+1)\hat{\sigma}_0^2}{(n-p-t_i^2)} \end{pmatrix} \tag{5.32c}$$

and

$$\hat{\boldsymbol{\phi}}_{\mathcal{A}ei(1)} = \hat{\boldsymbol{\phi}}_0 + \boldsymbol{\mathcal{I}}_{\mathcal{A}_{ei}}(\hat{\boldsymbol{\phi}}_0)^{-1}\boldsymbol{U}_i(\hat{\boldsymbol{\phi}}_0)$$

$$= \begin{pmatrix} \dfrac{4(n-p)(t_i^2-1)}{(1-v_i)[4t_i^2(n-p)-(t_i^2+1)^2]} \\ \dfrac{\hat{\sigma}_0^2[4t_i^2(n-p)-(t_i^2+1)^2]-2(t_i^4-1)\hat{\sigma}_0^2}{[4t_i^2(n-p)-(t_i^2+1)^2]} \end{pmatrix}. \tag{5.32d}$$

Figure 5.2 shows plots of $\log[(1-v_i)\hat{\omega}_i]$ (transformations of the REML shift variances) against $\log[(1-v_i)\hat{\omega}_i]$ for $n = 30$ and $p = 2$ assuming the model (5.1). The plots indicate that the one-step variance shift estimates produced by both the expected and approximate average information matrices are better approximations to the REML shift variances compared to those approximations obtained using the observed and exact average information matrices (Figure 5.2). The one-step estimates of σ^2 based on the expected information matrix coincide with the REML estimates of σ^2 (Figure 5.3). Both the exact and approximate average matrix produce one-step estimates of σ^2 which are consistently larger than their REML estimate counterparts, while the one-step updates of σ^2 based on the observed information matrix are always smaller than the REML estimates of σ^2 (Figure 5.3).

Figure 5.4 presents plots of $\log[(1-v_i)\hat{\omega}_{i(1)}]$ for the four different information matrices against t_i^2 values for $n = 30$ and $p = 2$ assuming the model (5.1). The figure shows, analytically, the behaviour of the one-step updates of the variance shift parameter under updating scheme A, step 1 for increasing values of t_i^2. The one-step updates are shown for $t_i^2 > 1$ since their analytical expressions (5.32a to 5.32d) are positive if and only if $t_i^2 > 1$. The region defined by $t_i^2 \leq 1$ does not concern us since we are interested in positive values of ω_i, i.e. for $t_i^2 \leq 1$ the one-step updates would be negative, and we suppress them. All the variance shift updates are increasing functions of t_i^2 with expected information matrix producing the largest one-step updates among the four information matrices, and the observed information and exact average information matrices giving similar one-step updates.

Figure 5.5 shows a corresponding plot for the one-step updates of σ^2 based on the four information matrices. All the different updates for σ^2 are decreasing functions of t_i^2. This relationship between $\sigma^2_{(1)}$ and t_i^2 reflects the role t_i^2 as an indicator of outlyingness in the data set. So the decreasing values of $\sigma^2_{(1)}$ are compensated by increasing variance shift estimates ($\omega_{i(1)}$) associated with increasing values of t_i^2. Figure 5.5 also exhibits an undesirable behaviour of the one-step updates based on the approximate average information matrix, i.e. the algorithm (updating scheme A, step 1) can produce invalid (negative) estimates of σ^2 for large values of t_i^2 when the approximate average information is used to perform the update. This property of the one-step update of σ^2 based on the average information matrix is evident from (5.32c), i.e.

$$\hat{\sigma}^2_{(1)} = \frac{(n-p-2t_i^2+1)\hat{\sigma}_0^2}{(n-p-t_i^2)}$$

which requires that $t_i^2 < (n-p+1)/2$ for $\hat{\sigma}^2_{(1)}$ to be positive. This constraint is stricter than the bound on the t_i^2, $t_i^2 \leq (n-p)$ given by Atkinson (1985).

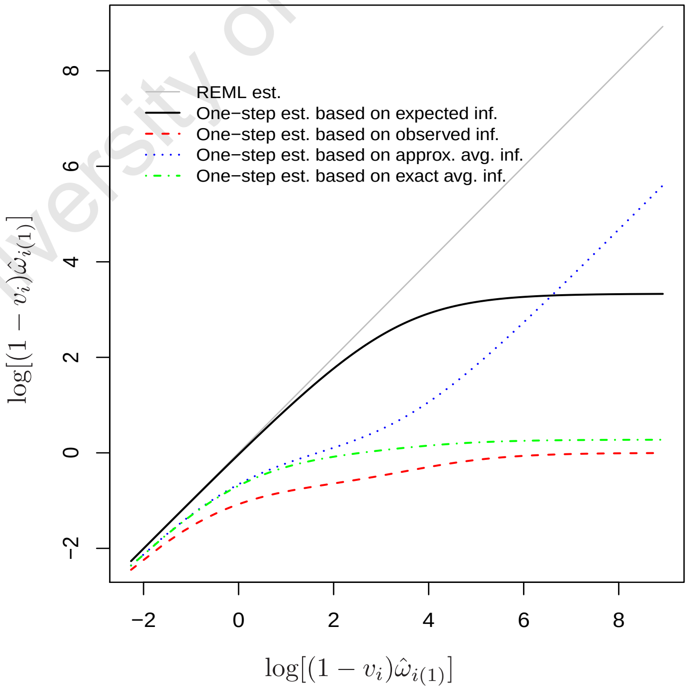

Figure 5.2: *Plots of* $\log[(1-v_i)\hat{\omega}_{i(1)}]$ *based on the four information matrices, against* $\log[(1-v_i)\hat{\omega}_i]$*, using updating scheme A, step 1: updating* σ^2 *and* ω_i *simultaneously: for* $n = 30$ *and* $p = 2$ *assuming the linear model 5.1.*

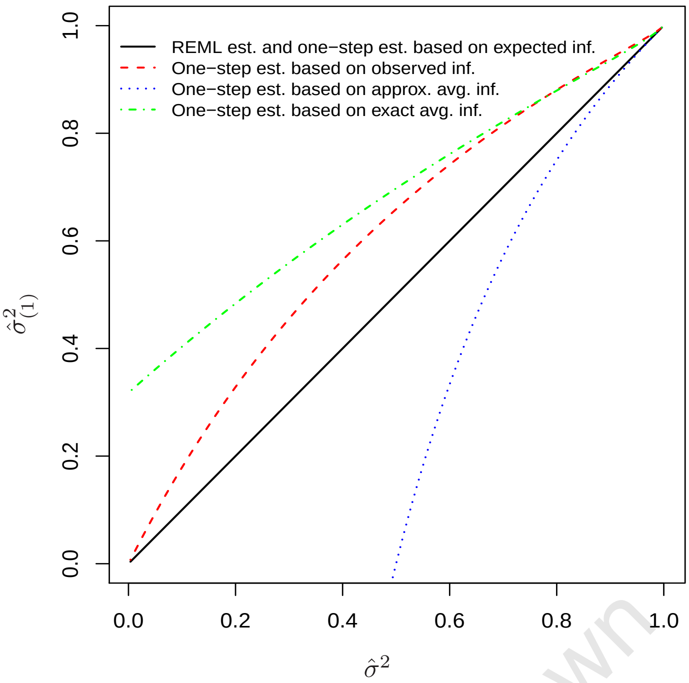

Figure 5.3: *Plots of one-step estimates $\hat{\sigma}^2_{(1)}$ based on the four information matrices, against REML estimates $\hat{\sigma}^2$, using updating scheme A, step 1: updating σ^2 and ω_i simultaneously: for $n = 30$ and $p = 2$ assuming the linear model 5.1.*

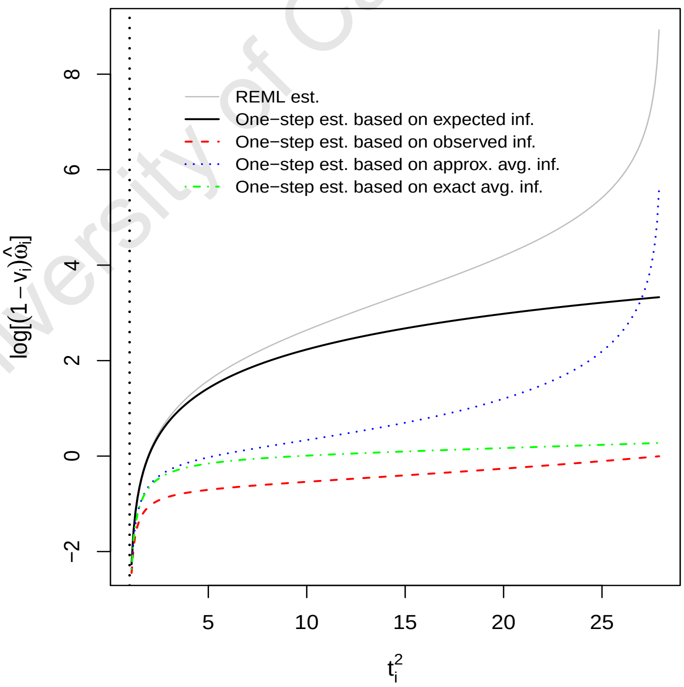

Figure 5.4: *Plots of* $\log[(1 - v_i)\hat{\omega}_i]$ *together with* $\log[(1 - v_i)\hat{\omega}_{i(1)}]$*, based on the four information matrices, against t_i^2 (with dotted line at $t_i^2 = 1$), using updating scheme A, step 1: updating σ^2 and ω_i simultaneously: for $n = 30$ and $p = 2$ assuming the linear model 5.1.*

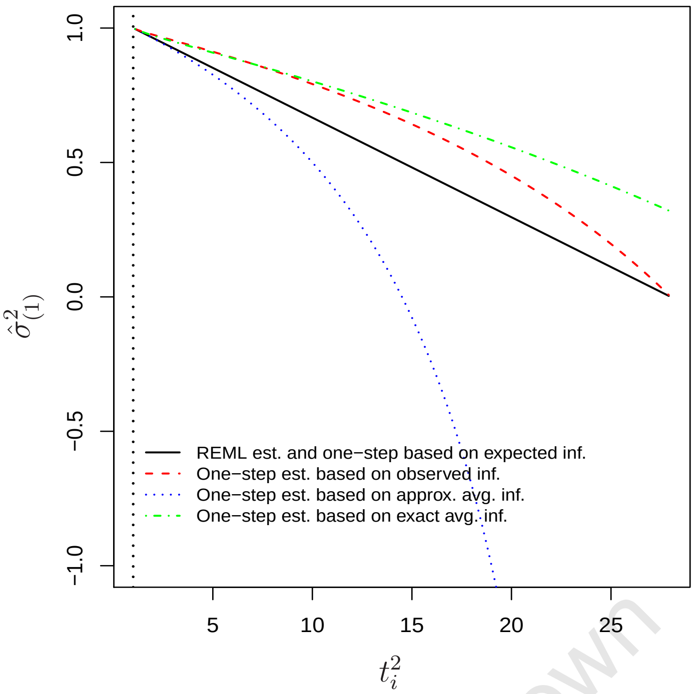

Figure 5.5: *Plots of one-step estimates* $\hat{\sigma}^2_{(1)}$ *based on the four information matrices, against* t_i^2 *(with dotted line at* $t_i^2 = 1$*), using updating scheme A, step 1: updating* σ^2 *and* ω_i *simultaneously: for* $n = 30$ *and* $p = 2$ *assuming the linear model 5.1.*

Using $\theta_1 = \sigma^2$ and $\theta_2 = [1+\omega_i(1-v_i)]\sigma^2$, as in Thompson (1985), and then replacing θ_1 and ω_i with respective one-step estimates (5.32a)-(5.32d) in the log-likelihood (5.2), gives corresponding one-step REML log-likelihood functions as

$$l_i(\hat{\boldsymbol{\phi}}_{\mathcal{O}i(1)}; \boldsymbol{y}) = -\frac{1}{2}\Bigg\{(n-p-1)\log\hat{\sigma}^2_{(1)} + \log|\boldsymbol{X}'\boldsymbol{X}| + \log\hat{\theta}_{2(1)} + \frac{(n-p-t_i^2)\hat{\sigma}_0^2}{\hat{\sigma}^2_{(1)}} + \frac{\hat{\sigma}_0^2 t_i^2}{\hat{\theta}_{2(1)}}\Bigg\}$$

$$= -\frac{1}{2}\Bigg\{(n-p)\log\hat{\sigma}_0^2 + \log|\boldsymbol{X}'\boldsymbol{X}| + (n-p) + \log\frac{[(n-p)(3t_i^2-2)-t_i^4]}{[(n-p)(2t_i^2-1)-t_i^4]} + (n-p)\log\frac{(n-p-t_i^2)(2t_i^2-1)}{[(n-p)(2t_i^2-1)-t_i^4]} + \frac{(n-p)^2 t_i^2(t_i^2-1)^2}{(n-p-t_i^2)(2t_i^2-1)[(n-p)(3t_i^2-2)-t_i^4]}\Bigg\},$$

$$l_i(\hat{\boldsymbol{\phi}}_{\mathcal{E}i(1)}; \boldsymbol{y}) = -\frac{1}{2}\Bigg\{(n-p)\log\hat{\sigma}_0^2 + \log|\boldsymbol{X}'\boldsymbol{X}| + (n-p)\log\frac{(n-p-t_i^2)}{(n-p-1)}$$

$$+ (n-p) + \log \frac{[(n-p)t_i^2 - 1]}{(n-p-1)} + \frac{(n-p)(t_i^2-1)^2}{(n-p-t_i^2)[(n-p)t_i^2-1]} \Bigg\},$$

$$l_i(\hat{\boldsymbol{\phi}}_{\mathcal{A}i(1)}; \boldsymbol{y}) = -\frac{1}{2}\Bigg\{ (n-p)\log\hat{\sigma}_0^2 + \log|\boldsymbol{X}'\boldsymbol{X}| + (n-p)$$
$$+ (n-p)\log\frac{(n-p-2t_i^2+1)}{(n-p-t_i^2)} + \log\frac{[(n-p)(2t_i^2-1)-t_i^4]}{t_i^2(n-p-t_i^2)}$$
$$+ \frac{(n-p)^2(t_i^2-1)^2}{(n-p-2t_i^2+1)[(n-p)(2t_i^2-1)-t_i^4]} \Bigg\} \tag{5.33a}$$

and

$$l_i(\hat{\boldsymbol{\phi}}_{\mathcal{A}ei(1)}; \boldsymbol{y}) = -\frac{1}{2}\Bigg\{ (n-p-1)\log\frac{2t_i^2(n-p-t_i^2-1)\hat{\sigma}_0^2}{[4t_i^2(n-p)-(t_i^2+1)^2]} + \log|\boldsymbol{X}'\boldsymbol{X}|$$
$$+ (n-p)\log\frac{[4t_i^2(n-p)-(t_i^2+1)^2]-2(t_i^4-1)}{[4t_i^2(n-p)-(t_i^2+1)^2]}$$
$$+ \log\frac{[4t_i^2(n-p)-(t_i^2+1)^2]}{[4t_i^2(n-p)-(t_i^2+1)^2+4(n-p)(t_i^2-1)]}$$
$$+ \frac{(n-p-t_i^2)[4t_i^2(n-p)-(t_i^2+1)^2]}{[4t_i^2(n-p)-(t_i^2+1)^2]-2(t_i^4-1)}$$
$$+ \frac{[4t_i^2(n-p)-(t_i^2+1)^2]^2 t_i^2}{[8t_i^2(n-p)-4(n-p)-(t_i^2+1)^2][4t_i^2(n-p)-(t_i^2+1)^2]-2(t_i^4-1)} \Bigg\}. \tag{5.33b}$$

Thus the one-step LRT statistics, based on the observed, expected, approximate and exact average information matrices, are given by

$$LRT_{\mathcal{O}i(1)} = (n-p)\log\frac{[(n-p)(2t_i^2-1)-t_i^4]}{(n-p-t_i^2)(2t_i^2-1)} - \log\frac{[(n-p)(3t_i^2-2)-t_i^4]}{[(n-p)(2t_i^2-1)-t_i^4]},$$
$$- \frac{(n-p)^2 t_i^2 (t_i^2-1)^2}{(n-p-t_i^2)(2t_i^2-1)[(n-p)(3t_i^2-2)-t_i^4]}, \tag{5.34a}$$

$$LRT_{\mathcal{E}i(1)} = (n-p)\log\frac{(n-p-1)}{(n-p-t_i^2)} - \log\frac{[(n-p)t_i^2-1]}{(n-p-1)}$$
$$- \frac{(n-p)(t_i^2-1)^2}{(n-p-t_i^2)[(n-p)t_i^2-1]}, \tag{5.34b}$$

$$
\begin{aligned}
LRT_{\mathcal{A}i(1)} = {} & (n-p)\log\frac{(n-p-t_i^2)}{(n-p-2t_i^2+1)} - \log\frac{[(n-p)(2t_i^2-1)-t_i^4]}{t_i^2(n-p-t_i^2)} \\
& - \frac{(n-p)^2(t_i^2-1)^2}{(n-p-2t_i^2+1)[(n-p)(2t_i^2-1)-t_i^4]}
\end{aligned} \tag{5.34c}
$$

and

$$
\begin{aligned}
LRT_{\mathcal{A}ei(1)} = {} & (n-p) + (n-p)\log\frac{[4t_i^2(n-p)-(t_i^2+1)^2]-2(t_i^4-1)}{[4t_i^2(n-p)-(t_i^2+1)^2]} \\
& - \log\frac{[4t_i^2(n-p)-(t_i^2+1)^2+4(n-p)(t_i^2-1)]}{[4t_i^2(n-p)-(t_i^2+1)^2]} \\
& - \frac{(n-p-t_i^2)[4t_i^2(n-p)-(t_i^2+1)^2]}{[4t_i^2(n-p)-(t_i^2+1)^2]-2(t_i^4-1)} \\
& - \frac{[4t_i^2(n-p)-(t_i^2+1)^2]^2 t_i^2}{[8t_i^2(n+p)-4(n-p)-(t_i^2+1)^2][4t_i^2(n-p)-(t_i^2+1)^2]-2(t_i^4-1)}.
\end{aligned} \tag{5.34d}
$$

One-step LRT statistics based on updates from step 2 can also be obtained in a similar way but are not given here. We note that the one-step LRT statistics based on the average information matrices (approximate and exact) and updates from step 2 will be identical because $\hat{\omega}_{\mathcal{A}i(1)}$ and $\hat{\omega}_{\mathcal{A}ei(1)}$ are equivalent since the terms $\mathcal{I}_{\mathcal{A}i}^{11}$ and $\mathcal{I}_{\mathcal{A}ei}^{11}$ are identical.

In the simple case of a single outlier, we have an explicit expression for each of the candidate LRTs (5.34a, 5.34b, 5.34c and 5.34d) as a function of t_i^2 and can evaluate their behaviour directly. When the one-step variance estimates are obtained from updating scheme A, both the one-step LRTs based on the expected information and the exact average information matrices increase as a function of t_i^2 while those based on the observed information and the approximate average information matrices again first increase and then decrease as t_i^2 increases (Figure 5.6). We do not want the LRT to decrease as t_i^2 increases, so the one-step LRTs based on the observed and approximate average information matrices are unacceptable. The one-step LRT based on the exact average information matrix behaves as expected. The one-step LRT based on the expected information matrix also behaves reasonably except in extreme cases which are very unlikely to occur even with very large outliers.

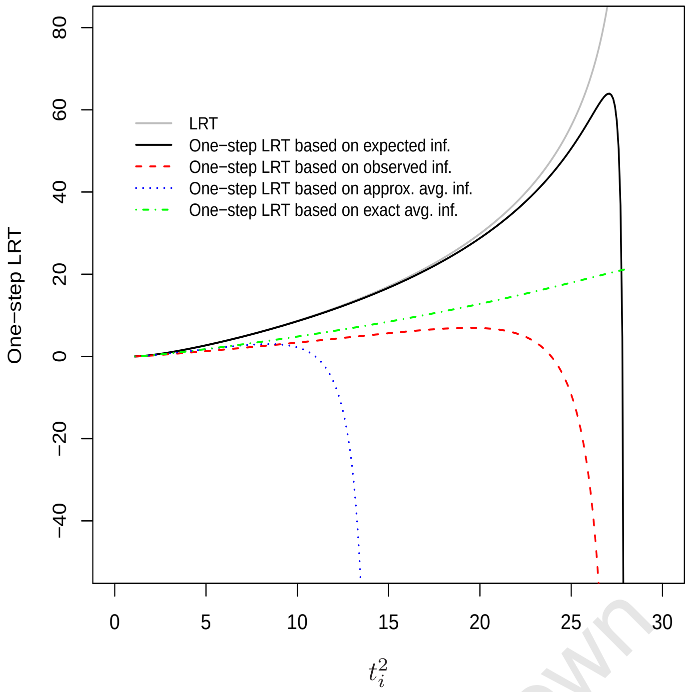

Figure 5.6: *Likelihood ratio test statistic together with one-step likelihood ratio test statistics as a function of t_i^2. One-step likelihood ratio test statistics are based on one-step updates of variance parameters obtained using updating scheme A, step 1: updating σ^2 and ω_i simu3ltaneously: for $n = 30$ and $p = 2$ assuming the linear model 5.1.*

Scheme B. Updating σ^2 first

Instead of updating σ^2 and ω_i simultaneously we consider updating σ^2 first before updating ω_i. We might expect this sequential updating to give better updates of ω_i.

Step 1 Update σ^2 directly using the score function and compute the update as

$$\hat{\sigma}^2_{(1)} = \frac{(n - p - t_i^2)\hat{\sigma}_0^2}{(n - p - 1)}, \tag{5.35}$$

with $\hat{\sigma}^2_{(1)}$ constrained to be positive since $t_i^2 \leq n - p$ and $\hat{\sigma}_0^2$ is strictly positive. Note that $\hat{\sigma}^2_{(1)}$ is equivalent to the REML error variance estimate $\hat{\sigma}^2$ (5.4), i.e. we can solve for $\hat{\sigma}^2$ in one-step. This update of σ^2 is also exactly equal to the error variance estimate when the ith unit is removed from the data set.

Step 2 Update ω_i given $\hat{\sigma}^2_{(1)}$ using

$$\hat{\omega}_i = \mathcal{I}^{11}(\omega_i = 0, \hat{\sigma}^2_{(1)}) U_i(\omega_i = 0, \hat{\sigma}^2_{(1)}),$$

with $\mathcal{I}^{11}$ replaced by the respective elements of the observed, expected, approximate average and exact average information matrices in (5.31a)-(5.31d).

In the following we give exact expressions for the one-step updates of ω_i based on the different information matrices. Given $\hat{\sigma}^2_{(1)}$ the score function takes the form

$$\begin{aligned} U_i &= \frac{(1-v_i)}{2}\left[\frac{(n-p-1)t_i^2}{(n-p-t_i^2)} - 1\right] \\ &= \frac{(n-p)(t_i^2-1)(1-v_i)}{2(n-p-t_i^2)}. \end{aligned}$$

Then the one-step updates of ω_i based on the different information matrices evaluated at $\hat{\sigma}^2_{(1)}$ are given by

$$\begin{aligned} \hat{\omega}_{\mathcal{O}i(1)} &= 0 + \mathcal{I}^{11}_{\mathcal{O}i} U_i \\ &= \begin{cases} \dfrac{(n-p)(t_i^2-1)}{(1-v_i)[(n-p)(2t_i^2-1)-t_i^2]}, & t_i^2 > 1 \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{5.36a}$$

$$\begin{aligned} \hat{\omega}_{\mathcal{E}i(1)} &= 0 + \mathcal{I}^{11}_{\mathcal{E}i} U_i \\ &= \begin{cases} \dfrac{(n-p)(t_i^2-1)}{(n-p-t_i^2)(1-v_i)}, & t_i^2 > 1 \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{5.36b}$$

and

$$\begin{aligned} \hat{\omega}_{\mathcal{A}i(1)} &= 0 + \mathcal{I}^{11}_{\mathcal{A}i} U_i \\ &= \begin{cases} \dfrac{(n-p)(t_i^2-1)}{(n-p-1)(1-v_i)t_i^2}, & t_i^2 > 1 \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{5.36c}$$

where $\mathcal{I}^{11}_i$ is the portion of the inverse of information matrix associated with ω_i evaluated at $\hat{\sigma}^2_{(1)}$. Again, similar to the updating scheme A, step 2, $\hat{\omega}_{\mathcal{A}i(1)}$ and $\hat{\omega}_{\mathcal{A}ei(1)}$ are equivalent since the terms $\mathcal{I}^{11}_{\mathcal{A}i}$ and $\mathcal{I}^{11}_{\mathcal{A}ei}$ are identical.

The one-step variance shift estimates ($\hat{\omega}_i$'s) based on the approximate average and observed information matrices are underestimates of the REML variance estimates (Figure 5.7). Note that under this algorithm (updating scheme B) the one-step variance shift estimates based on the expected information matrix coincide with the REML variance shift estimates, and the one-step updates based on the approximate average and exact average information matrices are identical.

Figure 5.8 shows the behaviour of the one-step updates of the variance shift parameter $\hat{\omega}_{i(1)}$ under updating scheme B for large values of t_i^2. Similar to updating scheme A, step 1, the one-step estimates based on the expected information are larger than those based on either the observed information or average information matrices.

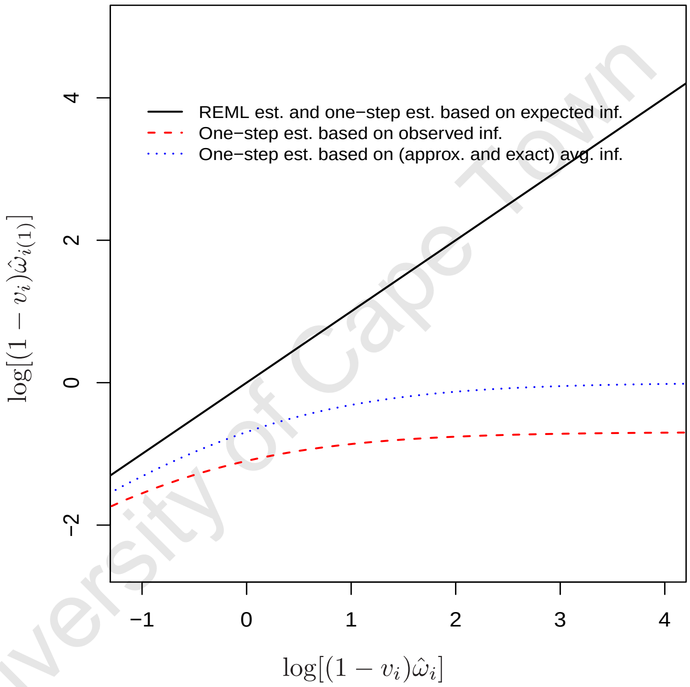

Figure 5.7: *Plots of* $\log[(1-v_i)\hat{\omega}_{i(1)}]$ *based on the three information matrices, against* $\log[(1-v_i)\hat{\omega}_i]$*, using updating scheme B: updating* σ^2 *first before updating* ω_i*: for* $n = 30$ *and* $p = 2$ *assuming the linear model 5.1.*

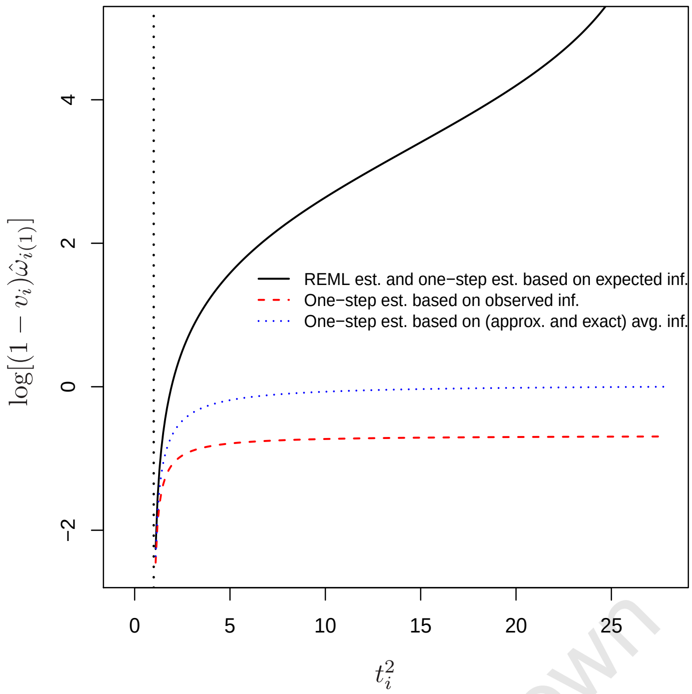

Figure 5.8: *Plots of* $\log[(1-v_i)\hat{\omega}_{i(1)}]$ *based on the three information matrices, against* t_i^2 *(with dotted line at* $t_i^2 = 1$*), using updating scheme B: updating* σ^2 *first before updating* ω_i*: for* $n = 30$ *and* $p = 2$ *assuming the linear model 5.1.*

Since the one-step updates based on the expected information matrix are identical to the REML estimates i.e. (5.4)=(5.35) and (5.3)=(5.36b) then the one-step LRT statistics based on the expected information and the LRT statistics are equivalent. The one-step LRTs based on the approximate average and exact average information matrices are also equivalent, since the information matrix elements corresponding to the variance shift parameter used in the updating are equivalent, as noted earlier.

Therefore we give expressions for the REML log-likelihood functions evaluated at the one-step updates obtained using the observed information (and approximate (or exact) average information matrices) only. The REML log-likelihood functions are obtained in a similar manner to those under updating scheme A, i.e. the respective one-step estimates (5.36a) and (5.36c) are plugged in the log-likelihood (5.2) to give the one-step REML log-likelihood functions

$$\begin{aligned} l_i(\hat{\boldsymbol{\phi}}_{\mathcal{O}i(1)}; \boldsymbol{y}) = & -\frac{1}{2}\Bigg\{(n-p-1)\log\hat{\sigma}^2_{(1)} + \log|\boldsymbol{X}'\boldsymbol{X}| + \log\hat{\theta}_{2(1)} \\ & \qquad + \frac{(n-p-t_i^2)\hat{\sigma}_0^2}{\hat{\sigma}^2_{(1)}} + \frac{\hat{\sigma}_0^2 t_i^2}{\hat{\theta}_{2(1)}}\Bigg\} \\ = & -\frac{1}{2}\Bigg\{(n-p-1)\log\frac{(n-p-t_i^2)\hat{\sigma}_0^2}{(n-p-1)} + \log|\boldsymbol{X}'\boldsymbol{X}| \end{aligned}$$

$$+ \log \frac{(n-p)(3t_i^2-2)-t_i^2}{(n-p)(2t_i^2-1)-t_i^2} + (n-p)$$
$$+ \left[\frac{t_i^2[(n-p-1)(2t_i^2-1)-t_i^2] - [(n-p-t_i^2)(3t_i^2-2)-t_i^2]}{(n-p-t_i^2)[(n-p)(3t_i^2-2)-t_i^2]} \right] \Bigg\} \tag{5.37a}$$

and

$$l_i(\hat{\boldsymbol{\phi}}_{\mathcal{A}i(1)}; \boldsymbol{y}) = -\frac{1}{2} \Bigg\{ (n-p-1) \log \frac{(n-p-t_i^2)\hat{\sigma}_0^2}{(n-p-1)} + \log |\boldsymbol{X}'\boldsymbol{X}| + \log \frac{(n-p-t_i^2)\hat{\sigma}_0^2}{(n-p-1)}$$
$$+ \log \frac{[(n-p)(2t_i^2-1)-t_i^2]}{(n-p-1)t_i^2} + (n-p)$$
$$+ \frac{(n-p-1)^2 t_i^4 - (n-p-t_i^2)[(n-p)(2t_i^2-1)-t_i^2]}{(n-p-t_i^2)[(n-p)(2t_i^2-1)-t_i^2]} \Bigg\} \tag{5.37b}$$

respectively.

The corresponding one-step LRT statistics are given by

$$LRT_{\mathcal{O}i(1)} = (n-p) \log \frac{(n-p-1)}{(n-p-t_i^2)} + \log \frac{[(n-p)(2t_i^2-1)-t_i^2]}{[(n-p)(3t_i^2-2)-t_i^2]}$$
$$- \frac{t_i^2[(n-p-1)(n-p)(2t_i^2-1)-t_i^2]}{(n-p-t_i^2)[(n-p)(3t_i^2-2)-t_i^2]}$$
$$- \frac{(n-p-t_i^2)[(n-p)(3t_i^2-2)-t_i^2]}{(n-p-t_i^2)[(n-p)(3t_i^2-2)-t_i^2]} \tag{5.38a}$$

and

$$LRT_{\mathcal{A}i(1)} = (n-p) \log \frac{(n-p-1)}{(n-p-t_i^2)} - \log \frac{[(n-p)(2t_i^2-1)-t_i^2]}{(n-p-1)t_i^2}$$
$$- \frac{(n-p-1)^2 t_i^4 - (n-p-t_i^2)[(n-p)(2t_i^2-1)-t_i^2]}{(n-p-t_i^2)[(n-p)(2t_i^2-1)-t_i^2]}. \tag{5.38b}$$

Figure 5.9 presents plots of the one-step LRTs, obtained using updating scheme B, against values of t_i^2. The plots show that whereas the LRT (which coincides with one-step LRT based on the expected information matrix under updating scheme B) increases as a function of t_i^2, both the one-step LRTs based on the observed and average information matrices first increase and then decrease as t_i^2 increases, becoming negative for large values of t_i^2.

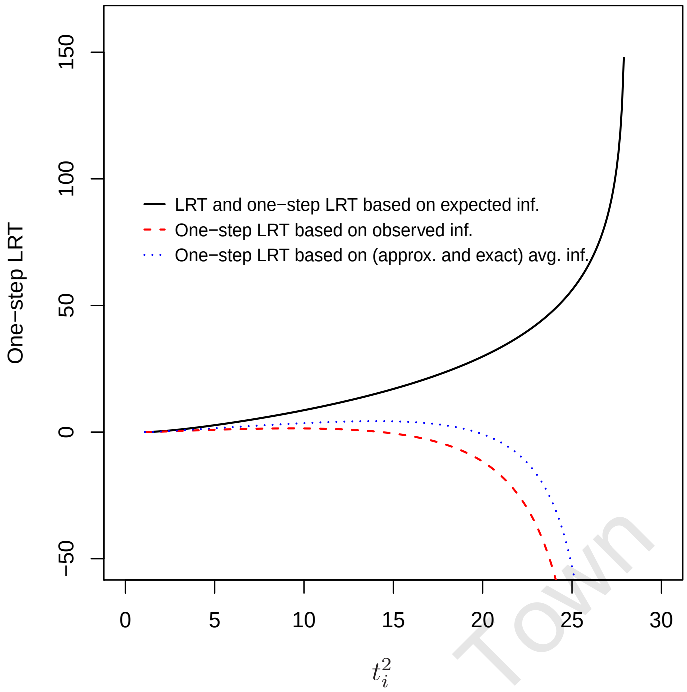

Figure 5.9: *One-step likelihood ratio test statistics as a function of* t_i^2*. One-step LRT statistics are based on one-step updates of variance parameters obtained using updating scheme B: updating* σ^2 *first before updating* ω_i*: for* $n = 30$ *and* $p = 2$ *assuming the linear model 5.1.*

Comparisons of one-step likelihood ratio tests obtained using the two updating schemes

Examination of the behaviour of the one-step updates $\hat{\omega}_{i(1)}$ against t_i^2, under each updating scheme (Figures 5.3 and 5.6) shows that the increase in the one-step updates for increasing values of t_i^2 is much smaller when the observed or average information matrix is used, suggesting that these matrices, which use the value $\hat{e}_i^2$, do not provide a good approximation to the likelihood surface for ω_i when the starting value (zero) is far from the final estimate. Correspondingly, Figure 5.4 shows drastic decreases in the one-step updates $\sigma^2_{(1)}$ for increasing values of t_i^2 when the approximate average information matrix is used. Conversely, the expected information matrix, which is independent of the $\hat{e}_i^2$, appears to provide a better update of ω_i (a better update of σ^2 in case of updating scheme A). Updating σ^2 first (scheme B) does not appear to have much impact in this context, i.e. behaviour of one-step updates based on the observed or average information matrix is not improved.

The decrease in the one-step LRT functions implies that these functions are inappropriate test statistics: large values of t_i^2, which should indicate outliers, can generate small (or even negative) values of the LRT.

5.3 Score tests for the variance shift parameter

The score test statistic (Cox and Hinkley, 1990, § 9.3) for testing $H_0 : \omega_i = 0$ against $H_A : \omega_i > 0$ takes the form

$$S_i(\omega_i = 0) = \begin{cases} U_i^2(\omega_i = 0)\mathcal{I}^{11} & U_i(\omega_i = 0) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.39}$$

where $\mathcal{I}^{11}$ is the portion of the inverse of information matrix associated with ω_i, evaluated under the null hypothesis. $\mathcal{I}^{11}$ is calculated as

$$\mathcal{I}^{11} = \left\{\mathcal{I}_{11} - \mathcal{I}_{21}\mathcal{I}_{22}^{-1}\mathcal{I}_{12}\right\}^{-1}. \tag{5.40}$$

$\mathcal{I}_{22}$, $\mathcal{I}_{12}$ and $\mathcal{I}_{21}$ are as defined in (5.28) and are evaluated at $\hat{\boldsymbol{\phi}}_0 = (\omega_i = 0, \hat{\sigma}_0^2)'$.

Similar to the one-step LRTs, we construct four types of score test statistics based on the four different information matrices. The standard score test statistic uses the expected information matrix and this option is the preferred information matrix in most statistical packages except for GenStat which uses the (approximate) average information matrix of Gilmour et al. (1995) for reasons of computational efficiency. The use of the observed information matrix, is common in constructing the score tests (for example Pawitan, 2001, pp. 245-250). We therefore consider score test statistics based on the expected, observed, approximate average information matrix, and on the exact average information matrix. We expect score test statistics based on the exact average information matrix to behave similarly to the score test based on the approximate average information matrix if the approximate average information matrix approximates the average of the observed and expected information matrices adequately.

The different score test statistics are constructed by replacing $\mathcal{I}^{11}$ in (5.39) with respective inverse observed, expected, approximate average and exact average information matrix elements, and $U_i^2(\omega_i = 0)$ by the square of (5.23) in (5.39) to obtain

$$S_{\mathcal{O}i}(\omega_i = 0) = U_i^2(\omega_i = 0)\mathcal{I}^{11}$$
$$= \begin{cases} \dfrac{(n-p)(t_i^2-1)^2}{2[(n-p)(2t_i^2-1)-t_i^4]}, & t_i^2 > 1 \\ 0 & \text{otherwise,} \end{cases} \tag{5.41a}$$

$$S_{\mathcal{E}i}(\omega_i = 0) = U_i^2(\omega_i = 0)\mathcal{I}^{11}$$
$$= \begin{cases} \dfrac{(n-p)(t_i^2-1)^2}{2(n-p-1)}, & t_i^2 > 1 \\ 0 & \text{otherwise,} \end{cases} \tag{5.41b}$$

$$S_{\mathcal{A}i}(\omega_i = 0) = U_i^2(\omega_i = 0)\mathcal{I}^{11}$$
$$= \begin{cases} \dfrac{(n-p)(t_i^2-1)^2}{2t_i^2(n-p-t_i^2)}, & t_i^2 > 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.41c}$$

and

$$S_{\mathcal{A}ei}(\omega_i = 0) = U_i^2(\omega_i = 0)\mathcal{I}^{11}$$
$$= \begin{cases} \dfrac{2(n-p)(t_i^2-1)^2}{[4t_i^2(n-p)-t_i^4-2t_i^2-1]}, & t_i^2 > 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5.41d}$$

Figure 5.10 plots the score test statistics (5.41a)-(5.41d) against values of t_i^2 for $n = 30$ and $p = 2$ assuming the model (5.1). The plot exhibits the behaviour of the score test statistics for increasing values of t_i^2. All the score test statistics increase as t_i^2 increases, as expected. The plot also illustrates the relative sizes of the score test statistics: both the exact average and expected information matrices appear to give larger score test statistics compared to the approximate average and observed information matrices.

The negativity of the observed information score test statistic has been discussed by

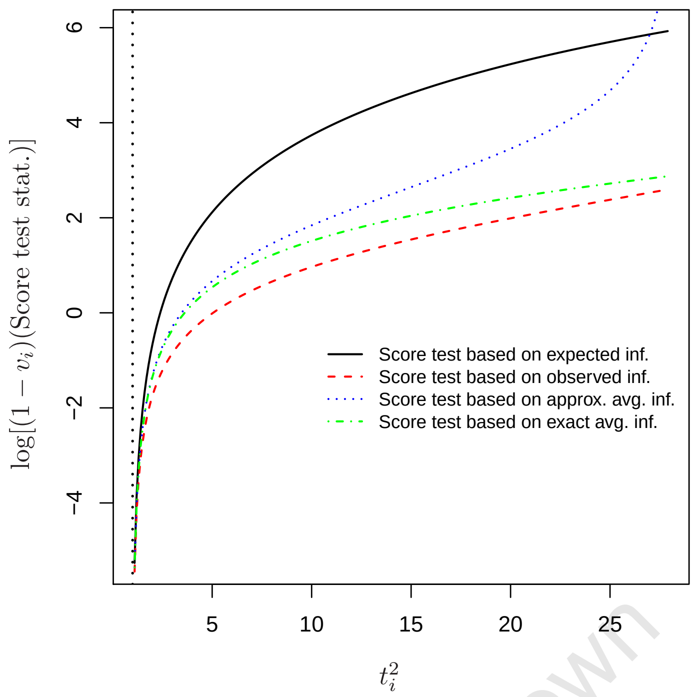

Figure 5.10: *Plots of* $\log[(1-v_i)(\textit{Score test stat.})]$, *for the four information matrices, against* t_i^2 *(with dotted line at* $t_i^2 = 1$*): for* $n = 30$ *and* $p = 2$ *assuming the linear model 5.1.*

several authors including Lawrance (1987), Godfrey and Orme (2001), and Yang and Abeysinghe (2003). These authors argue that this negativity of score test is caused by a poor fit of the null model. Morgan et al. (2007) give an example of this phenomenon for a zero-inflated Poisson model fitted to count data. Our context of using the score test statistic is different from that of Morgan et al. (2007) as we are using it for variance parameter testing in a linear mixed model framework. However, in more complex models such as a linear mixed VSOM, the argument for the poor fit of the model may be valid. We return to this point later in Chapter 6.

5.4 Assessing significance and multiple testing

5.4.1 Asymptotic null distributions of the likelihood ratio and score tests

The likelihood ratio test is known to have an approximate chi-squared distribution in large samples with degrees of freedom equal to the difference between the number of parameters in the two models (hypotheses). However, in the testing situation (5.15) the standard asymptotic theory no longer holds since the null hypothesis is on the boundary of the parameter space and regularity conditions are not met. The regularity conditions require that the score function exists and is differentiable in a small neighbourhood around the null hypothesis value. Using results of Self and Liang (1987), Stram and Lee (1994, 1995) showed that under certain conditions the asymptotic null distribution of the likelihood ratio test for testing the hypothesis of type (5.15) is a $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution, where χ_0^2 represents a distribution with a point mass at 0. Their results assumed that the data values are independently and identically distributed, or that the data set can be partitioned into a number of independent subsets such that the number of subsets increases with the size of the data set (see Crainiceanu and Ruppert, 2004). For a VSOM, these conditions clearly cannot be met, as we are selecting a single observation. Crainiceanu and Ruppert (2004) also showed the asymptotic approximations to be poor in the simple variance components model variance of which a VSOM can be regarded as a special case. Pinheiro and Bates (2000, pp. 84-87) simulated the distribution of the likelihood ratio test statistic and showed that the standard asymptotic distribution of the likelihood ratio test may underestimate the significance level for testing variance components in a linear mixed model. They reported that a $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution gave a reasonable approximation of the small sample distribution of the REMLRT for testing variance components in a linear mixed model.The absence of a satisfactory sampling distribution for the likelihood ratio test statistics for variance parameters in the linear mixed model prompted Crainiceanu and Ruppert (2004) to suggest a parametric bootstrap for the evaluation of the null distribution of the likelihood ratio test. Their method uses the spectral decomposition of the REMLRT statistics to derive the finite sample distributions. Scheipl et al. (2008) found that the finite sample distribution of the REMLRT statistic of Crainiceanu and Ruppert (2004) performed comparably to its counterpart, obtained by a parametric bootstrap method, in terms of both type I error and power, while being computationally faster. Greven et al. (2008) give two improvements of the finite sample distribution of the REMLRT statistic of Crainiceanu and Ruppert (2004) which use the method of moments and quantile regression. They showed that the improved finite sample distributions of the REMLRT statistic performed better than the finite sample distribution of the REMLRT statistic of Crainiceanu and Ruppert (2004) in terms of type I error.

The score test statistic also has an asymptotic chi-squared distribution under the null hypothesis, in line with the likelihood ratio test. Verbeke and Molenberghs (2003) develop one-sided score test statistics for variance parameter testing in linear mixed models and derive their null distributions. They showed that the results of Stram and Lee (1994; 1995) can be extended to the score test for one-sided alternatives for variance parameter testing in linear mixed models. They further generalized Silvapulle and Silvapulle (1995) theory for one-sided alternatives to variance parameter testing in the linear mixed model and proved that the likelihood ratio and score tests are equivalent under both one-sided and two-sided alternatives. Molenberghs and Verbeke (2007) revisit the problem of testing one-sided alternatives and review the likelihood ratio, score and Wald tests as these three tests are asymptotically equivalent. They also provide guidelines as to when each of the three tests is preferable. They propose the Wald tests for non-likelihood situations, such as models fitted using generalized estimating equations (GEE), and prefer the likelihood ratio test over the score test on computational grounds except when convergence is not achieved in fitting the alternative model. The use of the score test for variance parameter testing had also been investigated in the generalized linear mixed model context by several authors (Commenges and Jacqmin-Gadda, 1997; Lin, 1997; Hall and Praestgaard, 2001; Zhu and Zhang, 2006).

5.4.2 Finite sample distribution of the likelihood ratio and score test statistics for a VSOM

In the absence of a satisfactory reference distribution(s) for both the likelihood ratio and score tests, it behoves us to find the exact distributions of these test statistics. This task can be achieved by exploiting the analytical forms of the score tests since they all involve t_i^2, whose distribution is known (Cook and Weisberg, 1982, pp. 19).

A finite sample distribution of the likelihood ratio test statistic is difficult to obtain, although it can be evaluated by simulation. It can be established immediately (from $t_i^2/(n-p) \sim \text{Beta}(1/2, (n-p)/2)$) that the probability of a zero value for the LRT statistic LRT_i, under the null hypothesis with $\hat{\omega}_i = 0$, is

$$\Pr[t_i^2 \leq 1] = \Pr[t_i^2/(n-p) \leq 1/(n-p)]$$

which ranges from 0.707 for $n-p=2$ down to 0.68 as $(n-p) \to \infty$. This property of the LRT statistic gives us grounds for using the chi-squared mixture distribution, $0.68\chi_0^2 + 0.32\chi_1^2$ as a reference distribution for LRT_i. We evaluate this procedure in § 5.8.

Approximation to the finite sample distribution of the score test statistic

We could consider testing the null hypothesis $H_0 : \omega_i = 0$ against the alternative $H_A : \omega_i > 0$ using the score test statistic based on the expected information matrix given in equation (5.41b). This score test statistic is chosen because it has a simpler form compared to the other score test statistics. Define the test statistic

$$W_{ap} = \sqrt{S_{\mathcal{E}i}} = \begin{cases} \sqrt{\dfrac{(n-p)(t_i^2-1)^2}{2(n-p-1)}} \approx \dfrac{t_i^2-1}{\sqrt{2}}, & t_i^2 > 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5.42}$$

The approximate statistic W_{ap} gives a further simplification of the expected information score statistic. Hence we set out to derive the distribution of W_{ap}.

Let

$$\begin{aligned} \rho &= \Pr\left(t_i^2 \leq 1\right) \\ &= \Pr\left(\frac{t_i^2}{n-p} \leq \frac{1}{n-p}\right) \end{aligned}$$

with

$$\lim_{(n-p)\to\infty} \Pr\left(\frac{t_i^2}{n-p} \leq \frac{1}{n-p}\right) \approx 0.68.$$

For $W_{ap} > 0$ we have

$$\begin{aligned} \Pr(W_{ap} \leq w_{ap}) &= \Pr(t_i^2 \leq 1)\Pr(W_{ap} \leq w_{ap} \mid t_i^2 \leq 1) + \Pr\left(t_i^2 > 1\right)\Pr(W_{ap} \leq w_{ap} \mid t_i^2 > 1\big) \\ &= \rho + \Pr\left(W_{ap} \leq w_{ap} \cap t_i^2 > 1\right) \\ &= \rho + \Pr\left(\frac{t_i^2-1}{\sqrt{2}} \leq w_{ap} \cap t_i^2 > 1\right) \\ &= \rho + \Pr\left(1 \leq t_i^2 \leq w_{ap}\sqrt{2}+1\right). \end{aligned}$$

Let $W_{ap}^* = W_{ap}/(n-p)$, then using $t_i^2/(n-p)$ gives

$$\Pr(W_{ap}^* \leq w_{ap}^*) = \rho + \Pr\left(\frac{1}{n-p} \leq \frac{t_i^2}{n-p} \leq w_{ap}^*\sqrt{2}+1\right)$$

$$
\begin{aligned}
&= \rho + \mathcal{B}\left(w^*_{ap}\sqrt{2} + \frac{1}{n-p}\right) - F\left(\frac{1}{n-p}\right) \\
&= \rho + \mathcal{B}\left(w^*_{ap}\sqrt{2} + \frac{1}{n-p}\right) - \rho \\
&= \mathcal{B}\left(w^*_{ap}\sqrt{2} + \frac{1}{n-p}\right) \\
&\Rightarrow \Pr(W_{ap}) \leq w_{ap} = \mathcal{B}\left(\frac{w_{ap}\sqrt{2}+1}{n-p}\right)
\end{aligned}
$$

where $\mathcal{B}$ denotes the cumulative distribution function for the Beta$(1/2, (n-p)/2)$ distribution.

Hence

$$
\Pr(W^*_{ap} \leq w^*_{ap}) = \tau = \begin{cases} \mathcal{B}\left(w^*_{ap}\sqrt{2} + \frac{1}{n-p}\right), & w^*_{ap} > 0 \\ \rho & \text{otherwise.} \end{cases}
$$

Given τ, W_{ap} is determined as follows

$$
\begin{aligned}
&\mathcal{B}^{-1}(\tau) = w^*_{ap}\sqrt{2} + \frac{1}{n-p} \\
\Longleftrightarrow \quad & w^*_{ap} = \frac{\mathcal{B}^{-1}(\tau) - [1/(n-p)]}{\sqrt{2}}
\end{aligned}
$$

Thus

$$
w^*_{ap} = \begin{cases} \frac{(n-p)\mathcal{B}^{-1}(\tau) - 1}{\sqrt{2}}, & \tau > \rho \\ 0 & \text{otherwise} \end{cases} \tag{5.43}
$$

where $\rho \approx 0.68$ as $(n-p) \to \infty$.

Finite sample distribution of the score test statistic

Instead of using the approximate statistic W_{ap} we could consider the exact test statistic

$$W_f = \sqrt{S_{\mathcal{E}i}} = \begin{cases} \dfrac{\sqrt{(n-p)}(t_i^2 - 1)}{\sqrt{2(n-p-1)}} & t_i^2 > 1 \\ 0 & \text{otherwise.} \end{cases} \quad (5.44)$$

Following the development of the test statistic W_{ap} we obtain

$$W_f = \begin{cases} \dfrac{\sqrt{(n-p)}\,[(n-p)\mathcal{B}^{-1}(\tau) - 1]}{\sqrt{2(n-p-1)}}, & \tau > \rho \\ 0 & \text{otherwise} \end{cases} \quad (5.45)$$

where $\rho \approx 0.68$ as $(n-p) \to \infty$.

Having found the approximate or exact cumulative distribution functions of the expected information score test we can then obtain significance levels associated with the calculated score test statistics (those based on the expected information matrix).

Sampling distributions of likelihood ratio and score tests using the bootstrap

Since there is no satisfactory reference distribution(s) for either the LRT or the one-step LRTs or the score tests based on the observed, approximate average and exact average information matrices, like Crainiceanu and Ruppert (2004), we could use parametric bootstrap methods to simulate the null distributions for these test statistics.

The bootstrap was introduced by Efron (1979). Also see Efron and Tibshirani (1993). The original bootstrap is a distribution-free technique (also known as the non-parametric bootstrap) which permits the assessment of the variability of a random statistic using only the data at hand. The bootstrap distribution of the statistic is defined as its sampling distribution with the distribution function $\mathcal{F}$ replaced by the empirical distribution function $\widehat{\mathcal{F}}$, which puts probability mass $1/n$ at each observed data point. The bootstrap distribution is usually approximated by Monte Carlo methods by which samples (bootstrap samples) are repeatedly drawn from $\widehat{\mathcal{F}}$, i.e. from the set of data points.

An extension of the original bootstrap is the parametric bootstrap which entails drawing the bootstrap samples from a parametric model instead of $\widehat{\mathcal{F}}$. For the likelihood ratio test, the parametric bootstrap is implemented as follows. We first generate a data set of size n, fit the null model and obtain REML estimates of the variance parameters. We then draw a bootstrap sample from the null model in which all parameters have been replaced by their estimates, and then this pseudo-sample is analyzed once under H_0 and once under H_A, leading to a realized value of the likelihood ratio test. This process is repeated independently a number of times R, for R reasonable large, for example $R = 10000$, and the replicated values of the likelihood ratio test statistic give the null distribution of LRT_i. This distribution is then used to obtain an estimated significance level, corresponding to the specific value of LRT_i evaluated from the original sample. For the score test, we need only to fit the null model and then calculate the score statistic (5.23) and the inverses of the observed, expected, approximate average and exact average information matrices, evaluated under the null model, to obtain the score tests (5.41a)-(5.41d). The advantage of the parametric bootstrap compared to the non-parametric bootstrap is that it takes into account the uncertainty in estimation of the fixed effects as well variation in the random errors, through the re-estimation of all model parameters. Hence it mimics analysis of real data in which the true fixed effect values are unknown.

5.4.3 Handling multiple testing

In applying a VSOM, we can either calculate the score test statistic for any suspect observation(s), or fit the alternative model as a VSOM for the particular ith unit, $i = 1, \dots, n$ and then compute likelihood ratio test statistic. Then we wish to know how many and which observations have inflated variance in the data set. This question gives rise to the problem of multiple testing, i.e. simultaneously testing more than one hypothesis. This problem of multiple testing does not arise when the investigator picks out a potential outlier before the analysis.

The bootstrap methods of the previous section do not solve multiple testing problem since they give us a reference distribution for the selected test statistic with respect to a single observation, but we still wish to know how many observations are outlying. Multiple testing procedures for testing more than one hypothesis while controlling the overall or experiment-wise type I error rate were reviewed by Hochberg and Tamhane (1987). Benjamini and Hochberg (1995) introduced a procedure for testing m independent null hypotheses $H_1, H_2, \dots, H_m$ which is now widely used. Their procedure (the Benjamini-Hochberg or B-H multiple testing procedure) can be briefly described as follows.

Step 1 Obtain m observed p-values and order them as $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$

Step 2 Calculate $\hat{k} = \max\left\{k : p_{(k)} \leq \alpha k/m\right\}$ for a chosen value of α

Step 3 Reject null hypotheses corresponding to $p_{(1)}, p_{(2)}, \ldots, p_{(\hat{k})}$.

Storey (2002) argued that the B-H method does not give an error measure on $\hat{k}$ and hence may compromise the accuracy of the method. He further argued that multiple testing procedures (including the B-H procedure) do not use information in the data about the number of true null hypotheses in controlling the type I error rate. He proposed a multiple-testing method which fixes k and then estimates α instead of fixing α and then estimating k as in the B-H method. He showed that this procedure has greater power than the B-H method. A Bayesian extension of this procedure is given in Storey (2003).

Implementation of the B-H procedure in a VSOM approach would be problematic for three reasons. Firstly, the individual tests for $H_0 : \omega_i = 0$ may not be independent. Green and Diggle (2007) have attempted to address this problem of independence of the tests in the B-H procedure (also see Storey, 2007). Secondly, the B-H procedure also assumes the number of hypotheses m to be known in advance. In a VSOM approach, m is generally unknown if we are screening for outliers. Thirdly, the asymptotic distributions for the likelihood ratio and score tests are so poor that the p-values required as input in the B-H procedure are misleading. An alternative would be to use p-values obtained from simulated distributions of the likelihood ratio and score tests.

In this study we adopt a resampling method similar to that of Lin et al. (1993) to approximate the distribution of maximum score or likelihood ratio test statistic under the null hypothesis (see Rebai et al., 1994) in order to assess the significance of the score test statistics (or likelihood ratio test statistics). This approach handles the problem of multiple testing and ensures that the joint type I error rate is α across the n score test statistics (or likelihood ratio test statistics). Observations are then identified as outliers if their corresponding test statistics exceed the $100(1-\alpha)$th percentile of the simulated distribution, where α represents the desired overall type I error rate. The simulation required for the tests also allows calculation of a threshold for two (or r) outlying observations, based on the distribution of the second (rth) order statistic under the null hypothesis.

The resampling procedure is implemented as follows:

Step 1 For a given data set, fit an appropriate model (the null model).

Step 2 Generate the data vector $\boldsymbol{y}^*$ under the fitted model and fit the null model in step 1 to the simulated data.

Step 3 Computation of test statistics for the simulated data set.

Step 4 For each set of computed test statistics in step 3 compute the maximum statistic T^*. This statistic is the nth order statistic, so that the procedure allows for the calculation of the r order statistic whose sampling distribution can be used as a threshold for r outlying observations.

Step 5 Repeat steps 2 to 4 R times, for R reasonably large, for example $R = 10000$.

Step 6 Calculate the $100(1-\alpha)$th percentile of the R values of T^* for a given α.

The test statistics computed in step 3 can be any of the test statistics discussed earlier in this chapter, in § 5.2 and § 5.3.

Our resampling method is essentially a parametric bootstrap but differs from the method of Crainiceanu and Ruppert (2004) in three important respects. Firstly, their method was developed for the likelihood ratio statistic and does not apply directly to the score test statistic. Secondly, our method deals with the test statistic directly, for instance the LRT statistic (LRT_i), rather than working on its spectral decomposition as required in Crainiceanu and Ruppert's (2004) method. Finally, our procedure also deals with the problem of multiple testing which is a consequence of a VSOM approach to outlier detection.

To end this development, we suggest a procedure which could be followed in applying a VSOM approach to detecting single-case outliers in linear fixed effects analysis when there is no prior information about potential outliers. We implement this procedure in our examples below. A flow chart of this proposed procedure of a VSOM approach is given in Figure 5.11.

1 Fit the linear model (5.1) to a given data set and obtain the statistics t_i^2, $i = 1, \ldots, n$.

2 Fit a VSOM (model 5.5) for each unit with $t_i^2 > 1$, otherwise take the model fitted in the previous step as the final model, i.e. the model for all units with $t_i^2 \leq 1$.

3 Compute likelihood ratio and score test statistics for all units with $t_i^2 > 1$ to evaluate the evidence that $\omega_i > 0$. Units with $t_i^2 \leq 1$ all have zero likelihood ratio and score test statistics.

4 Generate sampling distributions of rth order statistics of the likelihood ratio and score tests statistics, e.g. the nth order statistics. Use these sampling distributions to assess the significance of the respective test statistics for a given significance level α.

5 Fit the final model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{D\delta} + \boldsymbol{e}$ where $\boldsymbol{D}$ is an $n \times r$ matrix composed of r vectors each with a value of 1 in the ith position and 0 elsewhere, $\boldsymbol{\delta}$ is an $r \times 1$ vector of random coefficients. r is the number of significant test statistics, i.e. the number of units identified as possible outliers. If none of the test statistics are significant take the fitted model in step 1 as the final model.

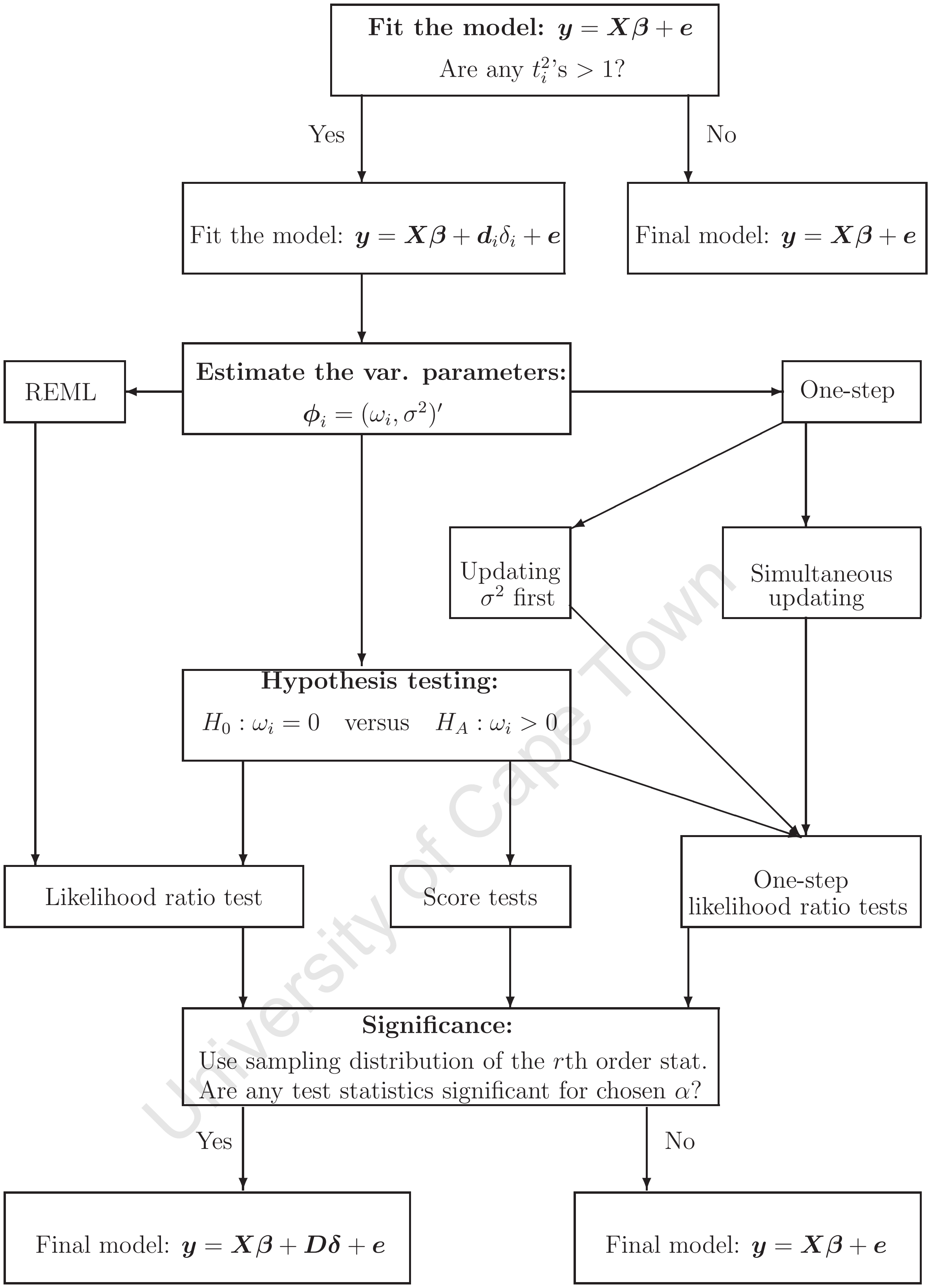

Figure 5.11: *Flow chart of a variance shift outlier model approach for detecting single-case outliers in linear regression with no prior information about potential outliers.*

In the next section a VSOM approach to detecting single-case outliers in linear regression will be illustrated using simulated data sets and for the orthodont data set we introduced in Chapter 2. For these data sets, a VSOM is fitted for each observation, and then likelihood ratio and score test statistics are computed and assessed for significance to determine whether particular observations are outliers.

5.5 Example: Simulated data

We generated data from the linear regression model

$$y_i = \mu + \beta x_i + e_i, \tag{5.46}$$

for $i = 1, \ldots, n; n = 30, 50$. We set $\mu = 20.5$ and $\beta = 0.25$. The values of x_i were drawn randomly from the uniform distribution on $[0, 10]$ and then sorted for $i = 1 \ldots n - 1$, with $x_n = 15$. The random errors e_i are simulated from a Gaussian distribution with mean zero and variance 1 for all units units except for $i = 10, n$, which used $e_i = 5$. This choice was intended to generate two outliers in each data set: one with high influence on the fitted line and one with low influence.

Figures 5.12 shows the scatter plots for the two simulated data sets together with the fitted regression lines and VSOMs for the inserted outliers. In both data sets the fitted line for a VSOM for unit 10 seems to be closer to the overall fitted linear regression line while the fitted line for a VSOM for the nth unit ($n = 30, 50$) has a different slope from that of the overall fitted linear regression line.

In the following we present the results for the simulation study. The ith VSOM was fitted for all units with $t_i^2 > 1$ giving the variance shift estimates $\hat{\omega}_i$ and $\hat{\sigma}^2$ as a by-product. Both $\hat{\omega}_i$ and $\hat{\sigma}^2$ could also be obtained directly using (5.3) and (5.4). Figure 5.12 also shows the fitted regression line together with the true underlying model and the fitted lines for VSOMs corresponding to units with $t_i^2 > 1$ for the sample sizes (a) $n = 30$ and (b) $n = 50$.

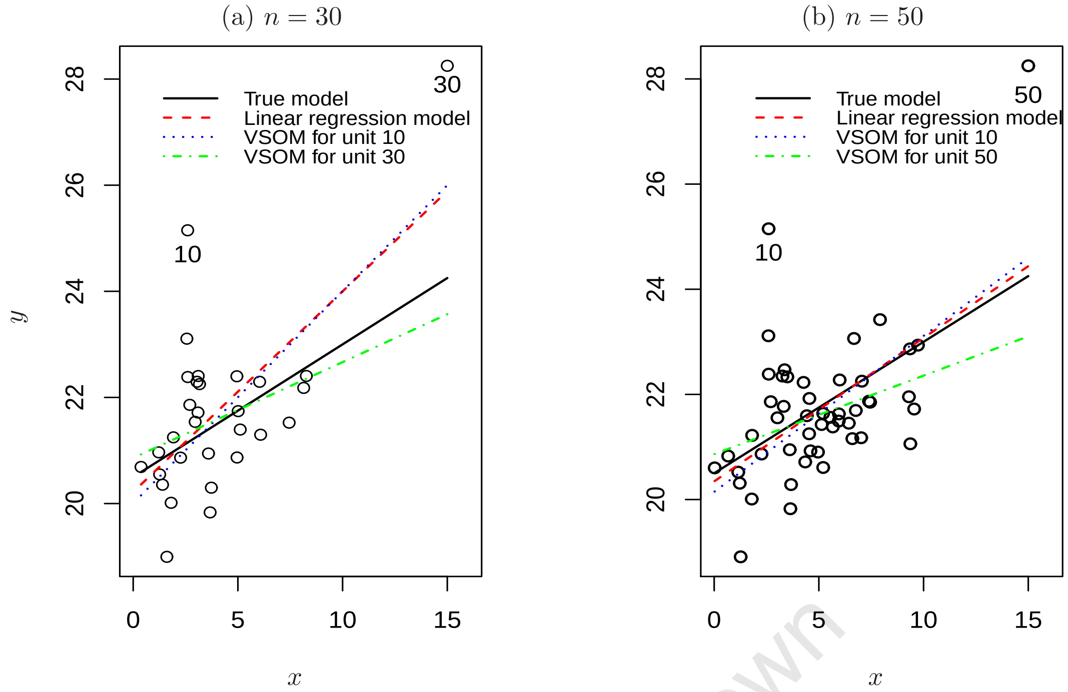

Figure 5.12: *Simulated data and fitted lines for simple linear regression models (a) $n = 30$: VSOMs for units 10 and 30 (b) $n = 50$: VSOMs for units 10 and 50.*

Table 5.1 shows summary statistics for the two data sets, $n = 30$ and $n = 50$. We show only REML estimates of the variance parameters for VSOMs fitted for the outlying units. One-step estimates based on average (approximate or exact) and observed information matrices are smaller than one-step updates based on the expected information matrix (not shown in Table 5.1). It must be noted that one-step estimates based on updating scheme B using the approximate average and exact average information matrix are equivalent. Furthermore, under updating scheme A, the one-step and the REML estimates of the error variance are equivalent when the expected information matrix is used (see equations equations (5.4) and (5.32b)). However, this equivalence does not hold in the case of the variance shift parameter ω_i. For instance, comparison of the expression in the REML estimate of ω_i (5.3) and the corresponding one-step estimates in (5.32b) shows that the denominator terms are respectively $(n-p-t_i^2)(1-v_i)$ and $(n-p-1)(1-v_i)$ (with the same numerator terms) so that the REML $\hat{\omega}_i$'s will always be smaller that their one-step counterparts when $t_i^2 > 1$.

The likelihood ratio and score test statistics, discussed earlier in this chapter, were calculated for each unit with $t_i^2 > 1$, for both data sets. The calculated values of t_i^2, variance parameter estimates, likelihood ratio and score test statistics are shown in Table 5.1 for the data sets with outliers inserted only, $n = 30, 50$. The proportions of

$t_i^2 > 1$ corresponds to the proportions of observations in each data set for which we fit VSOMs for individual observations.

In general the one-step LRT statistics based on the expected information were larger and closer to the LRT statistics than their observed and average information matrix counterparts. The expected information score test statistics were observed to be larger than those based on either the observed or average information matrices.

To implement the multiple testing resampling procedure of § 5.4.3, 10000 simulated data sets were generated from the fitted model under the null hypothesis. In each simulation, a VSOM was fitted for each observation and the three largest values of each test statistic were saved and used to generate the empirical distribution of the order statistics for each test. The 95th percentiles from the empirical distribution of the maximum and second largest value for each test statistic are shown in brackets in Table 5.1 under the value for each test statistic. As expected, for the data sets with no outliers inserted, no observation was detected as an outlier according the percentiles of the maximum and second largest value of each test statistic (calculated statistics not shown). However in the data sets with outliers inserted, all the outliers were identified as outliers on the basis of the 95th percentiles of the empirical distributions of the maximum and second largest test statistics (Table 5.1). It is important to note that even though the test statistics have different sizes, our resampling procedure allows us to detect outliers if they are present regardless of the size of the test statistic. For example for $n = 50$ both outliers (units 10 and 50) had LRT values of 10.94 and 13.32 being respectively larger than the 95th percentiles of the distributions of the second largest LRT and maximum LRT. The two percentiles are shown in brackets as 4.00 and 7.82, respectively. In practical situations, these units should be checked for identifiable errors, but if the anomalies cannot be explained, a point can be retained in the data set but down-weighted using the inflated variance $\hat{\sigma}^2(\hat{\omega}_i + 1)$ in contrast to $\hat{\sigma}^2$. This strategy might be considered preferable to omitting data in many contexts.

Table 5.1: *Values of t_i^2, variance parameter estimates, likelihood ratio and score test statistics for simulated data sets; for 10th and last observations: $n = 30, 50$. Figures in brackets are 95th percentiles of the empirical distributions for the first k order test statistics ($k = 1$, $k = 2$) under the null hypothesis.*

Statistic	$n = 30$		$n = 50$	
% of $t_i^2 > 1$	23.3		20	
	Obs. 10	Obs. 30	Obs. 10	Obs. 50
t_i^2	9.57	6.91	12.72	14.56
$\hat{\omega}_i$	13.58	16.56	16.52	26.62
$\hat{\sigma}^2$	1.16	1.33	1.02	0.97
LRT	8.05	4.74	10.94	13.32
	(6.98, 3.16)		(7.82, 4.00)	
One-step LRT: (Scheme A)				
Exp. inf.	7.98	4.72	10.90	13.26
	(6.93, 3.15)		(7.80, 3.99)	
Obs. inf.	3.18	2.05	4.25	5.03
	(2.82, 1.46)		(3.18, 1.79)	
Approx. avg. inf.	2.63	2.65	5.29	5.38
	(2.93, 1.97)		(4.34, 2.52)	
Exact avg inf.	4.55	2.89	6.12	7.27
	(4.03, 2.03)		(4.56, 2.53)	
One-step LRT: (Scheme B)				
Obs. inf.	1.84	1.51	2.96	3.18
	(1.77, 1.19)		(2.47, 1.57)	
Approx. avg. inf.	3.77	2.60	5.38	6.18
	(3.43, 1.89)		(4.16, 2.41)	
Score test				
Exp. inf.	38.09	18.14	70.16	93.87
	(30.78, 10.20)		(42.41, 15.56)	
Obs. inf.	2.47	1.57	3.26	3.88
	(2.19, 1.12)		(2.43, 1.375)	
Approx. avg. inf.	5.83	3.36	7.35	9.06
	(5.00, 2.25)		(5.20, 2.70)	
Exact avg. inf.	4.28	2.75	5.85	6.91
	(3.81, 1.94)		(4.38, 2.45)	

5.6 Example: The orthodont data

In order to illustrate the methods presented in this chapter we use the orthodont data set which was introduced in Chapter 2. Here we fit the simple linear regression model

$$y_i = \mu + \beta_1(age - 11) + \beta_2 sex + \beta_3(age - 11)gender + subject_i + e_i, \qquad (5.47)$$

where, for $i = 1, \ldots, n$; $n = 108$ is the total number of observations, y_i is distance of the ith subject, μ is the overall mean, $(age - 11)$ is the centred age, β_1 is effect of age, β_2 is the intercept shift for gender, β_3 is the interaction effect for age and gender, and subject$_i$ is the effect of child i compared to the first child. This model is fitted for illustrative purposes only since it does not represent the structure of the data, i.e. the repeated measurements on each subject violates the independence assumption that underlies the linear regression model. Pinheiro and Bates (2000, § 4.1) considered a simpler linear regression model for the orthodont data with only age as an explanatory variable. In this section will index the observations, which are assumed to be independent, as observations $1, \ldots, 64$ for boys and observations $65, \ldots, 108$ for girls.

After fitting the initial model, denoted $\mathcal{M}_0$, observations 34, 35, 49 and 52 stand out as possible outliers with all four observations giving relatively large values of t_i^2 (Figure 5.13). These observations were also identified as outliers in the previous chapter except for observation 52. The pair of observations 34 and 35 belong to boy 9 while the other pair (observations 49 and 52) belong to boy 13. Figure 5.13 also shows the estimates of the variance shift parameter, $\hat{\omega}_i$, and residual variance, $\hat{\sigma}^2$, for each observation under a VSOM; observations 34, 35, 49 and 52 have relatively large values of $\hat{\omega}_i$ together with decreased estimates of $\hat{\sigma}^2$.

The likelihood ratio and score test statistics were also calculated for each VSOM, and 10000 simulated data sets were generated from the fitted model under the null hypothesis. In each simulation, a VSOM was fitted for each observation and the four largest values of each test statistic were saved and used to generate the empirical distribution of the order statistics for each test. The test statistics from the original data are shown in Figure 5.14 together with 95th percentile from the empirical distribution of the maximum, second, third and fourth largest value for each test statistic. The test statistics for observation 35 are larger than the 95th percentile of the distribution of the maximum. The test statistics for observation 49 are larger than the 95th percentile of the distribution of the second largest value. For observation

34, the test statistics are also larger than the 95th percentile of the distribution of the third largest value. Observation 52 has the fourth highest test statistics and these values are larger than the 95th percentile of the distribution of the fourth largest value. We need to be cautious in making a conclusion, on the basis of this analysis, about the number of outliers present in the orthodont data set since we have fitted a linear regression model to repeated measures data. Nevertheless it appears there are four possible outliers in the orthodont data set, at observations 34, 35, 49 and 52.

Estimates under the initial model ($\mathcal{M}_0$) and the final model ($\mathcal{M}_1$), with VSOM terms for observations 34, 35, 49 and 52 are shown in Table 5.2. The value of the $-2\times$REML log-likelihood function (excluding constant terms) decreased from 220.38 in model $\mathcal{M}_0$ to 186.83 in model $\mathcal{M}_1$. Down-weighting these observations does not seem to have an effect on the estimates of the fixed effects but it does result in slight changes in the standard errors of the fixed effects estimates after fitting a VSOM. However, this reduction in the standard errors does not change the inference about the fixed effects.

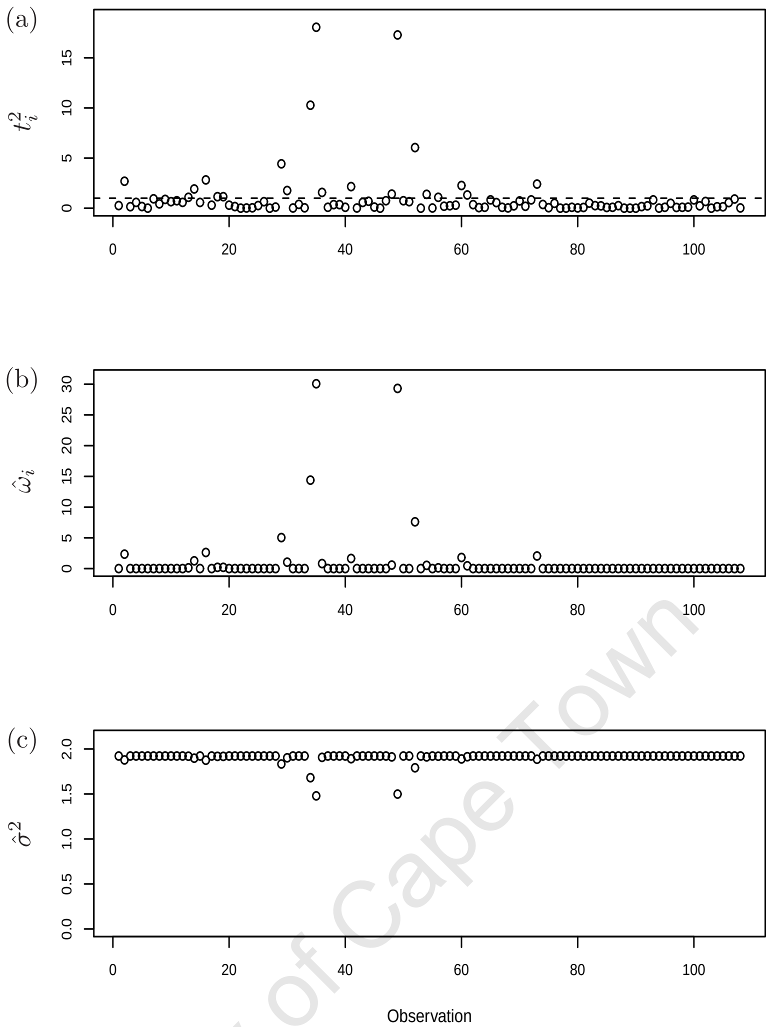

Figure 5.13: *Index plots of (a)* t_i^2 *(with dashed line at* $t_i^2 = 1$*), (b) REML variance shift estimates* $\hat{\omega}_i$ *and (c) REML error variance estimates* $\hat{\sigma}^2$ *for orthodont data.*

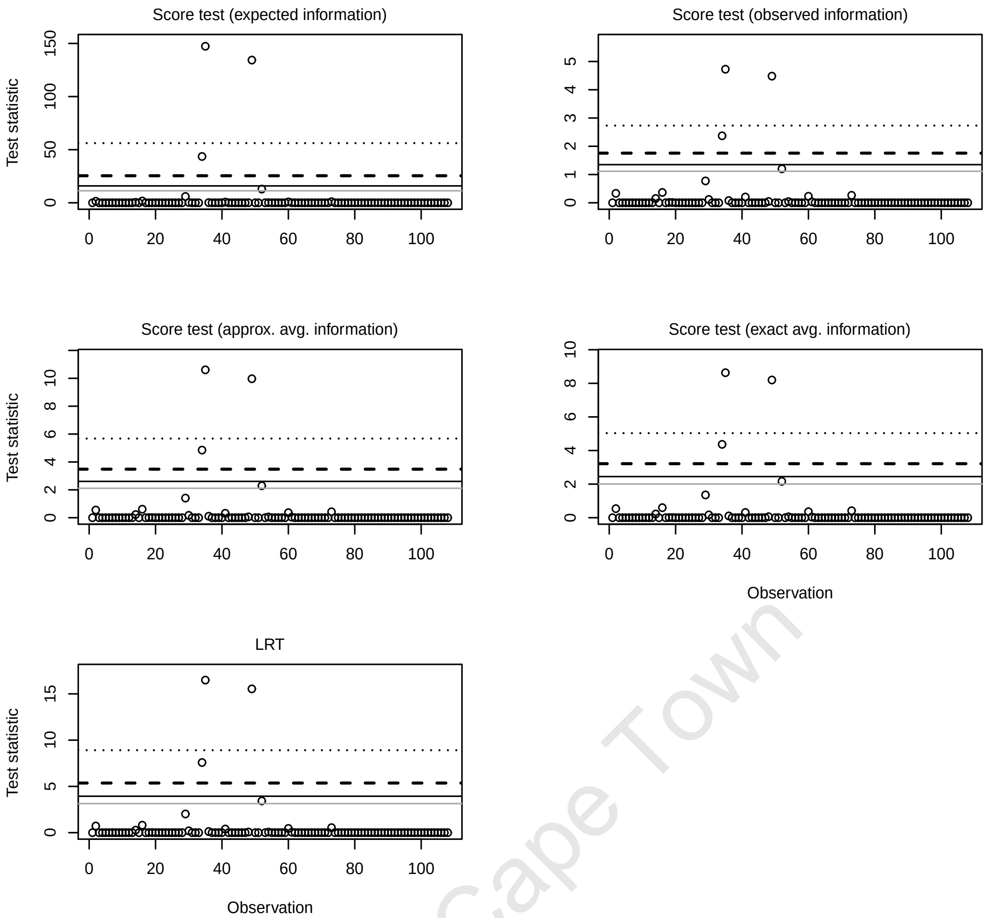

Figure 5.14: *Index plots of score test and likelihood ratio test statistics for the orthodont data, with 95th percentile of the empirical distribution under the null hypothesis shown for the first k order statistics for each test: $k = 1$ (dotted line), $k = 2$ (dashed line), $k = 3$ (solid line) and $k = 4$ (grey solid line).*

Table 5.2: *Estimated parameters for models fitted to orthodont data.*

Effect	Parameter♯	Model $\mathcal{M}_0$ Estimate (s.e.)	Model $\mathcal{M}_1$ Estimate (s.e.)
Fixed			
constant	μ	21.380 (0.693)	21.360 (0.492)
age	β_1	0.480 (0.093)	0.480 (0.066)
sex	β_2	6.375 (0.980)	6.375 (0.695)
sex.age	β_3	0.305 (0.121)	0.204 (0.087)
Random			
$\boldsymbol{d}_{34}$	$\omega_{34}\sigma^2$	-	10.089 (16.318)
$\boldsymbol{d}_{35}$	$\omega_{35}\sigma^2$	-	34.036 (50.174)
$\boldsymbol{d}_{49}$	$\omega_{49}\sigma^2$	-	39.626 (58.101)
$\boldsymbol{d}_{52}$	$\omega_{52}\sigma^2$	-	3.745 (7.385)
	σ^2	1.922 (0.306)	0.966 (0.158)

♯ Estimated subject effects not shown.

Comparison between VSOM in linear regression and case-deletion approaches

For comparison purposes we also considered the Cook's distance for variance parameters of Christensen et al. (1992a), which is defined as

$$D_i = (\hat{\boldsymbol{\phi}}_{(i)} - \hat{\boldsymbol{\phi}}_0)'\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\phi}}_0)(\hat{\boldsymbol{\phi}}_{(i)} - \hat{\boldsymbol{\phi}}_0), \tag{5.48}$$

where $\hat{\boldsymbol{\phi}}_0$ and $\hat{\boldsymbol{\phi}}_{(i)}$ are the variance parameter estimates under the null model and when the ith observation is deleted from the data set respectively, and $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\phi}}_0)$ is the information matrix for the variance parameters under the null model. This Cook's statistic is a measure of the change in the variance parameter estimates in the model when the ith observation is deleted from the data set, i.e. the influence of the ith observation on variance estimates. It is important to note that $\hat{\boldsymbol{\phi}}_0$ and $\hat{\boldsymbol{\phi}}_{(i)}$ are of the same dimension whereas in the case of a VSOM the two parameter vectors have different dimensions, for instance under VSOM in linear regression (5.5), $\hat{\boldsymbol{\phi}}_{(i)}$ has two parameters, namely ω_i and σ^2. Therefore D_i can not be used in a VSOM.

The statistic (5.48) is calibrated by using the χ^2_q distribution, where q is the dimension of $\hat{\boldsymbol{\phi}}_0$ (Christensen et al., 1992a). The basis for the use of the χ^2_q distribution as a reference distribution for D_i is questionable given that D_i uses estimates from two sets of data but only uses the information matrix from the fit of the null model, i.e. the Cook's statistic ignores the correlation between $\hat{\boldsymbol{\phi}}_0$ and $\hat{\boldsymbol{\phi}}_{(i)}$. Obenchain (1977) makes a similar point about the use of the Cook's distance for fixed effects as a test statistic for outliers in linear regression.

In the linear regression model, for instance our model (5.47), $\hat{\boldsymbol{\phi}}_0$ has only one parameter, σ^2 so that the Cook's distance becomes

$$D_i = \frac{2\hat{\sigma}_0^4(t_i^2 - 1)^2}{(n-p)(n-p-1)^2} \tag{5.49}$$

which should be distributed as χ^2_1 according to Christensen et al. (1992a). We note that D_i related to our score test statistic based on the expected information matrix (5.41), by the factor $(n-p)^2/[2\hat{\sigma}_0^4(n-p-1)]$.

We computed the statistic D_i for each observation for the orthodont data after fitting the null model (5.47). Observations, 34, 35, 49 and 52 were found to have relatively large Cook's statistic values compared to the rest of the data (Figure 5.15).

These observations are not picked out as outliers using the suggested χ_1^2 as a reference distribution. A sampling distribution for D_i could be obtained using bootstrap methods such as the resampling method we propose in § 5.4. We did not generate the sampling distribution of D_i in this study. Nevertheless for comparison purposes we deleted the 4 observations and fitted the model (5.47) to the reduced data set. The parameter estimates from the fitted model, denoted $\mathcal{M}_2$, are given in Table 5.3. The estimates and their standard errors were similar to those for a VSOM $\mathcal{M}_1$ (see Table 5.3). This is not surprising since a very large down-weighting under a VSOM implies that the case is essentially deleted and does not contribute to the analysis, i.e. case-deletion is the limiting case of a VSOM. However, we expect a VSOM to have a major advantage in situations in which it is not clear whether a case should be deleted or included in the analysis, since a VSOM allows for the down-weighting of the case. There is no corresponding case-deletion solution to this problem.

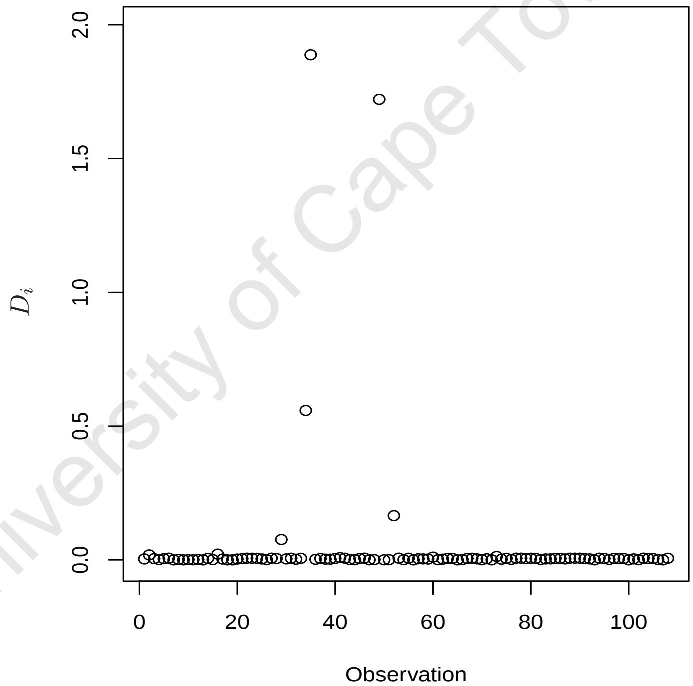

Figure 5.15: *Index plots of Cook's distance for the error variance for orthodont data.*

Table 5.3: *Estimated parameters for the case-deletion model*[†] *fitted to orthodont data.*

		Model $\mathcal{M}_2$
Effect	Parameter[♯]	Estimate (s.e.)
constant	μ	21.380 (0.492)
age	β_1	0.480 (0.066)
sex	β_2	6.375 (0.695)
sex.age	β_3	0.197 (0.087)
	σ^2	0.967 (0.158)

♯ Estimated subject effects not shown.
† Observations deleted: 34, 35, 49 and 52.

5.7 A VSOM for groups of outliers

In this section we briefly consider groups of outliers under a VSOM in linear regression. Since the focus of the thesis is the linear mixed model, we consider a VSOM for groups of outliers in detail later in Chapter 7 for models in which natural groups of outliers related to random effects may arise.

If a data set has multiple outliers, then the outliers may mask one another making outlier identification difficult. If masked outliers are not removed from the model as a group, their presence goes undetected. Consideration of all possible subsets to deal with masking is a computationally intensive task. The problems of detecting multiple outliers and masking has received wide-spread attention in the Bayesian literature on linear model diagnostics (for example Hoeting et al., 1996; Justel and Pena, 2001; Mohr, 2007). In this Section we attempt to deal with the problem of detecting groups of outliers and including them in the analysis when their inclusion can be justified.

The ith VSOM of the previous sections in this chapter adopted a one-at-a-time approach. In this section we extend the ith VSOM to a model that detects whether or not a specified group of observations is an outlying subset. The derivation for this model assumes that each affected unit has a separate outlier component (k random effects in $\boldsymbol{\delta}_I$) but with common general variance shift matrix $\boldsymbol{G}_I$. Other extensions with different assumptions are also possible. A VSOM for an arbitrary set of k VSOM outliers in the linear fixed effects model (VSOM$_k$) takes the form

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{D}_I\boldsymbol{\delta}_I + \boldsymbol{e} \\ &\sim N\left(\boldsymbol{X\beta}, \sigma^2(\boldsymbol{D}_I\boldsymbol{G}_I\boldsymbol{D}_I' + \boldsymbol{I})\right) \\ &\equiv N\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H}_I\right) \end{aligned} \tag{5.50}$$

which is the same as model (5.5) with the term $\boldsymbol{d}_i\delta_i$ replaced by $\boldsymbol{D}_I\boldsymbol{\delta}_I$ where I is an arbitrary subset of size k of $\{1, 2, \ldots, n\}$, $\boldsymbol{\delta}_I = \{\delta_i : i \in I\}$, $\boldsymbol{D}_I = \{\boldsymbol{d}_i : i \in I\}$ is an $n \times k$ matrix, $\boldsymbol{G}_I = \text{var}(\boldsymbol{\delta}_I)$ is $k \times k$ conformably with $\boldsymbol{\delta}_I$ and

$$\boldsymbol{H}_I = \boldsymbol{D}_I\boldsymbol{G}_I\boldsymbol{D}_I' + \boldsymbol{I}.$$

We note that $\boldsymbol{D}_I'\boldsymbol{D}_I = \boldsymbol{I}_k$, $\boldsymbol{H}_I$ is positive definite and $\boldsymbol{D}_I\boldsymbol{D}_I'$ is a diagonal matrix with 1 for $i \in I$ and 0 otherwise. $\boldsymbol{G}_I$ must be symmetric and positive definite for $\boldsymbol{G}_I^{-1}$ to exist but is otherwise unstructured in the general case and contains the parameters $\omega_{I1}, \ldots, \omega_{Im}$ where $m = k(k+1)/2$. Non-diagonal $\boldsymbol{G}_I$ allows for covariance among δ_i and hence y_i due to outliers, $i \in I$.

Note that for $\boldsymbol{G}_I$ positive definite (pd), we have

$$\begin{aligned} \boldsymbol{H}_I^{-1} &= \boldsymbol{I} - \boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{D}_I'\boldsymbol{D}_I)^{-1}\boldsymbol{D}_I' \\ &= \boldsymbol{I} - \boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{D}_I' \end{aligned}$$

and

$$\begin{aligned} (\boldsymbol{X}'\boldsymbol{H}_I^{-1}\boldsymbol{X})^{-1} &= [\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{D}_I'\boldsymbol{X}]^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}_I[\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k - \boldsymbol{V}_I]^{-1}\boldsymbol{D}_I'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &\qquad \text{using Result A.1,} \end{aligned}$$

where $\boldsymbol{V}_I = \boldsymbol{D}_I'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}_I$.

Hence

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{H}_I^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}_I^{-1}\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k - \boldsymbol{V}_I)^{-1}\boldsymbol{D}_I'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \end{aligned}$$

$$
\begin{aligned}
&- (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{D}_I'\boldsymbol{y} \\
&- (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k - \boldsymbol{V}_I)^{-1}\boldsymbol{V}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{D}_I'\boldsymbol{y} \\
= {}& \hat{\boldsymbol{\beta}}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k - \boldsymbol{V}_I)^{-1}\boldsymbol{D}_I'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&- (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{D}_I'\boldsymbol{y} \\
&- (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k - \boldsymbol{V}_I)^{-1}\boldsymbol{V}_I(\boldsymbol{G}_I^{-1} + \boldsymbol{I}_k)^{-1}\boldsymbol{D}_I'\boldsymbol{y} \\
= {}& \hat{\boldsymbol{\beta}}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}_I[\boldsymbol{V}_I - \boldsymbol{I}_k - \boldsymbol{G}_I^{-1}]^{-1}\hat{\boldsymbol{e}}_I,
\end{aligned}
$$

where $\hat{\boldsymbol{e}}_I = \boldsymbol{D}_I'(\boldsymbol{y} - \hat{\boldsymbol{y}}_0)$.

The solution for $\boldsymbol{\delta}_I$ is given by

$$\tilde{\boldsymbol{\delta}}_I = \boldsymbol{\omega}_I\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{y},$$

where $\boldsymbol{P}_I = \boldsymbol{H}_I^{-1} - \boldsymbol{H}_I^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}_I^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}_I^{-1}$.

The REML log-likelihood function for a VSOM$_k$ involves the matrix equivalents of the terms in the log-likelihood function (5.10) and is given by

$$l_I(\boldsymbol{\omega}_I, \sigma^2; \boldsymbol{y}) = -\frac{1}{2}\Big\{(n-p)\log\sigma^2 + \log|\boldsymbol{H}_I| + \log|\boldsymbol{X}'\boldsymbol{H}_I^{-1}\boldsymbol{X}| + \boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{y}/\sigma^2\Big\}, \quad (5.51)$$

with $\boldsymbol{H}_i$ and $\boldsymbol{P}_i$ replaced by $\boldsymbol{H}_I$ and $\boldsymbol{P}_I$ respectively.

Result 5.5 *The score function for ω_{Ii} for $i = 1, \ldots, k(k+1)/2$*

$$U_i(\omega_{Ii}) = -\frac{1}{2}\Big\{\mathrm{tr}(\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I') - \boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{y}/\sigma^2\Big\},$$

where $\dot{\boldsymbol{H}}_{I_i} = \frac{\partial \boldsymbol{H}_I}{\partial\omega_{Ii}} = \boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'$ and $\dot{\boldsymbol{G}}_{I_i} = \frac{\partial\boldsymbol{G}_I}{\partial\omega_{Ii}}$.

Result 5.6 *The score function for σ^2 is given by*

$$U(\sigma^2) = -\frac{1}{2}\Big\{\frac{(n-p)}{\sigma^2} - \frac{\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{y}}{\hat{\sigma}_0^4}\Big\}.$$

Result 5.7 *The elements of the observed information matrix for ω_{Ii} and σ^2 are*

$$\mathcal{I}_{\mathcal{O}}(\omega_{Ii},\omega_{Ij}) = -\frac{1}{2}\text{tr}\left(\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_j}\boldsymbol{D}_I'\right) + \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_j}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{y} \quad (5.52a)$$

$$\mathcal{I}_{\mathcal{O}}(\sigma^2,\omega_{Ii}) = \frac{1}{2\sigma^4}\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{y} \quad (5.52b)$$

$$\mathcal{I}_{\mathcal{O}}(\sigma^2,\sigma^2) = -\frac{(n-p)}{2\sigma^4} + \frac{\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{y}}{\sigma^6}, \quad (5.52c)$$

where $\ddot{\boldsymbol{H}}_{I_{ij}} = \partial\dot{\boldsymbol{H}}_{I_i}/\partial\omega_{Ij}$.

Result 5.8 *The elements of the expected information matrix for ω_{Ii} and σ^2 are*

$$\mathcal{I}_{\mathcal{E}}(\omega_{Ii},\omega_{Ij}) = \frac{1}{2}\text{tr}\left(\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_j}\boldsymbol{D}_I'\right) \quad (5.53a)$$

$$\mathcal{I}_{\mathcal{E}}(\sigma^2,\omega_{Ii}) = \frac{1}{2\sigma^2}\text{tr}\left(\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\right) \quad (5.53b)$$

$$\mathcal{I}_{\mathcal{E}}(\sigma^2,\sigma^2) = \frac{(n-p)}{2\sigma^4}. \quad (5.53c)$$

Result 5.9 *The elements of the approximate average information matrix for ω_{iI} and σ^2 are*

$$\mathcal{I}_{\mathcal{A}}(\omega_{Ii},\omega_{Ij}) = \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_j}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{y} \quad (5.54a)$$

$$\mathcal{I}_{\mathcal{A}}(\sigma^2,\omega_{Ii}) \approx \frac{1}{2\sigma^4}\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{y} \quad (5.54b)$$

$$\mathcal{I}_{\mathcal{A}}(\sigma^2,\sigma^2) = \frac{\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{y}}{2\sigma^6}. \quad (5.54c)$$

Result 5.10 *The elements of the exact average information matrix for the variance parameters are obtained by taking equally-weighted averages within the three pairs of terms in Result 5.7 and 5.8, and are given by*

$$\mathcal{I}_{\mathcal{A}e}(\omega_{Ii},\omega_{Ij}) = \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_j}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{y} \quad (5.55a)$$

$$\mathcal{I}_{\mathcal{A}e}(\sigma^2,\omega_{Ii}) = \frac{1}{4\sigma^4}\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\boldsymbol{P}_I\boldsymbol{y} + \frac{1}{4\sigma^2}\text{tr}\left(\boldsymbol{P}_I\boldsymbol{D}_I\dot{\boldsymbol{G}}_{I_i}\boldsymbol{D}_I'\right). \quad (5.55b)$$

$$\mathcal{I}_{\mathcal{A}e}(\sigma^2,\sigma^2) = \frac{\boldsymbol{y}'\boldsymbol{P}_I\boldsymbol{y}}{2\sigma^6}. \quad (5.55c)$$

Result 5.11 *An alternative version of the REML log-likelihood function for a VSOM for groups of outliers is the outlier vector version of the log-likelihood function* (5.13) *which is,*

$$l_I(\boldsymbol{\omega}_I, \sigma^2; \boldsymbol{y}) = l_0(\sigma^2; \boldsymbol{y}) - \frac{1}{2}\Big\{ \log |\boldsymbol{G}_I \boldsymbol{A}_I + \boldsymbol{I}| - \boldsymbol{e}_I'(\boldsymbol{G}_I\boldsymbol{A}_I + \boldsymbol{I})^{-1}\boldsymbol{G}_I\boldsymbol{e}_I/\sigma^2 \Big\} \quad (5.56)$$

where $\boldsymbol{A}_I = \boldsymbol{D}_I'\boldsymbol{P}_{X^\perp}\boldsymbol{D}_I$*,* $\boldsymbol{e}_I = \boldsymbol{D}_I'\boldsymbol{P}_{X^\perp}\boldsymbol{y}$ *and* $\boldsymbol{P}_{X^\perp} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$*.* $\boldsymbol{G}_I$*,* $\boldsymbol{e}_I$ *and* $\boldsymbol{A}_I$ *are matrix equivalents of* ω_i*,* e_i *and* $1 - v_i$ *respectively.*

Proof. By inspection, the log-likelihood function (5.56) is an outlier vector version of the log-likelihood function (5.13). It can also be obtained via the expansion of the terms involving $\boldsymbol{H}_I$ in the log-likelihood function (5.51). ■

Analogous to the single-case VSOM, the advantage of the log-likelihood function (5.56) over (5.51) is that we only need to maximize the second term of (5.56) in order to obtain an estimate of $\boldsymbol{\omega}_I$. The score function for ω_{I_i} is given by

$$\begin{aligned} U_I(\omega_{I_i}) &= \frac{\partial l_I(\boldsymbol{\omega}_I, \sigma^2)}{\partial \omega_{I_i}} \\ &= -\frac{1}{2}\Big\{ \mathrm{tr}(\boldsymbol{A}_I(\boldsymbol{I} + \boldsymbol{G}_I\boldsymbol{A}_I)^{-1}\dot{\boldsymbol{G}}_{\omega_{I_i}}) - \boldsymbol{e}_I'(\boldsymbol{I} + \boldsymbol{G}_I\boldsymbol{A}_I)^{-1}\dot{\boldsymbol{G}}_{\omega_{I_i}}(\boldsymbol{I} + \boldsymbol{G}_I\boldsymbol{A}_I)^{-1}\boldsymbol{e}_I/\sigma^2 \\ &\qquad + \boldsymbol{e}_I'(\boldsymbol{I} + \boldsymbol{G}_I\boldsymbol{A}_I)^{-2}\dot{\boldsymbol{G}}_{\omega_{I_i}}\boldsymbol{A}_I\boldsymbol{G}_I\boldsymbol{e}_I/\sigma^2 \Big\} \end{aligned}$$

where $\dot{\boldsymbol{G}}_{\omega_{I_i}} = \dfrac{\partial \boldsymbol{G}_I}{\partial \omega_{I_i}}$ is a $k \times k$ zero matrix with a 1 in the i position. Analytical expressions for the ω_{I_i}'s do not exist and so they must be estimated as part of the iterative estimation procedure for the variance parameters in VSOM.

5.7.1 Special case I: A VSOM for group of outliers with a common variance shift

The simplest case of model (5.50) is when $\boldsymbol{G}_I = \omega_I \boldsymbol{I}_k$ so that

$$\boldsymbol{H}_I = \omega_I \boldsymbol{D}_I \boldsymbol{D}_I' + \boldsymbol{I}.$$

This simplest case of the model assumes that the individual δ_i's are independently and identically distributed with common variance $\mathrm{var}(\delta_i) = \sigma^2\omega_I$ for all i. In this case one variance shift parameter needs to be estimated for the whole group with k random effects for the affected observations in the group.

5.7.2 Special case II: A VSOM for group of correlated outliers

A second special case of model (5.50) is when $\boldsymbol{D}_I$ isolates all of the observations belonging to the subset indexed by I. The derivation for this model assumes that the group of outliers are correlated by sharing a common 'outlier' effect such that $\boldsymbol{G}_I = \omega_I \boldsymbol{J}$ with $\boldsymbol{J} = \mathbf{1}\mathbf{1}'$, where $\mathbf{1}$ is a column vector of ones. Then

$$\begin{aligned}\omega_I \boldsymbol{D}_I \boldsymbol{G}_I \boldsymbol{D}_I' &= \omega_I \boldsymbol{D}_I \boldsymbol{J} \boldsymbol{D}_I' \\ &= \omega_I \boldsymbol{D}_I \mathbf{1}\mathbf{1}' \boldsymbol{D}_I' \\ &= \omega_I \boldsymbol{D}_I^* \boldsymbol{D}_I^{*\prime},\end{aligned}$$

where $\boldsymbol{D}_I^* = \boldsymbol{D}_I \mathbf{1}$. Thus the model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}_I^* \delta_I^* + \boldsymbol{e} \tag{5.57}$$

which is similar to (5.50) with the term $\boldsymbol{D}_I \delta_I$ replaced by $\boldsymbol{D}_I^* \delta_I^*$, where δ_I^* is a scalar with variance $\sigma^2 \omega_I$ and $\boldsymbol{H}_I = \boldsymbol{D}_I^* \boldsymbol{D}_I^{*\prime} \omega_I + \boldsymbol{I}$. One variance shift parameter needs to be estimated for the whole group with only one random effect estimate for the term δ_I^*. An explicit expression for ω_I has the same form as that of ω_i in a VSOM for the linear fixed effects linear model (§ 5.1).

These special cases of a VSOM$_k$ assumes that the specified outlying observations indexed by I have a common increase in variance. This assumption may be questionable especially when the subset is arbitrary rather than based on scientific grounds. The choice of grouping for a VSOM$_k$ is also an issue for case-deletion methods. In our illustration of a VSOM$_k$ below (see § 5.7.3) we use a grouping suggested by the structure of the data. We extend a VSOM$_k$ later in Chapter 7 where the ω_I models the increased variance associated with a higher-level random effect in a linear mixed model.

Likelihood ratio and score test statistics

Parameter estimation and construction of likelihood ratio test statistics for a VSOM$_k$ (and the special cases mentioned above) follows the same form as for the linear fixed effects variance shift outlier model (ith VSOM). For instance for model (5.57) the LRT statistic is given by

$$LRT_I = -2\Big\{l_0(\sigma_0^2; \boldsymbol{y}) - l_I(\omega_I, \sigma^2; \boldsymbol{y})\Big\}, \tag{5.58}$$

where $l_I(\omega_I, \sigma^2; \boldsymbol{y})$ is the REML log-likelihood function for $\boldsymbol{y}$ of the form (5.56). The corresponding score test statistic(s) can be constructed as

$$S_I(\omega_I = 0) = \begin{cases} U_I^2(\omega_I = 0)\mathcal{I}^{11} & U_I(\omega_I = 0) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.59}$$

where $\mathcal{I}^{11}$ is the portion of the inverse of information matrix associated with ω_I, evaluated under the null hypothesis. $\mathcal{I}^{11}$ is calculated as before using (5.40) with the information matrix terms of the forms in Results 5.7 to 5.10.

Again the resampling algorithm is used to obtain the sampling distributions of the test statistics and to deal with the problem of multiple testing. However, for a VSOM$_k$ there are no analytic forms of the ω_I. This common parameter must then be estimated using an iterative algorithm such as Newton-Raphson, Fisher Scoring or Average Information algorithm.

5.7.3 Example: The orthodont data

The observations 34 and 35 both belong to male subject 9 and the observations 49 and 52 both belong to male subject 13. A natural grouping in the orthdont data set is an individual subject, so that a VSOM for groups of outliers model for male subjects 9 and 13 could be considered. However, it is not possible to fit a VSOM for groups of outlier model (see § 5.7) for either male subject 9 or male subject 13 because the null model (5.47) has subjects as fixed effects so that subject terms would be aliased.

In Chapter 2, § 2.5, we observed that the data for boys were more variable than for girls. We modelled the extra variation among the boys by fitting a VSOM for groups model (5.57), special case I which assumes separate random effect for each boy but a common variance variance shift. The fitted model, denoted $\mathcal{M}_3$, gave a $-2\times$REML log-likelihood value of 130.60 compared to the $-2\times$REML log-likelihood value of 134.93 for the combined VSOMs for units 34, 35, 49 and 52 (model $\mathcal{M}_1$ in Table 5.2). Based on the difference between the $-2\times$REML log-likelihoods it appears we take model $\mathcal{M}_3$ as our final model. Parameter estimates under model $\mathcal{M}_3$ are shown in Table 5.4. Figure 6.16 presents plots of the raw residuals against the fitted values from the same model and shows the variation is now homogeneous across the two gender groups (compared to the pattern in Figure 2.3) with observation 73 (measurement for girl number 3 at age 8) possibly outlying. We will revisit the issue of groups of outliers in Chapter 7 in the context of the linear mixed model in which natural groups of observations arise.

Table 5.4: *Estimated parameters for a VSOM for groups model fitted to orthodont data.*

Effect	Parameter♯	Model $\mathcal{M}_3$ Estimate (s.e.)
Fixed		
constant	μ	21.380 (0.880)
age	β_1	0.480 (0.053)
sex	β_2	6.375 (1.104)
sex.age	β_3	0.206 (0.083)
Random		
$\boldsymbol{d}_{34}$	$\omega_{34}\sigma^2$	9.823 (16.504)
$\boldsymbol{d}_{35}$	$\omega_{35}\sigma^2$	34.059 (50.760)
$\boldsymbol{d}_{49}$	$\omega_{49}\sigma^2$	39.860 (58.989)
$\boldsymbol{d}_{52}$	$\omega_{52}\sigma^2$	3.417 (7.499)
$\boldsymbol{D}_{\text{boys}}$	$\omega_{\text{boys}}\sigma^2$	0.622 (0.306)
	σ^2	0.608 (0.152)

♯ Estimated subject effects not shown.

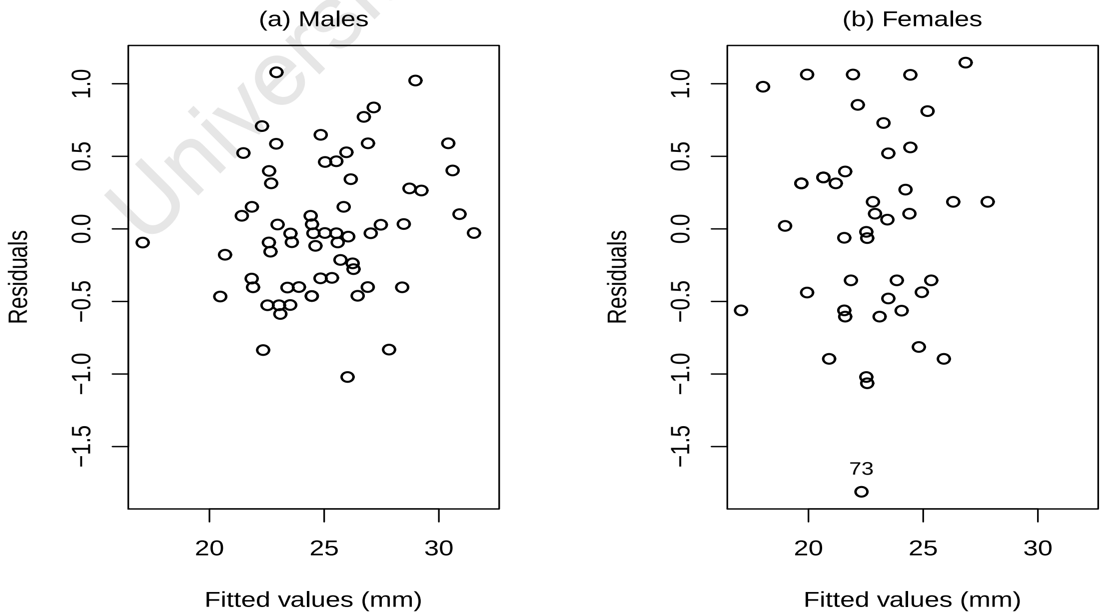

Figure 5.16: *Scatter plots of raw residuals against fitted values (from fitting model $\mathcal{M}_3$) by gender for orthodont data.*

5.8 Performance of likelihood ratio and score test statistics

In this section we compare the performance of the different likelihood ratio test and score test statistics in terms of computing time, type I error and power. The performance of the different tests in terms of type I errors is assessed using the adherence of the empirical type I errors to nominal type I errors. For assessing the power of the different tests to detect an outlier when there is none in the data set, we compute empirical power estimates for nominal type I errors for outliers of different sizes. The power is expected to increase with the increase of the size of the outliers.

5.8.1 Computational efficiency

The test statistics (likelihood ratio and score test statistics) can be computed directly, for a given data set, using the respective formulae we have given in this chapter. Therefore none of the test statistics has a computational advantage over the other. However, using the formulae is efficient than fitting the alternative model, for instance, computing the distribution of the maximum LRT statistic using (5.19) for $n = 100$ (with a single outlier of size 4 inserted in first observation) and 5000 simulations took 11 minutes and 18 seconds on a 3.4 GHz Pentium processor compared to 30 minutes and 5 seconds when the distribution of the maximum LRT statistic involved the fitting of model (5.5) for the same sample size and number of simulations.

5.8.2 Type I error

A simulation study for a simple linear regression model was conducted to assess performance of the various proposed test statistics in terms of type I error (i.e. how often do the test statistics suggest an outlier when there is none in the data set) with respect to both the standard asymptotic distributions and in terms of the empirical distribution generated by a parametric bootstrap procedure.

For each combination of parameters, 500 data sets were generated. The jth simulated data set was generated in a similar manner to the data set generated using model (5.46), i.e.

$$\boldsymbol{y}_j = \mu \mathbf{1}_n + \beta \boldsymbol{x}_j + \boldsymbol{e}_j,$$

for $n = 30, 50, 100$, where the elements of $\boldsymbol{x}_j$ are drawn from the uniform distribution on $[0, 10]$ and then sorted, $\boldsymbol{e}_j \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, $\mathbf{1}_n = (1 \dots 1)'$ is a vector of length n,

$\mu = 20.25$, $\beta = 0.25$ and $\sigma^2 = 1$.

For each simulated data set, several VSOM test statistics were calculated for the first observation: the LRT, partial LRT, one-step LRTs (based on updating scheme A or B) and the score test statistics based on the four information matrices. The reason for using the empirical distribution of the first observation was to avoid the computational burden associated with obtaining empirical type I errors based on the empirical distribution of order statistics. The use of other observations, other than the first observation gave comparable results. The test statistics were compared to the 90th, 95th, 97.5th and 99th percentiles of the standard asymptotic distribution of a $0.5\chi^2_0 + 0.5\chi^2_1$ mixture of chi-squared distributions on 0 and 1 degrees of freedom, respectively. In addition, motivated by the known distribution in a VSOM in regression, test statistics were compared to the same percentiles of a $0.68\chi^2_0 + 0.32\chi^2_1$ mixture of chi-squared distributions on 0 and 1 degrees of freedom. To generate an empirical distribution of the test statistic under the null hypothesis, data sets for $k = 1, \ldots, 2500$ were simulated as

$$\boldsymbol{y}_{jk} = \hat{\mu}_j \mathbf{1}_n + \hat{\beta}_j \boldsymbol{x}_j + \boldsymbol{e}^*_{jk},$$

where $\hat{\mu}_j$ and $\hat{\beta}_j$ are the estimates of μ and β from $\boldsymbol{y}_j$ respectively, $\boldsymbol{e}^*_{jk} \sim N(\mathbf{0}, \hat{\sigma}^2_j \boldsymbol{I}_n)$ with $\hat{\sigma}^2_j$ estimated from $\boldsymbol{y}_j$. All tests were performed for a VSOM for the first observation of each simulated data set $\boldsymbol{y}_{jk}$, $k = 1 \ldots 2500$, and the 90th, 95th, 97.5th and 99th percentiles from the empirical distribution of each test statistic were used as threshold values for the test statistics observed on the original data set $\boldsymbol{y}_j$.

The 500 data sets were used to obtain estimates of type I errors associated with the percentiles from the standard and empirical sampling distributions i.e. the type I error estimate for a given test statistic and level of significance (α) was calculated as the number of data sets (out of 500) for which the test statistic exceeded the $100(1-\alpha)$ percentile of the empirical distribution. A test statistic which produces a type I error which is greater than the nominal value (α) is regarded as anti-conservative, and as conservative if the reverse is true.

Results

Tables 5.5 shows the mean value and standard deviation (over 500 data sets) of the 95th percentile of the empirical distribution of the test statistic under the null hypothesis.

Ignoring issues of multiple testing, this percentile is the value that an observed test statistic would have to exceed in order to be regarded as an outlier. The different test statistics have quite different thresholds: the LRT statistics have much lower thresholds than the score test based on the expected information matrix, but higher thresholds than score tests based on the observed information matrix. In all cases the thresholds are stable across the different sample sizes.

Table 5.5: *95th percentiles (mean (std. deviation)) from empirical distributions of* t_i^2*, likelihood ratio test and score test statistics for the first unit based on 2500 simulations of 500 data sets for* $n = 30, 50, 100$.

Statistic	$n = 30$	$n = 50$	$n = 100$
t_i^2	0.887 (1.207)	1.050 (1.382)	1.009 (1.357)
LRT	1.606 (0.119)	1.559 (0.109)	1.527 (0.111)
One-step LRT: (Scheme A)			
Exp. inf.	1.602 (0.119)	1.558 (0.109)	1.527 (0.110)
Obs. inf.	0.833 (0.051)	0.816 (0.047)	0.804 (0.048)
Approx. avg. inf.	1.125 (0.072)	1.115 (0.068)	1.098 (0.070)
Exact avg. inf.	1.129 (0.073)	1.109 (0.068)	1.094 (0.070)
One-step LRT: (Scheme B)			
Obs. inf.	0.742 (0.040)	0.767 (0.042)	0.781 (0.046)
Approx. avg. inf.	1.085 (0.068)	1.086 (0.065)	1.084 (0.068)
Score test			
Exp. inf.	4.020 (0.406)	4.036 (0.390)	4.046 (0.412)
Obs. inf.	0.639 (0.039)	0.624 (0.036)	0.614 (0.036)
Approx. avg. inf.	1.183 (0.082)	1.125 (0.072)	1.087 (0.071)
Exact avg. inf.	1.081 (0.070)	1.069 (0.066)	1.061 (0.068)

Table 5.6 reports the empirical type I errors for thresholds derived from the empirical distribution under the null hypothesis for the different test statistics for $\alpha = 0.05$ and $\alpha = 0.01$. The empirical type I error estimates are generally close to the nominal α values for all of the test statistics. Results for one-step likelihood ratio test statistics based on either updating scheme A or B were similar, but are not shown. This behaviour of the tests is expected since all the tests are functions of t_i^2 and hence highly correlated with each other (Table 5.7). It must be noted that the calculated correlations assume that there is neither additive bias (estimated intercept parameter

being non-zero) nor multiplicative bias (estimated slope parameter being non-zero) in a simple linear regression model for any pair of our test statistics.

Table 5.6: *Empirical type I errors of LRT statistics and score test statistics, based on the expected information matrix, for the first unit, based on 2500 simulations of 500 data sets for samples size* $n = 30, 50, 100$.

	LRT		Score test using Exp. inf.	
Sample	Nominal probability of rejection (α)			
size	0.05	0.01	0.05	0.01
30	0.060	0.006	0.060	0.006
50	0.034	0.002	0.034	0.002
100	0.052	0.010	0.052	0.010

Table 5.7: *Pearson's correlation coefficients between t_i^2, likelihood ratio test and score test statistics for the first unit for 500 data sets: $n = 50$.*

		t_i^2	LRT	$\text{LRT}_{(1)}$: Scheme A				$\text{LRT}_{(1)}$: Scheme B		Score test			
				Exp.	Obs.	Approx. avg.	Exact avg.	Obs.	Approx. avg.	Exp.	Obs.	Approx. avg.	Exact avg.
	t_i^2	1.000											
	LRT	0.908	1.000										
$\text{LRT}_{(1)}$: Scheme A	Exp. inf.	0.909	1.000	1.000									
	Obs. inf.	0.940	0.993	0.993	1.000								
	Approx. avg.	0.940	0.986	0.987	0.997	1.000							
	Exact avg.	0.932	0.996	0.996	0.999	0.996	1.000						
$\text{LRT}_{(1)}$: Scheme B	Obs.	0.953	0.974	0.974	0.993	0.997	0.990	1.000					
	Approx. avg.	0.939	0.992	0.992	1.000	0.999	0.999	0.995	1.000				
Score test	Exp.	0.823	0.974	0.973	0.942	0.923	0.950	0.897	0.936	1.000			
	Obs.	0.941	0.993	0.993	1.000	0.997	0.999	0.993	1.000	0.942	1.000		
	Approx. avg.	0.917	1.000	1.000	0.996	0.989	0.998	0.979	0.994	0.969	0.996	1.000	
	Exact avg.	0.932	0.996	0.996	1.000	0.997	1.000	0.990	0.999	0.949	0.999	0.998	1.000

Table 5.8 reports the empirical type I errors for thresholds derived from the mixtures of chi-squared distributions for the likelihood ratio and score tests based on expected information. These two test statistics were chosen as those with well-developed asymptotic theory, for which standard distributions might be expected to hold. Both the $0.5\chi_0^2+0.5\chi_1^2$ mixture and the $0.68\chi_0^2+0.32\chi_1^2$ mixture performed poorly in both cases, being conservative for the likelihood ratio test and anti-conservative for the score test. Empirical type I errors based on exact distribution of score test statistic do not appear to be any better than those obtained using the asymptotic distributions.

Table 5.8: *Empirical type I errors, based on asymptotic distributions, of the LRT statistics and expected information score test statistics (*$100(1-\alpha)$*th percentile) for the first unit based on 2500 simulations of 500 data sets for samples size* $n = 30, 50, 100$*. Also shown are empirical type I errors based on exact distribution of score test statistic when the expected information matrix is used.*

Sample size	LRT				Score test using exp. inf.					
	Asymptotic reference distribution								Exact distr.	
	$0.5\chi_0^2 + 0.5\chi_1^2$		$0.68\chi_0^2 + 0.32\chi_1^2$		$0.5\chi_0^2 + 0.5\chi_1^2$		$0.68\chi_0^2 + 0.32\chi_1^2$			
	Nominal probability of rejection: α (Quantile)									
	0.05 (2.71)	0.01 (5.41)	0.05 (2.01)	0.01 (4.64)	0.05 (2.71)	0.01 (5.41)	0.05 (2.01)	0.01 (4.64)	0.05 (1.87)	0.01 (3.53)
30	0.020	0.004	0.044	0.004	0.086	0.044	0.096	0.050	0.098	0.068
50	0.004	0.002	0.010	0.002	0.046	0.014	0.054	0.026	0.056	0.036
100	0.012	0.002	0.032	0.004	0.070	0.038	0.086	0.046	0.090	0.056

5.8.3 Power

In order to investigate the power of the different tests, an outlier of size $\lambda = 1, 2, 4, 8, 16$ or 32 units was introduced into the first observation in the linear regression model. The model parameters and sample sizes remained as in the evaluation of type I error. For each combination of parameters, 100 data sets were generated. The jth simulated data set was generated as

$$\boldsymbol{y}_j = \mu \mathbf{1}_n + \beta \boldsymbol{x}_j + \lambda \boldsymbol{v}_1 + \boldsymbol{e}_j,$$

for $\lambda = 1, 2, 4, 8, 16$ or 32 and $\boldsymbol{v}_1 = (1\ 0\ \ldots 0)'$ is a vector of length n with value 1 in unit 1 and zero elsewhere. Test statistics and their empirical distributions were calculated as for type I error above.

Results

The proportion of likelihood ratio test statistics exceeding the $100(1-\alpha)$ percentiles of the simulated distribution under the null hypothesis are shown in Figure 5.17 (updating scheme A) and Figure 5.18 (updating scheme B) for $\alpha = 0.05$. Corresponding empirical power estimates for selected sample size combinations for $\alpha = 0.05$ are given in Appendix C, § C.2. As expected the power of the likelihood ratio test statistic increases as the size of the displacement, λ, increases. However, the one-step likelihood ratio tests (based on either updating scheme A or B), when the observed or approximate average information matrix is used, have poor power for large displacements. This outcome does not occur for the one-step likelihood ratio test based on the expected information or the exact average information matrix. The behaviour of the powers of the one-step LRTs with respect to outlier size, depicted in Figures 5.18 and 5.19, confirm the undesirable behaviour of the one-step LRTs discussed earlier in § 5.2.2 (see Figures 5.6 and 5.9). There were no differences between the different score test statistics in terms of power thus we present results for the score test statistic based on the expected information matrix (Figure 5.19). As expected, the power increases as the size of the displacement, λ, increases. Again this outcome confirms the behaviour of the different score test statistics with respect to t_i^2 discussed earlier in § 5.3 (see Figure 5.10).

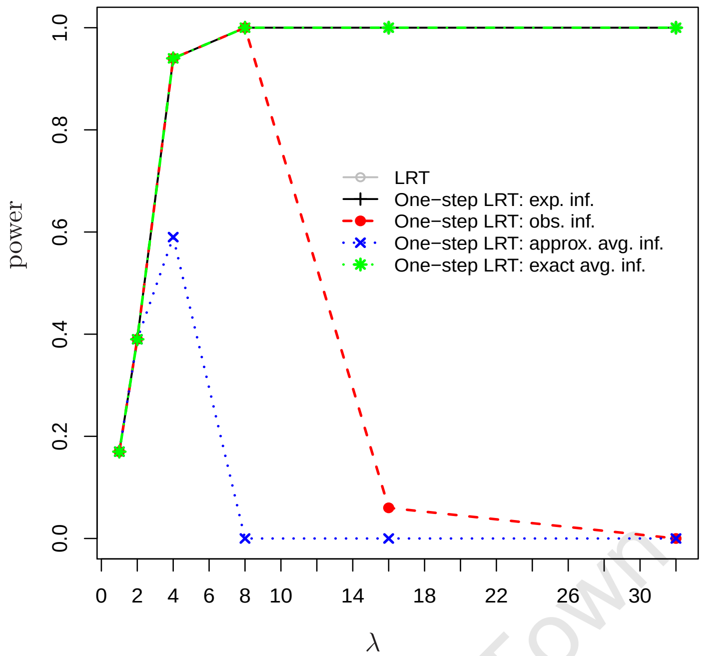

Figure 5.17: *Empirical power of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets for samples size* $n = 30$. *One-step likelihood ratio test statistics are based on updating scheme A.*

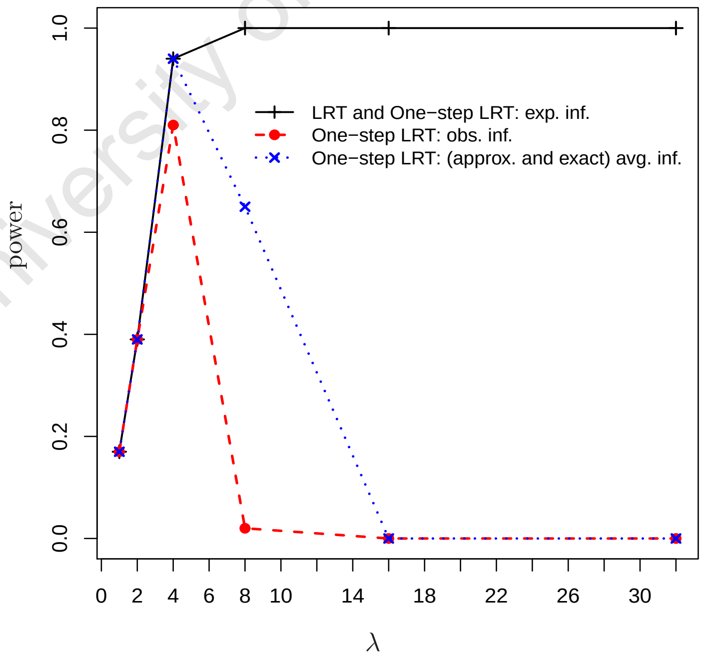

Figure 5.18: *Empirical power of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets for samples size* $n = 30$. *One-step likelihood ratio test statistics are based on updating scheme B.*

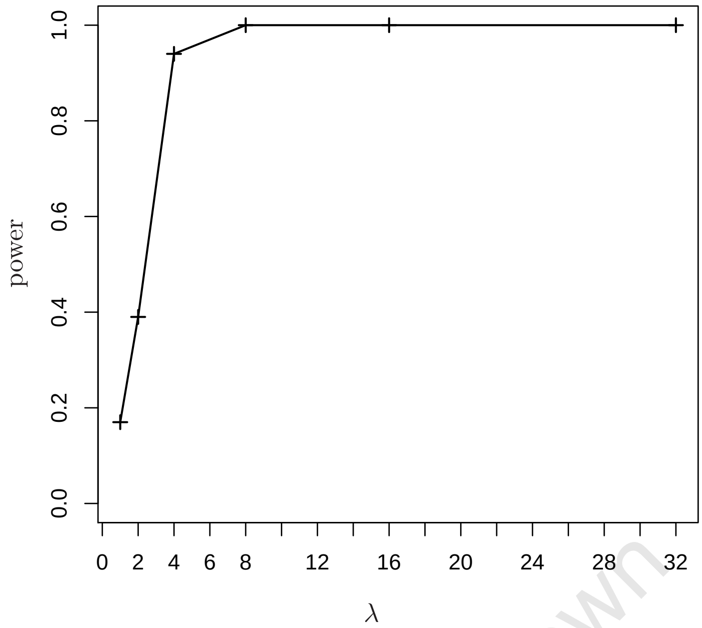

Figure 5.19: *Empirical power of the score test statistic based on expected information matrix for the first unit based on 2500 simulations of 500 data sets for samples size* $n = 30$.

5.9 Summary

In this chapter we have reviewed a VSOM and described parameter estimation in the model. Likelihood ratio and score test statistics were developed as objective measures assessing the size and consequentiality of the variance shift estimates. The dependence of the one-step likelihood ratio tests, through the one-step updates of the variance parameters, and the score tests on the second derivative of the log-likelihood function was also examined. Analytical expressions that permit direct calculation of the likelihood ratio and score test statistics were derived. We also discussed the distributional properties of the likelihood ratio and score test statistics. A parametric bootstrap-based algorithm for significance testing and handling the problem of multiple testing for outliers was also presented. A VSOM for detecting outliers one-at-a-time in linear regression was extended to detecting multiple outliers (more-than-one-at-a-time).

From the results in this chapter we draw the following conclusions:

- The likelihood ratio and score test statistics for a VSOM can be used to detect unknown outliers in a data set, using the resampling algorithm we have proposed. When observations are identified as possible outliers and their presence in a data set cannot be explained, a VSOM offers an opportunity to include them in the analysis, down-weighted by using the variance shift estimates. This down-weighting might be considered preferable to omitting data points (as in case-deletion methods).

- A VSOM procedure is easy to implement with standard software packages which fit linear mixed models.

- Both the likelihood ratio and score test statistics are easy and quick to evaluate so that none of the test statistics has a computational advantage over the other in terms of computing time. However, the computational gains from computing test statistics not requiring the full fit of the alternative model (one-step LRT or score test statistics) may be substantial especially in linear mixed models with complex covariance structures. We return to this point later in Chapter 6.

- In the real data set example we used the likelihood ratio and score test statistics under a VSOM and identified the same outliers as case-deletion methods. A VSOM and the case-deletion model also gave similar parameter estimates and standard errors which reflects the fact that the variance shift estimates which are used to down-weight the detected outliers are so large that the down-weighting becomes equivalent to deleting the observations.

- The resampling procedure we have proposed for handling multiple testing and assessing the significance of both the likelihood ratio and score test statistics is easy to implement and fast enough for interactive data analysis.

- From Figure 5.4 we noted that all the four different types of score test statistics are increasing functions of t_i^2. On the other hand, the decrease in the one-step LRT functions based on the observed and approximate average information matrices implies that these functions are inappropriate test statistics: large values of t_i^2 which should indicate outliers can generate small (or even negative) values of the LRT. This area warrants further research, but suggests that one-step LRTs for a VSOM should be treated with some caution, especially in more complex models.

- Both the likelihood ratio and score test statistics also consistently detect observations with inflated error variance. These proposed tests also performed equally well both in terms of type I errors and power of detecting outliers of moderate size but one-step LRT based on the observed and approximate average informations matrices had very low power for large outliers.

- For detecting groups of outliers using a VSOM in linear regression (as for case-deletion methods), the choice of grouping is an issue. In our illustrations we used a grouping suggested by the structure of the data.

Areas of further research on a VSOM in linear regression are given in Chapter 8.

CHAPTER 6

A variance shift outlier model for linear mixed effects analysis

The purpose of this chapter is to extend the linear regression VSOM results to the linear mixed model. A linear mixed VSOM can be considered to be an extension of either the linear mixed model, by the addition of a random shift covariate or the linear regression VSOM (discussed in Chapter 5) by an additional random term for the random effects, to account for dependence in the data. In this chapter we will focus on a VSOM for individual observations; we dedicate the next chapter to the discussion of VSOMs for groups of observations. 3 The likelihood ratio (LRT or one-step) and score test statistics under a linear mixed VSOM do not have analytical expressions as a consequence of the variance parameters being estimated iteratively. The extension of the linear regression VSOM to a linear mixed VSOM also comes with computational challenges; for instance the model(s) may not converge when fitted so that the tests can not always be computed. These computational problems make it difficult to the implement the resampling algorithm for obtaining the sampling distribution of the test statistics. We will discuss some strategies for handling these computational challenges including the use of approximations based on the results from Chapter 5.

The outline of the chapter is as follows. In Section 6.1 we present a formulation of a linear mixed VSOM for a single observation in a given data set. Section 6.2 deals with parameter estimation in the model with a focus on the estimation of the variance shift parameter. Here we also give the score function for the variance shift parameter and the information matrices which are building blocks for the calculation of the score test statistics and one-step variance estimates needed for the calculation of the one-step LRT statistics. In Sections 6.3 to 6.4 we extend the likelihood ratio and score tests, for testing the significance of the variance shift estimate, derived in Chapter 5, to a linear mixed VSOM. We also give computing schemes for obtaining the one-step variance estimates when some of the variance components in the model are near zero.

A modification of the resampling algorithm of Chapter 5 for significance testing in a linear mixed VSOM is given in Section 6.5. In Sections 6.6 and 6.7 we illustrate the usefulness of the proposed tests in identifying possible outliers, using a simulated data set and two real data sets (the orthodont and aerosol data sets). In these examples we also compare a VSOM approach (which down-weights observations suspected to be outliers) with the case-deletion approach. The chapter concludes with an investigation of the performance of the proposed test statistics in terms of computing time, and type I and type II errors, using simulations.

The new contributions introduced in this chapter are:

- general forms of the LRT, one-step LRTs and score tests for a VSOM in linear mixed effects analysis,
- approximations of the LRT, one-step LRTs and score tests for a VSOM in linear mixed effects analysis and,
- properties of one-step updates of the variance parameters under VSOM in linear mixed effects analysis.

6.1 The model and log-likelihood function

6.1.1 The model

A VSOM for the ith observation in the linear mixed model is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{d}_i\delta_i + \boldsymbol{e} \\ &\sim N\left(\boldsymbol{X\beta}, \sigma^2(\boldsymbol{ZGZ}' + \omega_i \boldsymbol{d}_i\boldsymbol{d}_i' + \boldsymbol{R})\right) \\ &\equiv N\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H}_i\right) \end{aligned} \tag{6.1}$$

where $\boldsymbol{d}_i$ is a column vector of length n with a single non-zero entry whose value is 1 in the ith position and

$$\begin{aligned} \boldsymbol{H}_i &= \boldsymbol{ZGZ}' + \omega_i \boldsymbol{d}_i\boldsymbol{d}_i' + \boldsymbol{R} \\ &= \boldsymbol{H} + \omega_i \boldsymbol{d}_i\boldsymbol{d}_i'. \end{aligned} \tag{6.2}$$

$\boldsymbol{H} = \boldsymbol{ZGZ}' + \boldsymbol{R}$ is the variance-covariance matrix of $\boldsymbol{y}$ for a model excluding $\boldsymbol{d}_i\delta_i$ and $\text{var}(\delta_i) = \sigma^2\omega_i$. The variance-covariance matrix under the model (6.1) becomes

$$\text{var}(\boldsymbol{y}) = \sigma^2(\,\boldsymbol{ZGZ}' + \omega_i\boldsymbol{d}_i\boldsymbol{d}_i' + \boldsymbol{R}\,) = \sigma^2\boldsymbol{H}_i\ ,$$

The variance for unit i is inflated as $\sigma^2(z_i'\boldsymbol{G}z_i + \omega_i + R_{ii})$, where z_i' is the ith row of $\boldsymbol{Z}$ and R_{ii} is the ith diagonal element of $\boldsymbol{R}$. This shift has an impact on the variance of unit i, but not on its covariances with other units and so in general may not support or model all relevant deviations from the variance model. We return to this point later in Chapter 8.

6.1.2 The log-likelihood function for the model

The REML log-likelihood function, $l_i(\omega_i, \boldsymbol{\kappa}, \sigma^2; \boldsymbol{y})$, for a VSOM in the mixed model is given by

$$l_i(\omega_i, \boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) = -\frac{1}{2}\left\{(n-p)\log\sigma^2 + \log|\boldsymbol{H}_i| + \log|\boldsymbol{X}'\boldsymbol{H}_i^{-1}\boldsymbol{X}| + \frac{\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{y}}{\sigma^2}\right\}, \quad (6.3)$$

where

$$\boldsymbol{H}_i^{-1} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{H}^{-1}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{H}^{-1}$$

and

$$\boldsymbol{P}_i = \boldsymbol{H}_i^{-1} - \boldsymbol{H}_i^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}_i^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}_i^{-1}$$

The log-likelihood function (6.3) is analogous to the likelihood function (2.19) for the linear mixed model (see Chapter 2 § 2.3).

The REML log-likelihood function can be rewritten as

$$l_i(\omega_i, \boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) = -\frac{1}{2}\left\{(n-p)\log\sigma^2 + \log|\boldsymbol{R}| + \log|\boldsymbol{G}_i| + \log|\boldsymbol{C}_i| + \frac{\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{y}}{\sigma^2}\right\} \quad (6.4)$$

where

$$\boldsymbol{C}_i = \begin{bmatrix} \boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{d}_i + \frac{1}{\omega_i} & \boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i & \boldsymbol{C} \end{bmatrix}$$

and

$$\boldsymbol{G}_i = \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \omega_i \end{bmatrix}.$$

6.1.3 Joint estimation of fixed and random effects

The mixed model equations for model (6.1) are analogous to the MMEs for the general linear mixed model (2.7), i.e.

$$\boldsymbol{C}_i \begin{bmatrix} \tilde{\delta}_i \\ \tilde{\boldsymbol{\psi}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}, \tag{6.5}$$

where $\tilde{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}', \tilde{\boldsymbol{u}}')'$. Solutions for $\tilde{\boldsymbol{\psi}}$ follow from Lemma 2.1 and the solution for δ_i is similar to (5.8) with $\boldsymbol{H}_i$ replaced by equation (6.2).

6.2 Variance parameter estimation

Since variance component estimation in the linear mixed model was discussed in Chapter 2, this section focuses on estimation of the variance shift parameter ω_i in model (6.1). We first give the score function and information matrices required for estimating the variance parameters, $\boldsymbol{\phi} = (\omega_i, \boldsymbol{\kappa}', \sigma^2)'$, in the model. Note that this vector of variance parameters is an extension of $\boldsymbol{\phi}$ defined in § 2.3 and has an extra parameter ω_i.

The following result gives an alternative expression of the REML log-likelihood function (6.4) which is in terms of the variance shift parameter ω_i and is convenient for obtaining approximate estimates of ω_i and σ^2.

Result 6.1 *The REML log-likelihood function* (6.4) *can be written alternatively as*

$$l_i(\omega_i, \boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) = -\frac{1}{2}\Big\{(n-p)\log\sigma^2 + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \log|\boldsymbol{C}| + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^2} - \frac{(\boldsymbol{y}'\boldsymbol{P}\boldsymbol{d}_i\omega_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i\omega_i+1)^{-1}\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{y})}{\sigma^2} + \log(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i\omega_i+1)\Big\}. \tag{6.6}$$

Proof. The proof relies on expanding the terms $\boldsymbol{C}_i$, $\boldsymbol{G}_i$ and $\boldsymbol{P}_i$ in (6.4).

Corollary 6.1 *An alternative formula for* $\boldsymbol{P}_i$*, which follows from Result A.5 of Appendix A, is*

$$\begin{aligned}\boldsymbol{P}_i &= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}_i\boldsymbol{C}_i^{-1}\boldsymbol{W}_i'\boldsymbol{R}^{-1} \\ &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{P}.\end{aligned} \tag{6.7}$$

Proof. The proof utilizes the matrix results Results A.2 and A.3 from Appendix A by first letting $\boldsymbol{C} = \boldsymbol{A}$ so that if

$$\begin{aligned}\boldsymbol{Q} &= \boldsymbol{C}_{i11} - \boldsymbol{C}_{i12}\boldsymbol{C}_{i22}^{-1}\boldsymbol{C}_{i21} \\ &= \boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{d}_i + 1/\omega i - \boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i \\ &= \boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i,\end{aligned}$$

then

$$\boldsymbol{C}_i^{-1} = \begin{bmatrix}\boldsymbol{C}_i^{11} & \boldsymbol{C}_i^{21} \\ \boldsymbol{C}_i^{12} & \boldsymbol{C}_i^{22}\end{bmatrix}$$

where

$$\begin{aligned}\boldsymbol{C}_i^{11} &= (\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1} \\ \boldsymbol{C}_i^{21} &= -(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1} \\ \boldsymbol{C}_i^{12} &= -\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i'(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1} \\ \boldsymbol{C}_i^{22} &= \boldsymbol{C}^{-1} + \boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}.\end{aligned}$$

Hence

$$\begin{aligned}
\boldsymbol{W}_i\boldsymbol{C}_i^{-1}\boldsymbol{W}_i' = {} & \boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i' - \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i' \\
& -\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}' \\
& +\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}' + \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'.
\end{aligned}$$

Thus

$$\begin{aligned}
\boldsymbol{P}_i &= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}_i\boldsymbol{C}_i^{-1}\boldsymbol{W}_i'\boldsymbol{R}^{-1} \\
&= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\Big\{\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i' - \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i' \\
&\qquad - \boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}' + \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}' \\
&\qquad + \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\Big\}\boldsymbol{R}^{-1} \\
&= \boldsymbol{P} - \boldsymbol{R}^{-1}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1} \\
&\quad + \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1} \\
&\quad + \boldsymbol{R}^{-1}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1} \\
&\quad - \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1} \\
&= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1} + \boldsymbol{P}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1} \\
&= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{P}.
\end{aligned}$$

■

Note that

$$\begin{aligned}
|\boldsymbol{C}_i| &= |\boldsymbol{C}||\boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{d}_i + 1/\omega_i - \boldsymbol{d}_i'\boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{d}_i| \\
&= |\boldsymbol{C}||(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i|),
\end{aligned}$$

$$\log|\boldsymbol{C}_i| = \log|\boldsymbol{C}| + \log(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i), \qquad \text{using Result } A.4$$

and

$$\log|\boldsymbol{G}_i| = \log|\boldsymbol{G}| + \log(\omega_i).$$

Hence

$$|\boldsymbol{G}_i| + |\boldsymbol{C}_i| = |\boldsymbol{G}| + |\boldsymbol{C}| + \log(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i) + \log(\omega_i). \tag{6.8}$$

Thus replacing $\boldsymbol{P}_i$, and $\boldsymbol{G}_i$ and $\boldsymbol{C}_i$ with (6.7) and (6.8), respectively, in the log-likelihood function (6.6), proves the result. ■

The REML log-likelihood function (6.6) can also be expressed as

$$\begin{aligned} l_i(\omega_i, \boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) &= l_0(\boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) - \frac{1}{2}\left\{ \log(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i\omega_i + 1) - \frac{(\boldsymbol{y}'\boldsymbol{P}\boldsymbol{d}_i(\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i + 1/\omega_i)^{-1}\boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{y})}{\sigma^2} \right\} \\ &= l_0(\boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) - \frac{1}{2}\left\{ \log(a_{ii}\omega_i + 1) - \frac{\tilde{e}_i^2\omega_i}{\sigma^2(a_{ii}\omega_i + 1)} \right\} \end{aligned} \tag{6.9}$$

where $\tilde{e}_i = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{d}_i$ is the conditional residual for unit i under the null model, $a_{ii} = \boldsymbol{d}_i'\boldsymbol{P}\boldsymbol{d}_i$ and

$$l_0(\boldsymbol{\kappa}, \sigma^2; \boldsymbol{y}) = -\frac{1}{2}\left\{ (n-p)\log\sigma^2 + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \log|\boldsymbol{C}| + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^2} \right\}$$

is the log-likelihood function for the null model, the linear mixed model (2.19). The log-likelihood function (6.9) is a similar form to that for a VSOM in linear regression (5.13), with a_{ii} as a generalisation of $(1 - v_{ii})$.

The score function for ω_i is given by

$$U_i(\omega_i) = \frac{1}{2}\left\{ \frac{\tilde{e}_i^2}{\sigma^2(a_{ii}\omega_i + 1)^2} - \frac{a_{ii}}{(a_{ii}\omega_i + 1)} \right\}. \tag{6.10}$$

Both $\tilde{e}_i$ and a_{ii} depend on all other estimated variance components in the model through $\boldsymbol{P}$.

Under the null hypothesis, the score function for ω_i is

$$U_i(\omega_i = 0) = \frac{1}{2}\left\{ a_{ii}\left(\frac{\tilde{e}_i^2}{a_{ii}\sigma^2} - 1 \right) \right\}. \tag{6.11}$$

The score function for σ^2 is given by

$$\frac{\partial l_i(\omega_i, \boldsymbol{\kappa}, \sigma^2)}{\partial \sigma^2} = -\frac{(n-p)}{2\sigma^2} + \frac{(\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y})}{2\sigma^4} - \frac{\tilde{e}_i^2\omega_i}{2\sigma^4(a_{ii}\omega_i + 1)}.$$

In general, analytic forms of the variance parameter estimates are not available. Gogel (1997) showed that an analytic form of the estimates for ω_i and σ^2 can be obtained if the vector $\boldsymbol{\kappa}$ of all other variance parameters in the model) is held fixed at the baseline model estimate, $\hat{\boldsymbol{\kappa}}_0$. This approach assumes that outliers are more likely to affect σ^2 than the other variance components in the model, and hence can be less reliable in situations where outliers do in fact also affect the other variance parameters. If we write

$$\begin{aligned}\frac{\tilde{e}_i^2\omega_i}{(a_{ii}\omega_i+1)} &= \frac{\tilde{e}_i^2(a_{ii}\omega_i+1-1)}{a_{ii}(a_{ii}\omega_i+1)} \\ &= \frac{\tilde{e}_i^2}{a_{ii}} - \frac{\tilde{e}_i^2}{a_{ii}(a_{ii}\omega_i+1)} \\ &= s_i^2 - \frac{s_i^2}{(a_{ii}\omega_i+1)},\end{aligned}$$

where $s_i^2 = \tilde{e}_i^2/a_{ii}$, then the REML log-likelihood function (6.9) becomes

$$\begin{aligned}l_a(\omega_i, \hat{\boldsymbol{\kappa}}_0, \sigma^2; \boldsymbol{y}) = -\frac{1}{2}\Big\{&(n-p-1)\log\sigma^2 + \log|\boldsymbol{H}| + \log|\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}| \\ &+ \log\sigma^2(a_{ii}\omega_i+1) + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y} - s_i^2}{\sigma^2} + \frac{s_i^2}{\sigma^2(a_{ii}\omega_i+1)}.\Big\}\end{aligned} \quad (6.12)$$

Solving the score equations from (6.12) gives estimates of ω_i and σ^2 as

$$\hat{\omega}_i(\hat{\boldsymbol{\kappa}}_0) = \begin{cases} \dfrac{(n-p)(t_i^2-1)}{a_{ii}(n-p-t_i^2)} & t_i^2 > 1, \\ 0 & \text{otherwise}, \end{cases} \quad (6.13)$$

$$\hat{\sigma}^2(\hat{\boldsymbol{\kappa}}_0) = \begin{cases} \dfrac{(n-p-t_i^2)\hat{\sigma}_0^2}{(n-p-1)} & t_i^2 > 1 \\ \hat{\sigma}_0^2 & \text{otherwise}. \end{cases} \quad (6.14)$$

where now $t_i^2 = \tilde{e}_i^2/(\hat{\sigma}_0^2 a_{ii})$, which depends on $\boldsymbol{\kappa}$ through $\boldsymbol{P}$, with $\boldsymbol{P}$ evaluated at $\hat{\boldsymbol{\kappa}}_0$. We will refer to the estimates obtained using (6.13) and (6.14) as the partial variance estimates to reflect that they are approximations to the REML estimates.

Again an index plot of $\hat{\omega}_i(\hat{\boldsymbol{\kappa}}_0)$ values or t_i^2 values, as in the previous chapter, can be used to highlight units with inflated error variances. In the next two sections we present objective measures for determining the size of the estimated variance shift estimates: likelihood ratio and score test statistics.

6.3 Likelihood ratio tests for the variance shift parameter

6.3.1 Likelihood ratio test

With full re-estimation of the variance parameters under the ith VSOM, the likelihood ratio test for a linear mixed VSOM is given by

$$LRT_i = \begin{cases} -2\Big\{l_0(\hat{\boldsymbol{\phi}}_0; \boldsymbol{y}) - l_i(\hat{\boldsymbol{\phi}}_i; \boldsymbol{y})\Big\}, & \hat{\omega}_i > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6.15}$$

where $\hat{\boldsymbol{\phi}}_0 = (0, \hat{\boldsymbol{\kappa}}_0', \hat{\sigma}_0^2)'$ and $\hat{\boldsymbol{\phi}}_i = (\hat{\omega}_i, \hat{\boldsymbol{\kappa}}', \hat{\sigma}^2)'$ are the variance estimates for null model and the ith VSOM, respectively.

In general the quantity LRT_i has no analytic form, but the log-likelihood for the null and alternate models can be obtained using standard software packages. With partial re-estimation of the variance parameters under the alternative model, the likelihood ratio test has the form

$$\begin{aligned} LRT_i^* &= \begin{cases} -2\Big\{l_0(\hat{\boldsymbol{\phi}}_0; \boldsymbol{y}) - l_i(\hat{\boldsymbol{\phi}}_i^*; \boldsymbol{y})\Big\}, & \hat{\omega}_i(\hat{\boldsymbol{\kappa}}_0) > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} (n-p-1)\log\dfrac{(n-p-1)}{n-p-t_i^2} - \log t_i^2, & t_i^2 > 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{6.16}$$

where $\hat{\boldsymbol{\phi}}_i^* = (\hat{\omega}_i(\hat{\boldsymbol{\kappa}}_0), \hat{\boldsymbol{\kappa}}_0', \hat{\sigma}^2(\hat{\boldsymbol{\kappa}}_0))'$ are the variance estimates under the ith VSOM, $\hat{\omega}_i(\hat{\boldsymbol{\kappa}}_0)$ and $\hat{\sigma}^2(\hat{\boldsymbol{\kappa}}_0)$ are estimated given the null model estimates of $\boldsymbol{\kappa}$, $\hat{\boldsymbol{\kappa}}_0$. LRT_i^* is analogous to the LRT of (5.19) for a linear regression VSOM and we will refer to it as the partial likelihood ratio test statistic (partial LRT). Similarly to (5.19), LRT_i^* can be calculated directly given the squared Studentized residuals t_i^2. However it ignores the impact of outliers on other random terms in the model since it uses $\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}_0$.

6.3.2 One-step likelihood ratio tests

The one-step LRT statistic offers another measure for evaluating the size of the estimated ω_i and generally requires fewer computations than the LRT statistic of the previous section. Given the one-step variance estimates, the one-step likelihood ratio test statistics, based on the four information matrices, can be constructed as

$$LRT_{i(1)} = \begin{cases} -2\Big\{l_0(\hat{\boldsymbol{\phi}}_0;\boldsymbol{y}) - l_i(\hat{\boldsymbol{\phi}}_{i(1)};\boldsymbol{y})\Big\}, & \hat{\omega}_{i(1)} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6.17}$$

where $\hat{\boldsymbol{\phi}}_0 = (0, \hat{\boldsymbol{\kappa}}_0', \hat{\sigma}_0^2)'$ and $\hat{\boldsymbol{\phi}}_{i(1)} = (\hat{\omega}_{i(1)}, \hat{\boldsymbol{\kappa}}_{(1)}', \hat{\sigma}_{(1)}^2)'$ are the variance estimates for null model and one-step variance estimates for the ith VSOM, respectively. The one-step LRT statistics are similar to those presented in § 5.2.2 except that they use the variance estimates $\hat{\boldsymbol{\phi}}_0 = (0, \hat{\boldsymbol{\kappa}}_0', \hat{\sigma}_0^2)'$ and $\hat{\boldsymbol{\phi}}_{i(1)} = (\hat{\omega}_{i(1)}, \hat{\boldsymbol{\kappa}}_{(1)}', \hat{\sigma}_{(1)}^2)'$. However, these one-step LRT statistics do not have analytical expressions since the variance estimates under the null model are obtained iteratively. The one-step variance estimates used to obtain the one-step LRTs are computed as

$$\hat{\boldsymbol{\phi}}_{i(1)} = \hat{\boldsymbol{\phi}}_0 + \boldsymbol{\mathcal{I}}_i(\hat{\boldsymbol{\phi}}_0)^{-1}\boldsymbol{U}_i(\hat{\boldsymbol{\phi}}_0), \tag{6.18}$$

$\boldsymbol{U}_i(\hat{\boldsymbol{\phi}}_0)$ and $\boldsymbol{\mathcal{I}}_i(\hat{\boldsymbol{\phi}}_0)$ are the score vector and information matrix respectively, evaluated at $\hat{\boldsymbol{\phi}}_0$. Replacing $\boldsymbol{\mathcal{I}}_i(\hat{\boldsymbol{\phi}}_0)$ in (6.18) with respective observed, expected, approximate average and exact average information matrix estimates results in one-step estimates based on the observed, expected, approximate average and exact average information matrices respectively.

In the following we describe the information matrices needed for calculation of the one-step estimates of the variance parameters and hence the one-step LRT statistics. The information matrices are partial derivatives of the REML log-likelihood function (6.3) with respect to the variance parameters $(\omega_i, \boldsymbol{\kappa}, \sigma^2)$. They could also be obtained using the log-likelihood function (6.6). These expressions for the information matrix terms are similar to those given in Chapter 2 (Results 2.4 to 2.7).

The score functions for κ_j, $j = 1, \ldots, r+s$ and σ^2, and their corresponding elements in the information matrices are the same as those presented in Chapter 2 (Results 2.2 to 2.3) with $\boldsymbol{H}$ and $\boldsymbol{P}$ replaced by $\boldsymbol{H}_i$ and $\boldsymbol{P}_i$ respectively. Hence we give the information matrix elements that involve ω_i only, and those involving ω_i along with the other variance parameters in the model ($\boldsymbol{\kappa}$ and σ^2).

Result 6.2 *The elements of the observed information matrix for ω_i and the variance parameters, κ_j, $j = 1, \ldots, r+s$, and σ^2 are*

$$\mathcal{I}_{\mathcal{O}}(\omega_i, \omega_i) = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i\right) + \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y} \tag{6.19a}$$

$$\mathcal{I}_{\mathcal{O}}(\kappa_j, \omega_i) = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}_i\dot{\boldsymbol{H}}_j\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\right) + \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}_i\dot{\boldsymbol{H}}_j\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y} \tag{6.19b}$$

$$\mathcal{I}_{\mathcal{O}}(\sigma^2, \omega_i) = \frac{1}{2\sigma^4}\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y}. \tag{6.19c}$$

Result 6.3 *The elements of the expected information matrix for ω_i and the variance parameters κ_j, $j = 1, \ldots r+s$, and σ^2 are*

$$\mathcal{I}_{\mathcal{E}}(\omega_i, \omega_i) = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i\right) = \frac{1}{2}(\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i)^2 \tag{6.20a}$$

$$\mathcal{I}_{\mathcal{E}}(\kappa_j, \omega_i) = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}_i\dot{\boldsymbol{H}}_j\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\right) = \frac{1}{2}(\boldsymbol{d}_i'\boldsymbol{P}_i\dot{\boldsymbol{H}}_j\boldsymbol{P}_i\boldsymbol{d}_i) \tag{6.20b}$$

$$\mathcal{I}_{\mathcal{E}}(\sigma^2, \omega_i) = \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\right) = \frac{1}{2\sigma^2}\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i. \tag{6.20c}$$

Result 6.4 *The elements of the approximate average information matrix for ω_i and the variance parameters κ_j, $j = 1, \ldots, r+s$, and σ^2 are*

$$\mathcal{I}_{\mathcal{A}}(\omega_i, \omega_i) = \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y} \tag{6.21a}$$

$$\mathcal{I}_{\mathcal{A}}(\kappa_j, \omega_i) \approx \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}_i\dot{\boldsymbol{H}}_j\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y} \tag{6.21b}$$

$$\mathcal{I}_{\mathcal{A}}(\sigma^2, \omega_i) \approx \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y}. \tag{6.21c}$$

Result 6.5 *The elements of the exact average information matrix for the variance parameters are given by*

$$\mathcal{I}_{\mathcal{A}e}(\omega_i, \omega_i) = \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y} \tag{6.22a}$$

$$\mathcal{I}_{\mathcal{A}e}(\kappa_j, \omega_i) = \frac{1}{4\sigma^2}\boldsymbol{y}'\boldsymbol{P}_i\dot{\boldsymbol{H}}_j\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y} \tag{6.22b}$$

$$\mathcal{I}_{\mathcal{A}e}(\sigma^2, \omega_i) = \frac{1}{4\sigma^4}\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{d}_i\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{y} + \frac{1}{4\sigma^2}\boldsymbol{d}_i'\boldsymbol{P}_i\boldsymbol{d}_i. \tag{6.22c}$$

Now the information matrices can be written in the form of (5.28), with the notation for the type of information matrix suppressed, as

$$\boldsymbol{\mathcal{I}}_i = \begin{bmatrix} \boldsymbol{\mathcal{I}}_{11} & \boldsymbol{\mathcal{I}}_{12} \\ \boldsymbol{\mathcal{I}}'_{12} & \boldsymbol{\mathcal{I}}_{22} \end{bmatrix}, \tag{6.23}$$

where $\mathcal{I}_{11} = \mathcal{I}(\omega_i, \omega_i)$ is the information matrix element for ω_i, $\boldsymbol{\mathcal{I}}_{12}$ corresponds to the information matrix elements involving ω_i and the joint variance parameters $\boldsymbol{\kappa}$ and σ^2 as

$$\boldsymbol{\mathcal{I}}_{12} = \begin{bmatrix} \boldsymbol{\mathcal{I}}(\omega_i, \boldsymbol{\kappa}') & \boldsymbol{\mathcal{I}}(\omega_i, \sigma^2) \end{bmatrix}$$

and $\boldsymbol{\mathcal{I}}_{22}$ corresponds to the information matrix elements for $\boldsymbol{\kappa}$ and σ^2 (the information matrix for the variance parameters under the null model (2.1)):

$$\boldsymbol{\mathcal{I}}_{22} = \begin{bmatrix} \boldsymbol{\mathcal{I}}(\boldsymbol{\kappa}, \boldsymbol{\kappa}') & \boldsymbol{\mathcal{I}}(\boldsymbol{\kappa}, \sigma^2) \\ \boldsymbol{\mathcal{I}}(\sigma^2, \boldsymbol{\kappa}') & \boldsymbol{\mathcal{I}}(\sigma^2, \sigma^2). \end{bmatrix}$$

Updating schemes for one-step estimates of variance parameters

In this section we describe updating schemes for obtaining the one-step updates of the variance parameters. These schemes are extensions of updating schemes described earlier, in Chapter 5 § 5.2.2.

Scheme A. Simultaneous updating of variance estimates

Step 1 Obtain one-step estimates of variance parameters using (6.18) with null model estimates $\hat{\boldsymbol{\phi}}_0 = (0, \hat{\boldsymbol{\kappa}}_0, \hat{\sigma}_0^2)'$ as initial values. Under the null hypothesis the score vector given by

$$\boldsymbol{U}_i\left(\hat{\boldsymbol{\phi}}_0\right) = \begin{pmatrix} \dfrac{a_{ii}(t_i^2 - 1)}{2} \\ \mathbf{0} \\ 0 \end{pmatrix} \tag{6.24}$$

and the information matrices consist of the terms in Results 6.2 to 6.5 evaluated under the null hypothesis. Then the one-step updates are

$$\hat{\boldsymbol{\phi}}_{i(1)} = \begin{pmatrix} \hat{\omega}_{i(1)} \\ \hat{\boldsymbol{\kappa}}_{(1)} \\ \hat{\sigma}^2_{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{\boldsymbol{\kappa}}_0 \\ \hat{\sigma}^2_0 \end{pmatrix} + \boldsymbol{\mathcal{I}}_i^{-1} \begin{pmatrix} \frac{(a_{ii})(t_i^2 - 1)}{2} \\ \mathbf{0} \\ 0 \end{pmatrix} \tag{6.25}$$

where $\boldsymbol{\mathcal{I}}_i$ is as defined in with (6.23) with $\boldsymbol{\mathcal{I}}_{22}$ and t_i^2 evaluated at $\hat{\boldsymbol{\phi}}_0 = (0, \hat{\boldsymbol{\kappa}}_0, \hat{\sigma}^2_0)'$.

If $\hat{\omega}_{i(1)} < 0$, then no updating is required and the one-step estimates of the variance parameters are replaced by their null model estimates, i.e. $\hat{\omega}_{i(1)} = 0$, $\hat{\boldsymbol{\kappa}}_{(1)} = \hat{\boldsymbol{\kappa}}_0$ and $\hat{\sigma}^2_{(1)} = \hat{\sigma}^2_0$.

If $\hat{\omega}_{i(1)} < 0$ or $\hat{\sigma}^2_{(1)} < 0$, then the update has failed and the one-step updates of the variance parameters can not be computed.

Step 2 Calculation of updates if $\hat{\boldsymbol{\kappa}}_{(1)} \leq \mathbf{0}$ from step 1:

(i) If some elements of $\hat{\boldsymbol{\kappa}}_{(1)}$ are less than zero, set the corresponding elements of $\hat{\boldsymbol{\kappa}}_{(1)}$ to zero and redo the update as

$$\hat{\boldsymbol{\phi}}_{i(1)} = \begin{pmatrix} \hat{\omega}_{i(1)} \\ \hat{\boldsymbol{\kappa}}^*_{(1)} \\ \hat{\sigma}^2_{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{\boldsymbol{\kappa}}^*_0 \\ \hat{\sigma}^2_0 \end{pmatrix} + \boldsymbol{\mathcal{I}}_i^{-1} \begin{pmatrix} \frac{(a_i)(t_i^2 - 1)}{2} \\ \mathbf{0} \\ 0 \end{pmatrix} \tag{6.26}$$

where $\boldsymbol{\mathcal{I}}_i$ is the information matrix involving the variance parameters ω_i, $\boldsymbol{\kappa}^*$, where $\boldsymbol{\kappa}^*$ contains variance parameters which had updates greater than zero in step 1. $\boldsymbol{\mathcal{I}}_i$ is evaluated together with a_{ii}, and t_i^2 at $\hat{\boldsymbol{\phi}}_0 = (0, \hat{\boldsymbol{\kappa}}^*_0, \hat{\sigma}^2_0)'$.

If after step 2 (i) $\hat{\boldsymbol{\kappa}}_{(1)} < \mathbf{0}$ or $\hat{\sigma}^2_{(1)} < 0$, then the update has failed and the one-step updates of the variance parameters can not be computed.
If $\hat{\omega}_{i(1)} < 0$, then set $\hat{\omega}_{i(1)} = 0$, $\hat{\boldsymbol{\kappa}}_{(1)} = \hat{\boldsymbol{\kappa}}_0$ and $\hat{\sigma}^2_{(1)} = \hat{\sigma}^2_0$.

(ii) If all elements of $\hat{\boldsymbol{\kappa}}_{(1)}$ are less than zero, set $\hat{\boldsymbol{\kappa}}_{(1)} = \mathbf{0}$ and redo the update as

$$\begin{pmatrix} \hat{\omega}_{i(1)} \\ \\ \hat{\sigma}^2_{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ \\ \hat{\sigma}^2_0 \end{pmatrix} + \boldsymbol{\mathcal{I}}_i^{-1} \begin{pmatrix} \frac{(1-v_i)(t_i^2-1)}{2} \\ \\ 0 \end{pmatrix} \tag{6.27}$$

where $\boldsymbol{\mathcal{I}}_i$ is in form of (5.28) (Chapter 5 pp. 5-15) and is evaluated together with t_i^2 at $\omega_i = 0$, $\hat{\sigma}^2 = \hat{\sigma}^2_0$. Note that the score function is also changed, i.e. a_{ii} is equivalent to $(1 - v_i)$ with $\boldsymbol{P}$ replaced by $\boldsymbol{P}_{X^\perp}$ ($\boldsymbol{P}$ evaluated at $\boldsymbol{\kappa} = 0$).

Under this step, the implied null model fitted to the data is the linear regression model and no longer the linear mixed model. This situation may arise in simulating the distributions of one-step LRT statistics for a linear mixed VSOM which we discuss later in § 6.5 and § 6.6. This step allows for the computation of the test statistic when the $\boldsymbol{\kappa}_{(1)} = 0$ from step 1.

If after step 2 (ii) $\hat{\omega}_{i(1)} < 0$ or $\hat{\sigma}^2_{(1)} < 0$, then the update has failed. Then set $\hat{\omega}_{i(1)} = 0$, $\hat{\boldsymbol{\kappa}}_{(1)} = \hat{\boldsymbol{\kappa}}_0$ and $\hat{\sigma}^2_{(1)} = \hat{\sigma}^2_0$.

Scheme B. Updating σ^2 first

Step 1 Update σ^2 directly using its score function and define the update as

$$\hat{\sigma}^2_{(1)} = \frac{(n-p-t_i^2)\hat{\sigma}^2_0}{(n-p-1)}$$

Step 2 Update ω_i and $\boldsymbol{\kappa}$ given $\sigma^2_{(1)}$ as

$$\hat{\boldsymbol{\phi}}_{i(1)} = \begin{pmatrix} \hat{\omega}_{i(1)} \\ \\ \hat{\boldsymbol{\kappa}}_{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ \\ \hat{\boldsymbol{\kappa}}_0 \end{pmatrix} + \boldsymbol{\mathcal{I}}_i^{-1} \begin{pmatrix} \frac{a_{ii}(t_i^2-1)}{2} \\ \\ \boldsymbol{0} \end{pmatrix} \tag{6.28}$$

where

$$\boldsymbol{\mathcal{I}}_i = \begin{bmatrix} \mathcal{I}(\omega_i, \omega_i) & \boldsymbol{\mathcal{I}}(\omega_i, \boldsymbol{\kappa}) \\ \boldsymbol{\mathcal{I}}(\boldsymbol{\kappa}', \omega_i) & \boldsymbol{\mathcal{I}}(\boldsymbol{\kappa}, \boldsymbol{\kappa}) \end{bmatrix}, \tag{6.29}$$

is the information matrix for the variance parameters ω_i and $\boldsymbol{\kappa}$. $\boldsymbol{\mathcal{I}}_i$ is evaluated together with a_{ii}, and t_i^2 at $(\omega = 0, \hat{\boldsymbol{\kappa}}_0, \hat{\sigma}^2_{(1)})'$.

If $\hat{\omega}_{i(1)} < 0$, then no updating is required and the one-step estimates of the variance parameters are replaced by their null model estimates, i.e. $\hat{\omega}_{i(1)} = 0$, $\hat{\boldsymbol{\kappa}}_{(1)} = \hat{\boldsymbol{\kappa}}_0$ and $\hat{\sigma}^2_{(1)} = \hat{\sigma}^2_0$.

Step 3 Calculation of updates if $\hat{\boldsymbol{\kappa}}_{(1)} \leq \mathbf{0}$ from step 2:

(i) If some elements of $\hat{\boldsymbol{\kappa}}_{(1)}$ are less than zero, set the corresponding elements of $\hat{\boldsymbol{\kappa}}_{(1)}$ to zero and redo the update using a modification of step 2 (i) under updating Scheme A which updates only those non-zero elements of $\hat{\boldsymbol{\kappa}}_{(1)}$ together with ω_i, i.e.

$$\hat{\boldsymbol{\phi}}_{i(1)} = \begin{pmatrix} \hat{\omega}_{i(1)} \\ \\ \hat{\boldsymbol{\kappa}}^*_{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ \\ \hat{\boldsymbol{\kappa}}^*_0 \end{pmatrix} + \boldsymbol{\mathcal{I}}_i^{-1} \begin{pmatrix} \dfrac{(1-v_i)(t_i^2-1)}{2} \\ \\ \mathbf{0} \end{pmatrix} \tag{6.30}$$

If after step 3 (i) $\hat{\omega}_{i(1)} < 0$ or $\hat{\boldsymbol{\kappa}}^*_0 < 0$, then the update has failed. Then set $\hat{\omega}_{i(1)} = 0$, $\hat{\boldsymbol{\kappa}}_{(1)} = \hat{\boldsymbol{\kappa}}_0$ and $\hat{\sigma}^2_{(1)} = \hat{\sigma}^2_0$.

(ii) If all elements of $\hat{\boldsymbol{\kappa}}_{(1)}$ are less than zero, set $\hat{\boldsymbol{\kappa}}_{(1)} = \mathbf{0}$ and only update ω_i as (see Chapter 5 5-24)

$$\hat{\omega}_{i(1)} = \mathcal{I}^{11}(\omega_i = 0, \hat{\sigma}^2_{(1)}) U_i(\omega_i = 0, \hat{\sigma}^2_{(1)}),$$

Again this step implies that the null model fitted to the data is the linear regression model and no longer the linear mixed model. However, it does allow for the calculation of the one-step likelihood ratio statistic when the $\boldsymbol{\kappa}_{(1)} = 0$ from step 1.

If after step 3 (ii) $\hat{\omega}_{i(1)} < 0$, then set $\hat{\omega}_{i(1)} = 0$, $\hat{\boldsymbol{\kappa}}_{(1)} = \hat{\boldsymbol{\kappa}}_0$ and $\hat{\sigma}^2_{(1)} = \hat{\sigma}^2_0$.

6.4 Score tests for variance shift parameter

The score test statistics for the linear regression VSOM, discussed in the previous chapter (§ 5.3), can be extended to the linear mixed model VSOM using the score statistic (6.11) and the respective information matrices consisting of the terms in Results 6.2 to 6.5.

The score test statistic for ω_i in the linear mixed model VSOM is similar to (5.39) and takes the form

$$S_i(\omega_i = 0) = U_i^2(\omega_i = 0)\mathcal{I}^{11}, \tag{6.31}$$

where U_i is score statistic (6.11) and $\mathcal{I}^{11}$ is the portion of the inverse of information matrix associated with ω_i evaluated under the null hypothesis. $\mathcal{I}^{11}$ is calculated as

$$\mathcal{I}^{11} = \left\{\mathcal{I}_{11} - \boldsymbol{\mathcal{I}}_{12}\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{12}'\right\}^{-1}, \tag{6.32}$$

where $\boldsymbol{\mathcal{I}}_{22}$, and $\boldsymbol{\mathcal{I}}_{12}$ are as defined in (6.23) and are evaluated under the null model with $\omega_i = 0$, $\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}_0$ and $\sigma^2 = \hat{\sigma}_0^2$.

Some elements of $\hat{\boldsymbol{\kappa}}_0$ from the fitted null model (2.1) may be very small, say $\approx 10^{-5}$, leading to a negative variance for the score statistic i.e. $\mathcal{I}^{11} < 0$. Below we describe a computing algorithm for obtaining $\mathcal{I}^{11}$ to circumvent this problem. The description follows the approach adopted in the updating schemes for the variance components, i.e. the simple variance components model is used to describe the algorithm.

Step 1 If all elements of $\hat{\boldsymbol{\kappa}}_0$ are greater than 10^{-5} then compute the null model information matrix using Results 6.2 to 6.5 (either observed, expected, approximate average and exact average) and write the matrix in form of (6.23). Then compute $\mathcal{I}^{11}$ using (6.32).

Step 2 If some elements of $\hat{\boldsymbol{\kappa}}_0$ are less than or equal to 10^{-5} then set corresponding elements of $\boldsymbol{\mathcal{I}}_{12}$ to zero (the information matrix elements involving ω_i and $\boldsymbol{\kappa}$) and then compute $\mathcal{I}^{11}$ as in step 1.

Step 3 If all elements of $\hat{\boldsymbol{\kappa}}_0$ are less than or equal to 10^{-5} then $\boldsymbol{\mathcal{I}}_{12}$ contains only the information matrix element involving ω_i and σ^2 as in Chapter 5, a VSOM in linear regression. Again $\mathcal{I}^{11}$ is computed as in step 1.

Analogous to the likelihood ratio test, we can consider a score test calculated whilst holding the parameters $\boldsymbol{\kappa}$ fixed. The score test then takes the form

$$S_i(\omega_i = 0) = U_i^2(\omega_i = 0) \left\{ \mathcal{I}(\omega_i, \omega_i) - \boldsymbol{\mathcal{I}}(\omega_i, \sigma^2) \boldsymbol{\mathcal{I}}^{-1}(\sigma^2, \sigma^2) \boldsymbol{\mathcal{I}}(\omega_i, \sigma^2) \right\}^{-1}$$

In the case where the expected information matrix is used to calculate this statistic, the partial score test is evaluated as

$$S_i(\omega_i = 0) = \begin{cases} \dfrac{(n-p)}{(n-p-1)}(t_i^2 - 1)^2 & t_i^2 > 1 \\ 0 & \text{otherwise,} \end{cases}$$

with the information matrix evaluated at null model estimates $\omega_i = 0$, $\sigma^2 = \hat{\sigma}_0^2$ and $\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}_0$. This score test statistic is an extension of the score test statistic (5.41b) to a linear mixed VSOM. As with the partial likelihood ratio test statistic (6.16), this score test ignores any perturbation of elements of $\boldsymbol{\kappa}$ due to down-weighting the ith observation. The partial LRT and this partial score test take similar values for small values of t_i^2, i.e. $t_i^2 < 1.5$, but the partial score test increases faster as t_i^2 increases.

We can also extend the partial score test based on the expected information matrix to partial score tests based on the observed, approximate average and exact average information matrices. These partial score tests could be obtained using the formulae in (5.41a), (5.41c) and (5.41d).

6.5 Assessing significance and multiple testing

The same multiple testing issues arise for a linear mixed VSOM as for the linear regression VSOM. A modified version of the resampling procedure described in 5.4.3 is used to handle the problem of multiple testing. The procedure is modified as follows:

Step 1 For a given data set, fit an appropriate model (the null model).

Step 2 Generate the data vector $\boldsymbol{y}^*$ under the fitted model and fit the null model in step 1 to the simulated data with the

(i) variance parameters fixed at their null model estimates or

(ii) variance ratios fixed and error variance re-estimated or

(ii) all variance parameters re-estimated.

Steps 2 (ii) and 2 (iii) take into account the variation in the variance estimates whereas step 2 (i) does not. In our illustration with simulated data later (in § 6.6) we re-estimate all the variance parameters as this re-estimation mimics the fitting of a linear mixed model to an appropriate real data set.

Step 3 Computation of test statistics for the simulated data set

Step 4 For each set of computed test statistics in steps 3 compute the maximum statistic T^*.

Step 5 Repeat steps 2 to 4 R times, for R reasonably large, for example $R = 10000$.

Step 6 Calculate the $100(1-\alpha)$th percentile of the R values of T^* for a given α.

The test statistics computed in step 3 can be any of the test statistics discussed earlier in this chapter, in § 6.3 and § 6.4.

In the following three sections we illustrate the use of a linear mixed VSOM to detect single-case outliers using simulated data sets and two real data sets.

6.6 Example: Simulated data

We generated data from the simple one-way random effects model as

$$\boldsymbol{y} = \mu \mathbf{1}_n + (\boldsymbol{I}_g \otimes \mathbf{1}_r)\boldsymbol{u} + 4\boldsymbol{v}_5 + 8\boldsymbol{v}_{25} + \boldsymbol{e},$$

for $g = 5, 10$, $r = 5, 10$ with $n = gr$, where g and r represent the number of groups and the number of observations per group respectively, $\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\gamma\boldsymbol{I}_g)$, $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$, $\mathbf{1}_n = (1\ldots1)'$ is a vector of length n and $\boldsymbol{v}_1$ and $\boldsymbol{v}_{25}$ are vectors of length n each with a value 1 in the 5th and 25th elements and zero elsewhere, respectively, $\mu = 0$, $\sigma^2 = 1$ and $\gamma = 0.5$. The purpose of the terms $4\boldsymbol{v}_5$ and $8\boldsymbol{v}_{25}$ is to insert outliers of different sizes at the observations 5 and 25 in each data set, so that there are at least two outliers in each data set. The observations will be denoted as 1.5 and 5.5.

We fit an initial model to the simulated data (with the two imposed outliers) and obtain values of t_i^2, for $i = 1, \ldots, g \times r$. We then fit the ith VSOM for units with $t_i^2 > 1$ only and obtain variance estimates for the ith VSOM which include the variance shift estimate $\hat{\omega}_i$. We also calculated the partial variance estimates for $\hat{\omega}_i$ and σ^2 using the analytical expressions (6.13) and (6.14). The estimates we obtained were linearly correlated with REML variance shift estimates for the different sample sizes with correlation coefficients of over 0.99 for the different sample sizes. This indicates strong agreement indicates that the partial variance parameter estimates are good approximations to the REML variance estimates in this context. Again the estimated correlation coefficients assume that there is neither additive bias (estimated intercept parameter being non-zero) nor multiplicative bias (estimated slope parameter being non-zero) in a simple linear regression model for any pair of our test statistics.

Figures 6.1 presents the one-step variance shift estimates against REML variance estimates for the different information matrices with $n = 50, g = 10, r = 5$, based on updating scheme A, with the inserted outliers in the top right-hand corner of each scatter plot. The one-step updates based on the expected information matrix were a good approximation to the REML variance shift estimates whereas both the observed, approximate average and exact average information matrices gave poor approximations to the REML variance estimates. On the other hand one-step updates of the error variance were a good approximation to the REML variance estimates for all information matrices (Figures 6.2). The one-step updates of the variance parameter estimates obtained using updating scheme B exhibit similar behaviour to those estimates obtained using updating scheme A (Figures 6.3 and 6.4).

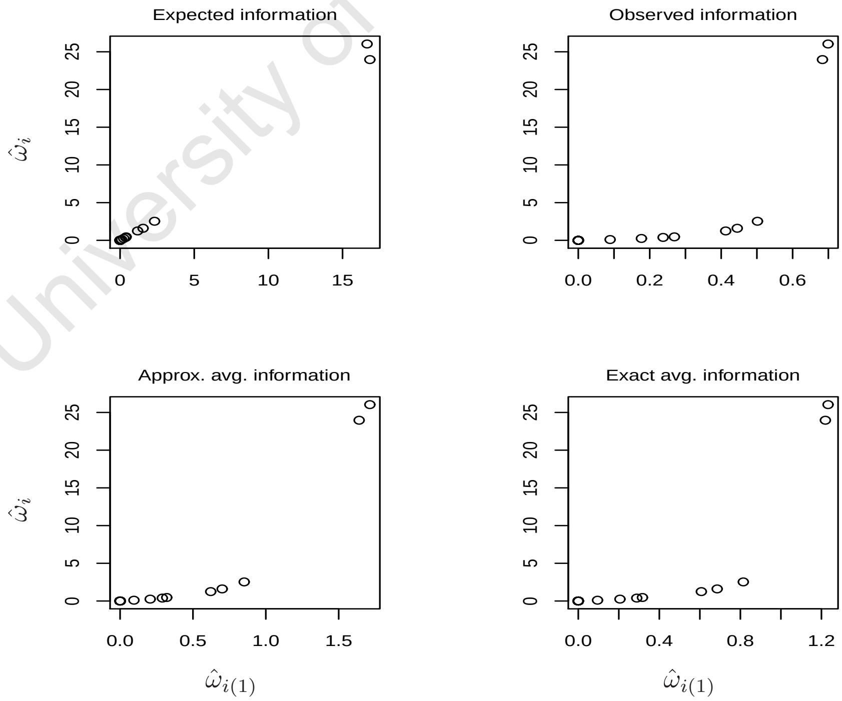

Figure 6.1: *Plots of REML variance shift estimates $\hat{\omega}_i$ against one-step variance shift estimates $\hat{\omega}_{i(1)}$ based on updating scheme A for the four different information matrices: $n = 50, g = 10, r = 5$.*

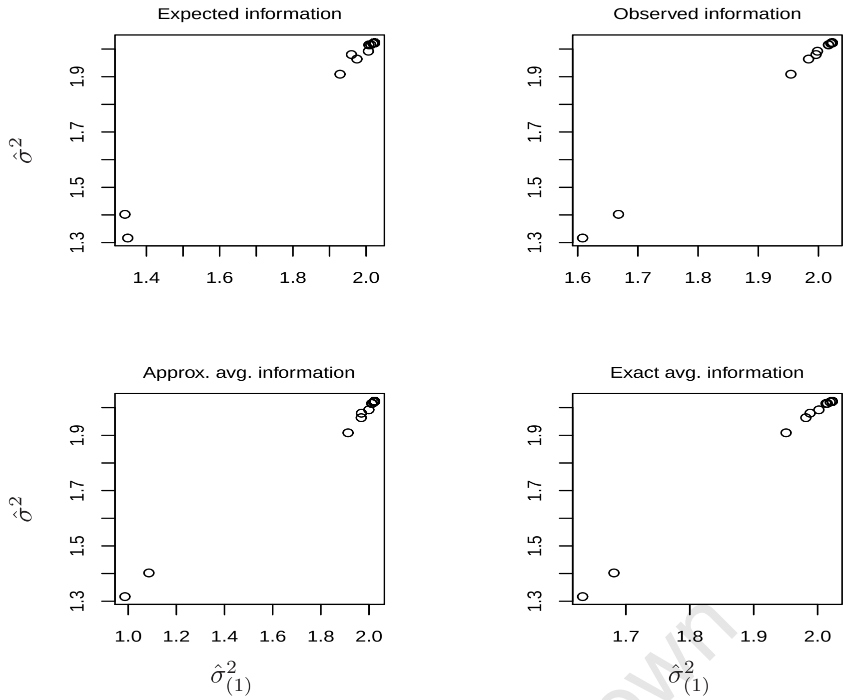

Figure 6.2: *Plots of REML variance estimates $\hat{\sigma}^2$ against one-step variance estimates $\hat{\sigma}^2_{(1)}$ based on updating scheme A for the four different information matrices: $n = 50, g = 10, r = 5$.*

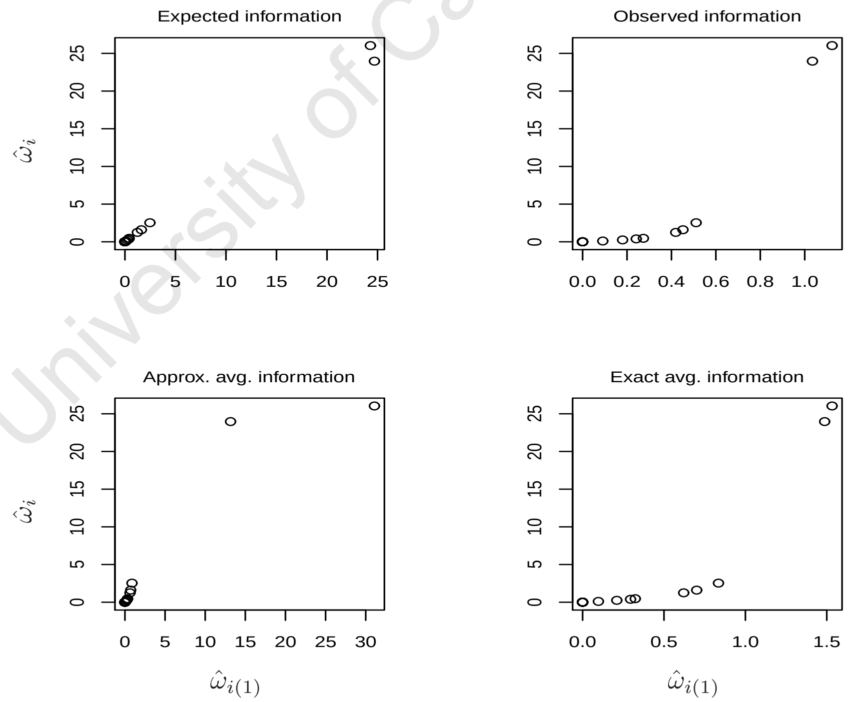

Figure 6.3: *Plots of REML variance shift estimates $\hat{\omega}_i$ against one-step variance shift estimates $\hat{\omega}_{i(1)}$ based on updating scheme B for the four different information matrices: $n = 50, g = 10, r = 5$.*

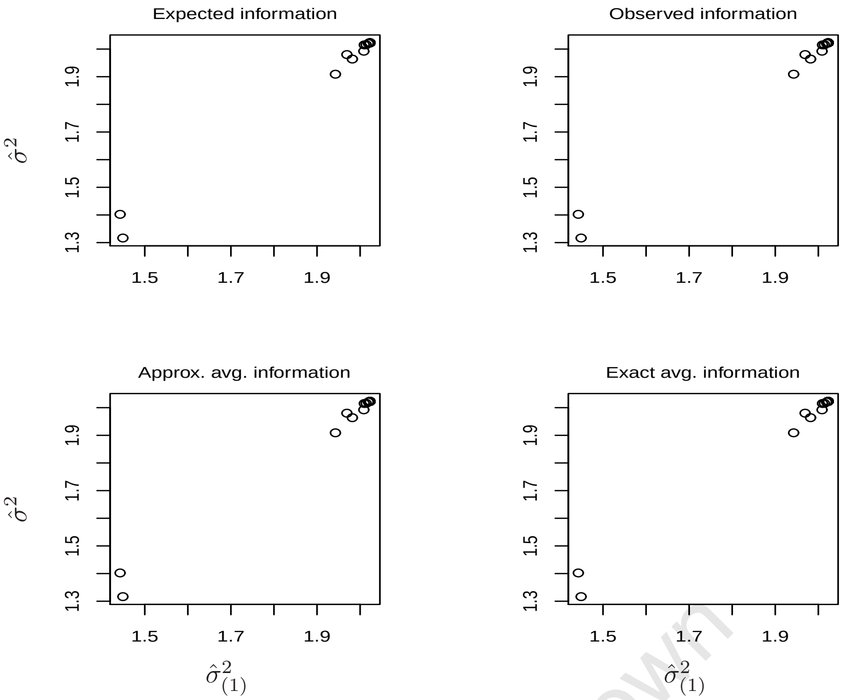

Figure 6.4: *Plots of REML variance estimates* $\hat{\sigma}^2$ *against one-step variance estimates* $\hat{\sigma}^2_{(1)}$ *based on updating scheme B for the four different information matrices:* $n = 50, g = 10, r = 5$*.*

Figure 6.5 shows index plots of the REML variance parameter estimates together index plots of t_i^2 for the different sample sizes. In all the sample sizes, observations 1.5 and 5.5 (5th and 25th observations) have relatively large values of t_i^2 with correspondingly large values of $\hat{\omega}_i$ and with decreased estimates of $\hat{\sigma}^2$. This pattern indicates that these observations are consistently suggested as possible outliers.

The likelihood ratio and score test statistics described in § 6.3 and 6.4 were then computed for observations with $t_i^2 > 1$. The partial LRT statistics were good approximations to the LRT statistics with Pearson's correlation coefficients of close to 1 for the different sample size (figures not shown). This observation was consistent with the relationship observed between the partial variance shift estimates and the REML variance shift estimates in (figures not shown). Similarly there was reasonable agreement between the partial score test statistics based on the four different information matrices (observed, expected, approximate average and exact average) and their full counterparts (figures not shown). For all the different versions of the likelihood ratio and score test statistics, both observation 1.5 and 5.5 had relative higher values of the calculated test statistic relative to other observations in the data set. For example taking $n = 50, g = 10, r = 5$, the LRT statistics for observation 1.5 and 5.5 were respectively 2.648 and 11.554 (Figure 6.6); the large values confirm that these observations may be possible outliers given fitted linear mixed model. As expected, both these observations were picked out as possible outliers on the basis of the 95th percentile of the largest LRT, partial LRT and the score test statistics or the 95th percentile of the second largest LRT, partial LRT and the score test statistics (Figure 6.6).

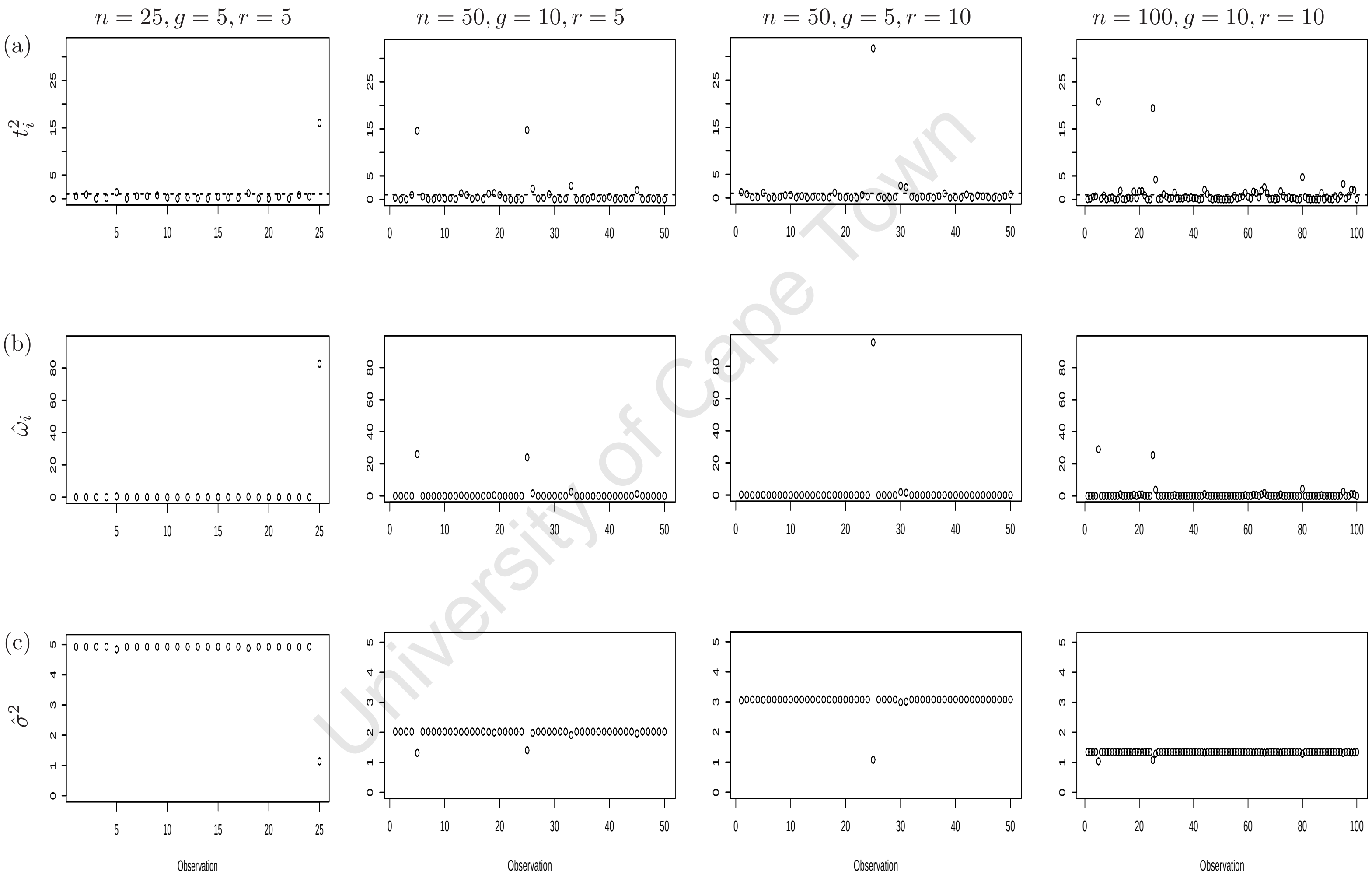

Figure 6.5: Index plots of (a) t_i^2 (with dashed line at $t_i^2 = 1$); (b) REML variance shift estimates, $\hat{\omega}_i$; and (c) error variance estimates $\hat{\sigma}^2$ for simulated data sets of different sizes.

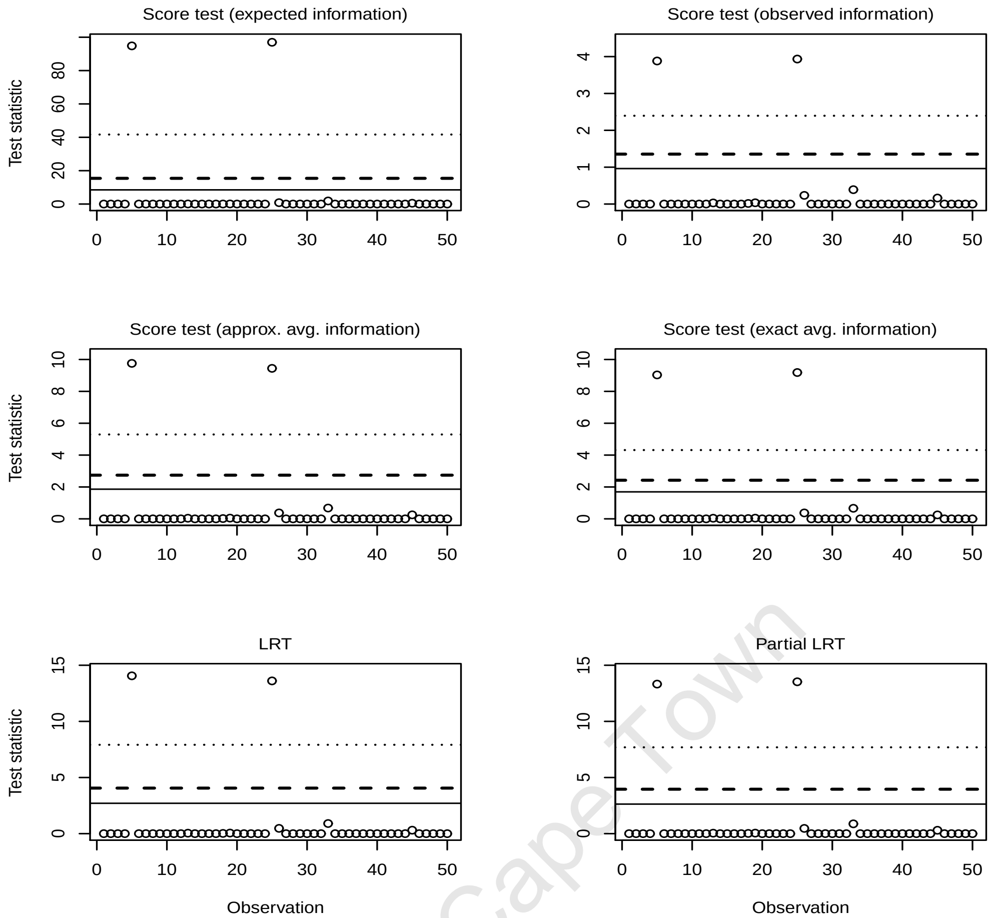

Figure 6.6: *Index plots of score test and likelihood ratio test statistics for the simulated data, with 95th percentiles of the empirical distributions under the null hypothesis (10 000 simulations) shown for the first k order statistics for each test:* $k = 1$ *(dotted line),* $k = 2$ *(dashed line) and* $k = 3$ *(solid line):* $n = 50, g = 10, r = 5$.

6.7 Example: The orthodont data

In this section we give an illustration of a linear mixed VSOM using the orthodont data set introduced in Chapter 2. In the previous chapter we fitted a linear regression model to the orthodont data set. However, the repeated measurements on each subject at the different ages violates the independence assumption that underlies the linear regression model fitted to the data. Therefore in this section we reconsider the random coefficient model as fitted in Chapters 2 and 4. Here we extend the random coefficient model (2.31) to account for the extra variability among the boys by fitting VSOM terms for common boys variation (the model given in § 5.7.1). The model allows for separate random effects for each boy but assumes a common variance shift for the data belonging to boys. The fitted model is similar to the model $\mathcal{M}_3$ of the previous chapter (see § 5.7.3).

For convenience we will index the observations, observation number $1, \ldots, 64$ for boys and observation number $65, \ldots, 108$ for girls, as in the previous chapter.

Figure 6.7 shows index plots of t_i^2 after fitting the null model (the model with no VSOM terms for boys but no VSOM terms individual observations) for $i = 1, \ldots, 108$. Observations 34, 35, 49 and 73 have relatively large values of t_i^2 compared to the remainder of the data and appear to be outliers. These observations were also identified as outliers in Chapters 4 and 5, except for unit 73. Observation 52 or 13.4 (the measurement for boy 13 at age 14) no longer appears to be an outlier. This change in the status of this observatin is due to the fact that the model fitted accounts for the extra variation among the data for boys (in contrast to model fitted in Chapters 2 and 4).

For each observation with $t_i^2 > 1$, a VSOM was fitted and variance estimates were obtained. Figure 6.7 also presents index plots of selected variance estimates ($\hat{\omega}_i$ and $\hat{\sigma}^2$); observations 34, 35, 49 and 73 have relatively large values of $\hat{\omega}_i$ together with decreased estimates of $\hat{\sigma}^2$. The different likelihood ratio and score test statistics were also calculated for each observation with $\hat{\omega}_i > 0$. Again the observations 34, 35, 49 and 73 have elevated test statistics compared to the rest of the data (Figure 6.8). The sampling distributions of the first, second, third and fourth order statistics of the different test statistics were constructed using 10000 simulated data sets. The 95th percentiles of the first, second, third and fourth order statistics of the sampling distributions of the different test statistics picked out observations 35, 49, 34 and 73 as outliers (Figure 6.8). We then fitted a combined VSOM for observations 35, 49, 34 and 73 as a final model. The estimated parameters from this model are given in Table 6.1 under the model denoted $\mathcal{M}_2$. The standard errors for the variance shift parameters are large relative to the variance shift estimates because the variance shift estimates are obtained from a single observation. The value of $-2\times$REML log-likelihood function (excluding constant terms) decreased from 220.38 in model $\mathcal{M}_1$ to 186.83 in model $\mathcal{M}_2$. Both the fixed effects estimates and their standard errors under $\mathcal{M}_2$ were different from those obtained from $\mathcal{M}_1$. However, the inferences regarding the fixed effects were the same under both models.

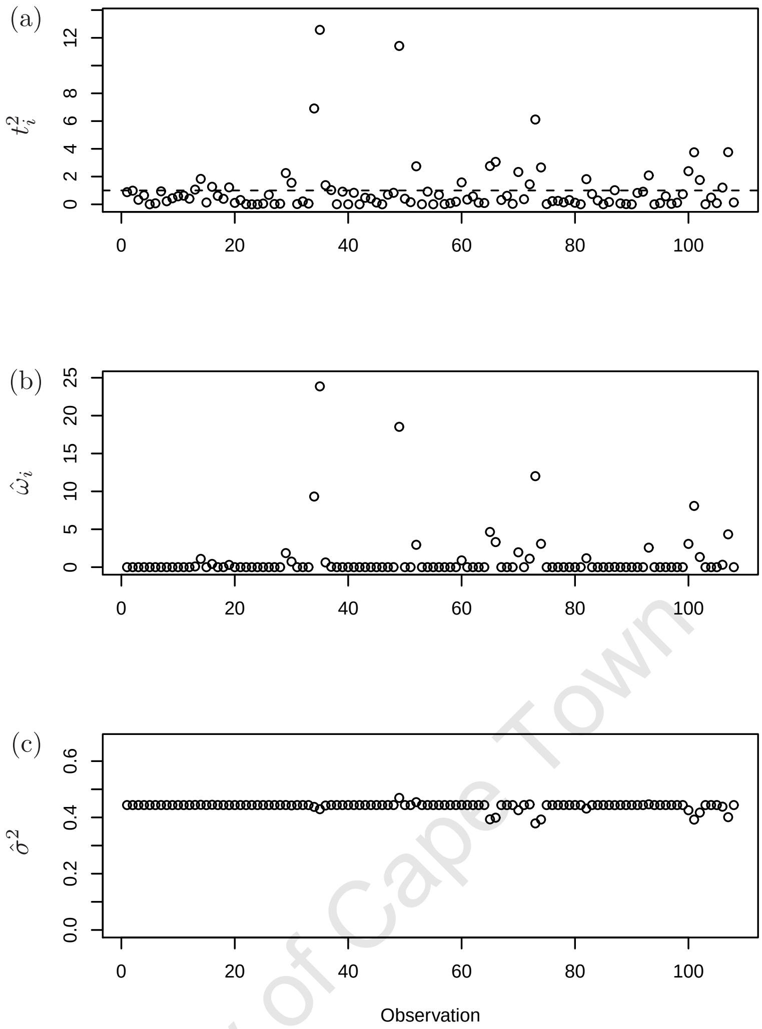

Figure 6.7: *Index plots of (a) t_i^2 (with dashed line at $t_i^2 = 1$), (b) REML variance shift estimates $\hat{\omega}_i$ and (c) error variance estimates $\hat{\sigma}^2$, for orthodont data.*

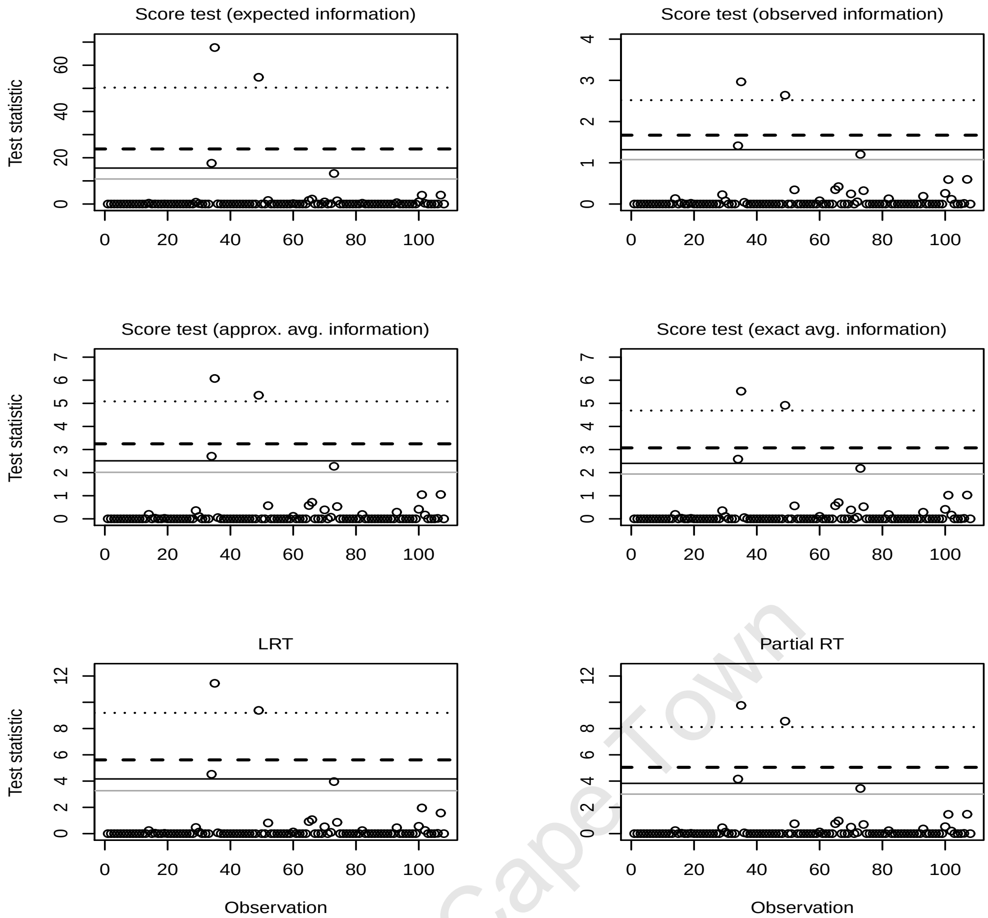

Figure 6.8: *Index plots of score test and likelihood ratio test statistics for the orthodont data, with 95th percentiles of the empirical distributions under the null hypothesis (10 000 simulations) shown for the first k order statistics for each test:* $k = 1$ *(dotted line),* $k = 2$ *(dashed line),* $k = 3$ *(solid line) and* $k = 4$ *(grey solid line).*

Table 6.1: *Estimated parameters for models fitted to orthodont data.*

Effect	Parameter	Model $\mathcal{M}_0$ Estimate (s.e.)	Model $\mathcal{M}_1$ Estimate (s.e.)	Model $\mathcal{M}_2$ Estimate (s.e.)
Fixed				
constant	μ	22.650 (0.586)	22.650 (0.568)	22.690 (0.586)
age	β_1	0.480 (0.104)	0.480 (0.065)	0.452 (0.057)
sex	β_2	2.321 (0.761)	2.231 (1.671)	2.333 (0.770)
sex.age	β_3	0.305 (0.135)	0.305 (0.119)	0.246 (0.090)
Random				
$\boldsymbol{d}_{34}$	$\omega_{34}\sigma^2$	-	-	10.332 (16.992)
$\boldsymbol{d}_{35}$	$\omega_{35}\sigma^2$	-	-	33.676 (50.063)
$\boldsymbol{d}_{49}$	$\omega_{49}\sigma^2$	-	-	41.526 (61.128)
$\boldsymbol{d}_{73}$	$\omega_{73}\sigma^2$	-	-	4.534 (7.344)
$\boldsymbol{D}_{\text{boys}}$	σ^2_{boys}	-	2.212 (0.586)	0.802 (0.288)
subject	$\sigma^2_{\text{subject}}$	3.350 (1.072)	3.441 (1.084)	3.679 (1.101)
subject.age	$\sigma^2_{\text{subject.age}}$	0.033 (0.037)	0.025 (0.021)	0.015 (0.015)
correlation	$\sigma^2_{\text{corr.}}$	0.068 (0.134)	0.114 (0.108)	0.110 (0.090)
	σ^2	1.716 (0.330)	0.444 (0.133)	0.381 (0.115)

6.8 Example: The aerosol data

For an additional illustration we re-analyze the aerosol data set from Beckman and Nachtsheim (1987) which were also analyzed by Christensen et al. (1992a). The data involve high-efficiency particulate air filter cartridges that are used in commercial respirators to prevent or reduce the respiration of toxic fumes, radionuclides, dusts, mists, and other particulate matter. The goals of the analysis are to determine factors that contribute most to the variability in penetration of the filters and to determine whether the standard aerosol can be replaced by an alternative aerosol in quality assurance testing. Two types of aerosol were tested on filters from two manufacturers. Within each manufacturer, three filters were used to evaluate the penetration of the two aerosols. Two observations from filter 5 (observation numbers 13 and 14) have elevated percent penetrations and were identified as outliers by Beckman and Nachtsheim (1987) and Christensen et al. (1992a). We use this data here to verify that a VSOM method identifies these same outliers.

We consider the following the model for the data (the same model fitted by Beckman and Nachtsheim, 1987)

$$y_{jklm} = \mu + \alpha_j + \beta_k + \eta_{kl} + e_{jklm}, \tag{6.33}$$

where, for $j = 1, 2$, $k = 1, 2$, $l = 1, 2, 3$ and $m = 1, 2, 3$, y_{jklm} is the percentage penetration, α_j is a fixed effect for the jth aerosol, β_k is a fixed effect for the kth filter manufacturer, η_{kl} is a random effect for the lth filter nested within the kth manufacturer, and e_{jklm} is the random error associated with the mth replicate in the jklth cell. Corner-point constraints are used to ensure identifiability of the fixed terms, with $\alpha_1 = 0$ and $\beta_1 = 0$. Random effects are assumed to be independent and normally distributed with $\text{var}(\eta_{kl}) = \sigma_\eta^2$ and $\text{var}(e_{jklm}) = \sigma^2$.

After fitting the initial model, denoted $\mathcal{N}_0$, the 2 observations from filter 5 (observations numbers 13 and 14) stand out as possible outliers with both observations giving large values of t_i^2 (Figure 6.9). Figure 6.9 also shows the estimates of the variance shift parameter, $\hat{\omega}_i$, and residual variance, $\hat{\sigma}^2$, for each observation, under a VSOM with single outliers; observations 13 and 14 have relatively large values of $\hat{\omega}_i$ together with decreased estimates of $\hat{\sigma}^2$. The likelihood ratio and score test statistics were also calculated for each VSOM, and 10000 simulated data sets were generated from the fitted model under the null hypothesis (no outliers present). In each simulation, a VSOM was fitted for each observation and the three largest values of each test statistic were saved and used to generate the empirical distribution of the order statistics for each test. The test statistics from the original data are shown in Figure 6.10 together with 95th percentile from the empirical distribution of the maximum, second and third largest value for each test statistic. The test statistic for observation 14 is larger than the 95th percentile of the distribution of the maximum, and the test statistic for observation 13 is larger than the 95th percentile of the distribution of the second largest value. The third highest test statistic is smaller than the 95th percentile of the distribution of the third largest value, and so we conclude that this data set contains only two outliers, at observations 13 and 14, in accordance with previous analyses of this data set.

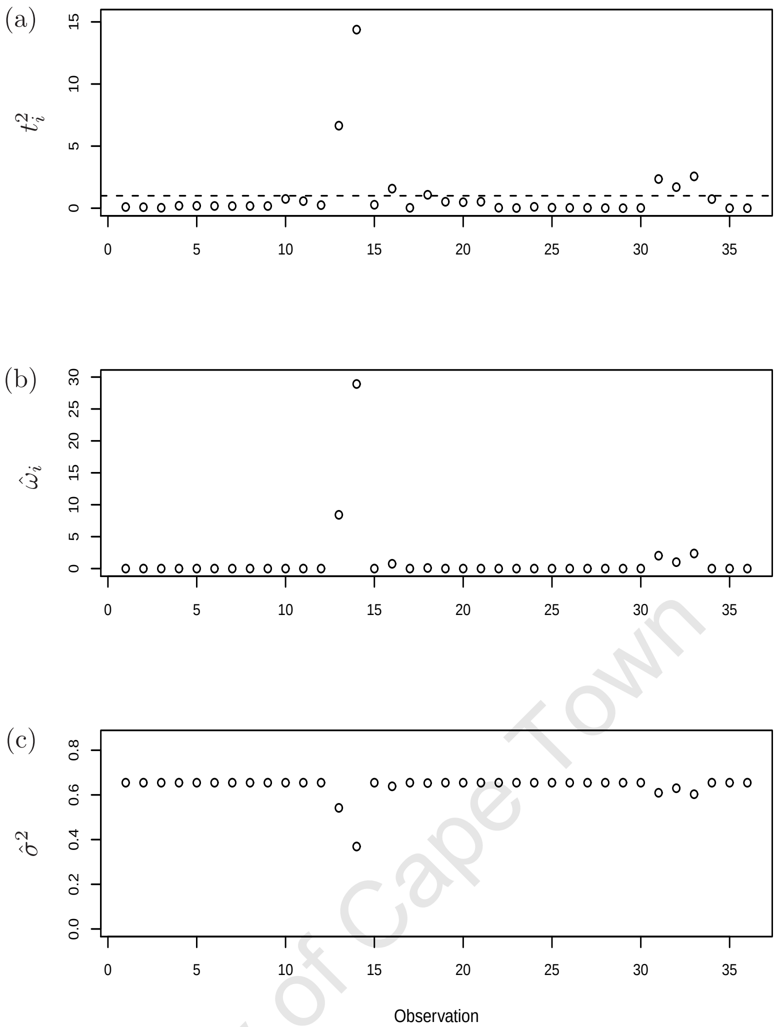

Figure 6.9: *Index plots of (a) t_i^2 (with dashed line at $t_i^2 = 1$); (b) REML variance shift estimates, $\hat{\omega}_i$; and (c) error variance estimates $\hat{\sigma}^2$ for aerosol data.*

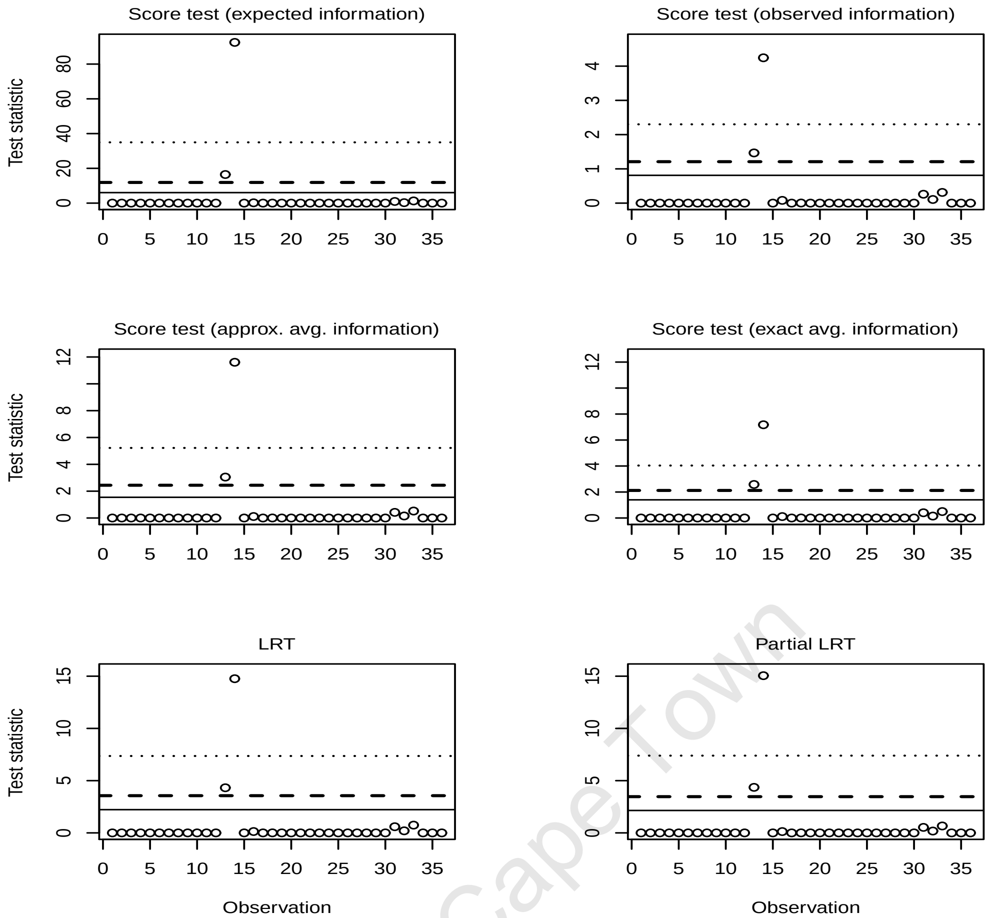

Figure 6.10: *Index plots of score test and likelihood ratio test statistics for the the aerosol data, with 95th percentiles of the empirical distributions under the null hypothesis (10 000 simulations) shown for the first k order statistics for each test: $k = 1$ (dotted line), $k = 2$ (dashed line) and $k = 3$ (solid line).*

Estimates under the initial model ($\mathcal{N}_0$) and the final model ($\mathcal{N}_1$, with VSOM terms for observations 13 and 14) are shown in Table 6.2. The value of $-2\times$REML log-likelihood (excluding constant terms) function decreased from 31.80 in model $\mathcal{N}_0$ to -12.67 in model $\mathcal{N}_1$. Down-weighting these observations has a large effect on estimates of the fixed effects (in particular the effect of aerosol 2 compared to aerosol 1) and changes the conclusions of the analysis. Some consideration must then be given to whether these outliers represent errors in data or measurements that should be down-weighted, whether they indicate a problem with the assumptions behind the analysis (as seems possible in this case) or whether other explanations are plausible; the correct action will depend on the aims and context of the analysis.

Table 6.2: *Estimated parameters for models fitted to aerosol data.*

Effect	Parameter	Model $\mathcal{N}_0$ Estimate (s.e.)	Model $\mathcal{N}_1$ Estimate (s.e.)
Fixed			
constant	μ	0.592 (0.373)	0.592 (0.297)
aerosol	α_2	-0.394 (0.270)	-0.005 (0.115)
manufacturer	β_2	1.194 (0.492)	0.805 (0.413)
Random			
filter	σ_η^2	0.253 (0.258)	0.236 (0.181)
$\boldsymbol{d}_{13}$	$\omega_{13}\sigma^2$	-	9.438 (13.546)
$\boldsymbol{d}_{14}$	$\omega_{14}\sigma^2$	-	15.800 (22.544)
	σ^2	0.655 (0.172)	0.109 (0.030)

Comparison between VSOM in linear mixed effects analysis and case-deletion approaches

For comparison purposes we also considered Cook's distance for variance parameters of Christensen et al. (1992a), (5.48). Note here the vector of variance parameters under the null model is $\hat{\boldsymbol{\phi}}_0$ where the null model is the linear mixed model (2.1). The criticism against the use of Cook's distance for variance parameters, mentioned in the previous chapter in § 5.6 (see pp. 5-51), also applies here., i.e. D_i is based on the information matrix for the variance parameters from the fit of the null model and does not account for the correlation between $\hat{\boldsymbol{\phi}}_0$ and $\hat{\boldsymbol{\phi}}_{(i)}$.

Under the null hypothesis for a VSOM (6.1), $\hat{\boldsymbol{\phi}}_0$ had only one parameter value ($\hat{\sigma}_0^2$), now $\hat{\boldsymbol{\phi}}_0$ has more than one variance parameter value, i.e. all the variance parameter values from fitting model (2.1). An alternative version of the statistic (5.48), which is quicker computationally, is obtained by using the one-step estimate $\hat{\boldsymbol{\phi}}_{(1)}$ instead of the REML estimate $\hat{\boldsymbol{\phi}}$ (Christensen et al., 1992a). We did not consider this alternative Cook's statistic in this research.

Example: The orthodont data

Figure 6.11 shows an index plot of Cook's statistics (5.48) for the orthodont data set after fitting the model $\mathcal{M}_0$. According to these Cook's statistics, observations 34, 35, 49, 73 and 101 appear to be outliers. The fixed effects estimates when these 5 observations were deleted are shown in Table 6.3 under the model denoted $\mathcal{M}_3$. The estimates and their standard errors differed slightly from those obtained from fitting VSOM (model $\mathcal{M}_2$). However, the inferences regarding the fixed effects were the same under both models.

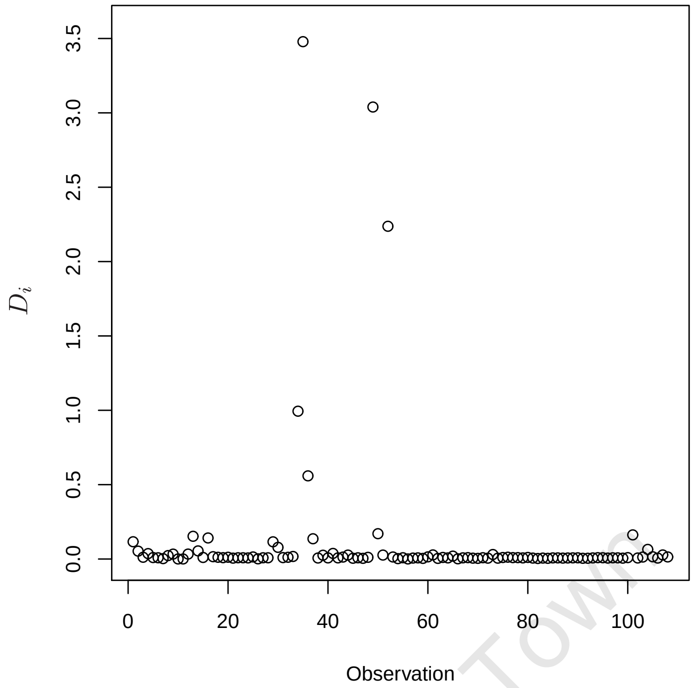

Figure 6.11: *Index plots of Cook's distance for variance parameters for orthodont data.*

Table 6.3: *Estimated parameters for the case-deletion model*† *fitted to orthodont data.*

Effect	Parameter	Model $\mathcal{M}_3$ Estimate (s.e.)
Fixed		
constant	μ	22.650 (0.587)
age	β_1	0.480 (0.075)
sex	β_2	2.379 (0.766)
sex.age	β_3	0.209 (0.100)
Random		
subject	$\sigma^2_{\text{subject}}$	3.583 (1.080)
subject.age	$\sigma^2_{\text{subject.age}}$	0.019 (0.020)
correlation	$\sigma^2_{\text{corr.}}$	0.110 (0.101)
	σ^2	0.850 (0.168)

† Observations deleted: 34, 35, 36, 49, and 52.

Example: The aerosol data

In the preceding example (the orthodont data set example), both a VSOM and case-deletion do not change the inferences on the fixed effects even though the approaches treat possible outliers differently; a VSOM down-weights the possible outliers using the variance shift estimates whereas the case-deletion removes the observations completely from the analysis. In Table 6.4 we also reproduced Christensen et al.'s case-deletion results, the model denoted $\mathcal{N}_2$ (model estimates with observations 13 and 14 deleted); this model gave a positive sign for the effect of aerosol even though this effect was not statistically significant. This example demonstrates that a VSOM approach can identify the same outliers as case-deletion but can also change the conclusions of an analysis.

Table 6.4: *Estimated parameters for the case-deletion model† fitted to aerosol data.*

Effect	Parameter	Model $\mathcal{N}_2$ Estimate (s.e.)
Fixed		
constant	μ	0.395 (0.298)
aerosol	α_2	0.001 (0.115)
manufacturer	β_2	0.799 (0.414)
Random		
filter	σ^2_η	0.238 (0.182)
	σ^2	0.109 (0.030)

† Observations deleted: 13 and 14.

6.9 Performance of likelihood ratio and score test statistics

In this section we compare the performance of the different likelihood ratio test and score test statistics in terms of computing time, type I error and power.

6.9.1 Computational efficiency

We compared the performance of selected likelihood ratio and score test statistics in terms of computing time when the resampling algorithm is implemented. We simulated data from the simple one-way random effects model as

$$\boldsymbol{y} = \mu \boldsymbol{1}_n + (\boldsymbol{I}_g \otimes \boldsymbol{1}_r)\boldsymbol{u} + 4\boldsymbol{v}_1 + \boldsymbol{e},$$

for $n = gr$, $g = 10$, $r = 5$, where g and r represent the number of groups and the number of observations per group respectively, $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\gamma\boldsymbol{I}_g)$, $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, $\boldsymbol{1}_n = (1\ldots1)'$ is a vector of length n and $\boldsymbol{v}_1 = (1\ 0\ \ldots 0)'$ is a vector of length n with value 1 in the first element and zero elsewhere, $\mu = 0$ and $\sigma^2 = 1$. The purpose of the term $4\boldsymbol{v}_1$ is to insert an outlier at the first observation, so that there at least one outlier in the data set.

The results of the simulation study for performance of the tests are shown in Table 6.5 where the abbreviation mins:ss refers to minutes and seconds of elapsed time. The simulation study was conducted on a 3.4 GHz Pentium processor. The computation of the partial LRT statistic was quicker compared to the LRT statistic since it can be evaluated directly using (6.16), i.e. it does not require the fitting of the alternative model; the same observation applies to the partial score test statistics. The one-step LRT statistic takes longer to compute because it requires the both calculation of the one-step updates of the variance parameters and evaluation of the log-likelihood function at the one-step updates. For the simple variance components example we considered, the score test statistics were quicker than all the versions of the likelihood ratio test statistics. There was little difference between the computing times of the full score test statistics and partial score test statistics.

Table 6.5: *Comparison of the performance of selected likelihood ratio and score test statistics when the resampling algorithm is implemented:* $n = 50, g = 10, r = 5$ *for 5000 simulations. The abbreviation mins:ss refers to minutes and seconds.*

Test statistic	Time (mins:ss)
Lik. test stat.	13:37
Partial lik. ratio test stat.	05:28
Test stat. based on the expected info. matrix	
Sore test stat.	06:27
Partial score score test stat.	05:28
One-step lik. ratio test stat.: updating scheme A	16:01

6.9.2 Type I error

A simulation study for a simple one-way random effects model was conducted to assess performance of the various proposed test statistics in terms of type I error (i.e. how often do the test statistics erroneously detect an outlier when there is none in the data set). The type I errors were computed for a single unit, with respect to both the standard asymptotic distributions and in terms of the empirical distribution generated by a parametric bootstrap procedure. Evaluation of the performance of the empirical distribution of order statistics was computationally impractical; this difficulty is discussed further in § 6.10.

For each combination of parameters, 500 data sets were generated. The jth simulated data set was generated as

$$\boldsymbol{y}_j = \mu \mathbf{1}_n + (\boldsymbol{I}_g \otimes \mathbf{1}_r)\boldsymbol{u}_j + \boldsymbol{e}_j,$$

for $n = gr$ and $j = 1, \ldots, 500$ where $\boldsymbol{u}_j \sim N(\mathbf{0}, \sigma^2\gamma \boldsymbol{I}_g)$, $\boldsymbol{e}_j \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, $\mathbf{1}_n = (1 \ldots 1)'$ is a vector of length n, $\mu = 0$ and $\sigma^2 = 1$. We consider 2 scenarios: (i) $n = 36$ with $g = 12, 6, 3$ and $r = 3, 6, 12$; and (ii) $n = 72$ with $g = 24, 12, 6, 3$ and $r = 3, 6, 12, 24$. The 2 scenarios are adopted to allow for a wider range of combinations of sample sizes (g and r). In both scenarios simulation data sets were generated for different values of the variance ratio, i.e. $\gamma = 0.1$, 1 and 10, representing small, medium and large group effects.

For each simulated data set, several VSOM test statistics were calculated for the first observation: the likelihood ratio test (LRT (6.15) or partial LRT (6.16) or one-step (6.17)) and the score test statistics (6.31) based on the four information matrices. The choice of the first observation was arbitrary; any observation could have been chosen. The test statistics were compared to the 90th, 95th, 97.5th and 99th percentiles of the standard asymptotic distribution of a $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution, respectively. In addition, motivated by the known distribution in a VSOM in linear regression, test statistics were compared to the same percentiles of a $0.68\chi_0^2 + 0.32\chi_1^2$ mixture distribution. To generate an empirical distribution of the test statistic under the null hypothesis, data sets for $k = 1 \ldots 2500$ were simulated as

$$\boldsymbol{y}_{jk} = \hat{\mu}_j \mathbf{1}_n + (\boldsymbol{I}_g \otimes \mathbf{1}_r)\boldsymbol{u}^*_{jk} + \boldsymbol{e}^*_{jk},$$

where $\hat{\mu}_j$ is the estimate of μ from $\boldsymbol{y}_j$, $\boldsymbol{u}^*_{jk} \sim N(0, \hat{\sigma}_j^2 \hat{\gamma}_i \boldsymbol{I}_g)$ and $\boldsymbol{e}^*_{jk} \sim N(0, \hat{\sigma}_j^2 \boldsymbol{I}_n)$ with $\hat{\sigma}_j^2$ and $\hat{\gamma}_j$ estimated from $\boldsymbol{y}_j$. All tests were performed for a VSOM for the first observation of each simulated data set $\boldsymbol{y}_{jk}$, $k = 1 \ldots 2500$, and the 90th, 95th, 97.5th and 99th percentiles from the empirical distribution of each test statistic were used as threshold values for the test statistics observed on the original data set $\boldsymbol{y}_j$.

The 500 data sets were used to obtain estimates of type I errors associated with the percentiles from the standard and empirical sampling distributions i.e. the type I error estimate for a given test statistic, and level of significance (α) was calculated as the number of data sets (out of 500) for which the test statistic exceeded the $100(1-\alpha)$ percentile of the empirical distribution.

Results

Similar to Table 5.5 in Chapter 5, Tables 6.6 and 6.7 show the mean value and standard deviation (over 500 data sets) of the 95th percentile of the empirical distribution, under the null hypothesis, for the LRTs and the score test statistics, respectively. This percentile is the value that an observed test statistic would have to exceed in order to be regarded as an outlier. It is clear from this table that the different test statistics have quite different thresholds: the LRT and partial LRT statistics have much lower thresholds than the score test based on the expected information matrix, but higher thresholds than score tests based on the observed or average information matrices. With the exception of the score test based on the expected information matrix, in all cases the thresholds decrease slightly as the sample size increases, although there is no clear pattern with respect to the number of groups or the variance ratio (γ). The partial LRT is on average only slightly smaller than the LRT, with similar standard deviation. These tables also illustrate that the standard deviation of the 95% thresholds using 2500 simulation is acceptable, at under 10% of the threshold value.

Table 6.6: *95th percentiles (mean (std. deviation)) from empirical distributions of t_i^2, likelihood ratio test statistics (LRT and one-step LRTs one-step (updating scheme A)) for the first unit based on 2500 simulations of 500 data sets with 36 and 72 units each.*

No. of groups	Obs./ group	γ	t_i^2	LRT	Partial LRT	One-step LRT using: Exp. inf.	Obs. inf.	Approx. avg. inf.	Exact avg. inf.
12	3	0.1	0.951 (1.208)	1.644 (0.114)	1.569 (0.107)	1.573 (0.108)	0.835 (0.047)	1.134 (0.066)	1.134 (0.067)
		1	0.964 (1.338)	1.658 (0.120)	1.574 (0.115)	1.595 (0.115)	0.846 (0.049)	1.135 (0.069)	1.144 (0.072)
		10	0.984 (1.339)	1.631 (0.117)	1.563 (0.110)	1.613 (0.115)	0.840 (0.049)	1.115 (0.066)	1.138 (0.071)
6	6	0.1	1.117 (1.518)	1.633 (0.118)	1.580 (0.113)	1.581 (0.113)	0.839 (0.049)	1.138 (0.071)	1.135 (0.071)
		1	0.959 (1.272)	1.618 (0.118)	1.571 (0.116)	1.580 (0.117)	0.835 (0.050)	1.131 (0.071)	1.130 (0.072)
		10	1.092 (1.548)	1.600 (0.114)	1.573 (0.111)	1.590 (0.112)	0.830 (0.049)	1.125 (0.069)	1.125 (0.070)
3	12	0.1	0.995 (1.425)	1.615 (0.122)	1.579 (0.119)	1.579 (0.119)	0.834 (0.052)	1.134 (0.074)	1.130 (0.074)
		1	0.989 (1.410)	1.608 (0.112)	1.579 (0.109)	1.580 (0.110)	0.833 (0.047)	1.131 (0.068)	1.128 (0.068)
		10	0.980 (1.418)	1.587 (0.108)	1.570 (0.106)	1.575 (0.107)	0.825 (0.046)	1.121 (0.066)	1.118 (0.066)
24	3	0.1	0.892 (1.395)	1.587 (0.112)	1.548 (0.108)	1.550 (0.109)	0.823 (0.047)	1.124 (0.069)	1.119 (0.069)
		1	0.864 (1.372)	1.583 (0.115)	1.539 (0.111)	1.552 (0.113)	0.823 (0.049)	1.122 (0.072)	1.117 (0.071)
		10	0.898 (1.369)	1.560 (0.109)	1.528 (0.106)	1.554 (0.108)	0.815 (0.047)	1.114 (0.068)	1.108 (0.068)
12	6	0.1	0.918 (1.338)	1.564 (0.110)	1.533 (0.106)	1.534 (0.106)	0.816 (0.047)	1.113 (0.068)	1.108 (0.068)
		1	0.923 (1.353)	1.557 (0.118)	1.535 (0.117)	1.541 (0.117)	0.815 (0.051)	1.112 (0.074)	1.106 (0.074)
		10	0.955 (1.396)	1.548 (0.113)	1.535 (0.112)	1.545 (0.113)	0.811 (0.049)	1.109 (0.071)	1.104 (0.071)
6	12	0.1	1.030 (1.551)	1.560 (0.109)	1.538 (0.109)	1.539 (0.109)	0.815 (0.047)	1.113 (0.069)	1.108 (0.068)
		1	1.003 (1.492)	1.550 (0.106)	1.535 (0.105)	1.538 (0.105)	0.812 (0.046)	1.109 (0.067)	1.104 (0.066)
		10	1.003 (1.471)	1.537 (0.110)	1.531 (0.109)	1.535 (0.110)	0.807 (0.047)	1.103 (0.069)	1.098 (0.069)
3	24	0.1	1.024 (1.648)	1.551 (0.108)	1.536 (0.106)	1.536 (0.107)	0.812 (0.047)	1.109 (0.068)	1.104 (0.067)
		1	1.022 (1.620)	1.543 (0.111)	1.532 (0.110)	1.533 (0.110)	0.809 (0.048)	1.105 (0.070)	1.100 (0.069)
		10	1.024 (1.638)	1.531 (0.109)	1.527 (0.109)	1.528 (0.109)	0.805 (0.047)	1.099 (0.069)	1.094 (0.068)

Table 6.7: *95th percentiles (mean (std. deviation)) from empirical distributions of score tests for the first unit based on 2500 simulations of 500 data sets with 36 and 72 units each.*

No. of groups	Obs./ group	γ	Score test using Exp. inf.	Obs. inf	Approx. avg. inf.	Exact avg. inf.
12	3	0.1	3.987 (0.374)	0.651 (0.037)	1.168 (0.075)	1.083 (0.066)
		1	4.020 (0.405)	0.653 (0.038)	1.191 (0.081)	1.089 (0.069)
		10	4.007 (0.389)	0.646 (0.038)	1.204 (0.082)	1.085 (0.068)
6	6	0.1	4.023 (0.396)	0.648 (0.039)	1.168 (0.078)	1.085 (0.068)
		1	3.998 (0.408)	0.643 (0.039)	1.169 (0.081)	1.080 (0.070)
		10	4.012 (0.391)	0.637 (0.037)	1.171 (0.078)	1.079 (0.068)
3	12	0.1	4.018 (0.417)	0.642 (0.040)	1.163 (0.082)	1.082 (0.072)
		1	4.018 (0.382)	0.640 (0.037)	1.163 (0.075)	1.081 (0.066)
		10	3.990 (0.371)	0.633 (0.036)	1.156 (0.073)	1.073 (0.064)
24	3	0.1	4.081 (0.401)	0.634 (0.036)	1.118 (0.072)	1.079 (0.068)
		1	4.059 (0.412)	0.630 (0.038)	1.122 (0.075)	1.074 (0.070)
		10	4.031 (0.393	0.624 (0.036)	1.123 (0.072)	1.068 (0.066)
12	6	0.1	4.025 (0.393)	0.626 (0.036)	1.106 (0.071)	1.068 (0.067)
		1	4.039 (0.431)	0.623 (0.039)	1.109 (0.077)	1.067 (0.072)
		10	4.042 (0.415)	0.621 (0.037)	1.110 (0.075)	1.066 (0.070)
6	12	0.1	4.042 (0.403)	0.625 (0.036)	1.107 (0.071)	1.069 (0.067)
		1	4.034 (0.388)	0.622 (0.035)	1.105 (0.069)	1.066 (0.065)
		10	4.021 (0.402)	0.618 (0.036)	1.101 (0.072)	1.062 (0.068)
3	24	0.1	4.033 (0.392)	0.622 (0.036)	1.103 (0.070)	1.066 (0.066)
		1	4.021 (0.407)	0.619 (0.037)	1.100 (0.072)	1.063 (0.068)
		10	4.002 (0.400)	0.616 (0.036)	1.095 (0.071)	1.058 (0.067)

Table 6.8-6.9 reports the empirical type I errors for thresholds derived from the empirical distribution under the null hypothesis for the different likelihood ratio test (LRT or one-step LRT) and score test statistics for $\alpha = 0.05$ and $\alpha = 0.01$. The empirical type I error estimates are generally close to the nominal α values for all of the test statistics, i.e. there are no apparent differences in the empirical type I errors between the tests including the partial LRT. This behaviour of the tests may be due to the fact that the tests are highly correlated with each other (Table 6.10). Table 6.11 reports the empirical type I errors for thresholds derived from the mixtures of chi-squared distributions for the likelihood ratio tests and score test based on expected information. These two test statistics were chosen as those with well-developed asymptotic theory, for which standard distributions might be expected to hold. The $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution performed poorly in both cases, being conservative for the likelihood ratio test and anti-conservative for the score test. The $0.68\chi_0^2 + 0.32\chi_1^2$ mixture distribution gave a reasonable, if slightly conservative, approximation to the nominal type I error rate for the likelihood ratio test, but the same mixture distribution was anti-conservative for the score test.

Table 6.8: *Empirical type I errors of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets with 36 and 72 units each.*

			LRT		Partial LRT		One-step LRT using							
							Exp. inf.		Obs. inf.		Approx. avg. inf.		Exact. avg. inf.	
No. of groups	Obs./ group	γ	Nominal probability of rejection (α)											
			0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
12	3	0.1	0.054	0.012	0.052	0.012	0.052	0.010	0.052	0.012	0.050	0.016	0.050	0.012
		1	0.038	0.006	0.042	0.008	0.040	0.006	0.040	0.006	0.040	0.008	0.040	0.006
		10	0.044	0.014	0.046	0.014	0.044	0.014	0.044	0.014	0.044	0.012	0.044	0.014
6	6	0.1	0.048	0.018	0.050	0.018	0.050	0.018	0.048	0.018	0.048	0.018	0.048	0.018
		1	0.036	0.008	0.036	0.008	0.036	0.008	0.036	0.008	0.036	0.008	0.036	0.008
		10	0.040	0.002	0.040	0.002	0.040	0.002	0.040	0.002	0.040	0.002	0.040	0.002
3	12	0.1	0.054	0.006	0.052	0.004	0.054	0.004	0.054	0.006	0.054	0.004	0.054	0.004
		1	0.048	0.010	0.048	0.010	0.048	0.010	0.048	0.010	0.048	0.010	0.048	0.010
		10	0.064	0.008	0.064	0.008	0.064	0.008	0.064	0.008	0.064	0.008	0.064	0.008
24	3	0.1	0.054	0.014	0.056	0.010	0.056	0.008	0.054	0.010	0.054	0.010	0.054	0.010
		1	0.062	0.014	0.060	0.016	0.060	0.014	0.062	0.014	0.062	0.014	0.062	0.014
		10	0.040	0.006	0.042	0.006	0.040	0.006	0.040	0.006	0.040	0.006	0.040	0.006
12	6	0.1	0.052	0.012	0.050	0.014	0.054	0.014	0.052	0.014	0.052	0.014	0.052	0.012
		1	0.054	0.014	0.054	0.014	0.054	0.016	0.054	0.014	0.054	0.014	0.054	0.014
		10	0.048	0.008	0.048	0.008	0.048	0.008	0.048	0.008	0.048	0.008	0.048	0.008
6	12	0.1	0.062	0.010	0.066	0.010	0.066	0.010	0.064	0.010	0.066	0.010	0.066	0.010

Continued on next page

Table 6.8 – continued from previous page

			LRT		Partial LRT		One-step LRT using							
							Exp. inf.		Obs. inf.		Approx. avg. inf.		Exact. avg. inf.	
No. of	Obs./		Nominal probability of rejection (α)											
groups	group	γ	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
		1	0.060	0.012	0.060	0.012	0.060	0.012	0.060	0.012	0.060	0.012	0.060	0.012
		10	0.054	0.014	0.054	0.014	0.054	0.014	0.054	0.014	0.054	0.014	0.054	0.014
3	24	0.1	0.038	0.010	0.038	0.010	0.038	0.010	0.038	0.010	0.038	0.010	0.038	0.010
		1	0.062	0.014	0.062	0.014	0.062	0.014	0.062	0.014	0.062	0.014	0.062	0.014
		10	0.034	0.008	0.034	0.008	0.034	0.008	0.034	0.008	0.034	0.008	0.034	0.008

Table 6.9: *Empirical type I errors of the full score test statistics for the first unit based on 2500 simulations of 500 data sets with 36 and 72 units each.*

			Score test using							
			Exp. inf.		Obs. inf.		Approx. avg. inf.		Exact avg. inf.	
No. of groups	Obs./ group	γ	Nominal probability of rejection (α) 0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
12	3	0.1	0.052	0.012	0.052	0.012	0.052	0.010	0.050	0.012
		1	0.042	0.008	0.038	0.006	0.040	0.006	0.042	0.006
		10	0.046	0.014	0.044	0.014	0.044	0.014	0.046	0.014
6	6	0.1	0.050	0.018	0.048	0.018	0.050	0.018	0.048	0.018
		1	0.036	0.008	0.036	0.008	0.036	0.008	0.036	0.008
		10	0.040	0.002	0.040	0.002	0.040	0.002	0.040	0.002
3	12	0.1	0.052	0.004	0.054	0.006	0.054	0.004	0.054	0.004
		1	0.048	0.010	0.048	0.010	0.048	0.010	0.048	0.010
		10	0.064	0.008	0.064	0.008	0.064	0.008	0.064	0.008
24	3	0.1	0.056	0.010	0.054	0.010	0.054	0.008	0.056	0.010
		1	0.060	0.016	0.062	0.014	0.060	0.014	0.060	0.014
		10	0.042	0.006	0.040	0.006	0.040	0.006	0.040	0.006
12	6	0.1	0.050	0.014	0.052	0.012	0.052	0.014	0.052	0.014
		1	0.054	0.014	0.054	0.014	0.054	0.014	0.054	0.014
		10	0.048	0.008	0.048	0.008	0.048	0.008	0.048	0.008
6	12	0.1	0.066	0.010	0.064	0.010	0.066	0.010	0.066	0.010
		1	0.060	0.012	0.060	0.012	0.060	0.012	0.060	0.012
		10	0.054	0.014	0.054	0.014	0.054	0.014	0.054	0.014
3	24	0.1	0.038	0.010	0.038	0.010	0.038	0.010	0.038	0.010
		1	0.062	0.014	0.062	0.014	0.062	0.014	0.062	0.014
		10	0.034	0.008	0.034	0.008	0.034	0.008	0.034	0.008

Table 6.10: *Pearson's correlation coefficients between t_i^2, likelihood ratio test and score test statistics for the first unit for 500 data sets: $n = 72, p = 12, r = 6, \gamma = 1$.*

		t_i^2	LRT	Partial LRT	LRT$_{(1)}$: Scheme A				Score test			
					Exp.	Obs. avg.	Approx. avg.	Exact	Exp.	Obs.	Approx. avg.	Exact avg.
	t_i^2	1.000										
	LRT	0.916	1.000									
	Partial LRT	0.917	1.000	1.000								
LRT$_{(1)}$: Scheme A	Exp. inf.	0.916	1.000	1.000	1.000							
	Obs. inf.	0.944	0.994	0.994	0.994	1.000						
	Approx. avg.	0.937	0.996	0.996	0.996	1.000	1.000					
	Exact avg.	0.936	0.997	0.997	0.997	1.000	1.000	1.000				
Score test	Exp. inf.	0.810	0.957	0.955	0.957	0.922	0.928	0.931	1.000			
	Obs. inf.	0.935	0.998	0.998	0.998	0.998	0.999	0.999	0.942	1.000		
	Approx. avg.	0.909	0.998	0.998	0.998	0.988	0.990	0.992	0.970	0.995	1.000	
	Exact avg.	0.927	0.999	0.999	0.999	0.997	0.998	0.999	0.949	0.999	0.997	1.000

Table 6.11: *Empirical type I errors, based on asymptotic distributions, of likelihood ratio and expected information score test statistics (*$100(1-\alpha)$*th percentile) for the first unit based on 2500 simulations of 500 data sets with 36 and 72 units each.*

			LRT				Score test using exp. inf.			
			Asymptotic reference distribution							
			$0.5\chi_0^2+0.5\chi_1^2$		$0.68\chi_0^2+0.32\chi_1^2$		$0.5\chi_0^2+0.5\chi_1^2$		$0.68\chi_0^2+0.32\chi_1^2$	
No. of groups	Obs./ group	γ	Nominal probability of rejection (α)							
			0.05 (2.71)	0.01 (5.41)	0.05 (2.01)	0.01 (4.64)	0.05 (2.71)	0.01 (5.41)	0.05 (2.01)	0.01 (4.64)
12	3	0.1	0.038	0.008	0.044	0.010	0.072	0.046	0.094	0.048
		1	0.016	0.002	0.028	0.006	0.050	0.030	0.068	0.034
		10	0.026	0.012	0.034	0.014	0.060	0.034	0.068	0.038
6	6	0.1	0.026	0.004	0.034	0.008	0.070	0.036	0.086	0.042
		1	0.024	0.006	0.026	0.006	0.062	0.030	0.070	0.036
		10	0.006	0.002	0.016	0.002	0.056	0.018	0.062	0.024
3	12	0.1	0.018	0.002	0.028	0.002	0.060	0.032	0.076	0.046
		1	0.014	0.004	0.026	0.008	0.068	0.030	0.074	0.042
		10	0.026	0.004	0.044	0.008	0.090	0.050	0.098	0.058
24	3	0.1	0.020	0.006	0.034	0.008	0.082	0.040	0.098	0.048
		1	0.030	0.004	0.048	0.004	0.072	0.058	0.098	0.060
		10	0.022	0.002	0.030	0.006	0.064	0.032	0.076	0.032
12	6	0.1	0.020	0.006	0.036	0.008	0.086	0.038	0.094	0.044
		1	0.026	0.000	0.032	0.010	0.068	0.036	0.086	0.042
		10	0.014	0.004	0.020	0.006	0.064	0.036	0.074	0.046
3	24	0.1	0.024	0.004	0.032	0.008	0.060	0.034	0.068	0.036
		1	0.022	0.006	0.038	0.008	0.080	0.044	0.088	0.054
		10	0.020	0.002	0.022	0.006	0.050	0.026	0.064	0.028

6.9.3 Power

In order to examine the relative sensitivity of the different tests, an outlier of size λ =1, 2, 4, 8, 16 or 32 units was introduced into the first observation in the one-way random effects model. The model parameters remained as in the evaluation of type I error. For each combination of parameters, 100 data sets were generated. The jth simulated data set was generated as

$$\boldsymbol{y}_j = \mu \mathbf{1}_n + (\boldsymbol{I}_g \otimes \mathbf{1}_r)\boldsymbol{u}_j + \lambda \boldsymbol{v}_1 + \boldsymbol{e}_j,$$

for $j = 1, \ldots, 100$ where $\lambda = 1$, 2, 4, 8, 16 or 32 and $\boldsymbol{v}_1 = (1\ 0\ \ldots 0)'$ is a vector of length n with value 1 in unit 1 and zero elsewhere. Test statistics and their empirical distributions were calculated as for the type I errors above.

Results

The proportion of test statistics exceeding the $100(1-\alpha)$ percentiles of the simulated distribution under the null hypothesis are shown in Figures 6.12 and 6.13 for selected sample size combinations of 36 units, $\gamma = 0.1, 1$ or 10 and $\alpha = 0.05$. Figures 6.14 and 6.15 show corresponding empirical power estimates for selected sample size combinations of 72 units, $\gamma = 0.1, 1$ or 10 and $\alpha = 0.05$. Corresponding empirical power estimates for selected sample size combinations for $\alpha = 0.01$ are given in Appendix D, § D.2.

As expected, the power of the likelihood ratio test statistic (the partial likelihood ratio) increases as the size of the displacement, λ, increases (Figures 6.12 and 6.14). The power of the one-step likelihood ratio test when the exact average information matrix is used, increases as λ increases in line with the LRT. However, the power of the one-step likelihood ratio tests (based on updating scheme A) when the observed or approximate average information matrix is used, decreases for moderately large and very large values of the displacement, λ. This behaviour of the one-step LRTs in terms of power is similar to the behaviour of the one-step LRTs in a linear regression VSOM observed earlier in Chapter 5. Unlike in a linear regression VSOM, the power of the one-step likelihood ratio test when the expected information matrix is used, decreases for large displacement values, especially when the sample size is small, for example when $n = 36$ (Figures 6.12) . There was no detectable difference between the power of the different score tests at the specified values of λ for selected sample size combinations (Figures 6.13 and 6.15).

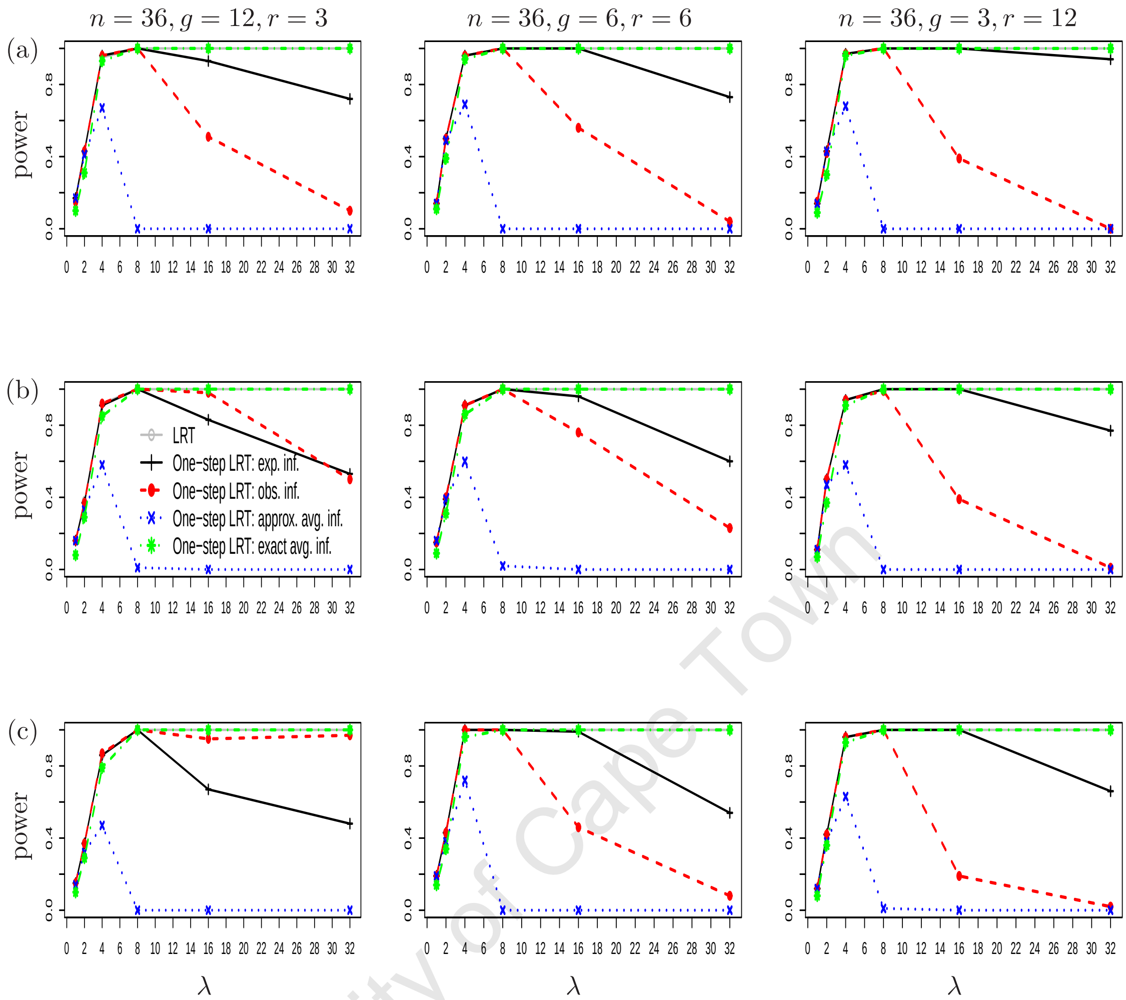

Figure 6.12: *Empirical power of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets with 36 units each for (a)* $\gamma = 0.1$*, (b)* $\gamma = 1$ *and (c)* $\gamma = 10$*:* $\alpha = 0.05$.

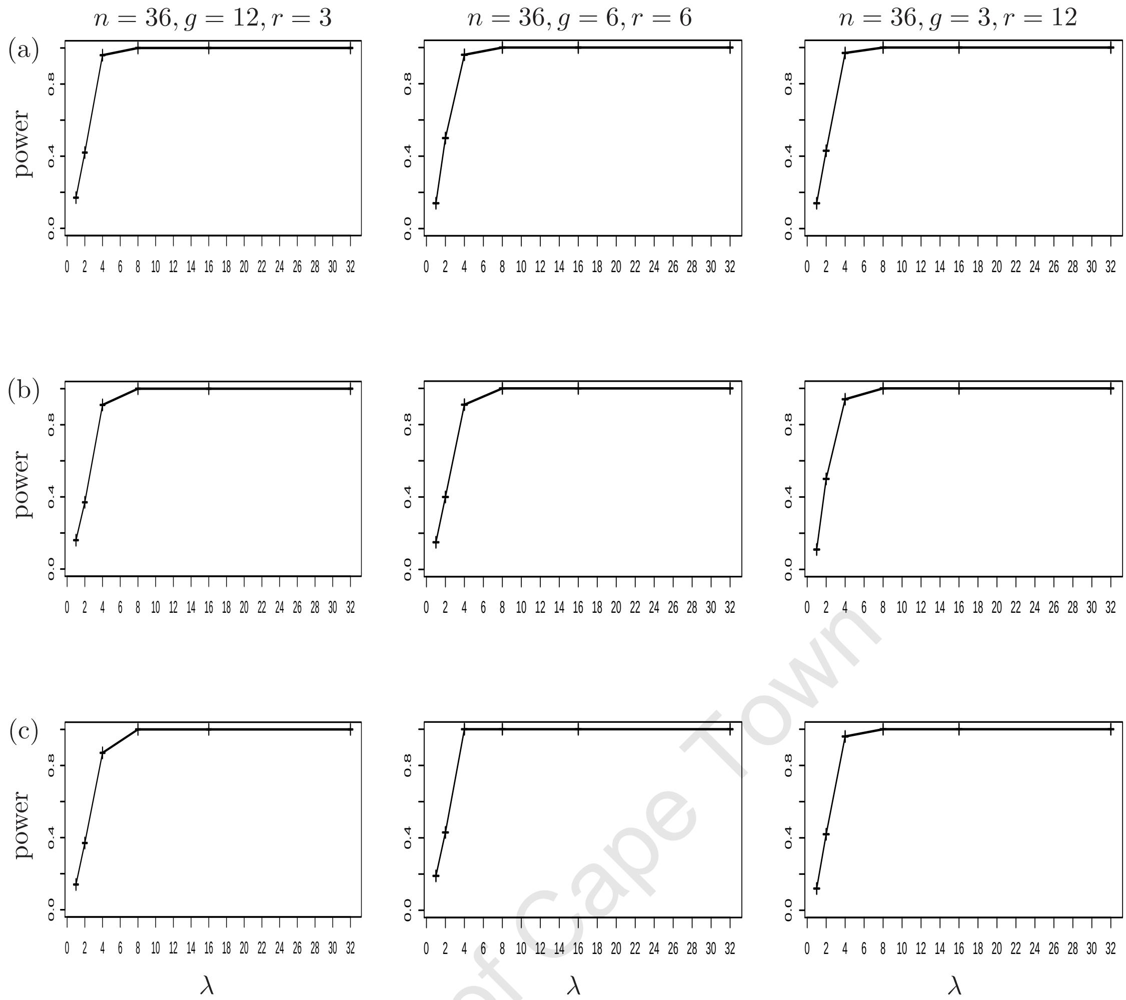

Figure 6.13: *Empirical power of the score test statistics for the first unit based on 2500 simulations of 500 data sets with 36 units each for for (a)* $\gamma = 0.1$*, (b)* $\gamma = 1$ *and (c)* $\gamma = 10$*:* $\alpha = 0.05$.

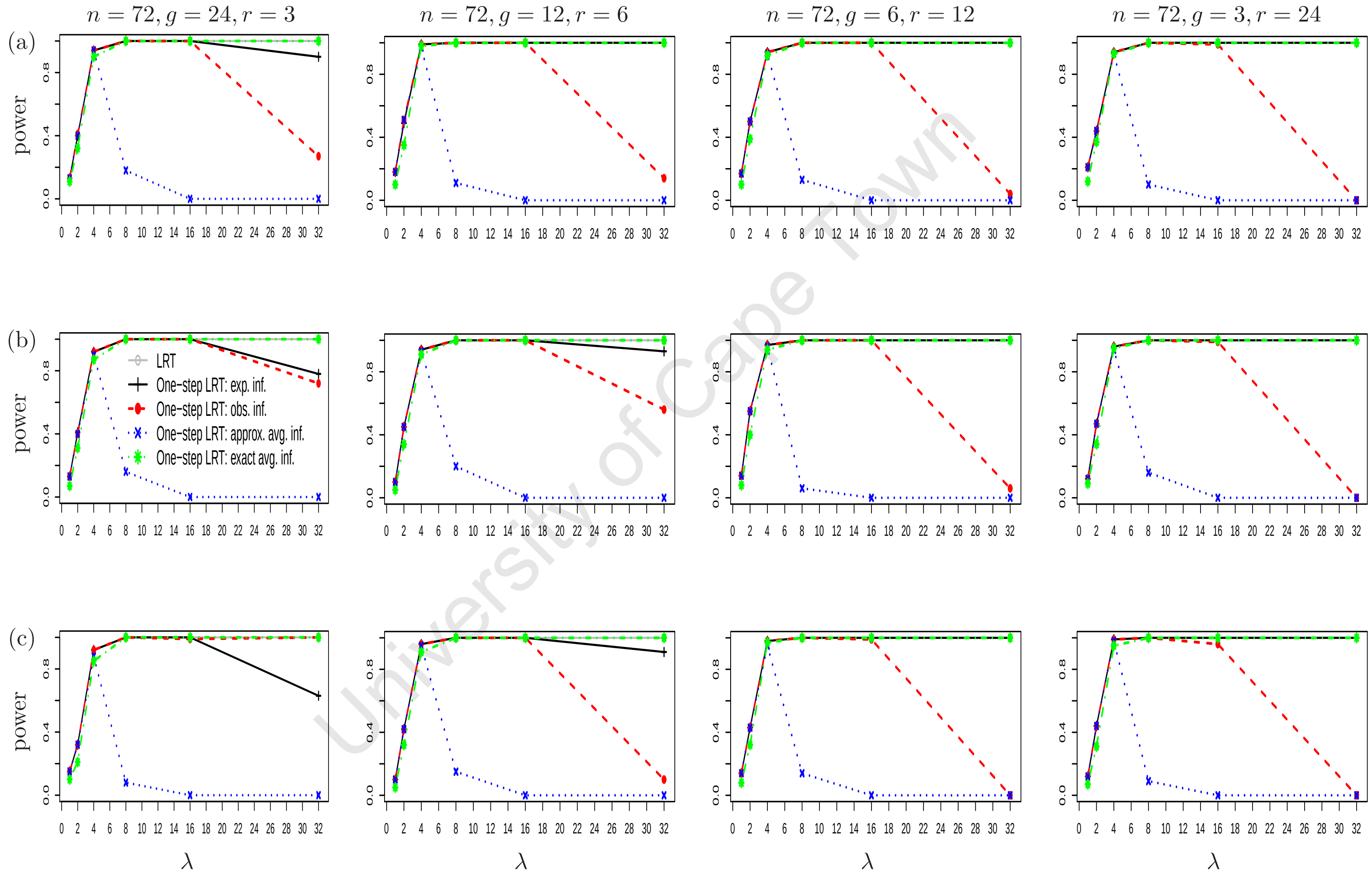

Figure 6.14: *Empirical power of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets with 72 units each for for (a) $\gamma = 0.1$, (b) $\gamma = 1$ and (c) $\gamma = 10$: $\alpha = 0.05$.*

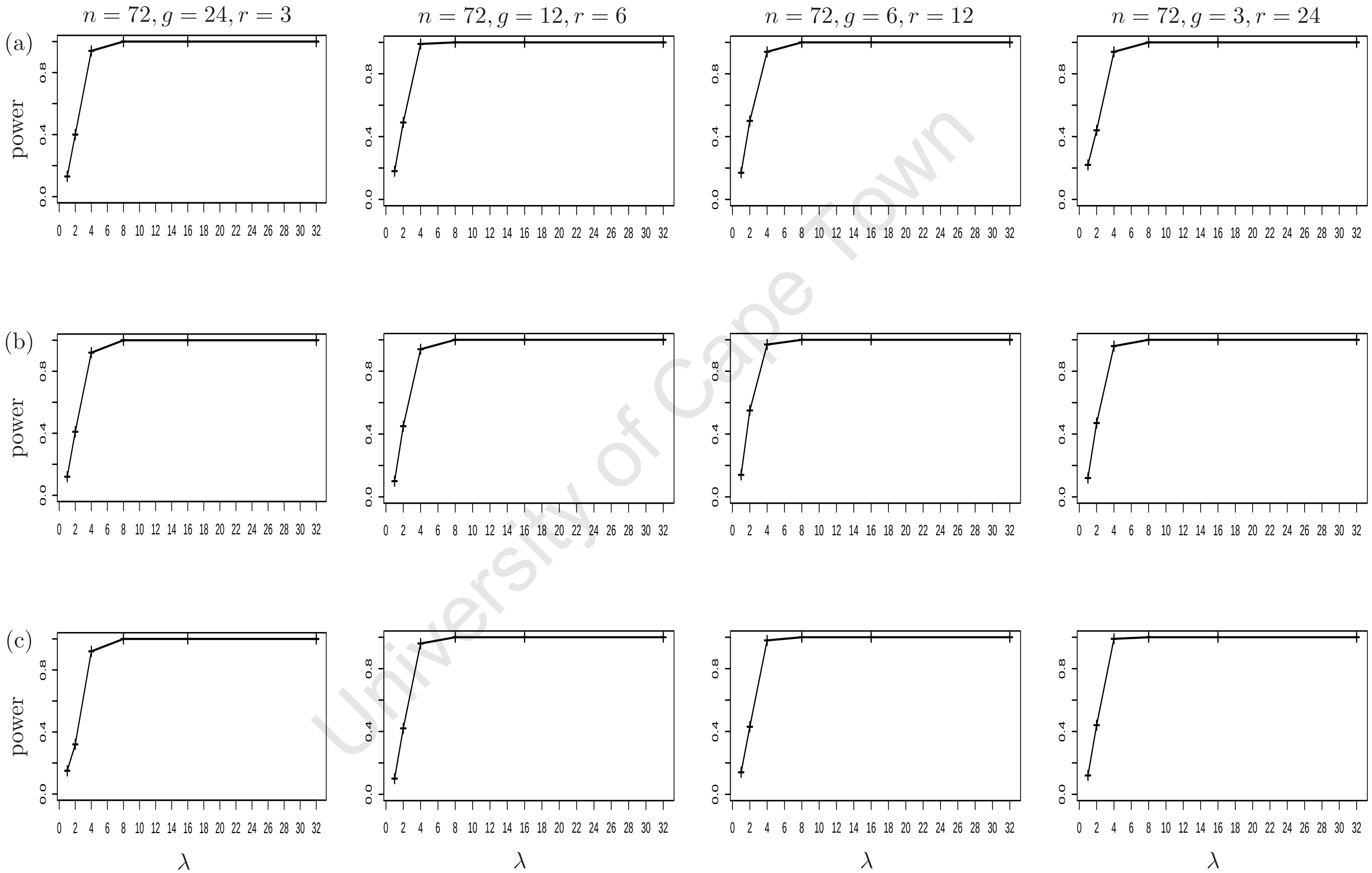

Figure 6.15: *Empirical power of the score test statistics for the first unit based on 2500 simulations of 500 data sets with 72 units each for (a) $\gamma = 0.1$, (b) $\gamma = 1$ and (c) $\gamma = 10$: $\alpha = 0.05$.*

6.9.4 Comparison between LRT and partial LRT in a linear mixed VSOM

In the previous two sections we observed that the LRT and the partial LRT did not differ in terms of type I error and power. This finding indicates that, for the simulation settings we considered, the partial LRT and hence t_i^2 captures the same information as the LRT about the outlyingness of the imposed outlier. To investigate this finding further we conducted a small simulation study. The purpose of this simulation study was to determine whether the equivalence between the LRT and the partial LRT holds for data sets of particular structures.

We generated 500 data sets with the jth data set simulated as follows

$$\boldsymbol{y}_j = \mu \mathbf{1}_n + (\boldsymbol{I}_g \otimes \mathbf{1}_r)\boldsymbol{u}_j + \boldsymbol{e}_j,$$

for $n = gr$ and $j = 1, \ldots, 200$ where $\boldsymbol{u}_j \sim N(\mathbf{0}, \sigma^2\gamma \boldsymbol{I}_g)$, $\boldsymbol{e}_j \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, $\mathbf{1}_n = (1 \ldots 1)'$ is a vector of length n, $n = 18$, $g = 6$, $r = 3$, $\mu = 0$, $\gamma = 10$ and $\sigma^2 = 1$. For each data set we inserted an outlier in the most extreme observation of the most extreme group (subject) as follows

Step 1 Find the group (subject) with the largest deviation from the overall mean,

Step 2 Get deviations of observations from their group means,

Step 3 Find the maximum deviation within the group (subjects) which was identified in step 1,

Step 3 If the deviation of the observation in step 3 is positive add 5, substract 5 otherwise.

For each simulated data set, both a baseline model (simple variance components model) and a VSOM for the inserted outlier were fitted, and the likelihood ratio test (LRT (6.15) or partial LRT (6.16) were calculated. Figure 6.16 presents a scatter plot of the LRTs against the partial LRTs for the 200 simulated data sets. The LRT and partial LRT are linearly related with a correlation coefficient of 0.851 (in contrast to the perfect positive correlation shown in Table 6.10). The magnitude of this correlation shows that there are sometimes differences between the two tests. This contrast may be caused by a change in the group (subjects) variance under a VSOM which is not captured by either t_i^2 or the partial test. This finding suggests that caution should be taken when partial VSOM tests are used in models with complex variance structures.

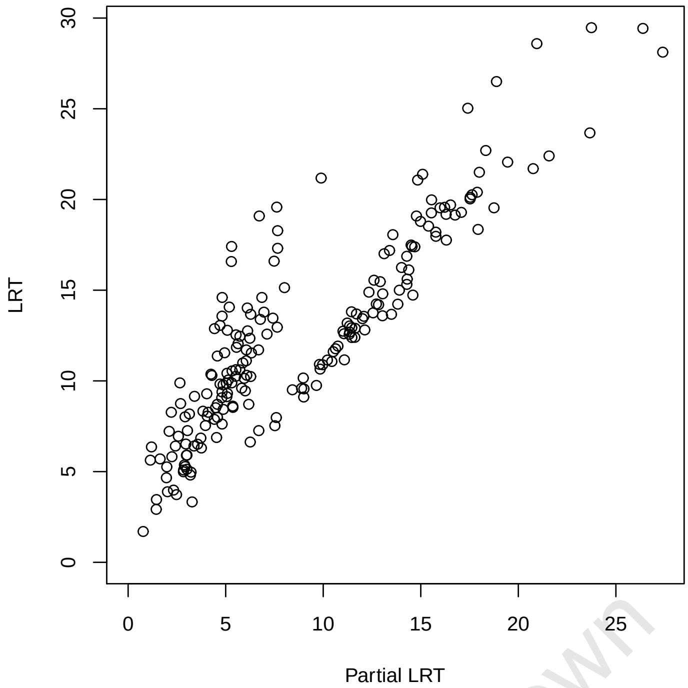

Figure 6.16: *Scatter plot of likelihood ratio test statistic against partial likelihood ratio test statistic for 200 simulated data sets :* $n = 18, g = 6, r = 3, \gamma = 10$.

6.10 Summary

In this chapter the linear fixed effects VSOM was extended to linear mixed effects VSOM for detecting outliers one-at-a-time. This extension is natural, as a VSOM is implemented within the linear model as a linear mixed model. However, the objective evaluation of a VSOM within the mixed model framework is more complex, as the distribution of the test statistic t_i^2 is no longer known. Therefore we extended the likelihood ratio and score test statistics for a VSOM in linear fixed effects analysis to a linear mixed VSOM. These tests no longer have analytical expressions since the variance parameters have to be estimated iteratively. We gave computing schemes for the one-step updates of the variance parameters (based on the observed, expected, approximate average and expected average information matrices) required for computing the one-step LRT statistics. We demonstrated the use of the proposed likelihood ratio and score test statistics in both a simple variance components model and a random coefficient model with the latter having a more complex variance structure, for the case of single outliers. In the simulation study and real data examples that we considered, the test statistics were effective in identifying unusual observations one-at-a-time. We also assessed the performance of the likelihood ratio and score test statistics in terms of computing speed, type I errors and power. The test statistics performed quite well both in terms of computing time and type I errors, for instance the empirical type I errors adhere to the nominal levels for all the tests. However, one-step LRTs based on the observed and approximate average information matrices performed poorly in terms of power.

The results in this chapter lead us into the following conclusions:

- We believe that a major advantage of a VSOM approach is the estimation of the shift variance for each observation, which can be used to down-weight the observation if required. This down-weighting gives an objective compromise between including and omitting the point where the status of the observation as a correct or erroneous point cannot be adequately resolved. The method can easily be extended to deal with other random effects or correlated data, as discussed below, and so gives a unified framework for detection of and adjustment to outliers at all levels of the random model.

- The major disadvantage of the method is the associated computation. The score tests are an appealing approach as, unlike the LRTs, they do not require a VSOM to be fitted, although elements of the relevant information matrix associated with the variance shift parameter, ω_i, must be calculated. However, we have demonstrated that the score tests do not conform to standard asymptotic distributions, and so simulation is required to generate their empirical distribution under the null hypothesis. This simulation is a reasonable task for a single observation, requiring 2500 simulations with a VSOM assessed on the observation of interest (i.e. 2500 assessments). We can reduce this simulation effort by assessing a VSOM only for observations with $t_i^2 > 1$, on average 32% of the observations, reducing the number of assessments to ~ 800. However, in the situation of multiple testing, the interest is in order statistics across the full set of observations, and so not only would a larger number of simulated data sets, say 10000, be required (to detect extremes of the distribution of the largest value under the null hypothesis), but also the null model has to be fitted each time (10000 fits) and a VSOM has to be assessed on all observations ($\sim 3200n$ assessments using an initial screen for $t_i^2 > 1$ each time). This computational effort is the reason why we did not assess type I and type II error of the empirical distributions of order statistics in our simulations.

- The LRT requires a VSOM to be refitted for each observation with $t_i^2 > 1$. However, our results showed that the LRT gave a reasonable approximation to a $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution, as motivated by comparison with the linear model case. If this result holds more widely, then this approach might give a reasonable and quick approximation for thresholds for individual observations. However, as the conditional residuals, and hence the LRT statistics, for individual observations are not independent, simulation is still required to generate thresholds for the order statistics. This simulation requires a VSOM to be re-fitted for each observation with $t_i^2 > 1$, i.e. $\sim 10000+3200n$ fits to establish thresholds for order statistics. The difficulty of this procedure is further increased in more complex models, where some variance parameters may be difficult to estimate, resulting in failure to fit some VSOMs. We have experienced this difficulty in random coefficient regression models.

  One way to reduce computation is by use of the partial LRT or partial score tests. We prefer the partial LRT as it is close to the LRT, and so also gives a reasonable approximation to the $0.68\chi_0^2 + 0.32\chi_1^2$ mixture distribution. In this case, there is no further computation once t_i^2 has been calculated from the baseline model and so a large number of simulations becomes more feasible. Further work is required to establish conditions under which the partial LRT gives a good approximation to the LRT, although we are encouraged by the fact that it works well in the simulations here even for small group sizes. An alternative approach would be the use of one-step approximations to the LRT, although the reduction in computation would be smaller. We might also reduce the need for simulation by use of false discovery rate (FDR) procedures for multiple testing (see e.g. Benjamini and Hochberg, 1995).

The models we have considered in this chapter addressed single-case outliers only. In linear mixed models or random coefficient models, subjects may have outlying profiles with intercepts and slopes that are inconsistent with those statistics of the remaining subjects in the data set. This contrast could be detected by implementing a VSOM at the level of random subject effects, i.e. allowing inflated variance for each subject in turn. This shift approach can be applied for any set of random effects in the model. We address the issue of outlying subject profiles in the next chapter.

Areas of further research on a linear mixed VSOM are given in Chapter 8.

CHAPTER 7

Extensions of a variance shift outlier model for linear mixed effects analysis

In this chapter we briefly present possible extensions of a variance shift outlier model in linear mixed effects analysis. Some of these extensions were considered earlier in Chapter 5 in the context of a variance shift outlier model in linear fixed effects analysis. Some of these models were also discussed by Gogel (1997) in the context of spatial mixed models. In addition, we briefly discuss influence on the fixed effects under a VSOM.

In the previous chapter we focused on single-case outliers; the models we introduce in this chapter deal with the identification of groups of outliers. In the context of the linear mixed model, natural groups of outliers are the subjects (which are related to the random effects) so that identification of group of outliers is equivalent to the identification of subjects with outlying profiles, i.e. subjects with random intercepts or random slopes that are different from the remaining subjects. The contribution of this thesis is the identification of outlying random effects and their recognition as natural groups of outliers in the linear mixed model. This approach differs from the identification of arbitrary groups of outliers by case-deletion methods.

The outline of the chapter is as follows. In the first three sections we give three extensions of a VSOM for the linear mixed effects analysis. In Sections 7.4 and 7.5 we illustrate the use of the extended VSOMs using two simulated data sets and a real data set (the orthodont data set). Next we discuss influence on the fixed effects under a VSOM. An example of influence analysis on the fixed effects under a VSOM is also given.

7.1 A VSOM for groups of outliers

A VSOM for groups of outliers in linear fixed effects analysis was considered in Chapter 5 § 5.7. Here model (5.50) is generalised to detect a group of outliers as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{D}_I\boldsymbol{\delta}_I + \boldsymbol{e} \\ &\sim N\left(\boldsymbol{X\beta}, \sigma^2(\boldsymbol{ZGZ}' + \boldsymbol{D}_I\boldsymbol{G}_I\boldsymbol{D}_I' + \boldsymbol{R})\right) \\ &\equiv N\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H}_I\right) \end{aligned} \tag{7.1}$$

where I is an arbitrary subset of size k, $\boldsymbol{D}_I$ is an $n \times k$ matrix, $\boldsymbol{G}_I = \text{var}(\boldsymbol{\delta}_I)$ and

$$\begin{aligned} \boldsymbol{H}_I &= \boldsymbol{ZGZ}' + \boldsymbol{D}_I\boldsymbol{G}_I\boldsymbol{D}_I + \boldsymbol{R} \\ &= \boldsymbol{H} + \boldsymbol{D}_I\boldsymbol{G}_I\boldsymbol{D}_I. \end{aligned}$$

$\boldsymbol{H} = \boldsymbol{ZGZ}' + \boldsymbol{R}$ is the variance-covariance matrix excluding $\boldsymbol{D}_I\boldsymbol{\delta}_I$.

The model (7.1) can also be viewed as an extension of the model (5.50) to the linear mixed model.

7.1.1 Special case I: A linear mixed VSOM for group of outliers with a common variance shift

We may consider a special case of (7.1) where $\boldsymbol{G}_I = \omega_I\boldsymbol{I}_k$ so that

$$\boldsymbol{H}_I = \boldsymbol{H} + \omega_I\boldsymbol{D}_I\boldsymbol{D}_I'.$$

This simplification of the model is similar to a VSOM for groups of outliers in linear regression presented in 5.7.2.

7.1.2 Special case II: A linear mixed VSOM for group of correlated outliers

The model (5.57) can be extended to the linear mixed model as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{D}_I^*\delta_I^* + \boldsymbol{e} \tag{7.2}$$

where $\boldsymbol{D}_I^*\delta_I^*$ and δ_I^* are as defined in (5.57) and $\boldsymbol{H}_I = \boldsymbol{H} + \boldsymbol{D}_I^*\boldsymbol{D}_I^{*\prime}\omega_I$. Analogous to model (5.57), the above model isolates all observations belonging to the subset indexed by I (all observations belonging to the jth subject) and requires one variance shift parameter to be estimated for the whole group with only one random effect estimate for the term δ_I^*.

7.2 A VSOM for random effects other than measurement error

In this section we consider a VSOM for random effects in $\boldsymbol{u}$ in which the jth random effect, u_j has inflated variance. In the general linear mixed model (2.1), $\text{var}(u_j) = \sigma^2\gamma_{u_j}$ is the jth diagonal element of $\boldsymbol{G}$. Then we can write a VSOM for the jth random effect as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Zd}_{u_j}^*\delta_{u_j} + \boldsymbol{e} \\ &\sim N\left(\boldsymbol{X\beta}, \sigma^2(\boldsymbol{ZGZ}' + \omega_{u_j}\boldsymbol{d}_{u_j}\boldsymbol{d}_{u_j}' + \boldsymbol{R})\right) \end{aligned} \tag{7.3}$$

where $\boldsymbol{d}_{u_j}^*$ is a $q \times 1$ vector with single non-zero element 1 in the jth position.

The model can also be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{d}_{u_j}\delta_{u_j} + \boldsymbol{e} \\ &\sim N\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H}_j\right) \end{aligned} \tag{7.4}$$

where $\boldsymbol{d}_{u_j} = \boldsymbol{Zd}_{u_j}^*$ is an $n \times 1$ vector with its only non-zero entries being ones corresponding to the jth level of $\boldsymbol{u}$ and

$$\begin{aligned} \boldsymbol{H}_j &= \boldsymbol{ZGZ}' + \omega_{u_j}\boldsymbol{d}_{u_j}\boldsymbol{d}_{u_j}' + \boldsymbol{R} \\ &= \boldsymbol{H} + \omega_{u_j}\boldsymbol{d}_{u_j}\boldsymbol{d}_{u_j}'. \end{aligned}$$

$\boldsymbol{H} = \boldsymbol{ZGZ}' + \boldsymbol{R}$ is the variance-covariance matrix excluding $\omega_{u_j}\boldsymbol{d}_{u_j}\boldsymbol{d}_{u_j}'$.

This model is equivalent to model (7.2) when $\boldsymbol{D}_I$ isolates the entire set of observations for a single subject (also see § 5.7.2, model (5.57)).

An approximate solution for $\hat{\omega}_{u_j}$, which is analogous to (6.13), is given by

$$\hat{\omega}_{u_j}(\hat{\boldsymbol{\kappa}}_0) = \frac{(n-p)(t_j^2 - 1)}{a_{u_{jj}}(n - p - t_j^2)}$$

where $t_j^2 = \tilde{e}_{u_j}^2/(\hat{\sigma}_0^2 a_{u_{jj}})$ is the squared standardized residual for the jth random effect and $a_{u_{jj}} = \boldsymbol{d}'_{u_j}\boldsymbol{P}\boldsymbol{d}_{u_j}$. Note that, similar to t_i^2 in (6.13), t_j^2 depends on $\boldsymbol{\kappa}$ through $\boldsymbol{P}$, with $\boldsymbol{P}$ evaluated at $\hat{\boldsymbol{\kappa}}_0$ (the estimates of the variance ratios under model (2.1)).

7.3 A VSOM in random coefficient regression analysis

The general linear mixed model in random coefficient regression analysis can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Z}_I\boldsymbol{u}_I + \boldsymbol{Z}_S\boldsymbol{u}_S + \boldsymbol{e} \\ &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Z}_{rc}\boldsymbol{u}_{rc} + \boldsymbol{e} \\ &\sim N\left(\boldsymbol{X\beta}, \sigma^2(\boldsymbol{ZGZ}' + \boldsymbol{Z}_{rc}\boldsymbol{G}_{rc}\boldsymbol{Z}'_{rc} + \boldsymbol{R})\right) \end{aligned} \tag{7.5}$$

where $\boldsymbol{Zu}$ is as defined in (2.1), the terms $\boldsymbol{u}_I(\boldsymbol{u}_S)$ carry the random intercept (slope) adjustments for each subject, $\boldsymbol{Z}_I(\boldsymbol{Z}_S)$ are the corresponding design matrices, $\boldsymbol{u}_{rc} = \begin{bmatrix} \boldsymbol{u}_I \\ \boldsymbol{u}_S \end{bmatrix}$ and $\boldsymbol{Z}_{rc} = \begin{bmatrix} \boldsymbol{Z}_I & \boldsymbol{Z}_S \end{bmatrix}$

Then the random coefficient VSOM for the jth individual can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Z}_I\boldsymbol{u}_I + \boldsymbol{Z}_S\boldsymbol{u}_S + \boldsymbol{Z}_I\boldsymbol{d}_j^*\delta_{I_j} + \boldsymbol{Z}_S\boldsymbol{d}_j^*\delta_{S_j} + \boldsymbol{e} \\ &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Z}_{rc}\boldsymbol{u}_{rc} + \boldsymbol{d}_{I_j}\delta_{I_j} + \boldsymbol{d}_{S_j}\delta_{S_j} + \boldsymbol{e} \\ &\sim N\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{H}_{rc_j}\right) \end{aligned} \tag{7.6}$$

where $\boldsymbol{d}_j^*$ is an $n_S \times 1$ zero vector with 1 in the jth position, $\boldsymbol{d}_{I_j}(\boldsymbol{d}_{S_j}) = \boldsymbol{Z}_I\boldsymbol{d}_j^*(\boldsymbol{Z}_S\boldsymbol{d}_j^*)$ and

$$\begin{aligned} \boldsymbol{H}_{rc_j} &= \boldsymbol{ZGZ}' + \boldsymbol{Z}_{rc}\boldsymbol{G}_{rc}\boldsymbol{Z}'_{rc} + \omega_{I_j}\boldsymbol{d}_{I_j}\boldsymbol{d}'_{I_j} + \omega_{I_j}\boldsymbol{d}_{S_j}\boldsymbol{d}'_{S_j} + \boldsymbol{R} \\ &= \boldsymbol{H} + \omega_{I_j}\boldsymbol{d}_{I_j}\boldsymbol{d}'_{I_j} + \omega_{I_j}\boldsymbol{d}_{S_j}\boldsymbol{d}'_{S_j}. \end{aligned}$$

$\boldsymbol{H} = \boldsymbol{ZGZ}' + \boldsymbol{Z}_{rc}\boldsymbol{G}_{rc}\boldsymbol{Z}'_{rc} + \boldsymbol{R}$ is the variance-covariance matrix excluding $\omega_{I_j}\boldsymbol{d}_{I_j}\boldsymbol{d}'_{I_j} + \omega_{I_j}\boldsymbol{d}_{S_j}\boldsymbol{d}'_{S_j}$.

A VSOM for individual observations discussed in the previous chapter can be combined with either of VSOMs (7.1)-(7.6) for the purpose of identifying individual observations or natural groups of observations with inflated error variances, and down-weighting them if necessary. In the next two sections we give two examples which illustrate the use of VSOMs (7.3) and (7.5).

7.4 Example: Simulated data

A VSOM analyses in the previous chapter dealt with individual observations having inflated measurement error variance, i.e. single-case outliers. In a repeated measures setting, subjects may have outlying profiles, i.e. profiles with intercepts and (or) slopes that are different from those of the remaining subjects. In this section we first conduct a small-scale simulation study to illustrate the use of VSOMs presented in this chapter to identify subjects with unusual profiles.

We generated data from the linear mixed model

$$\boldsymbol{y}_j = (\mu + u_{0j})\mathbf{1}_5 + (\beta + u_{1j})\boldsymbol{x} + \boldsymbol{e}_j, \tag{7.7}$$

where, for $j = 1, \ldots, 10$; $\boldsymbol{y}_j$ is a vector of length 5 of responses for the jth subject, $\boldsymbol{x} = (1, 2, \ldots, 5)'$, u_{0j} and u_{1j} are the random intercept and random slope of the jth subject, respectively, and $\boldsymbol{e}_j$ is the random error vector for subject j. The random effects vector for the jth subject $\boldsymbol{u}_j' = (u_{0j}, u_{1j})'$ is assumed to be Gaussian distributed with mean $\mathbf{0}$ and variance matrix $\gamma\boldsymbol{I}_2$ and the corresponding error vector $\boldsymbol{e}_j$ is assumed to have a Gaussian distribution with mean $\mathbf{0}$ and variance matrix $\sigma^2\boldsymbol{I}_5$. We set $\mu = 5$, $\beta = 0.3$, $\gamma = 0.5$ and $\sigma^2 = 1$. To investigate the presence of a single outlying subject profile we consider 2 scenarios: (i) the random intercept for the first subject is drawn from a Gaussian distribution with mean 4 and variance 0.5 and the observations for first subject are regenerated according to the above linear mixed model (7.7) and (ii) both random intercept and slope for the first subject are independently drawn from Gaussian distributions, each with mean 4 and variance 0.5 and the observations for first subject are regenerated according to the above linear mixed model. The simulated data for both scenarios (i) and (ii) are shown in Figure 7.1.

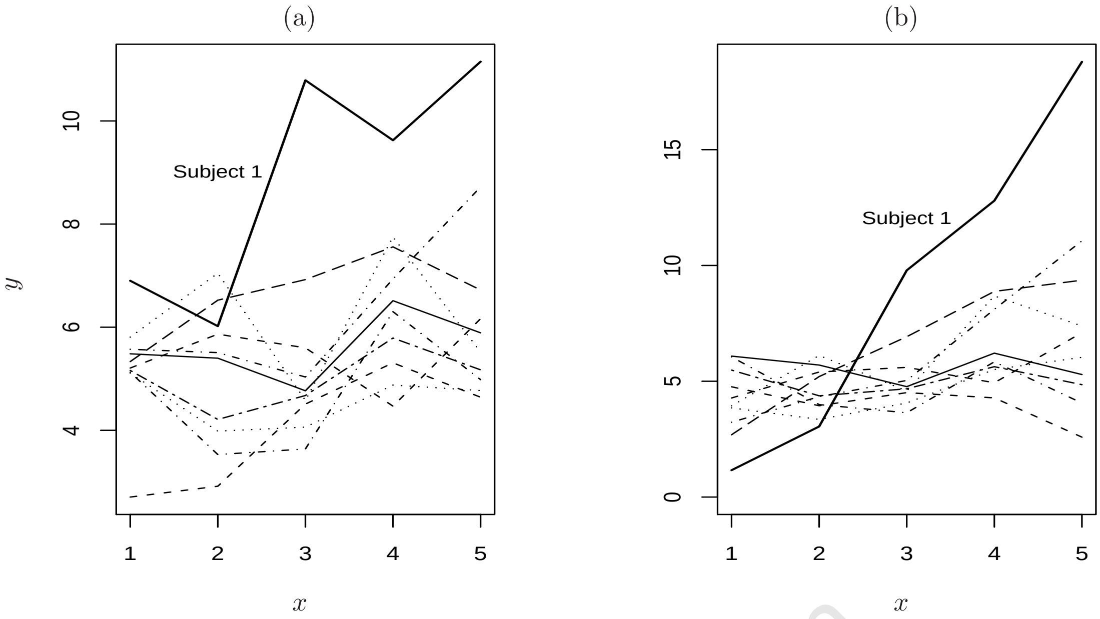

Figure 7.1: *Plots of y against x for simulated data according to (a) scenario (i) and (b) scenario (ii).*

For scenario (i), the model (7.4) was fitted for each subject using the model defined by (7.7) as the null model. Figures 7.2 shows the index plots of (a) variance shift estimates, (b) REML error variance estimate and (c) LRTs from the fitted VSOMs. As expected, the first subject has a relatively larger variance shift estimate and a significantly larger LRT value than the rest of the subjects, which confirms that it has an outlying random intercept.

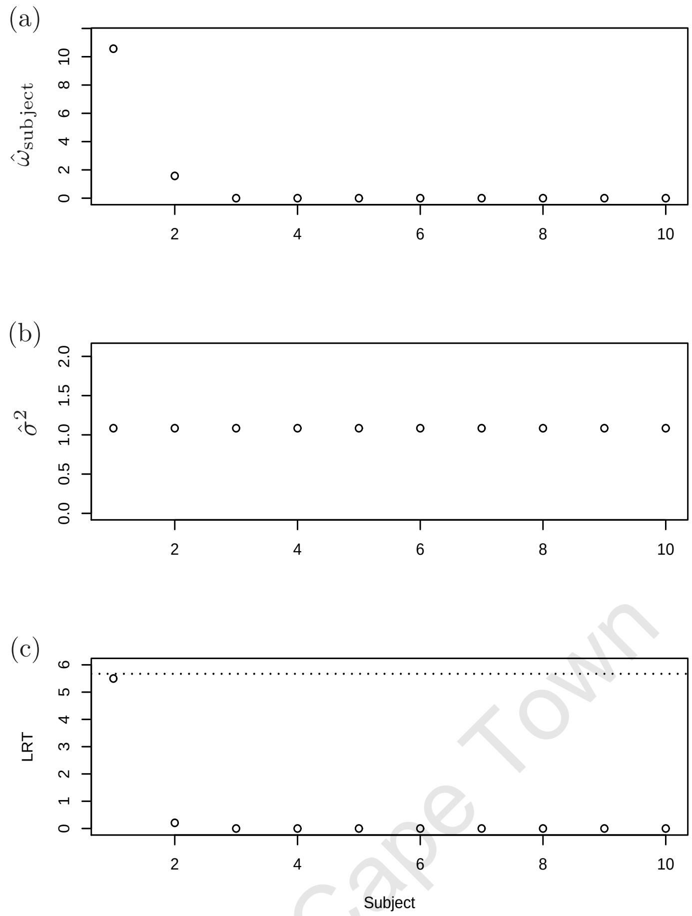

Figure 7.2: *Index plots of (a) REML variance shift estimates for random intercepts $\hat{\omega}_{subject}$ and (b) REML error variance estimates $\hat{\sigma}^2$ and (c) LRT with 95th percentile (dotted line) of the empirical distribution under the null hypothesis for the largest test statistic, scenario (i).*

For the second scenario we assess whether subjects have outlying intercepts or slopes separately. We first establish which subjects have outlying intercepts and then assess whether their slopes are also outlying. Another possible strategy is to establish whether subjects have both outlying intercepts and slopes, simultaneously; this approach is discussed further in § 7.7. The variance shift estimates $\hat{\omega}_{\text{subject}}$ associated with the subjects' intercepts and their corresponding estimated variances $\hat{\sigma}^2$, and LRTs are shown in Figure 7.3. Again the first subject was found to have an unusual intercept based on the calculated LRT value of 6.09 which was greater than the 95th percentile of the distribution of the maximum LRT.

To establish whether subjects have outlying random slopes we fit the model (7.6) for each subject, with $\boldsymbol{Z}_{rc}$ and $\boldsymbol{u}_{rc}$ having reduced dimensions, i.e. $\boldsymbol{Z}_{rc} = \boldsymbol{Z}_S$ and $\boldsymbol{u}_{rc} = \boldsymbol{u}_S$, where $\boldsymbol{Z}_S$ is the design matrix for the random slopes associated with the covariate x. This fitting process entails fitting an additional covariate, the individual columns of the matrix $\boldsymbol{Z}_S$, to the null model as defined by (7.7). Figure 7.4 presents the variance shift estimates associated with the subjects' slopes and their corresponding LRTs. The first subject had an LRT value of 7.15 and was found to have an outlying random slope on the basis of 95th percentile of the distribution of the maximum LRT.

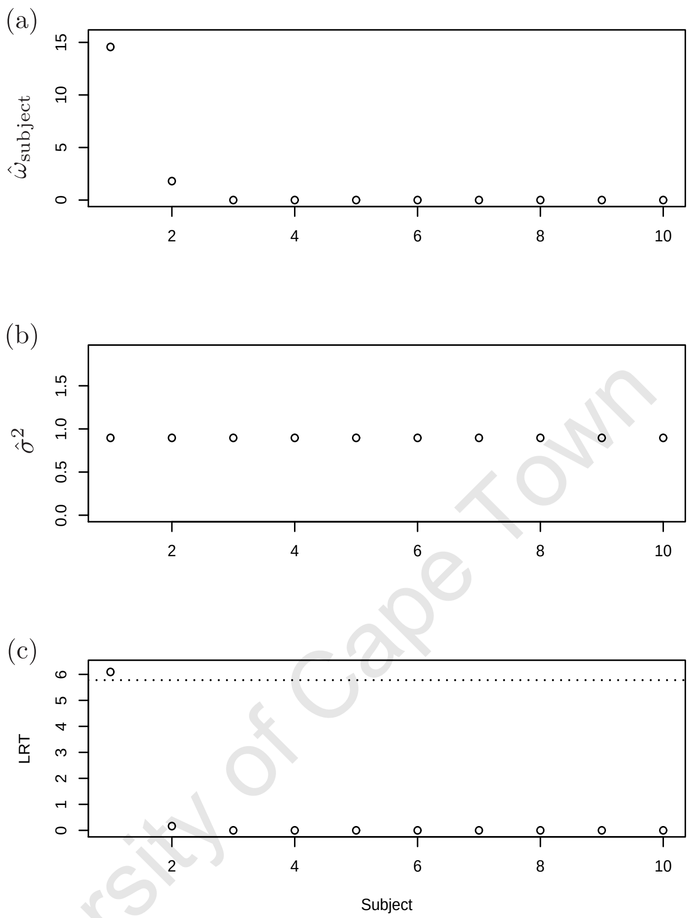

Figure 7.3: *Index plots of (a) REML variance shift estimates for random intercepts $\hat{\omega}_{subject}$ and (b) REML error variance estimates $\hat{\sigma}^2$ and (c) LRT with 95th percentile (dotted line) of the empirical distribution under the null hypothesis for the largest test statistic, scenario (ii).*

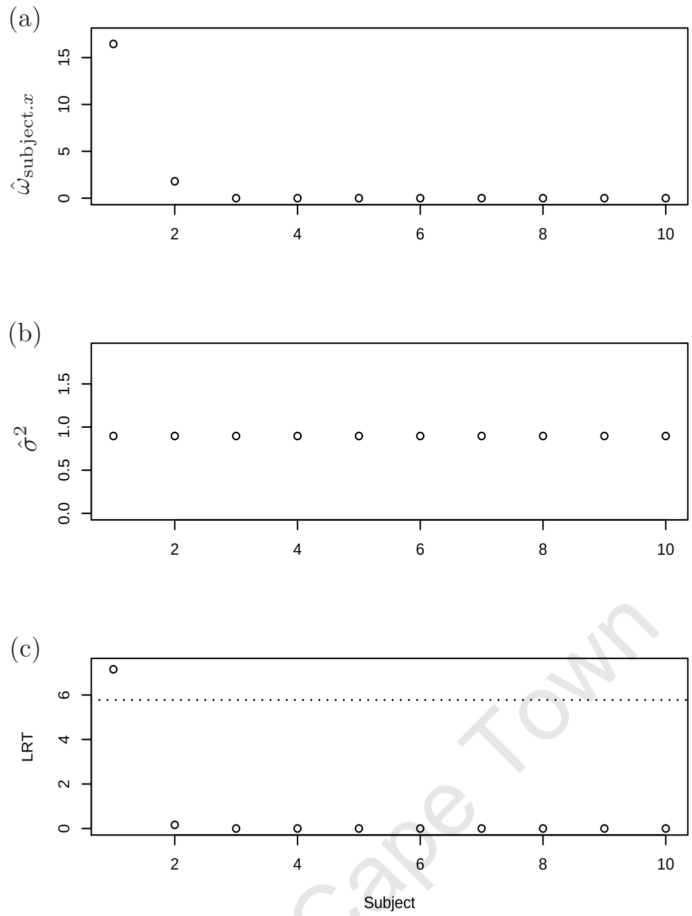

Figure 7.4: *Index plots of (a) REML variance shift estimates for random slopes $\hat{\omega}_{subject.x}$ and (b) REML error variance estimates $\hat{\sigma}^2$ and (c) LRT with 95th percentile (dotted line) of the empirical distribution under the null hypothesis for the largest test statistic, scenario (ii).*

7.5 Example: The orthodont data

VSOMs in the previous chapter addressed the issue of individual outliers, i.e. individual observations with inflated error variance and did not identify outlying subject profiles. In this section we extend a VSOM analyses of the orthodont data set conducted in the previous chapter, to VSOMs for groups which address outliying subject profiles. For convenience we will index the subjects in data set as subject number $1, \ldots, 27$: $1, \ldots, 16$ for boys and $17, \ldots, 27$ for girls.

We first consider VSOMs for groups of outliers (model (7.4)) for all subjects for the purpose identifying subjects with outlying random intercepts. As noted earlier a VSOM for groups of outliers which form a subset of all observations belonging to a single subject coincides with a VSOM for modelling the extra random variation in individual random intercepts for the subjects. Each of these models can be fitted by adding a variance shift covariate for the jth subject. A VSOM for group of outliers (7.4) was then fitted for each subject using the model $\mathcal{M}_1$ as the null model (the combined VSOM for single-case outliers in the previous chapter, see Table 6.1). The variance shift estimates, random intercept variance estimates and error variance estimates for a VSOM for each subject are shown in Figure 7.5. Subjects number 10 (boy number 10), number 26 and number 27 (girls number 10 and 11) have elevated variance shift estimates relative to the remainder of the subjects in the data set. Figure 7.5 also shows the LRT for each VSOM for groups of outliers and suggests that both subjects number 10, number 26 and number 27 are possibly outlying. Thus we fitted a single model with VSOM terms for subjects number 10 and number 26 with model $\mathcal{M}_1$ as the baseline model. The estimated parameters from this single model which included VSOM terms for the 3 groups of outliers (3 subjects), denoted $\mathcal{M}_3$, are shown in Table 7.1.

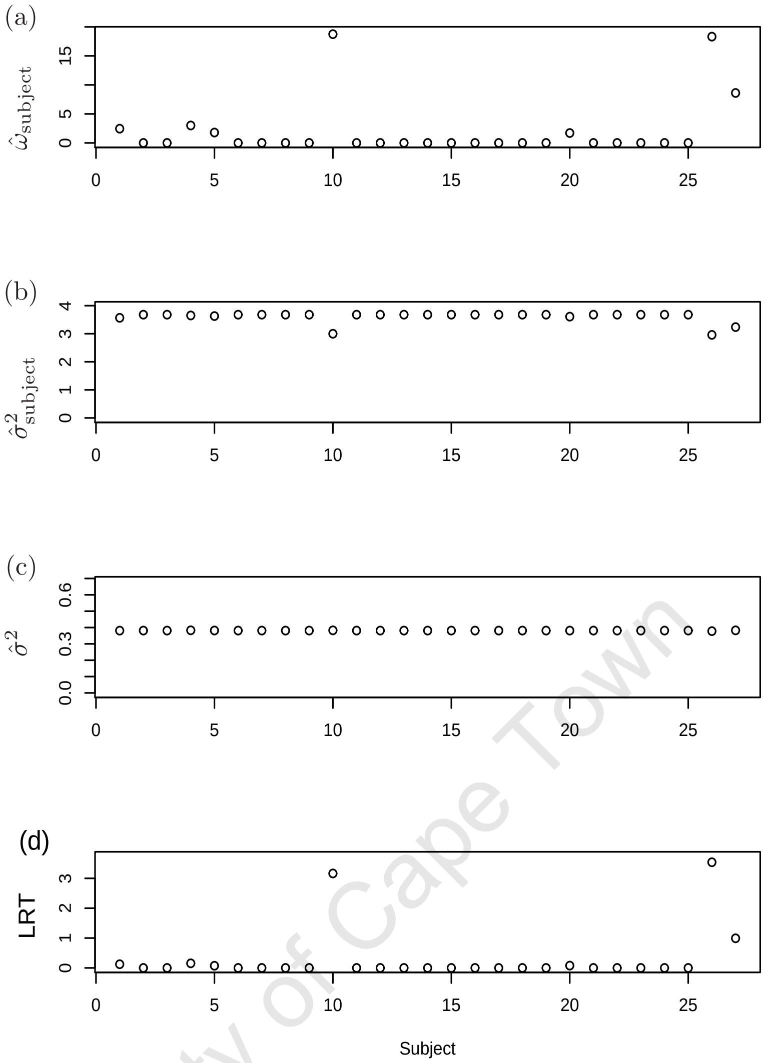

Figure 7.5: *Index plots of (a) REML variance shift estimates for random intercepts* $\hat{\omega}_{subject}$*, (b) random intercept REML variance estimates* $\hat{\sigma}^2_{subject}$*, (c) REML error variance estimates* $\hat{\sigma}^2$ *and (d) LRT for orthodont data.*

Next we investigate the presence of subjects with outlying random slopes. This investigation is conducted by fitting reduced forms of a VSOM (7.6) for each subject, i.e. $\boldsymbol{Z}_{rc} = \boldsymbol{Z}_S$, where $\boldsymbol{Z}_S$ is the design matrix for the random slopes associated with the covariate age. These models allow for extra variation in the random slopes for each subject. The model fitting process entails fitting an additional shift covariate to the null model $\mathcal{M}_1$ (the combined VSOM for single-case outliers in the previous chapter, see Table 6.1). Figure 7.6 presents the variance shift estimates, the random slope REML variance estimates, the error variance estimates and the LRTs for each fitted VSOM. Subjects 4, 18 and 24 appear to have outlying random slopes according to the LRT. Thus the final model fitted to the orthodont data, denoted $\mathcal{M}_4$, with 2 additional random terms associated with the age for the subjects 4 and 18 (boy number 4 and girl number 2). The value of $-2\times$REML log-likelihood function (excluding constant terms) decreased from 186.83 in model $\mathcal{M}_1$ to 170.63 in model $\mathcal{M}_4$. The estimated parameters for the final model $\mathcal{M}_4$ are shown in Table 7.1.

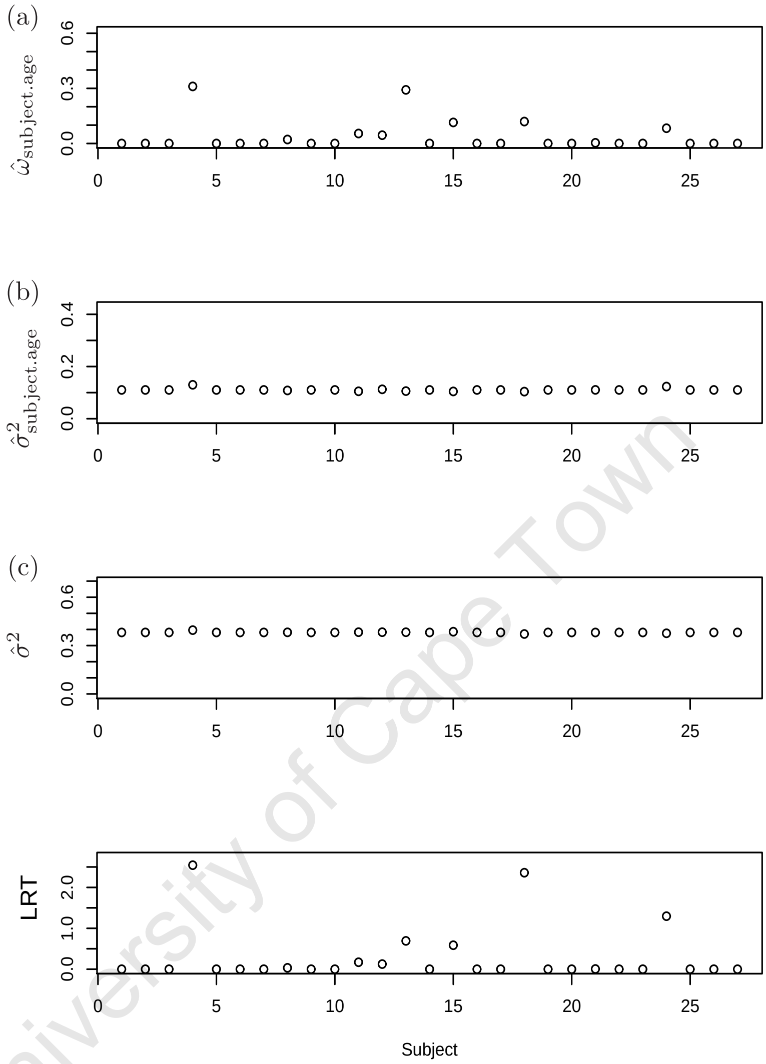

Figure 7.6: *Index plots of (a) REML variance shift estimates for random slopes $\hat{\omega}_{subject.age}$, (b) random slope REML variance estimates $\hat{\sigma}^2_{subject}$, (c) REML error variance estimates $\hat{\sigma}^2$ and (d) LRT for orthodont data.*

Table 7.1: *Estimated parameters for models fitted to orthodont data.*

Effect	Parameter	Model $\mathcal{M}_1$ Estimate (s.e.)		Model $\mathcal{M}_3$ Estimate (s.e.)		Model $\mathcal{M}_4$ Estimate (s.e.)	
Fixed							
constant	μ	22.690	(0.586)	22.820	(0.447)	22.940	(0.452)
age	β_1	0.452	(0.057)	0.454	(0.057)	0.421	(0.048)
sex	β_2	2.233	(0.770)	1.936	(0.582)	1.812	(0.591)
sex.age	β_3	0.246	(0.090)	0.246	(0.090)	0.315	(0.080)
Random							
$\boldsymbol{d}_{34}$	$\omega_{34}\sigma^2$	10.332	(16.992)	10.157	(16.504)	9.446	(15.486)
$\boldsymbol{d}_{35}$	$\omega_{35}\sigma^2$	33.676	(50.062)	33.973	(50.071)	33.633	(49.823)
$\boldsymbol{d}_{49}$	$\omega_{49}\sigma^2$	41.526	(61.128)	39.004	(57.487)	39.838	(58.232)
$\boldsymbol{d}_{73}$	$\omega_{73}\sigma^2$	4.533	(7.344)	4.051	(6.650)	4.725	(7.328)
$\boldsymbol{d}_{\text{boy 10}}$	$\omega_{\text{boy 10}}\sigma^2$		-	20.151	(31.516)	20.455	(31.894)
$\boldsymbol{d}_{\text{girl 10}}$	$\omega_{\text{girl 10}}\sigma^2$		-	17.064	(26.801)	19.712	(30.165)
$\boldsymbol{d}_{\text{girl 11}}$	$\omega_{\text{girl 11}}\sigma^2$		-	7.792	(13.726)	3.563	(7.474)
$\boldsymbol{d}_{\text{boy 4.age}}$	$\omega_{\text{boy 4.age}}\sigma^2$		-		-	0.383	(0.624)
$\boldsymbol{d}_{\text{girl 2.age}}$	$\omega_{\text{girl 2.age}}\sigma^2$		-		-	0.126	(0.205)
$\boldsymbol{D}_{\text{boys}}$	σ^2_{boys}	0.804	(0.289)	0.814	(0.287)	0.734	(0.257)
subject	$\sigma^2_{\text{subject}}$	3.679	(1.101)	1.782	(0.610)	1.891	(0.646)
subject.age	$\sigma^2_{\text{subject.age}}$	0.015	(0.015)	0.015	(0.014)	0.004	(0.011)
correlation	$\sigma^2_{\text{corr.}}$	0.110	(0.090)	0.080	(0.070)	0.113	(0.063)
	σ^2	0.381	(0.115)	0.381	(0.114)	0.388	(0.117)

For the simulated data and orthodont data set we assessed whether subjects had outlying intercepts or slopes separately. For the orthodont data set we also assessed whether subjects had both outlying intercepts and slopes, simultaneously. The variance shift estimates for the random intercepts and random slopes, and the LRTs from this investigation are shown in Figure 7.7. Subjects 10, 26 and 27 appear to have outlier intercepts (Figure 7.7, (a)), and subjects 4 and 18 appear to have outlier slopes (Figure 7.7, (b)). However subjects 4, 10, 18, 24, 26 and 27 are identified as having both outlier intercepts and slopes (Figure 7.7, (c)). The combined outlier intercept and slope model for subject number 20 did not converge. It is therefore unclear whether to consider a VSOM for intercepts and slopes simultaneously or separately. This dilemma, as highlighted by this example, needs further exploration through comprehensive simulation studies.

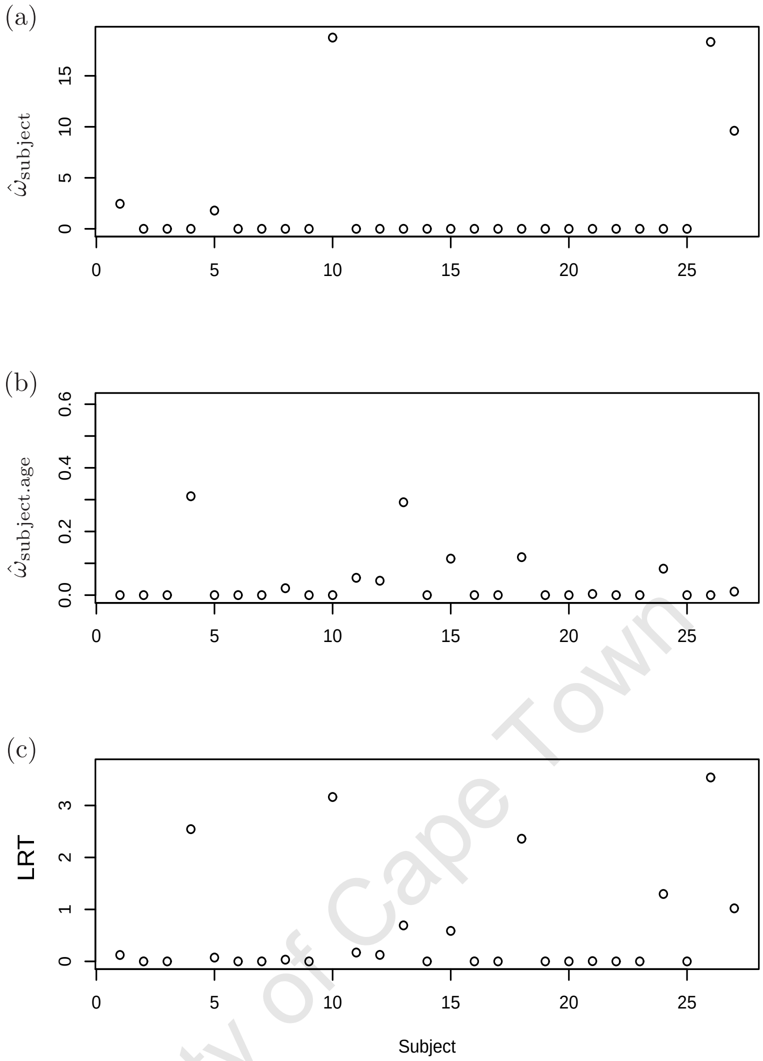

Figure 7.7: *Index plots of (a) REML variance shift estimates for random intercepts $\hat{\omega}_{subject}$ and (b) REML variance shift estimates for random slopes $\hat{\omega}_{subject.age}$ and (c) LRT for orthodont data.*

Comparison between VSOM in linear mixed effects analysis and case-deletion approaches

Christensen et al. (1992a) did not consider multiple-case deletion diagnostics. We therefore consider a multiple-case version of their Cook's statistic (5.48) which we use to detect a group of outlying observations e.g. all observations belonging to an individual subject. This Cook's statistic is defined as

$$D_I^s = (\hat{\boldsymbol{\phi}}_{(I)} - \hat{\boldsymbol{\phi}}_0)' \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\phi}}_0)(\hat{\boldsymbol{\phi}}_{(I)} - \hat{\boldsymbol{\phi}}_0), \tag{7.8}$$

where $\hat{\boldsymbol{\phi}}_0$ and $\hat{\boldsymbol{\phi}}_{(I)}$ are the variance parameter estimates under the null model and when the subset of observations indexed by I is deleted from the data set, respectively, and $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\phi}}_0)$ is the information matrix for the variance parameters under the null model. In our case the index I represents the subset of all observations belonging to the jth subject. This Cook's statistic is measure of the change in the variance parameter estimates in the model when the jth subject is deleted from the data set, i.e. the influence of the jth subject on variance estimates. Zewotir (2008) considers the statistic (7.8) but, for computational reasons, uses one-step estimates of the variance parameters instead of the REML estimates $\hat{\boldsymbol{\phi}}$.

Figure 7.8 shows an index plot of the Cook's statistics (7.8) for the orthodont data set after fitting the null model $\mathcal{M}_0$ given in Table 6.1. These statistics picked out subjects 9 and 13 as outliers which differ from the subjects picked out as outliers by a VSOM denoted $\mathcal{M}_3$ in Table 7.1. The estimated model parameters when observations belonging to subjects 9 and 13 were deleted together with observations 73 and 101, are given Table 7.2 under the model denoted $\mathcal{M}_5$. Note that in conducting this case-deletion analysis, observations 34 and 35 belong to subject 9 while observation 49 belongs to subject 13 so that these 3 observations are deleted as part of 2 groups of observations (subjects 9 and 13). The fixed effects estimates and their corresponding standard errors under the case-deletion model $\mathcal{M}_5$ were different from those a VSOM $\mathcal{M}_3$ (see Table 7.1). However, the inferences regarding the fixed effects were similar under both models.

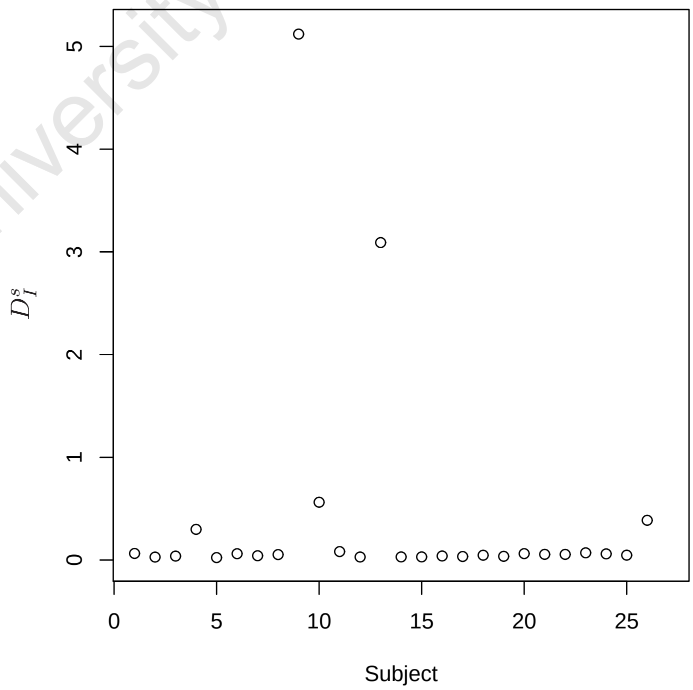

Figure 7.8: *Index plot of Cook's distance for variance parameters for orthodont data.*

Table 7.2: *Estimated parameters for the subject-deletion model*[†] *fitted to orthodont data.*

Effect	Parameter	Model $\mathcal{M}_5$ Estimate (s.e.)
Fixed		
constant	μ	22.650 (0.609)
age	β_1	0.480 (0.075)
sex	β_2	2.361 (0.814)
sex.age	β_3	0.208 (0.101)
Random		
subject	$\sigma^2_{\text{subject}}$	3.867 (1.205)
subject.age	$\sigma^2_{\text{subject.age}}$	0.019 (0.020)
correlation	$\sigma^2_{\text{corr.}}$	0.112 (0.108)
	σ^2	0.866 (0.173)

[†] Observations deleted:
33, . . . , 36; 49, . . . , 52 (subjects 9 and 13) .

7.6 Influence on the fixed effects under a VSOM.

A VSOM may not identify points that are influential with respect to the fixed model, i.e. a VSOM detects observations showing excess error variance given the fitted fixed model and the proposed variance model. To assess the influence of the ith unit given a VSOM in linear regression (5.5) for the unit we might consider the following modified Cook's distance

$$\begin{aligned} CD_i^a &= (\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_0)'\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_0)/p \\ &= \frac{1}{p}\left(\frac{\hat{\omega}_i e_i}{[1+\hat{\omega}_i(1-v_i)]}\right)^2 \boldsymbol{d}_i'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}[\hat{\sigma}_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d}_i \\ &= \frac{1}{p}\left(\frac{\hat{\omega}_i \hat{e}_i}{\hat{\sigma}_0[1+\hat{\omega}_i(1-v_i)]}\right)^2 \boldsymbol{d}_i'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d}_i \\ &= \begin{cases} \dfrac{v_i}{p}\left(\dfrac{\hat{\omega}_i \hat{e}_i}{\hat{\sigma}_0[1+\hat{\omega}_i(1-v_i)]}\right)^2 & \hat{\omega}_i > 0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{7.9}$$

where $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_i$ are the fixed effects estimates under the null and alternative model (VSOM for unit i), respectively. $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}}_0)$ is the expected information matrix, the inverse of the covariance matrix, for the fixed effects estimates under the null model. Note that the constraint $\hat{\omega}_i > 0$ implies that $t_i^2 > 1$. Furthermore as $\hat{\omega}_i \to \infty$ the statistic

(7.9) simplifies to the Cook's distance for fixed effects in linear regression of Cook (1977). The criticism against the use of the Cook's distance for variance parameters, mentioned in the Chapter 5, § 5.6 (see pp. 5-51) also applies to the Cook's distance for fixed effects of Cook (1977) and hence also to the Cook's statistic (7.9).

For illustration purposes we calculated the Cook's statistics for the linear model VSOMs (5.5) fitted to the orthodont data. Figure 7.9 presents an index plot of the calculated the Cook's statistics. According to this index plot, observations 34, 35, 49 and 52 had relatively large values of the Cook's statistic compared to the rest of the observations, indicating that these observations are possibly influential for the fixed effects. These observations were also found to have individual excess error variances in Chapter 5, § 5.6.

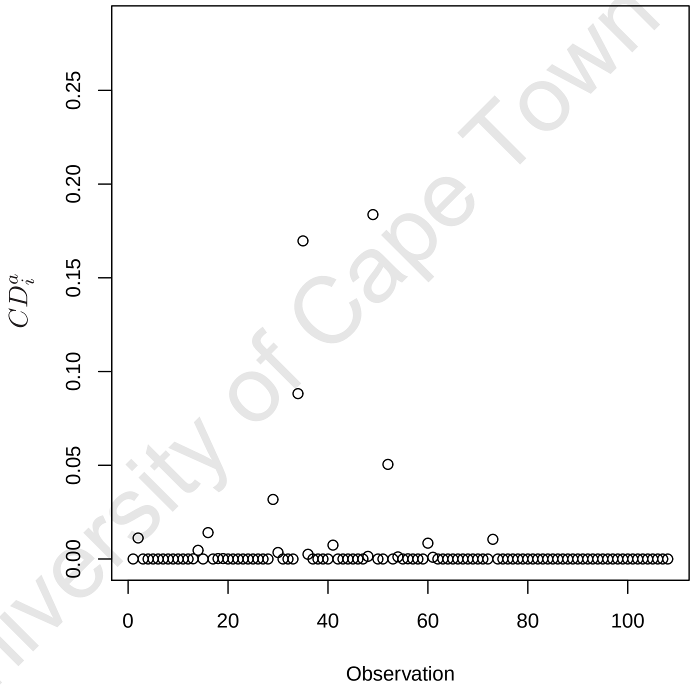

Figure 7.9: *Index plots of Cook's distance CD_i^a for influence on fixed effects for orthodont data.*

The Cook's statistic (7.9) can be extended to a linear mixed VSOM as

$$CD_i^{a*} = (\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_0)'\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_0)/p, \tag{7.10}$$

where $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_i$ are the fixed effects estimates under the null model (2.1) and alternative model (6.1), respectively, and p is the number of fixed effects parameters under the null model. $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}}_0)$ is the inverse of the estimated variance-covariance matrix of the fixed effects estimates under the null model. Figure 7.10 shows an index plot of the Cook's statistics for a linear mixed VSOMs (6.1) fitted to the orthodont data. The index plot indicates that observations 49, 73 and 101 are possibly influential for the fixed effects with observation 49 exerting the largest influence on the fixed model relative to observations 73 and 101. Observations 49 was also found to have an inflated variance in Chapter 6, § 6.7, and therefore outlier.

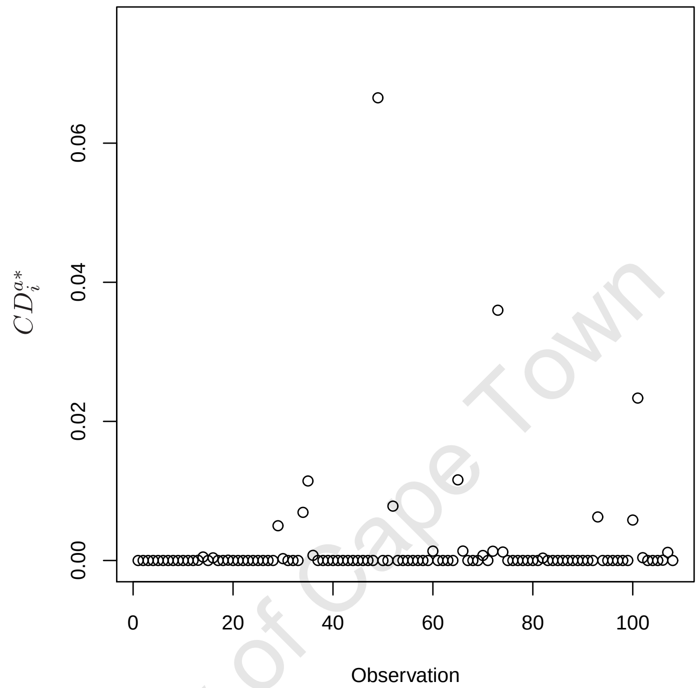

Figure 7.10: *Index plots of Cook's distance CD_i^{a*} for fixed effects for orthodont data.*

An extension of the Cook's statistic (7.10) for detecting influence of groups of observations on the fixed effects is given by

$$CD_I^{a*} = (\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}}_0)'\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}}_0)/p, \tag{7.11}$$

where $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_I$ are the fixed effects estimates under the null model (2.1) and alternative model (a VSOM for groups for example model (7.4)), respectively, $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}}_0)$ is the inverse of the estimated variance-covariance matrix of the fixed effects estimates under the null model, and the index I represents the subset of all observations belonging to the jth subject.

7.7 Summary

We have given some VSOMs for groups of outliers or random effects which are extensions of a linear mixed VSOM introduced in Chapter 6 for detecting single outliers. Two simulated data sets and a real data set were used to illustrate the use of these extended VSOMs. The models were successful in identifying subjects with outlying profiles given a fitted linear mixed model. However, the full-extent of their effectiveness in identifying outliers would require experience with a wider range of data sets than the two examples we have considered in this chapter.

It is interesting to note that in a repeated measures setting (for instance the orthodont data set), a VSOM allows both an observation within, say, subject j and all observations for subject j to have an inflated variance, i.e. one (or more) observations belonging to subject j may be outliers, and the subject itself may also have an outlying random intercept. In contrast under case-deletion, the influence of an individual observation within a subject and the influence all observations belonging to a same subject on the model parameters (fixed effects or variance parameters) need to be evaluated separately.

In some situations some of the models proposed in this chapter may not converge, for example, the model for extra variation in the random intercept for subject number 20 (girl number 4) in the orthodont data set, making it impossible to assess whether the random intercept and random slope for subjects 20 were both outlying.

A VSOM approach can be used in combination with a modified Cook's distance (based on fitting the ith VSOM) to detect outliers with respect to the fixed effects. However, the modified Cook's statistics (7.9) to (7.11) for assessing the influence of observations on the fixed model require further investigation. For instance the Cook's statistic (7.9) which uses the estimate $\hat{\boldsymbol{\beta}}_i$, which is based on the REML variance estimate $\hat{\boldsymbol{\phi}}_i$, could be modified by replacing $\hat{\boldsymbol{\beta}}_i$ with $\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{\phi}}_{i(1)})$ the fixed effects estimates evaluated at the one-step updates of the variance parameters with the array of one-step updates calculated using the different information matrices (observed, expected, approximate average and exact average). The distributional properties of the modified Cook's statistics are also not known.

CHAPTER 8

Conclusions and Future Research

8.1 Main conclusions and future research

In Chapter 4 we presented a review of the different types of residuals and leverages in the linear mixed model and explored their uses. The leverage matrix for the linear mixed model was decomposed into separate leverages for the fixed and random parts of the model. We also explored the use of the leverage-residual plot to identify possible high leverage observations or outliers or both. The following are possible areas of further research on residuals and leverages in the linear mixed model:

- The leverage measures and residuals we considered depend on the variance-covariance matrix $\sigma^2 \boldsymbol{H}$ and hence may be affected by misspecification of the variance-covariance structure. So there is a need to investigate the performance of the leverage measures and residuals under different covariance structures.

- The conditional residuals, joint leverages and random effects leverages depend on the distributional assumptions for the random effects. Finding residuals and leverages which are robust to the distributional assumptions is a matter of further research.

  - An alternative to the distribution of $\max|t_i|$ given by Zewotir and Galpin (2007) is the sampling distribution of $\max|t_i|$ obtained using bootstrap methods (parametric or non-parametric). However, there will also be a need to investigate the sensitivity of the resulting sampling distribution of the test statistic to the Gaussian assumption for the random effects, especially if parametric bootstrap methods are used. A comparison of the sampling distributions obtained using either parametric and non-parametric bootstrap methods may also be useful.

  - A possible route would be the extension to the linear mixed model of Lee and

Nelder (2000) to define robust residuals which do not use any distributional assumptions for the random effects.

- The investigation of the statistical properties of the multivariate conditional residuals is an area of further study.

- There is need for further investigation of the properties of the leverage matrices, especially the distributional properties of the diagonal elements of the leverage matrices. The justification for the common threshold value of $2p/n$ for the leverages in linear regression is based on approximations of the distributions of the leverages given in Belsley et al. (1980, pp. 67-68) and Chatterjee and Hadi (1986, pp. 31). Chave and Thomson (2003) give the finite sampling distribution of the leverages in the linear regression but this result has not been extended to the linear mixed model.

- The threshold values in Figure 4.1 which classify observations as outlying or influential or both, could be obtained from the joint distribution of the leverages and the residuals or $\tilde{e}_i^2/\tilde{\boldsymbol{e}}'\tilde{\boldsymbol{e}}$. This distribution can generated using bootstrap methods, for example generating many samples from a known fitted linear mixed model and assessing the resulting simulated distribution of the residuals and leverages.

- The leverage measures we discussed are based on individual diagonal elements of the leverage matrices or on traces of the leverage matrices. Measures based on both diagonal and off-diagonal elements of the leverage matrices along the lines of Beréod and Morgenthaler (1997) is an area of further research.

- There is a need for leverage measures for arbitrary groups of observations in the linear mixed model. Possible solutions could be extensions of the method of Dodge and Hadi (1999) and the diagnostic measure proposed by Hadi (1992), to the linear mixed model. For the latter diagnostic Hadi (1992) suggests a plot which is similar to the leverage-residual of Gray (1986), but this plot has the advantage that it can be used for single-case outliers and multiple-case outliers in the linear regression model.

Chapter 5 presented a review of a VSOM and described parameter estimation in the model. Likelihood ratio and score test statistics were developed as objective measures for assessing the size and significance of the variance shift estimates. The dependence of the one-step updates of the variance parameters, the likelihood ratio and score test statistics was also examined. We also studied the distributional properties of the test statistics. Issues of multiple testing were also discussed. A VSOM for detecting outliers one-at-a-time in linear regression was extended to detecting multiple outliers (more-than-one-at-a-time). Below we give possible directions for further research on a VSOM in linear regression:

- The linear regression VSOM assumes that the data are Gaussian distributed. In addition the likelihood ratio and score test statistics used to evaluate a VSOM in linear regression assume a Gaussian distribution. The sensitivity of the tests to the Gaussian assumption needs further investigation. Such an investigation may proceed along the lines of Miyashita and Newbold (1983) who studied the sensitivity to the Gaussian assumption of the outlier test statistic $\max|t_i|$ in linear regression. More recently, Martin and Roberts (2006) have suggested the use of bootstrap Studentized residuals for detecting outliers in linear regression, a method which they contend is robust to non-Gaussian distributed errors.

- There is a need to extend the linear regression VSOM results we have given to a VSOM for linear regression with correlated errors. It is not clear what a VSOM for correlated data should be. We might consider 'external' outlier, i.e., gross measurement error or 'internal' outlier, e.g. disturbance to the AR(1) innovations. The formulations we have considered only apply to the 'external' outlier type whereas we would probably want to consider both types of outliers.

- For a VSOM for groups of outliers, the choice of grouping is an issue (as for case-deletion methods). How should this choice be made? In a repeated measures setting, a possible choice of grouping is the set of all observations belonging to a single subject. We considered this point in Chapter 7 of the thesis.

In Chapter 6 we extended a VSOM in linear regression to a VSOM in linear mixed effects analysis. The following issues regarding a VSOM in linear mixed effects analysis need further exploration:

- The extension of a VSOM to correlated models needs further research. The straightforward implementation of variance shift outlier model for one effect independently of others corresponds to the case of an additive outlier in the terminology of Fox (1972), who also introduces the concept of the innovative outlier in time series. The innovative outlier is an outlier present in the innovations of a time series, whose effect is transmitted to all later observations. Use of a VSOM to down-weight an innovation for one individual in an AR(p) process (or in general an antedependence process) would be possible, but the generalization to other processes, e.g. an unstructured covariance matrix, is not clear. However, consider the case of simple random coefficient regression, with correlated intercept and slope for each subject: it may be unrealistic to postulate a perturbation in the intercept without any effect on the slope, and so the simple VSOM may be inefficient in the detection of outlying subjects.

- Similar to a VSOM in linear regression, a linear mixed VSOM and the proposed the likelihood ratio and score test statistics also rely on the Gaussian assumption for the random errors and the random effects (including the random effect(s) for the suspected outlier(s)). Thus the sensitivity of the tests to the Gaussian assumption is an area of further research.

- There are sometimes differences between the values of the LRT and partial LRT for particular data structures. This phenomenon needs further study.

In Chapter 7 we considered extensions of a linear mixed VSOM to models for detecting groups of outliers. In the context of data which can be analyzed using a linear mixed model, natural groups of observations are observations belonging to subjects. Therefore the extensions of VSOMs could be used to detect subjects with outlying profiles, i.e. subjects with random intercepts and random slopes that are inconsistent with those of the remaining subjects in the data set.

Two real data sets were used to illustrate the use of these extensions to a VSOM for the linear mixed model. The models appear to work well in identifying unusual observations or randoms effects. However, the full-extent of their effectiveness in identifying outliers would require experience with a wider range of data sets than the ones we have considered in this thesis. For instance, it is unclear whether to consider a VSOM for random intercepts and random slopes simultaneously or separately. There is also a need to explore multiple testing issues when these models are used to identify outliers.

8.2 Other further research areas

- In the single outlier case ω_i is a variance parameter of interest and our null hypothesis $H_0 : \omega_i = 0$ is on the boundary of the parameter space, and our alternative is one-sided, i.e. $H_A : \omega_i > 0$. When the alternative hypothesis is one-sided, Verbeke and Molenberghs (2003) suggest that the score test should be modified as

  $$S^* = U_i^2 \mathcal{I}^{11} - \inf\{(U_i^2 \mathcal{I}^{11} - b); b > 0\},$$

  where both U_i and $\mathcal{I}^{11}$ are evaluated at $\omega_i = 0$ and $\hat{\sigma}_0^2$. The test statistic S^* has an asymptotic $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution (Stram and Lee, 1994, 1995). We have not explored the usefulness of this test statistic in this research.

- The likelihood ratio and score tests developed for a VSOM do not specifically identify points that are influential with respect to the fixed model. The Cook's statistics and the examples we gave in Chapter 7, § 7.6 illustrate this fact. This finding suggests one might be able to get influence as a side-effect of a VSOM, not as an intrinsic part of it, although this claim would need further investigating to verify it.

  The usefulness and properties of the modified Cook's distances (7.9) to (7.11) need a thorough investigation. For instance the Cook's statistic (7.10) uses fixed effect estimates $\hat{\boldsymbol{\beta}}_i$ which are based on the REML variance parameter estimates $\hat{\boldsymbol{\phi}}_i$. In this Cook's statistic $\hat{\boldsymbol{\beta}}_i$ could be replaced with $\hat{\boldsymbol{\beta}}_i(\hat{\boldsymbol{\phi}}_{i(1)})$, the fixed effects estimates evaluated at the one-step updates of the variance parameters, with the one-step updates calculated using the different information matrices (observed, expected, approximate average and exact average).

- The extensions of a VSOM in linear regression to the linear mixed model and the development likelihood ratio and score test statistics assume Gaussian distributions for the responses and the random effects. Much work can still be done in extending these models and the proposed tests to non-Gaussian and dependent data.

Bibliography

Aitkin, M. and Wilson, G.T. (1980). Mixture models, outliers and the EM algorithm. *Technometrics*, **2**2, 325–331.

Altham, P.M.E. (1984). Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society, Series B*, **4**6, 118–119.

Anderson, R.L. and Bancroft, T.A. (1952). *Statistical Theory in Research.* McGraw-Hill, New York.

Andrews, D.F. and Pregibon, D. (1978). Finding outliers that matter. *Journal of the Royal Statistical Society, Series B*, **4**0, 85–93.

Anscombe, F.J. (1960). Rejection of outliers. *Technometrics*, **2**, 123–147.

Atkinson, A.C. (1985). *Plots, Transformations and Regression.* Clarendon Press, Oxford.

Banerjee, M. (1994). *Influence diagnostics in longitudinal models.* PhD thesis, Graduate School, University of Wisconsin-Madison.

Banerjee, M. (1998). Cook's distance in linear longitudinal models. *Communications in Statistics: Theory and Methods*, **2**7, 2973–2983.

Banerjee, M. and Frees, E.W. (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association*, **9**2, 999–1005.

Barndoff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **7**0, 343–365.

Barnett, V. and Lewis, T. (1995). *Outliers in Statistical Data.* Wiley, New York, Third edition.

Bates, D.M. and Watts, D.G. (1988). *Nonlinear regression analysis and its applications.* Wiley, New York.

Beckman, R.J. and Cook, R.D. (1983). Outlier...s. *Technometrics*, **2**5, 119–149, 161–163.

Beckman, R.J. and Nachtsheim, C.J. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, **2**9, 413–426.

Beckman, R.J. and Trussell, H.J. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association*, **6**9, 199–201.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: apractical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **5**7, 289–300.

Beréod, A.C. and Morgenthaler, S. (1997). A closer look at the hat matrix. *Student*, **2**, 1–12.

Box, G.E.P and Tiao, G.C. (1968). A Bayesian approach to some outlier problems. *Biometrika*, **5**5, 119–129.

Brown, P. and Prescott, R. (1999). *Applied Mixed Models in Medicine*. John Wiley, Chichester, United Kingdom.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data analysis Methods*. Sage Publications, New York.

Burns, P.J. (1992). Winsorised REML estimates of variance components. Technical report, Statistical Sciences Inc.

Bush, C.A. and MacEachern, J.B. (1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika*, **8**3, 275–285.

Butler, S. M. and Louis, T. A. (1992). Random effects model with nonparametric priors. *Statistics in Medicine*, **1**1, 1981–2000.

Callanan, T.P. and Harville, D.A. (1991). Some new algorithms for computing restricted maximum likelihood estimates of variance components. *Journal of Statistical Computation and Simulation*, **3**8, 239–259.

Chatterjee, S. and Hadi, A.S. (1986). Influential observations, high leverage points and outliers in linear regression. *Statistical Science*, **3**, 379–416.

Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*. Wiley, New York.

Chave, A.D. and Thomson, D.J. (2003). A bounded influence regression estimator based on the statistics of the hat matrix. *Journal of the Royal Statistical Society Series C*, **5**2, 307–322.

Christensen, R. (2001). *Advanced Linear Modelling: multivariate, time series and spatial data; nonparametric regression and response surface maximization*. Springer-Verlag, New York, Second edition.

Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear models*. Springer, New York, Third edition.

Christensen, R., Pearson, L.M., and Johnson, W. (1992a). Case-deletion diagnostics for mixed models. *Technometrics*, **3**4, 38–45.

Christensen, R., Pearson, L.M., and Johnson, W. (1992b). Prediction diagnostics for spatial linear models. *Biometrika*, **7**9, 583–591.

Commenges, D. and Jacqmin-Gadda, H. (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society, Series B*, **5**9, 157–171.

Cook, R.D. (1977). Detection of influential variables in linear regression. *Technometrics*, **1**9, 15–18.

Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, **4**8, 133–169.

Cook, R.D., Holschuh, N., and Weisberg, S. (1982). A note on an alternative outlier model. *Journal of the Royal Statistical Society, Series B*, **4**4, 370–376.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Corbeil, R.R. and Searle, R. (1976a). A comparison of variance components estimators. *Biometrics*, **3**2, 779–791.

Corbeil, R.R. and Searle, R. (1976b). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, **1**8, 31–38.

Cox, D.R. and Hinkley, D.V. (1990). *Theoretical Statistics*. Chapman and Hall, London.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Series B*, **4**9, 1–39.

Crainiceanu, C.M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B*, **6**6, 165–185.

Cressie, N.A.C. (1991). *Statistics for spatial data.* John Wiley and Sons, New York.

Crowder, M. J. and Hand, D.J. (1990). *Analysis of Repeated Measures.* Chapman and Hall, London.

DeGruttola, V., Ware, J.H., and Louis, T.A. (1987). Influence analysis of generalized least squares estimators. *Journal of the American Statistical Association*, **8**8, 911–917.

Demidenko, E. (2004). *Mixed models theory and applications.* Wiley, New York.

Demidenko, E. and Stukel, T.A. (2005). Influence analysis for linear mixed-effects models. *Statistics in medicine*, **2**4, 893–909.

Dempster, A.P., Selwyn, M.R., Patel, C.M., and Roth, A.J. (1984). Statistical and computational aspects of mixed model analysis. *Applied Statistics*, **3**3, 203–214.

Diggle, P.J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, **4**4, 959–971.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

Dillane, D. (2006). *Deletion Diagnostics for the Linear Mixed Model.* PhD thesis, Department of Statistics, TCD, Trinity College, Dublin.

Dodge, Y. and Hadi, A.S. (1999). Simple graphs and bounds for the elements of the hat matrix. *Journal of Applied Statistics*, **2**6, 817–823.

Efron, B. (1979). Boostrap methods: Another look at the jacknife. *Annals of Statistics*, **7**, 1–26.

Efron, B and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, **6**5, 457–487.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Ellenberg, J.H. (1973). The joint distribution of studentized least squares residuals from a general linear regression. *Journal of the American Statistical Association*, **6**8, 941–943.

Fellner, W.H. (1986). Robust estimation of variance components. *Technometrics*, **2**8, 51–60.

Fitzmaurice, G., Laird, N., and Ware, J. (2004). *Applied longitudinal analysis*. Wiley, Hoboken, N.J.

Fox, A.J. (1972). Outliers in time series. *Journal of the Royal Statistical Society, Series B*, **3**4, 350–363.

Fraccaro, R., Hyndman, R.J., and Veevers, A. (2000). Residual diagnostic plots for checking model misspecification in time series regression. *Australian and New Zealand Journal of Statistics*, **4**2, 463–477.

Fung, W.K., Zhu, Z., Wei, B.C., and He, X. . (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society, Series B*, **6**4, 565–579.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*. Chapman Hall, New York, second edition.

Ghosh, M and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, **9**, 55–93.

Gilmour, A.R., Cullis, B.R., and Verbyla, A.P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, **5**1, 269–273.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S., and Thompson, R. (2002). *ASReml User Guide Release 1.0*. VSN International, Waterhouse St, Hemel Hempstead, HP1 1ES UK.

Gilmour, A.R., Thompson, R., and Cullis, B.R. (1995). Average Information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **5**1, 1440–1450.

Godfrey, L.G. and Orme, C.D. (2001). On improving the robustness and reliability of Rao's score test. *Journal of Statistical Planning and Inference*, **97**, 153–176.

Gogel, B.J. (1997). *Spatial analysis of multi-environment variety trials*. PhD thesis, Department of Statistics, University of Adelaide.

Goldberger, A.S. (1962). Best linear unbiased prediction in general linear regression. *Journal of the American Statistical Association*, **57**, 369–375.

Goldstein, H. (1995). *Multilevel statistical models*. Edward Arnold, London, Second edition.

Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, Second edition.

Goodnight, J.H. (1978). Computing MIVQUEO estimates of variance components. Technical Report 105, SAS Institute Inc., Cary, NC.

Gray, J.B. (1986). A simple graphic for assessing influence in regression. *Journal of Statistical Computing and Simulation*, **24**, 121–134.

Gray, J.B. and Ling, R.F. (1984). K-clustering as a detection tool for influential subsets in regression. *Technometrics*, **26**, 305–330.

Green, G.H. and Diggle, P.J. (2007). On the operational characteristics of the Benjamini and Hochberg false discovery rate procedure. *Statistical Applications in Genetics and Molecular Biology*, **6** (1).

Greven, S., Crainiceanu, C.M., Küchenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, **17**, 870–891.

Grizzle, J.E. and Allen, D.M. (1969). Analysis of growth and response curves. *Biometrics*, **25**, 357–381.

Hadi, A.S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics and Data Analysis*, **14**, 1–27.

Hall, D. B. and Praestgaard, J. T. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika*, **88**, 739–751.

Hall, G., Rogers, W.H., and Pregibon, D. (1982). Outliers matter in survival analysis. Technical Report P-6761, Rand, Santa Monica.

Hand, D. and Crowder, M. (1996). *Practical Longitudinal Data Analysis*. Chapman and Hall, London.

Hartley, R. and Rao, J.N.K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, **5**4, 93–108.

Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **6**1, 383–385.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **7**2, 320–338.

Haslett, J. (1999). A simple derivation of deletion diagnostic results for the general linear model with correlated errors. *Journal of the Royal Statistical Society, Series B*, **6**1, 603–609.

Haslett, J. and Dillane, D. (2004). Application of 'delete=replace' to deletion diagnostics for variance component estimation in the linear mixed model. *Journal of the Royal Statistical Society, Series B*, **6**6, 131–143.

Haslett, J. and Haslett, S.J. (2007). The three basic types of residuals for a linear model. *International Statistical Review*, **7**5, 1–24.

Haslett, J. and Hayes, K. (1998). Residuals for the linear model with general covariance structure. *Journal of the Royal Statistical Society, Series B*, **6**0, 201–215.

Hemmerle, W. and Hartley, H. (1973). Computing maximum likelihood estimates for the linear A.O.V. model using W-transformation. *Technometrics*, **1**5, 819–831.

Henderson, C.R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics*, **2**1, 309–310.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226–252.

Henderson, C.R., Kempthorne, O., Searle, S.R., and Von Krosig, C.N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **1**5, 192–218.

Hilden-Minton, J.A. (1995). *Multilevel diagnostics for mixed and hierarchical models.* PhD thesis, Department of Mathematics, University of California, Los Angeles.

Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, **3**2, 17–22.

Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparison Procedures.* John Wiley and Sons, New York.

Hocking, R.R. (1985). *The Analysis of Linear Models.* Brooks Cole, California.

Hodges, J.S. (1998). Some algebra and geometry for hierarchical models applied to diagnostics (with discussion). *Journal of the Royal Statistical Society, Series B*, **6**0, 497–536.

Hodges, J.S. and Sargent, D.J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika*, **8**8, 367–379.

Hoeting, J., Raftery, A.E., and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis*, **2**2, 251–270.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression.* John Wiley, New York.

Houseman, A.E., Ryan, L.M., and Coull, B.A. (2004). Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association*, **9**9, 383–394.

Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.

Hurtardo, G.I. (1993). *Detection of influential observations in linear mixed models.* PhD thesis, Department of Statistics, University of North Carolina, North Carolina.

Insightful Corp. (2001). *S-PLUS 6 Guide to Statistics for Windows*, volume 1-2. Insightful Corporation, 1700 Westlake Ave. N., Suite 500Seattle.

Jaffrézic, F., White, I.M.S., and Thompson, R. (2003). Use of the score test as a goodness of fit measure of the covariance structure in genetic analysis of longitudinal data. *Genetics, Selection and Evolution*, **3**5, 185–198.

Jennrich, R.I. and Sampson, P.F. (1976). Newton-Raphson and related algorithms for maximum likelihood estimation of variance components. *Technometrics*, **1**8, 11–18.

Jennrich, R.I. and Schluchter, M. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **4**2, 805–820.

Jiang, J. (2001). Goodness-of-fit tests for mixed model diagnostics. *Annals of Statistics*, **2**9, 1137–1164.

Johnson, D. and Thompson, R. (1995). Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and a new quasi-Newton procedure. *Journal of Dairy Science*, **7**8, 449–456.

Jones, R.H. (1993). *Longitudinal data with serial correlation. A State-space Approach*. Chapman and Hall, London.

Judge, G.G., Griffiths, W.E., Hill, R.C., and Lee, T.-C. (1980). *The Theory and Practice of Econometrics*. Wiley, New York.

Justel, A. and Pena, D. (2001). Bayesian unmasking in linear models. *Computational Statistics and Data Analysis*, **3**6, 69–84.

Kenward, M. and Roger, J. (1997). Small sample inferences for fixed effects from restricted maximum likelihood. *Biometrics*, **5**3, 983–997.

Khatri, G.C. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals of Inst. Statistical Mathematics*, **1**8, 75–78.

Khuri, A.I. (1989). Testing a covariance matrix structure in a mixed model with no empty cells. *Journal of Statistical Planning and Inference*, **2**2, 117–125.

Kleinman, K.P. and Ibrahim, J.G. (1998). A semi-parametric Bayesian approach to the random effects model. *Biometrics*, **5**4, 921–938.

Laird, N.M., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, **8**2, 97–105.

Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **3**8, 963–974.

Lange, K.I., Little, R.J.A., and Taylor, J.M.G. (1989). Robust statistical modelling using the t distribution. *Journal of the American Statistical Association*, **8**4, 881–896.

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *Annals of Statistics*, **1**7, 624–642.

Langford, I.H. and Lewis, T. (1998). Outliers in multilevel data (with discussion). *Journal of the Royal Statistical Society, Series A*, **1**61, 121–160.

Lawrance, A.J. (1987). The score statistic for regression transformation. *Biometrika*, **7**4, 275–279.

Lawrence, A.J. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society, Series B*, **5**7, 181–189.

Lee, J.C. (1991). Tests and model selection for the growth curve model. *Biometrics*, **4**7, 147–159.

Lee, J., Nelder, J.A., and Pawitan, Y. (2006). *Generalized linear models with random effects: Unified analysis via H-likelihood.* Chapman and Hall, London.

Lee, S.-Y. and Xu, L. (2004). Influence analysis of nonlinear mixed-effects models. *Computational Statistics and Data Analysis*, **4**5, 321–341.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **5**8, 619–678.

Lee, Y. and Nelder, J.A. (1998). Generalized linear models for the analysis of quality-improvement experiments. *Canadian Journal of Statistics*, **2**6, 95–105.

Lee, Y. and Nelder, J.A. (2000). Two ways of modelling over-dispersion in non-normal data. *Journal of the Royal Statistical Society, Series C*, **4**9, 491–498.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **5**4, 570–582.

Lill, W. J., Gleeson, A. C., and Cullis, B. R. (1988). Relative accuracy of a neighbour model for field trials. *Journal of Agricultural Science, Cambridge*, **1**1, 339–346.

Lin, C.Y. and McAllister, A.J. (1984). Monte Carlo comparisons of four methods for estimation of genetic parameters in the univariate case. *Journal of Dairy Science*, **6**7, 2389–2398.

Lin, D.Y., Wei, L.J., and Ying, Z. (1993). Checking the Cox model using cumulative sums of martingale-based residuals. *Biometrika*, **8**0, 557–572.

Lin, X. (1997). Variance component testing in generalised linear modelswith random effects. *Biometrika*, **8**4, 309–326.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, **3**4, 1–41.

Lindstrom, M.J. and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association*, **8**3, 1014–1022.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models.* SAS Institute, Cary, NC.

Longford, N.T. (1993). *Random Coefficient Models.* Clarendon Press, Oxford.

Longford, N.T. (2001). Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society, Series A*, **1**64, 259–273.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (2003). *Multivariate Analysis.* Academic Press, London.

Marks, R. G. and Rao, P.V. (1979). An estimation procedure for data containing outliers with a one-directional shift in the mean. *Journal of the American Statistical Association*, **7**4, 614–620.

Marshall, E.C. and Spiegelhalter, D.J. (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, **2**, 409–444.

Martin, M.A. and Roberts, S. (2006). An evaluation of bootstrap methods for outlier detection in least squares regression. *Journal of Applied Statistics*, **3**3, 703–720.

Martin, R.J. (1992). Leverage, influence, residuals in regression models when observations are correlated. *Communications in Statistics: Theory and Methods*, **2**1, 1183–1212.

McCulloch, C.E. and Meeter, D. (1983). Discussion. *Technometrics*, **2**5, 152–155.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. John Wiley and Sons, New York.

Miyashita, H. and Newbold, P. (1983). On the sensitivity to non-normality of a test for outliers in linear models. *Communications in Statistics - Theory and Methods*, **1**2, 1413–1419.

Mohr, D.L. (2007). Bayesian identification of clustered outliers in multiple regression. *Computational Statistics and Data Analysis*, **5**1, 3955–3967.

Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score and Wald tests in a constrained parameter space. *The American Statistician*, **6**1, 22–27.

Morgan, B.J.T., Palmer, K.J., and Ridout, M.S. (2007). Negative score test statistic. *The American Statistician*, **6**1, 285–288.

Morrell, C.H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, **5**4, 1560–1568.

Naes, T. (1986). Detection of multivariate outliers in linear mixed models. *Communications in Statistics: Theory and Methods*, **1**5, 33–47.

Nobre, J.S. and Singer, J.M. (2007). Residual analysis for linear mixed models. *Biometrical Journal*, **4**9, 1–13.

Obenchain, R.L. (1977). Letter to the editor. *Technometrics*, **1**9, 348–349.

Oman, S. (1995). Checking the assumptions in mixed-model analysis of variance: a residual analysis approach. *Computational Statistics and Data Analysis*, **2**0, 309–330.

Pan, J. and Fang, K. (2002). *Growth Curve Models and Statistical Diagnostics*. Springer, New York.

Park, T. and Lee, S.-Y. (2004). Model diagnostic plots for repeated measures data. *Biometrical Journal*, **4**6, 441–452.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **5**8, 545–554.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Clarendon Press, Oxford.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-effects Models in S and S-plus*. Springer-Verlag, New York.

Pinheiro, J.C., Liu, C., and Wu, Y.N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, **10**, 249–276.

Poon, W-Y. and Poon, Y.S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **61**, 51–61.

Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful for growth curve problems. *Biometrika*, **51**, 313–326.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, **9**, 705–724.

Puterman, M.L. (1988). Leverage and influence in autocorrelated regression models. *Journal of the Royal Statistical Society, Series C*, **37**, 76–86.

R Development Core Team (2005). *R: A Language and Environmentfor Statistical Computing*. R Foundation for StatisticalComputing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, C.R. (1971). Estimation of variance and covariance components -MINQUE Theory. *Journal of Multivariate Analysis*, **1**, 257–275.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. John Wiley and Sons Inc., New York, Second edition.

Rebai, A., Goffinet, B., and Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138**, 235–240.

Richardson, A.M. (1997). Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association*, **92**, 154–161.

Robinson, G.K. (1991). That BLUP is a Good Thing: The estimation of random effects (with discussion). *Statistical Science*, **6**, 15–51.

Rocke, D.M. (1983). Robust statistical analysis of interlaboratory studies. *Biometrika*, **70**, 421–431.

Rocke, D.M. (1991). Robustness and balance in the mixed model. *Biometrics*, **47**, 303–309.

Rousseeuw, P.J. (1984). Least median squares regression. *Journal of the American Statistical Association*, **7**9, 871–880.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley, New York.

Rousseuw, P.J. and Van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **8**5, 633–651.

Rubin, D.B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **1**2, 1151–1172.

Ruppert, D. (2004). *Statistics and Finance: An Introduction*. Springer-Verlag, New York.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110–114.

Schabenberger, O. (2004). Mixed model influence diagnostics. Technical Report SGU 29: Paper 188-29, SAS Institute Inc., Cary, NC.

Schall, R. and Dunne, T.T. (1988). A unified approach to outliers in the general linear model. *Sankhya B: The Indian Journal of Statistics, Series B*, **5**0, 157–167.

Schall, R. and Dunne, T.T. (1992). A note on the relationship between parameter collinearity and local influence. *Biometrika*, **7**9, 399–404.

Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics and Data Analysis*, **5**2, 3283–3299.

Searle, R. (1995). An overview of variance component estimation. *Metrika*, **4**2, 215–230.

Searle, S.R. (1971). *Linear Models*. John Wiley and Sons Inc., New York.

Searle, S.R. (1982a). *Linear models for unbalanced data*. John Wiley and Sons Inc., New York.

Searle, S.R. (1982b). *Matrix algebra useful for statistics*. John Wiley and Sons Inc., New York.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components*. John Wiley and Sons Inc., New York.

Seely, J. (1971). Quadratic subspaces and completeness. *Annals of Mathematical Statistics*, **4**2, 710–721.

Self, S.G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **8**2, 605–610.

Sharples, L.D. (1990a). Identification and accommodation of outliers in general hierarchical models. *Biometrika*, **7**7, 445–453.

Sharples, L.D. (1990b). A semi-parametric Bayesian model for randomized block designs. *Biometrika*, **7**7, 445–453.

Shi, L. and Chen, G. (2008). Detection of outliers in multilevel models. *Journal of Statistical Planning and Inference*, **1**38, 3189–3199.

Shi, L. and Chen, G. (2009). Influence measures for general linear models with correlated errors. *The American Statistician*, **6**3, 40–42.

Silvapulle, M. J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*, **9**0, 342–349.

Solomon, P.J. (1985). Transformations for components of variance and covariance. *Biometrika*, **7**2, 233–239.

SPSS Inc. (1999). *SPSS for Windows, Rel.11.01. 2001.* SPSS Inc., Chicago IL.

StataCorp. (2005). *Statistical Software: Release 9.0.* Stata Corporation, College Station, TX.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random-effects models for serial observations with binary responses. *Biometrics*, **4**0, 961–971.

Storer, B. and Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihoods. *Journal of the American Statistical Association*, **8**0, 139–147.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **6**4, 479–498.

Storey, J.D. (2003). The positive false discovery rate : A Bayesian interpretation and the q-value. *Annals of Statistics*, **3**1, 2013–2035.

Storey, J.D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society, Series B*, **6**9, 347–368.

Stram, D. and Lee, J.W. (1994). Variance components testing in longitudinal mixed effects model. *Biometrics*, **5**0, 1171–1177.

Stram, D. and Lee, J.W. (1995). Correction to "Variance components testing in longitudinal mixed effects model". *Biometrics*, **5**1, 1196.

Swallow, W.H. and Monahan, J.F. (1984). Monte Carlo comparison of ANOVA, MINVQUE, REML, and ML estimators of variance components. *Technometrics*, **2**6, 47–57.

Swamy, P.A.V.B. (1971). *Statistical Inference in Random Coefficient Regression Models*. Springer, New York.

Tan, F.E.S., Ouwens, M.J.N, and Berger, M.P.F. (2001). Detection of influential observations in longitudinal mixed effects regression models. *The Statistician*, **5**0, 271–284.

Thisted, R. (1988). *Elements of Statistical Computing*. Chapman and Hall, New York.

Thompson, R. (1985). A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society, Series B*, **4**7, 53–55.

Vaida, F. and Blancard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, **9**2, 351–370.

Van der Merwe, A.J. and Pretorius, A.L. (2002). A Bayesian non-parametric approach to the mixed linear model with an application to animal breeding. *South African Statistical Journal*, **3**6, 1–38.

Venables, W.N. and Ripley, B.D. (1997). *Modern Applied Statistics with S-plus*. Springer-Verlag, New York.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, **2**3, 541–556.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, **5**9, 254–262.

Verbyla, A. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics*, **3**2, 227–230.

Verbyla, A. (1993). Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society, Series B*, **5**5, 493–508.

von Rosen, D. (1991). The growth curve model: a review. *Communications in Statistics: Theory and Methods*, **2**0, 2791–2822.

Vonesh, E.F. and Carter, R.L. (1987). Efficient inference for random-coefficient growth-curve models with unbalanced data. *Biometrics*, **4**3, 617–628.

Vonesh, E.F. and Chinchilli, V.M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. Marcel Dekker, New York.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **5**4, 426–482.

Wang, S.R. (1982). *Multivariate Statistical Analysis of Geological data*. Science Press, Beijing.

Ware, J.H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, **3**9, 95–101.

Waternaux, C., Laird, N.M., and Ware, J.H. (1989). Methods for analysis of longitudinal data: Blood-lead concentrations and cognitive development. *Journal of the American Statistical Association*, **8**4, 33–41.

Welham, S. and Thompson, R. (1997). Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society, Series B*, **5**9, 701–714.

Welham, S. and Thompson, R. (2000). *The Guide to Genstat, Part 2: Statistics. Ch. REML analysis of mixedmodels.* VSN International, Wilkinson House, Jordan Hill Road, Oxford, UK.

Welsh, A.H. and Richardson, A.M. (1997). Approaches to robust estimation of mixed models. In Maddala, G.S. and Rao, C.R., editors, *Handbook of Statistics*, pages 343–384. North-Holland, Amsterdam.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics: Simulation and Computation*, **2**2, 1079–1106.

Yang, Z. and Abeysinghe, T. (2003). A score test for Box-Cox functional form. *Economics Letters*, **7**9, 107–115.

Youden, W.J. (1975). Statistical techniques for collaborative tests. In *Statistical Manual of the Association of Official Analytical Chemists*. Association of Analytical Chemists, Arlington VA.

Zewotir, T. (2007). Infinitesimal model perturbation influence in the linear mixed model. *South African Statistical Journal*, **4**1, 105–126.

Zewotir, T. (2008). Multiple cases deletion diagnostics for linear mixed models. *Communications in Statistics - Theory and Methods*, **37**, 1071 – 1084.

Zewotir, T. and Galpin, J. (2005). Influence diagnostics for linear mixed models. *Journal of Data Science*, **3**, 153–177.

Zewotir, T. and Galpin, J. (2006). Evaluation of linear mixed model case deletion diagnostic tools by Monte Carlo simulation. *Communications in Statistics - Simulation and Computation*, **3**5, 645–682.

Zewotir, T. and Galpin, J. (2007). A unified approach to residuals, leverages and outliers in the linear mixed model. *Test*, **16**, 58–75.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions for longitudinal data. *Biometrics*, **5**7, 795–802.

Zhu, H. and Lee, S. (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B*, **6**3, 111–126.

Zhu, H., Lee, S., Wei, B., and Zhou, J. S. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, **8**8, 727–737.

Zhu, H. and Zhang, H. (2006). Generalized score test of homogeneity for mixed effects models. *Annals of Statistics*, **3**4, 1545–1569.

APPENDIX A

Useful (matrix) results and identities

Below is a summary of known results in matrix algebra which we use in the thesis. Also included are some fundamental statistical results which are used in the derivation of some of the proofs in the thesis. These results can be found in Mardia et al. (2003) and Searle (1982b).

Definition 1

The square matrix $\boldsymbol{A}$ *is idempotent if*

$$\boldsymbol{A}^2 = \boldsymbol{A} \quad \Rightarrow \quad (\boldsymbol{I} - \boldsymbol{A})^2 = \boldsymbol{I} - \boldsymbol{A}.$$

Definition 2 *A quadratic form of the vector* $\boldsymbol{y}$ *as given by* $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ *for some symmetric matrix* $\boldsymbol{A}$ *is said to be non-negative definite if* $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} \geq 0$ *for all* $\boldsymbol{y} \in \boldsymbol{\mathcal{R}}^n$. *Two forms of non-negative definiteness are:*

(ii) positive definite if $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} > 0$ *for all* $\boldsymbol{y}$ *other than* $\boldsymbol{y} = \boldsymbol{0}$. $\boldsymbol{A} = \boldsymbol{A}'$ *is also a positive definite matrix.*

(iii) positive semidefinite if $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} \geq 0$ *for all* $\boldsymbol{y}$ *and* $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ *for some* $\boldsymbol{y} \neq \boldsymbol{0}$. $\boldsymbol{A} = \boldsymbol{A}'$ *is also a positive semidefinite matrix.*

Result A.1 *For matrices* $\boldsymbol{B}^{p\times n}$ *and* $\boldsymbol{D}^{n\times p}$ *and for non-singular matrices* $\boldsymbol{C}^{n\times n}$ *and* $\boldsymbol{A}^{p\times p}$*, from Rao (1973, pp. 33) we have*

$$(\boldsymbol{A} + \boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}.$$

Result A.2 *Consider the matrix* $\boldsymbol{A}^{m\times m}$ *of full rank which is partitioned as*

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \tag{A.1}$$

Then the inverse of $\boldsymbol{A}$ is partitioned conformably with $\boldsymbol{A}$ as

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{bmatrix}$$
$$= \begin{bmatrix} (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}^{22} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}^{11} & (\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1} \end{bmatrix},$$

provided $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are non-singular.

Result A.3 *If $\boldsymbol{A}$ is symmetric and $\boldsymbol{A}_{22}$ and $\boldsymbol{Q} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$ are non-singular, then $\boldsymbol{A}^{-1}$ can be written as*

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{Q}^{-1} & -\boldsymbol{Q}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{22}^{-1} \\ \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}\boldsymbol{Q}^{-1} & \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}\boldsymbol{Q}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{22}^{-1} \end{bmatrix}.$$

Result A.4 *Using the definition of $\boldsymbol{A}$ in* (A.1)*, the determinant of $\boldsymbol{A}$ can be expressed as*

$$|\boldsymbol{A}| = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}| = |\boldsymbol{A}_{22}||\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}|,$$

for matrices $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ non-singular. The notation $|\boldsymbol{A}|$ denotes the determinant of the matrix $\boldsymbol{A}$.

Result A.5 *For matrices $\boldsymbol{B}^{p\times n}$ and $\boldsymbol{C}^{n\times p}$, and for non-singular matrix $\boldsymbol{A}^{p\times p}$,*

$$|\boldsymbol{A} + \boldsymbol{BC}| = |\boldsymbol{A}||\boldsymbol{I}_p + \boldsymbol{A}^{-1}\boldsymbol{BC}| = |\boldsymbol{A}||\boldsymbol{I}_n + \boldsymbol{CA}^{-1}\boldsymbol{B}|.$$

Result A.6 ***Matrix differentiation***

The derivative of $f(\boldsymbol{X})$ with respect to $\boldsymbol{X}^{n\times p}$ is given by

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} = \left(\frac{\partial f(\boldsymbol{X})}{\partial x_{ij}}\right).$$

For $\boldsymbol{A}^{n\times n}$ symmetric we have the following results:

(i) $\dfrac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$

(ii) $\dfrac{\partial \boldsymbol{x}'\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}, \qquad \dfrac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}, \qquad \dfrac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{y}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{y}$

Also when $\boldsymbol{B}^{p\times p} = \boldsymbol{B}(\boldsymbol{\kappa}) = \{b_{ij}\}$ *is non-singular with elements which are functions of* $\boldsymbol{\kappa}$, $\boldsymbol{x}^{p\times 1} = \boldsymbol{x}(\boldsymbol{\kappa})$ *and* $\boldsymbol{A}^{p\times p}$ *is a symmetric matrix, denoting the trace of a square matrix* $\boldsymbol{C}$ *as* $\mathrm{tr}(\boldsymbol{C})$, *we have*

(i) $\dfrac{\partial \log|\boldsymbol{B}|}{\partial \boldsymbol{\kappa}_k} = \mathrm{tr}\left(\boldsymbol{B}^{-1}\dot{\boldsymbol{B}}_k\right)$, *where* $\dot{\boldsymbol{B}}_k = \dfrac{\partial \boldsymbol{B}}{\partial \boldsymbol{\kappa}_k}$

(ii) $\dfrac{\partial \boldsymbol{B}^{-1}}{\partial \boldsymbol{\kappa}_k} = -\boldsymbol{B}^{-1}\dot{\boldsymbol{B}}_k\boldsymbol{B}^{-1}$

(iii) $\dfrac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{\kappa}_k} = 2\dfrac{\partial \boldsymbol{x}'}{\partial \boldsymbol{\kappa}_k}(\boldsymbol{A}\boldsymbol{x})$

(iv) *If* $\boldsymbol{B}$ *is symmetric then*

$$\mathrm{tr}\left(\frac{\partial \boldsymbol{B}\boldsymbol{A}}{\partial \boldsymbol{B}}\right) = 2\boldsymbol{A} - \mathrm{diag}(\boldsymbol{A}).$$

Result A.7 *Let* $\boldsymbol{y}$ *be multivariate Gaussian, with mean* $\boldsymbol{\mu}$ *and variance matrix* $\boldsymbol{\Sigma}$. *Partitioning* $\boldsymbol{y}$, $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *conformably as*

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

the multivariate Gaussian normal distribution can be written as

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right).$$

Then the conditional distribution of $\boldsymbol{y}_1$ *given* $\boldsymbol{y}_2$ *is also Gaussian and*

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

Result A.8 ***Quadratic forms***

A quadratic form of the vector $\boldsymbol{y}$ *is given by* $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ *for some symmetric matrix* $\boldsymbol{A}$. *If* $\boldsymbol{A}$ *is not symmetric then the quadratic form is given by*

$$\boldsymbol{y}'\boldsymbol{B}\boldsymbol{y} = \boldsymbol{y}'\left(\frac{\boldsymbol{A}'}{2} + \frac{\boldsymbol{A}}{2}\right)\boldsymbol{y}$$

If $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *then*

(i) $\mathrm{E}(\boldsymbol{y}'\boldsymbol{\Sigma}\boldsymbol{y}) = \mathrm{tr}[\boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')] = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$.

(ii) $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} \sim \chi^2_b(\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu})$ *if and only if* $\boldsymbol{A}\boldsymbol{\Sigma}$ *is idempotent, i.e.* $(\boldsymbol{A}\boldsymbol{\Sigma})^2 = \boldsymbol{A}\boldsymbol{\Sigma}$*), where* $b = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) = \mathrm{rank}(\boldsymbol{A})$ *since* $\boldsymbol{\Sigma}$ *is non-singular and* $\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$ *is the non-centrality parameter.*

(iii) $\mathrm{var}(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}) = 2\mathrm{tr}[(\boldsymbol{A}\boldsymbol{\Sigma})^2] + 4\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu}$.

(iv) $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ *and* $\boldsymbol{y}'\boldsymbol{B}\boldsymbol{y}$ *are independent if and only if* $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$.

(v) $\mathrm{cov}(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}, \boldsymbol{y}'\boldsymbol{B}\boldsymbol{y}) = 2\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma})$.

The following lemmas hold for the linear mixed model (2.1) namely,

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \\ &\sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H}\right). \end{aligned}$$

These lemmas assume that $\boldsymbol{G}$, $\boldsymbol{H}$ and $\boldsymbol{R}$ are positive definite and that design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are of full rank. A convenient reparametrization of $\boldsymbol{\beta}$ can change $\boldsymbol{X}$ to say, $\boldsymbol{X}^*$ so that $\boldsymbol{X}^*$ is also of full-column rank.

Lemma A.1 *Using Result A.1 the inverse of the variance-covariance matrix* $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$ *is given by*

$$\boldsymbol{H}^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}. \tag{A.2}$$

Lemma A.2 *Let* $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$ *then* $\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}$.

Proof. Since $\boldsymbol{H}\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$ is idempotent,

$\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{H}\boldsymbol{P}$, premultiplying by $\boldsymbol{H}^{-1}$ gives

$$\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}.$$

■

Lemma A.3 *The partial derivative of* $\boldsymbol{P}$ *with respect to the variance parameters* $\phi_j \in \boldsymbol{\phi}$ *is given by*

$$\frac{\partial \boldsymbol{P}}{\partial \phi_j} = -\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P},$$

where $\dot{\boldsymbol{H}}_j = \dfrac{\partial \boldsymbol{H}}{\partial \phi_j}$ *and* $\boldsymbol{\phi}$ *is the vector of variance parameters contained in* $\boldsymbol{P}$ *through* $\boldsymbol{H}$.

Proof. Using Result A.6 we obtain

$$\begin{aligned}
\frac{\partial \boldsymbol{P}}{\partial \phi_j} &= -\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1} + \boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} \\
&\quad - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} \\
&\quad + \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1} \\
&= -\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{P} + \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{P} \\
&= -\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}
\end{aligned}$$

which proves the Lemma. ■

Lemma A.4 *It can be shown that*

$$\boldsymbol{C}^{-1} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & -(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \\ -\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & \boldsymbol{S} + \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \end{bmatrix},$$

where

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{XX} & \boldsymbol{C}_{XZ} \\ \boldsymbol{C}_{ZX} & \boldsymbol{C}_{ZZ} \end{bmatrix} \qquad \text{(A.3)}$$

is assumed to be of full rank, i.e. $\boldsymbol{X}$ *is of full column rank, and*

$$\boldsymbol{S} = (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}.$$

Proof.

If we let $\boldsymbol{C} = \boldsymbol{A}$ ($\boldsymbol{A}$ as defined in Result A.3) then

$$\begin{aligned}\boldsymbol{Q} &= \boldsymbol{C}_{XX} - \boldsymbol{C}_{XZ}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX} \\ &= \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} \\ &= \boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}.\end{aligned}$$

Noting that

$$\begin{aligned}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\boldsymbol{G}\boldsymbol{Z}' &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{Z}' \\ &= \boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}) \\ &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{H}.\end{aligned}$$

Hence

$$(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}.$$

Thus

$$\begin{aligned}\boldsymbol{C}_{ZZ}^{-1}\boldsymbol{C}_{ZX} &= (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} \\ &= \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}.\end{aligned}$$

Therefore from Result A.3

$$\boldsymbol{C}^{-1} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & -(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \\ -\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & \boldsymbol{S} + \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G}. \end{bmatrix} \tag{A.4}$$

Using

$$\boldsymbol{H}^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1},$$

the lower right-hand matrix of $\boldsymbol{C}^{-1}$ can be simplified as follows

$$\begin{aligned}\boldsymbol{K} &= \boldsymbol{C}_{ZZ}^{-1} + \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \\ &= (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1} + \boldsymbol{G}\boldsymbol{Z}'(\boldsymbol{H}^{-1} - \boldsymbol{P})\boldsymbol{Z}\boldsymbol{G} \\ &= (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\end{aligned}$$

$$+ \boldsymbol{G}\boldsymbol{Z}'[\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}\boldsymbol{R}^{-1}]\boldsymbol{Z}\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}.$$

Writing $\boldsymbol{L} = \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}$ gives

$$\begin{aligned}
\boldsymbol{K} &= (\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1} + \boldsymbol{G}\boldsymbol{L}\boldsymbol{G} - \boldsymbol{G}\boldsymbol{L}(\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{L}\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\
&= (\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1} - \boldsymbol{G}\boldsymbol{L}(\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1}(\boldsymbol{L} + \boldsymbol{G}^{-1} - \boldsymbol{L})\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\
&= (\boldsymbol{I} + \boldsymbol{G}\boldsymbol{L})(\boldsymbol{L} + \boldsymbol{G}^{-1})^{-1} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \\
&= \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}.
\end{aligned}$$

Thus an alternative expression for the inverse of $\boldsymbol{C}$ is

$$\boldsymbol{C}^{-1} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & -(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \\ -\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \end{bmatrix}. \tag{A.5}$$

■

Lemma A.5 *An alternative expression for $\boldsymbol{P}$ is given by*

$$\boldsymbol{P} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}, \tag{A.6}$$

where $\boldsymbol{W} = [\boldsymbol{X}\ \boldsymbol{Z}]$.

Proof. We show that equation (A.6) is equivalent to $\boldsymbol{P}$ as given in Lemma A.2.

$$\begin{aligned}
\boldsymbol{P} &= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1} \\
&= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1} + \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1} \\
&\quad + \boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\
&\quad - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\
&= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} - (\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1} \\
&\quad + (\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\
&= \boldsymbol{H}^{-1} - (\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1}) \\
&= \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} \\
&= \boldsymbol{P},
\end{aligned}$$

where $\boldsymbol{S} = (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}$. ■

The proofs of the following results can be found in Gilmour et al. (1995).

Result A.9 *The score function for κ_j is given by*

$$U(\kappa_j) = \frac{\partial l_R(\boldsymbol{\phi};\boldsymbol{y})}{\partial \kappa_j} = -\frac{1}{2}\left\{\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) - \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}\right\}, \tag{A.7}$$

where $l_R(\boldsymbol{\phi};\boldsymbol{y})$ is the REML log-likelihood function (2.19)*, $\boldsymbol{\phi} = (\boldsymbol{\kappa}', \sigma^2)'$ and $\dot{\boldsymbol{H}}_j = \partial \boldsymbol{H}/\partial \kappa_j$; for $j = 1, \ldots, r+s$, where $r+s$ is the number of variance parameters in $\boldsymbol{\kappa}$. The number of variance parameters in the model including σ^2, i.e. the number of parameters in $\boldsymbol{\phi}$ is $t = r+s+1$.*

Proof. The derivative of the log determinant terms in (2.19) is given by

$$\begin{aligned}
&\frac{\partial \log|\boldsymbol{H}|}{\partial \kappa_j} + \frac{\partial \log|\boldsymbol{X}'\boldsymbol{H}\boldsymbol{X}|}{\partial \kappa_j} \\
&= \mathrm{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\right) + \mathrm{tr}\left((\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\frac{\partial \boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}}{\partial \kappa_j}\right) \\
&= \mathrm{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\right) + \mathrm{tr}\left((\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\boldsymbol{H}^{-1}\boldsymbol{X}\right) \\
&= \mathrm{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_j\right) \\
&= \mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right).
\end{aligned} \tag{A.8}$$

Then differentiating the sums of squares term in (2.19) we obtain

$$\boldsymbol{y}'\frac{\partial \boldsymbol{P}}{\partial \kappa_j}\boldsymbol{y} = -\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}, \tag{A.9}$$

using Lemma A.3.

Finally combining (A.8) and (A.9) gives (A.7) as required. ∎

Result A.10 *The score function for σ^2 is given by*

$$U(\sigma^2) = \frac{\partial l_R(\boldsymbol{\phi};\boldsymbol{y})}{\partial \sigma^2} = -\frac{1}{2}\left\{\frac{(n-p)}{\sigma^2} - \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^4}\right\}. \tag{A.10}$$

Result A.11 *The elements of the observed information matrix for the variance* *parameters, κ_j and σ^2 are*

$$\mathcal{I}_{\mathcal{O}}(\kappa_j, \kappa_k) = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) + \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{y} - \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P}\boldsymbol{y} \tag{A.11a}$$

$$\mathcal{I}_{\mathcal{O}}(\sigma^2, \kappa_j) = \frac{\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{2\sigma^4} \tag{A.11b}$$

$$\mathcal{I}_{\mathcal{O}}(\sigma^2, \sigma^2) = -\frac{(n-p)}{2\sigma^4} + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^6}. \tag{A.11c}$$

where $\ddot{\boldsymbol{H}}_{jk} = \partial^2\boldsymbol{H}/\partial\kappa_j\kappa_k$.

Proof. Taking negative second derivatives of (2.19) with respect to κ_j and σ^2 we obtain

$$\begin{aligned}\mathcal{I}_{\mathcal{O}}(\kappa_j, \kappa_k) &= -\frac{\partial^2 l_R(\boldsymbol{\phi};\boldsymbol{y})}{\partial\kappa_j\partial\kappa_k} \\ &= \frac{1}{2}\left\{\frac{\partial\mathrm{tr}\boldsymbol{P}\dot{\boldsymbol{H}}_j}{\partial\kappa_k} - \frac{1}{\sigma^2}\frac{\partial\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{\partial\kappa_k}\right\}\end{aligned} \tag{A.12}$$

The first term of (A.12) is given by

$$\begin{aligned}\frac{\partial\mathrm{tr}\boldsymbol{P}\dot{\boldsymbol{H}}_j}{\partial\kappa_k} &= \mathrm{tr}\left(\frac{\partial\boldsymbol{P}}{\partial\kappa_k}\dot{\boldsymbol{H}}_j + \boldsymbol{P}\frac{\partial\dot{\boldsymbol{H}}_j}{\partial\kappa_k}\right) \\ &= \mathrm{tr}\left(-\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\dot{\boldsymbol{H}}_j + \boldsymbol{P}\frac{\partial\dot{\boldsymbol{H}}_j}{\partial\kappa_k}\right) \qquad \text{using Lemma A.3} \\ &= \mathrm{tr}\left(-\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\dot{\boldsymbol{H}}_j + \boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right)\end{aligned} \tag{A.13}$$

and the second term is given by

$$\begin{aligned}\frac{\partial\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{\partial\kappa_k} &= \boldsymbol{y}'\left(\frac{\partial\boldsymbol{P}}{\partial\kappa_k}\dot{\boldsymbol{H}}_j\boldsymbol{P} + \boldsymbol{P}\frac{\partial\dot{\boldsymbol{H}}_j}{\partial\kappa_k}\boldsymbol{P} + \boldsymbol{P}\dot{\boldsymbol{H}}_j\frac{\partial\boldsymbol{P}}{\partial\kappa_k}\right)\boldsymbol{y} \\ &= \boldsymbol{y}'\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P} - (\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P} + \boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P})\right)\boldsymbol{y} \\ &= \boldsymbol{y}'\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P} - 2\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\right)\boldsymbol{y}\end{aligned} \tag{A.14}$$

Substituting (A.13) and (A.14) into (A.12) gives (A.11a).

$$\begin{aligned}
\mathcal{I}_{\mathcal{O}}(\sigma^2, \kappa_j) &= -\frac{\partial^2 l_R(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \sigma^2 \partial \kappa_j} \\
&= \frac{1}{2}\frac{\partial}{\partial \kappa_j}\left\{\frac{(n-p)}{\sigma^2} - \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^4}\right\} \\
&= \frac{\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{2\sigma^4} \qquad \text{using Lemma A.3} \\
\mathcal{I}_{\mathcal{O}}(\sigma^2, \sigma^2) &= -\frac{\partial^2 l_R(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \sigma^2 \partial \sigma^2} \\
&= -\frac{(n-p)}{2\sigma^4} + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^6}
\end{aligned}$$

■

Result A.12 *The elements of the expected information matrix for the variance parameters,* κ_j *and* σ^2 *are*

$$\mathcal{I}_{\mathcal{E}}(\kappa_j, \kappa_k) = \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) \tag{A.15a}$$

$$\mathcal{I}_{\mathcal{E}}(\sigma^2, \kappa_j) = \frac{1}{2\sigma^2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \tag{A.15b}$$

$$\mathcal{I}_{\mathcal{E}}(\sigma^2, \sigma^2) = \frac{(n-p)}{2\sigma^4}. \tag{A.15c}$$

Proof. Taking the expected values of the terms in Result 2.4 using the expected of quadratic forms (Result A.8) we obtain

$$\begin{aligned}
\mathcal{I}_{\mathcal{E}}(\kappa_j, \kappa_k) &= \text{E}\left(-\frac{\partial^2 l_R(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \kappa_j \partial \kappa_k}\right) \\
&= \frac{1}{2}\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) - \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) \\
&\quad + \text{tr}\left(\boldsymbol{H}\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) + \frac{1}{\sigma^2}(\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} \\
&\quad - \frac{1}{2}\text{tr}\left(\boldsymbol{H}\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P}\right) - \frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} \\
&= \frac{1}{2}\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) - \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) + \text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) \\
&\quad - \frac{1}{2}\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) \\
&= \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right)
\end{aligned}$$

$$
\begin{aligned}
\mathcal{I}_{\mathcal{E}}(\sigma^2, \kappa_j) &= \mathrm{E}\left(-\frac{\partial^2 l_R(\kappa, \sigma^2; \boldsymbol{y})}{\partial \sigma^2 \partial \kappa_j}\right) \\
&= \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) + \frac{1}{\sigma^4}(\boldsymbol{X\beta})'\boldsymbol{PX\beta} \\
&= \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \\
\mathcal{I}_{\mathcal{E}}(\sigma^2, \sigma^2) &= \mathrm{E}\left(-\frac{\partial^2 l_R(\boldsymbol{\phi}; \boldsymbol{y})}{\partial \sigma^2 \partial \sigma^2}\right) \\
&= -\frac{(n-p)}{2\sigma^4} + \frac{1}{\sigma^4}\mathrm{tr}\left(\boldsymbol{PH}\right) + \frac{1}{\sigma^6}(\boldsymbol{X\beta})'\boldsymbol{PX\beta} \\
&= \frac{(n-p)}{2\sigma^4}
\end{aligned}
$$

■

Result A.13 *The elements of the approximate average information matrix for the variance parameters, κ_j and σ^2 are*

$$\mathcal{I}_{\mathcal{A}}(\kappa_j, \kappa_k) = \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{Py} \tag{A.16a}$$

$$\mathcal{I}_{\mathcal{A}}(\sigma^2, \kappa_j) = \frac{1}{2\sigma^4}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{Py} \tag{A.16b}$$

$$\mathcal{I}_{\mathcal{A}}(\sigma^2, \sigma^2) = \frac{\boldsymbol{y}'\boldsymbol{Py}}{2\sigma^6}. \tag{A.16c}$$

Proof. Approximate averages of the observed and expected information terms from Results A.11 and A.12 give

$$
\begin{aligned}
\mathcal{I}_{\mathcal{A}}(\kappa_j, \kappa_k) &= \frac{1}{2}\Bigg\{\frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right) + \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{Py} \\
&\qquad - \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{Py} + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\right)\Bigg\} \\
&= \frac{1}{2}\Bigg\{\frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) + \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{Py} - \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{Py}\Bigg\} \\
&\approx \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{Py},
\end{aligned}
$$

where we approximate $\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{Py}$ by its expectation, $\sigma^2\mathrm{tr}(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk})$.

$$\mathcal{I}_{\mathcal{A}}(\sigma^2, \kappa_j) = \frac{1}{2}\Bigg\{\frac{\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{Py}}{2\sigma^4} + \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right)\Bigg\}$$

$$\approx \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y},$$

where $\sigma^2 \text{tr}(\boldsymbol{P}\dot{\boldsymbol{H}}_j)$ is approximated by $\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}/\sigma^2$ since these terms are equal upon setting equation (A.7) to zero. This approximation ensures that the information matrix $\boldsymbol{\mathcal{I}}_{\mathcal{A}}$ is positive-semidefinite.

$$\begin{aligned}\mathcal{I}_{\mathcal{A}}(\sigma^2, \sigma^2) &= \frac{1}{2}\left\{ -\frac{(n-p)}{2\sigma^4} + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^6} + \frac{(n-p)}{2\sigma^4} \right\} \\ &= \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^6}.\end{aligned}$$

■

Result A.14 *The elements of the exact average information matrix for variance parameters are obtained by taking equally-weighted averages within the three pairs of terms in Result A.11 and A.12 and are given by*

$$\begin{aligned}\mathcal{I}_{\mathcal{A}e}(\kappa_j, \kappa_k) &= \frac{1}{4}\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\right) + \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_k\boldsymbol{P}\boldsymbol{y} \\ &\quad - \frac{1}{4\sigma^2}\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{jk}\boldsymbol{P}\boldsymbol{y} \qquad \text{(A.17a)}\end{aligned}$$

$$\mathcal{I}_{\mathcal{A}e}(\sigma^2, \kappa_j) = \frac{\boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{4\sigma^4} + \frac{1}{4\sigma^2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \qquad \text{(A.17b)}$$

$$\mathcal{I}_{\mathcal{A}e}(\sigma^2, \sigma^2) = \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^6}. \qquad \text{(A.17c)}$$

APPENDIX B

Chapter 4 proofs

It can be shown that $\boldsymbol{M}_3^* = \boldsymbol{M}_3$.

Proof.

$$\begin{aligned}\boldsymbol{M}_3^* &= \boldsymbol{I} - \boldsymbol{R}\boldsymbol{P} \\ &= \boldsymbol{I} - \boldsymbol{R}[\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}]\end{aligned}$$

Using $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$ we obtain

$$\boldsymbol{R}\boldsymbol{H}^{-1} = \boldsymbol{I} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}.$$

Hence

$$\begin{aligned}\boldsymbol{M}_3^* &= \boldsymbol{I} - \boldsymbol{R}(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}) \\ &= \boldsymbol{I} - [(\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}) - (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}] \\ &= (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1} \\ &= \boldsymbol{M}_3.\end{aligned}$$

■

It can also be shown that $\boldsymbol{M}_4 = \boldsymbol{M}_3 = \boldsymbol{M}_2$.

Proof.

$$\begin{aligned}\boldsymbol{M}_4 &= \boldsymbol{M}_2 \\ &= \boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1} - \boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1} \\ &\quad - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} + \boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\ &\quad + \boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} \qquad \text{using Lemma A.4 of Appendix A,}\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1}) - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\
&\quad + \boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} + \boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\
&= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1} \\
&\quad + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\
&= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1} \\
&\quad - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}'\boldsymbol{R}^{-1}) \\
&= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1} \\
&= (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1} \\
&= \boldsymbol{M}_3,
\end{aligned}
$$

where $\boldsymbol{S} = (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}$. ■

APPENDIX C

Chapter 5 examples

C.1 GenStat code for simulated data example

```
"Chapter 5: AOM in linear fixed effects analysis"
"Calculate LRTs and score tests for  initial data set"
"Runs sets of simulations to obtain empirical distr. of order (max). test statistics for determining thresholds for LRTs and score tests"
"Parameters: n = number of units
              p = number of fixed effects parameters
              s2 = initial value of sigma2
              sim = number of simulations to obtain empirical distr. of max. test statistic."
read [ch=2] n,p,s2,sim
close 2; filetype=input
prin  n,p,sigu,sim; dec=0,0,0,0; fie=6
scal  Chan1
"Genstat log file"
open 'c:/research/phd/output-graphics/simul-lmaom.DAT'; CHANNEL=Chan1; FILETYPE=OUTPUT
"Generate initial data set"
variate[;1...#n]obs
"set up s2, mu "
scal  beta0,beta1; VAL=20.5,0.25
print beta0,beta1
"Simple Linear Regression"
"Generate initial data set"
 vari [NVALUES=n] y,e,xij
 calc e= GRNORMAL(n;0;s2)
 calc xij= sort(GRUNIFORM(n;0;10))
 calc y=beta0+beta1*xij+e
 "Insert outliers for observations 10 and n"
 calc   y$[10]=(beta0+beta1*xij$[10])+4
 &      xij$[n]=15
 &      y$[n]=(beta0+beta1*xij$[n])+4
 calc xij1=xij-mean(xij)
 "Fit inital model: linear regression model"
 Model y
 Terms [FACT=9] xij1
 Fit [PRINT=model,summary,estimates; CONSTANT=estimate; FPROB=yes;TPROB=yes; FACT=9] xij1
 Rkeep [RMETHOD=simple] Y=y;Residuals=resm;Fitted=fitm;Leve=levm;Estimates=fixedeff;deviance=dv;df=df1
calc sige1=dv/df1
"Setting up the X mat (nxp)"
variate [NVALUES=#n;VALUES=(1)#n]ONES
variate [NVALUES=#n]fitm1
matrix[ROWS=#n;COL=2]X
calc X$[*;1...1]=ONES
calc X$[*;2...2]=xij1
    " Get PXp for gamma=0 "
  symm [r=n] PXp
  calc PXp = ident(n) - X*+inverse((t(X)*+X))*+t(X)
  vari [nval=n] PXpy
```

```
  calc PXpy = PXp*+y
   variate [NVALUES=#n]cii;values=diag(PXp)
   variate [NVALUES=#n]ci;values=(PXpy)

Diagonalmatrix [R=#n] II; VAL=IDENTITY(#n)
matrix[ROWS=#n;COL=#n]I2
calc I2=II
"Calculate ti2,score and variance estimates, and lik. ratio and score tests fro initi data"
variate [NVALUES=#n]lrt,elrt1,olrt1,alrt1,aelrt1,elrt2,olrt2,alrt2,aelrt2,score,escore,oscore,
ascore,aescore,ti,ti2,omega,sige1a
   variate[;1...#n]obs
   For[NTIMES=#n;INDEX=k]
         calc ti$[k]=ci$[k]/(sqrt(sige1)*sqrt(cii$[k]))
         calc ti2$[k]=ti$[k]**2
         calc score$[k]=((ci$[k]**2/sige1)-cii$[k])/2
Endfor
variate [NVALUES=#n]lrt,elrt1,olrt1,alrt1,aelrt1,elrt2,olrt2,alrt2,aelrt2,escore,oscore,ascore,aescore,omega,sige1a,omeghat,sig2hat,\
        vwe,vwo,vwa,vwe1,vwo1,vwa1,vwa11,sgb1,sgep,sgop,sgap,sga1p
For[NTIMES=#n;INDEX=k]
   calc aom2=I2$[*;k]
   If ti2$[k] .gt. 1
    " one-step LRT updating w and s2 simultaneoulsy"
    calc elrt1$[k] = (n-p)*log((n-p-1)/(n-p-ti2$[k])) - log( ((n-p)*ti2$[k]-1)/(n-p-1)) \
     - (n-p)*(ti2$[k]-1)*(ti2$[k]-1)/( (n-p-ti2$[k])*((n-p)*ti2$[k]-1) )
    calc olrt1$[k] = (n-p)*log(( (n-p)*(2*ti2$[k]-1)-(ti2$[k]*ti2$[k]) )/( (n-p-ti2$[k])*(2*ti2$[k]-1) )) \
                 - log(( (n-p)*(3*ti2$[k]-2)-(ti2$[k]*ti2$[k]) )/( (n-p)*(2*ti2$[k]-1)-(ti2$[k]*ti2$[k]) )) \
                 - (n-p)*(n-p)*ti2$[k]*(ti2$[k]-1)*(ti2$[k]-1)/( (n-p-ti2$[k])*(2*ti2$[k]-1)*((n-p)*(3*ti2$[k]-2)-(ti2$[k]*ti2$[k])) )
    calc alrt1$[k] = (n-p)*log(( (n-p-ti2$[k]) )/( (n-p-2*ti2$[k]+1) )) \
                 - log(( (n-p)*(2*ti2$[k]-1)-(ti2$[k]*ti2$[k]) )/( (n-p-ti2$[k])*ti2$[k] )) \
                 - (n-p)*(n-p)*(ti2$[k]-1)*(ti2$[k]-1)/( (n-p-2*ti2$[k]+1)*((n-p)*(2*ti2$[k]-1)-(ti2$[k]*ti2$[k])) )
    calc xx=4*(n-p)*ti2$[k]-(ti2$[k]+1)**2
    calc aelrt1$[k] = (n-p) + (n-p)*log(xx/(xx-2*((ti2$[k]*ti2$[k])-1))) \
                     - log((8*(n-p)*ti2$[k]-4*(n-p)-(ti2$[k]+1)**2)/(xx)) \
                     + (xx)*(((4*(n-p)*ti2$[k]*(ti2$[k]-1))/(8*(n-p)*ti2$[k]-4*(n-p)-(ti2$[k]+1)**2))-(n-p))/(xx-2*((ti2$[k]*ti2$[k])-1))
   " one-step LRT updating w after s2 "
    calc elrt2$[k] = (n-p-1)*log((n-p-1)/(n-p-ti2$[k])) - log(ti2$[k])
    calc olrt2$[k]= (n-p)*log((n-p-1)/(n-p-ti2$[k])) - log(( (n-p)*(3*ti2$[k]-2)-ti2$[k] )/( (n-p)*(2*ti2$[k]-1)-ti2$[k] )) \
                  - 2*(n-p)*(n-p)*(ti2$[k]-1)*(ti2$[k]-1)/( (n-p-ti2$[k])*((n-p)*(3*ti2$[k]-2)-ti2$[k]) )
    calc alrt2$[k]=(n-p)*log((n-p-1)/(n-p-ti2$[k])) - log(( (n-p)*(2*ti2$[k]-1)-ti2$[k] )/( (n-p-1)*ti2$[k] ))\
                  -((n-p-1)*(n-p-1)*(ti2$[k])*(ti2$[k])-(n-p-ti2$[k])*((n-p)*(2*ti2$[k]-1)-ti2$[k]))/( (n-p-ti2$[k])*((n-p)*(2*ti2$[k]-1)-ti2$[k]) )
    calc aelrt2$[k]=alrt2$[k]
    "Full-step LRT = one-step expected inf. LRT lrt1e2"
     calc lrt$[k]=(n-p-1)*log((n-p-1)/(n-p-ti2$[k]))-log(ti2$[k])
    "Score test statistics"
    calc escore$[k]=((n-p)*(ti2$[k]-1)**2)/((n-p-1)*(2))
    calc oscore$[k]=((n-p)*(ti2$[k]-1)**2)/(2*((n-p)*(2*ti2$[k]-1)-ti2$[k]**2))
    calc ascore$[k]=((n-p)*(ti2$[k]-1)**2)/((n-p-ti2$[k])*(2*ti2$[k]))
    calc aescore$[k]=(2*(n-p)*(ti2$[k]-1)**2)/(4*ti2$[k]*(n-p)-(ti2$[k]+1)**2)
    vcomp [FIXED=xij1;cadjust=none]aom2
    reml [PRINT=model,components,effects,waldTests,dev; PSE=differences;\
    Rmethod=final;METHOD=ai]y
    vkeep [vest=v11;vare=v21;dev=dev2]
    calc omeghat$[k]=((n-p)*(ti2$[k]-1))/(cii$[k]*(n-p-ti2$[k]))
    calc sig2hat$[k]=((n-p-ti2$[k])*sige1)/(n-3)
    "or
    calc omegahat$[k]=v11$[2]
    calc sig2hat$[k]=v11$[3]"
        calc vwe$[k] = (n-p)*(ti2$[k]-1)/(cii$[k]*(n-p-ti2$[k]) )
         calc vwo$[k] = (n-p)*(ti2$[k]-1)/(cii$[k]*( (n-p)*(2*ti2$[k]-1)-ti2$[k] ) )
         calc vwa$[k] = (n-p)*(ti2$[k]-1)/(cii$[k]*( (n-p-1)*ti2$[k] ) )
         calc vwe1$[k] = (n-p)*(ti2$[k]-1)/(cii$[k]*(n-p-1))
         calc vwo1$[k] = (n-p)*(ti2$[k]-1)/(cii$[k]*((n-p)*(2*ti2$[k]-1)-ti2$[k]*ti2$[k]))
```

```
          calc vwa1$[k] = (n-p)*(ti2$[k]-1)/((cii$[k]*(n-p-ti2$[k])*ti2$[k]  ))
          calc vwa11$[k] = 4*(n-p)*(ti2$[k]-1)/(cii$[k]*(4*(n-p)*ti2$[k]-(ti2$[k]+1)**2))
          calc sgb1$[k] =sige1*((#n-2-ti2$[k])/(#n-p-1))
          calc sgep$[k] =sige1*((#n-2-ti2$[k])/(#n-p-1))
          calc sgop$[k] =sige1*(#n-p-ti2$[k])*(2*ti2$[k]-1)/( ( (#n-p)*(2*ti2$[k]-1)-ti2$[k]*ti2$[k]) )
          calc sgap$[k] = sige1*(#n-p-2*ti2$[k]+1)/( (#n-p-ti2$[k]) )
          calc sga1p$[k] = (sige1*(4*(#n-p)*ti2$[k]-(ti2$[k]+1)**2)-2*(ti2$[k]*ti2$[k]-1)*sige1)/( ( 4*(n-p)*ti2$[k]-(ti2$[k]+1)**2 ) )
       else
         calc lrt$[k]=0
         calc elrt1$[k]=0
         calc olrt1$[k]=0
         calc alrt1$[k]=0
         calc aelrt1$[k]=0
         calc elrt2$[k]=0
         calc olrt2$[k]=0
         calc alrt2$[k]=0
         calc aelrt2$[k]=0
         calc escore$[k]=0
         calc oscore$[k]=0
         calc ascore$[k]=0
         calc aescore$[k]=0
         calc omeghat$[k]=0
         calc sig2hat$[k]=0
     endif
   Endfor
" make results filename "
prin [ch=tn; ip=*; sq=yes] n; dec=0
prin [ch=tp; ip=*; sq=yes] p; dec=0
prin [ch=tsim; ip=*; sq=yes] sim; dec=0
concat [new=file] 'results_',tn,'_',tp,'_',tsim,'.xls'; skip=-1
export[outfile=#file;method=add;sheetname='initial']data=obs,y,xij,xij1,\
ti,ti2,ci,cii,lrt,elrt1,olrt1,alrt1,aelrt1,elrt2,olrt2,alrt2,aelrt2,\
escore,oscore,ascore,aescore,omeghat,sig2hat

"Simulate data from initial model and generate distributions of order statistics for the tests"
matrix[ROWS=#n;COL=sim]LRT,OLRT1,ELRT1,ALRT1,AELRT1,OLRT2,ELRT2,ALRT2,AELRT2,\
            OSCORE,ESCORE,ASCORE,AESCORE
calc start=now(0)
For[NTIMES=sim;INDEX=C1]
calc start1=now(0)
      variate [NVALUES=#n] e
      variate [NVALUES=#n] y1
      calc e =GRNORMAL(n;0;sige1)
      calc y1=fitm+e
Model y1
Terms [FACT=9] xij1
Fit [PRINT=model,summary,estimates; CONSTANT=estimate; FPROB=yes;TPROB=yes; FACT=9] xij1
Rkeep [RMETHOD=simple] Y=y1;Residuals=resm;Fitted=fitm;Leve=levm;Estimates=fixedeff;deviance=dv2;df=df2
calc sige2=dv2/df2
  " Get PXp for gamma=0 "
  symm [r=n] PXp
  calc PXp = ident(n) - X*+inverse((t(X)*+X))*+t(X)
  vari [nval=n] PXpy
  calc PXpy = PXp*+y1
   variate [NVALUES=#n]cii;values=diag(PXp)
   variate [NVALUES=#n]ci;values=(PXpy)
variate [NVALUES=#n]lrt,elrt1,olrt1,alrt1,aelrt1,elrt2,olrt2,alrt2,aelrt2,\
        escore,oscore,ascore,aescore,ti,ti2,omega,sige1a
   variate[;1...#n]obs
   For[NTIMES=#n;INDEX=k]
         calc ti$[k]=ci$[k]/(sqrt(sige2)*sqrt(cii$[k]))
         calc ti2$[k]=ti$[k]**2
```

```
Endfor
variate [NVALUES=#n]lrt,elrt1,olrt1,alrt1,aelrt1,elrt2,olrt2,alrt2,aelrt2,\
        escore,oscore,ascore,aescore,omega,sige1a
For[NTIMES=#n;INDEX=k]
   calc aom2=I2$[*;k]
    If ti2$[k] .gt. 1
    " one-step LRT updating w and s2 simultaneoulsy"
    calc elrt1$[k] = (n-p)*log((n-p-1)/(n-p-ti2$[k])) - log( ((n-p)*ti2$[k]-1)/(n-p-1)) \
     - (n-p)*(ti2$[k]-1)*(ti2$[k]-1)/( (n-p-ti2$[k])*((n-p)*ti2$[k]-1) )
    calc olrt1$[k] = (n-p)*log(( (n-p)*(2*ti2$[k]-1)-(ti2$[k]*ti2$[k]) )/( (n-p-ti2$[k])*(2*ti2$[k]-1) )) \
                 - log(( (n-p)*(3*ti2$[k]-2)-(ti2$[k]*ti2$[k]) )/( (n-p)*(2*ti2$[k]-1)-(ti2$[k]*ti2$[k]) )) \
                 - (n-p)*(n-p)*ti2$[k]*(ti2$[k]-1)*(ti2$[k]-1)/( (n-p-ti2$[k])*(2*ti2$[k]-1)*((n-p)*(3*ti2$[k]-2)-(ti2$[k]*ti2$[k])) )
    calc alrt1$[k] = (n-p)*log(( (n-p-ti2$[k]) )/( (n-p-2*ti2$[k]+1) )) \
                 - log(( (n-p)*(2*ti2$[k]-1)-(ti2$[k]*ti2$[k]) )/( (n-p-ti2$[k])*ti2$[k] )) \
                 - (n-p)*(n-p)*(ti2$[k]-1)*(ti2$[k]-1)/( (n-p-2*ti2$[k]+1)*((n-p)*(2*ti2$[k]-1)-(ti2$[k]*ti2$[k])) )
    calc xx=4*(n-p)*ti2$[k]-(ti2$[k]+1)**2
    calc aelrt1$[k] = (n-p) + (n-p)*log(xx/(xx-2*((ti2$[k]*ti2$[k])-1))) \
                     - log((8*(n-p)*ti2$[k]-4*(n-p)-(ti2$[k]+1)**2)/(xx)) \
                     + (xx)*(((4*(n-p)*ti2$[k]*(ti2$[k]-1))/(8*(n-p)*ti2$[k]-4*(n-p)-(ti2$[k]+1)**2))-(n-p))/(xx-2*((ti2$[k]*ti2$[k])-1))
   " one-step LRT updating w after s2 "
    calc elrt2$[k] = (n-p-1)*log((n-p-1)/(n-p-ti2$[k])) - log(ti2$[k])
    calc olrt2$[k]= (n-p)*log((n-p-1)/(n-p-ti2$[k])) - log(( (n-p)*(3*ti2$[k]-2)-ti2$[k] )/( (n-p)*(2*ti2$[k]-1)-ti2$[k] )) \
                   - 2*(n-p)*(n-p)*(ti2$[k]-1)*(ti2$[k]-1)/( (n-p-ti2$[k])*((n-p)*(3*ti2$[k]-2)-ti2$[k]) )
    calc alrt2$[k]=(n-p)*log((n-p-1)/(n-p-ti2$[k])) - log(( (n-p)*(2*ti2$[k]-1)-ti2$[k] )/( (n-p-1)*ti2$[k] ))\
                   -((n-p-1)*(n-p-1)*(ti2$[k])*(ti2$[k])-(n-p-ti2$[k])*((n-p)*(2*ti2$[k]-1)-ti2$[k]))/( (n-p-ti2$[k])*((n-p)*(2*ti2$[k]-1)-ti2$[k]) )
    calc aelrt2$[k]=alrt2$[k]
    "Full-step LRT = one-step expected inf. LRT lrt1e2"
     calc lrt$[k]=(n-p-1)*log((n-p-1)/(n-p-ti2$[k]))-log(ti2$[k])
    "Score test statistics"
    calc escore$[k]=((n-p)*(ti2$[k]-1)**2)/((n-p-1)*(2))
    calc oscore$[k]=((n-p)*(ti2$[k]-1)**2)/(2*((n-p)*(2*ti2$[k]-1)-ti2$[k]**2))
    calc ascore$[k]=((n-p)*(ti2$[k]-1)**2)/((n-p-ti2$[k])*(2*ti2$[k]))
    calc aescore$[k]=(2*(n-p)*(ti2$[k]-1)**2)/(4*ti2$[k]*(n-p)-(ti2$[k]+1)**2)
      else
        calc lrt$[k]=0
        calc elrt1$[k]=0
        calc olrt1$[k]=0
        calc alrt1$[k]=0
        calc aelrt1$[k]=0
        calc elrt2$[k]=0
        calc olrt2$[k]=0
        calc alrt2$[k]=0
        calc aelrt2$[k]=0
        calc escore$[k]=0
        calc oscore$[k]=0
        calc ascore$[k]=0
        calc aescore$[k]=0
    endif
  Endfor
calc LRT$[*;C1]=lrt
calc ELRT1$[*;C1]=elrt1
calc OLRT1$[*;C1]=olrt1
calc ALRT1$[*;C1]=alrt1
calc AELRT1$[*;C1]=aelrt1
calc ELRT2$[*;C1]=elrt2
calc OLRT2$[*;C1]=olrt2
calc ALRT2$[*;C1]=alrt2
calc AELRT2$[*;C1]=aelrt2
calc ESCORE$[*;C1]=escore
calc OSCORE$[*;C1]=oscore
calc ASCORE$[*;C1]=ascore
calc AESCORE$[*;C1]=aescore
```

```
Endfor
calc end=now(0)
calc elapsed=end-start
print start,end,elapsed; drep=38
"Output data for generating distributions of order statistics for the tests"
" export results "
export[outfile=#file;method=add;sheetname='lrt']data=T(LRT$[*;*])
export[outfile=#file;method=add;sheetname='elrt1']data=T(ELRT1$[*;*])
export[outfile=#file;method=add;sheetname='olrt1']data=T(OLRT1$[*;*])
export[outfile=#file;method=add;sheetname='alrt1']data=T(ALRT1$[*;*])
export[outfile=#file;method=add;sheetname='aelrt1']data=T(AELRT1$[*;*])
export[outfile=#file;method=add;sheetname='elrt2']data=T(ELRT2$[*;*])
export[outfile=#file;method=add;sheetname='olrt2']data=T(OLRT2$[*;*])
export[outfile=#file;method=add;sheetname='alrt2']data=T(ALRT2$[*;*])
export[outfile=#file;method=add;sheetname='escore']data=T(ESCORE$[*;*])
export[outfile=#file;method=add;sheetname='oscore']data=T(OSCORE$[*;*])
export[outfile=#file;method=add;sheetname='ascore']data=T(ASCORE$[*;*])
export[outfile=#file;method=add;sheetname='aescore']data=T(AESCORE$[*;*])
```

C.2 Additional power graphics

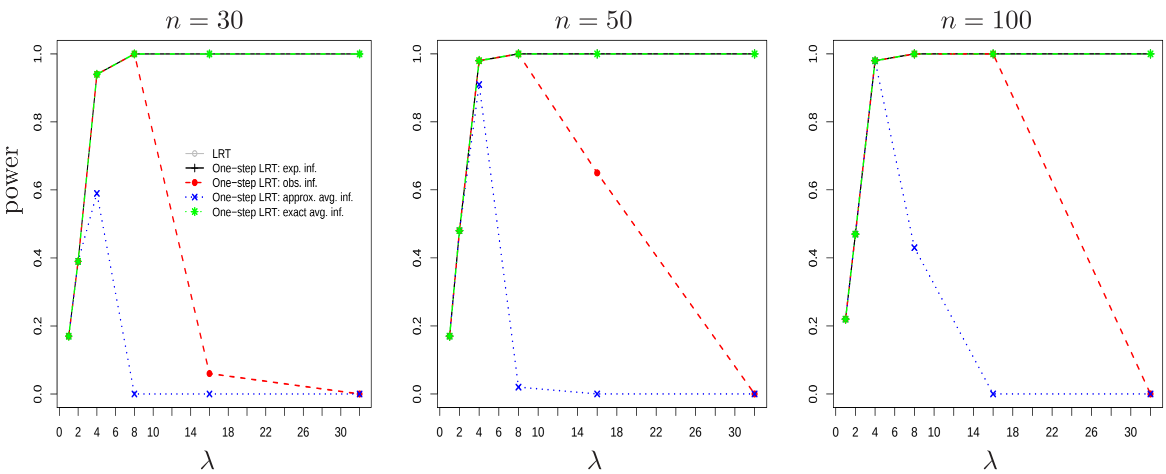

Figure C.1: *Empirical power of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets for samples sizes* $n = 30, 50, 100$. *One-step likelihood ratio test statistics are based on updating scheme A.*

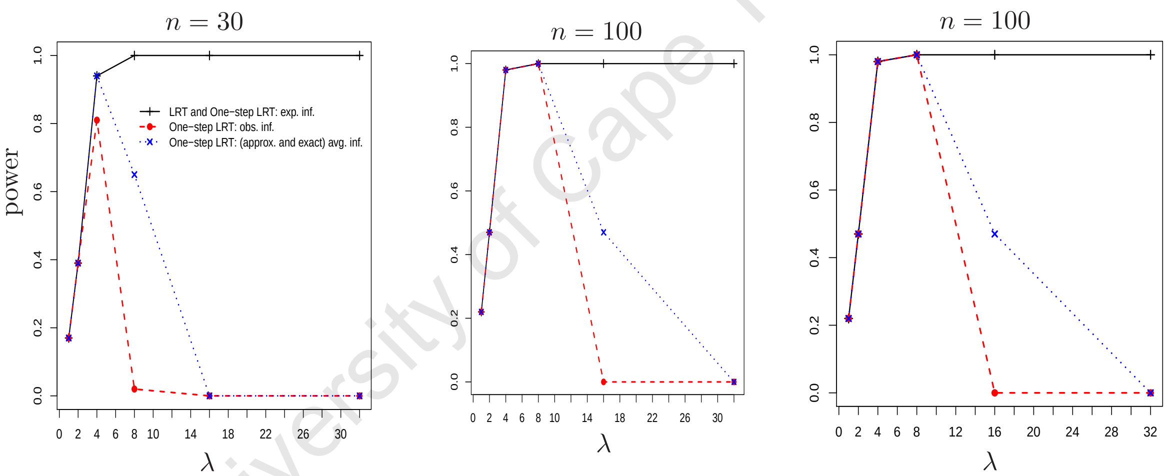

Figure C.2: *Empirical power of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets for samples sizes* $n = 30, 50, 100$. *One-step likelihood ratio test statistics are based on updating scheme B.*

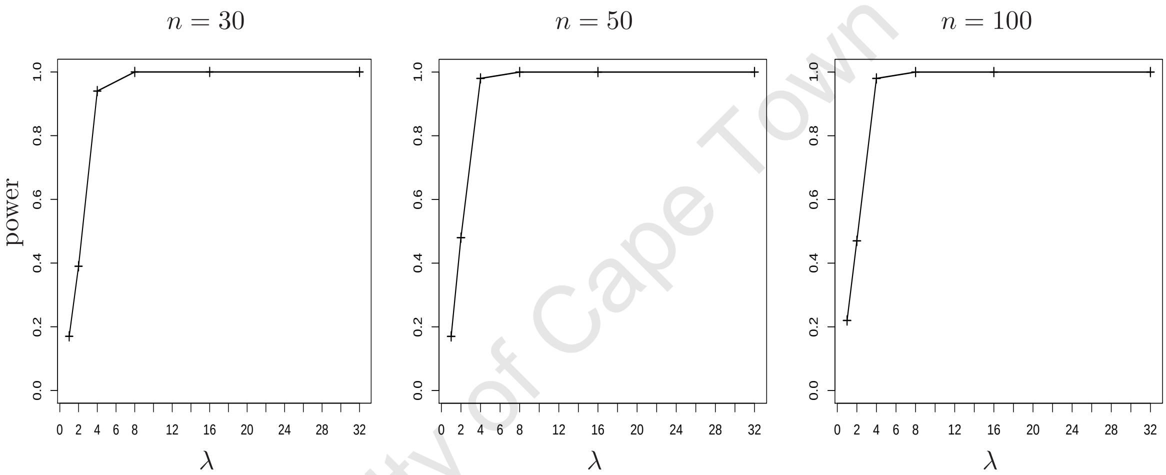

Figure C.3: *Empirical power of the score test statistic based on expected information matrix for the first unit based on 2500 simulations of 500 data sets for samples sizes* $n = 30, 50, 100$.

APPENDIX D

Chapter 6 examples

D.1 **GenStat** code for simulation study for power and type I errors

```
" Runs sets of simulations to evaluate type I error for LRTs and score tests for AOM for unit 1
 Factors: trt + CR layout - one-way random effects ANOVA
"
%CD ’C:/research/oneway_random’
"
  Parameters: n = number of units
              r = number of reps in each group
              p = number of groups
              g = gamma for groups
              nrep = number of data sets to simulate (usually 500)
              nit = number of simulations for each data set (usually 2500)
              out= outlier size (=0 fo type I errors simulations)"
read [ch=2] n,r,p,g,nrep,nit,out
prin  n,r,p,g,nrep,nit; dec=0,0,0,*,0,0; fie=6
" run across specified n, r, p, g and out combinations "
" get filenames for incremental output files "
print [ch=ttime; ip=*; sq=yes] now(0); drep=38
prin ttime
print [ch=out1; ip=*; sq=yes] ’f_’,ttime,’_1.out’; fie=0; skip=0; just=left
print [ch=out2; ip=*; sq=yes] ’f_’,ttime,’_2.out’; fie=0; skip=0; just=left
edit  [ch=!T(’(R/:/_/)2’)] out1,out2
print out1,out2
open #out1,#out2; filetype=output; channel=2,3; width=312
" header line for identification "
prin [ch=2; ip=*; sq=yes] n,r,p,g,nrep,nit,out; dec=0,0,0,*,0,0,0; fie=6
prin [ch=3; ip=*; sq=yes] n,r,p,g,nrep,nit,out; dec=0,0,0,*,0,0,0; fie=6
calc start=now(0)
" outer for loop to keep code together "
for
" set up s2, mu "
scal s2; val=1
scal mu; val=0
" various data structures "
symm [r=p] ZtPZ
symm [r=2] I11,Io11,Ia11,Iae11
matr [r=2; c=1] I12,Io12,Ia12,Iae12,i12,io12,ia12,iae12
scal I22,Io22,Ia22,Iae22,i22,ie22,io22,ia22,iae22
scal ogam,os2,ow
scal II11,II12,II22
matr [r=1; c=p] tdiPZ
" save structures for percentiles for each test "
pointer [nval=4] pc_lrt,pc_plrt,pc_score,pc_pscore
pointer [nval=4] pc_olrt,pc_oolrt,pc_oalrt,pc_oaelrt
```

```
pointer [nval=4] pc_oscore,pc_ascore,pc_aescore
pointer [nval=4] ind_lrt,ind_plrt,ind_score,ind_pscore
pointer [nval=4] ind_olrt,ind_oolrt,ind_oalrt,ind_oaelrt
pointer [nval=4] ind_oscore,ind_ascore,ind_aescore
variate [nval=nrep] pc_lrt[],pc_plrt[],pc_score[],pc_pscore[]
variate [nval=nrep] pc_olrt[],pc_oolrt[],pc_oalrt[],pc_oaelrt[]
variate [nval=nrep] pc_oscore[],pc_ascore[],pc_aescore[]
variate [nval=nrep] ind_lrt[],ind_plrt[],ind_score[],ind_pscore[]
variate [nval=nrep] ind_olrt[],ind_oolrt[],ind_oalrt[],ind_oaelrt[]
variate [nval=nrep] ind_oscore[],ind_ascore[],ind_aescore[]
" save structures for tests "
variate [nval=nrep] vlrt,vplrt,vscore,vpscore,volrt,voolrt,voalrt,voaelrt,voscore,vascore,vaescore
" cycle over nrep data sets "
for [ntimes=nrep; index=jj]
" where are we? "
  skip [file=out] 1
  prin [ip=*; sq=yes] '**** nrep =',jj; fie=0,6; j=left; skip=0,1; dec=0
" Step 1: generate design "
  factor [lev=p; val=(1...#p)#r] trt
  factor [lev=n; val=1...#n] unit
" Completely randomized design "
  randomize trt
" Generate data "
  variate [nval=n] y,e
  variate [nval=p] etrt
  calc etrt = grnormal(p;0;s2*g)
  calc e = grnormal(n;0;s2)
  calc y = mu + etrt$[trt] + e
  calc y$[1] = y$[1]+out
" Fit model "
  vcomp trt; con=pos
  reml [p=mon; meth=ai] y
" Keep results of analysis "
  vkeep [sigma2=es2; res=Py; dev=dev_0] 'Constant'+trt; eff=tc,tt; comp=*,vct
  scal ec; val=tc
  vari [nval=p] et; val=tt
  scal eg; val=vct/es2
" Vector of initial values for one-step updates: gamma, s2 "
  vari i1; val=!(eg,es2)
" Get design matrices "
  matrix [r=n; c=p] Z
  calc Z$[*;1...p]=trt.eq.1...p
  matrix [r=n; c=1; val=#n(1)] X
" Get H & P "
  symm [r=n] H,iH,P
  calc H = eg*Z*+T(Z) + ident(n)
  calc iH = inv(H)
  symm [r=1] XtiHX,iXtiHX
  calc XtiHX = T(X)*+iH*+X
  calc iXtiHX = inv(XtiHX)
  matr [r=n; c=1] iHX
  calc iHX = iH*+X
  calc P = iH - (iHX*+iXtiHX)*+T(iHX)
" Get PXp for gamma=0 "
  symm [r=n] PXp
  calc PXp = ident(n) - mat1(n;n)/n
  vari [nval=n] PXpy
  calc PXpy = PXp*+y
" Get information matrices "
  calc ZtPZ=(T(Z)*+P)*+Z
  calc ZtPy = T(Z)*+Py
  calc TrZtPZ = trace( ZtPZ )
```

```
  calc TrZtPZZtPZ = trace( ZtPZ*+ZtPZ )
" initialise to zero "
  calc I11,Io11,Ia11,Iae11=0
" expected information matrix "
  if eg.gt.1e-5
    calc I11$[1;1]=TrZtPZZtPZ/2
    calc I11$[2;1]=TrZtPZ/(2*es2)
  endif
  calc I11$[2;2]=(n-1)/(2*es2*es2)
" observed information matrix "
  if eg.gt.1e-5
    calc Io11$[1;1]=T(ZtPy)*+(ZtPZ*+ZtPy)/es2 - TrZtPZZtPZ/2
    calc Io11$[2;1]=T(ZtPy)*+ZtPy/(2*es2*es2)
  endif
  calc Io11$[2;2]=t(y)*+Py/(es2**3)-(n-1)/(2*es2*es2)
" approx average information matrix "
  if eg.gt.1e-5
    calc Ia11$[1;1]=T(ZtPy)*+(ZtPZ*+ZtPy)/(2*es2)
    calc Ia11$[2;1]=T(ZtPy)*+ZtPy/(2*es2*es2)
  endif
  calc Ia11$[2;2]=t(y)*+Py/(2*es2**3)
" exact average information matrix "
  calc Iae11=0.5*(I11+Io11)
" initialise I12 matrices "
  calc I12,Io12,Ia12,Iae12=0
" Do AOM LRT and score test for each unit "
  vari [nval=n] di
  scal lrt,plrt,score,pscore,tsqrd,olrt,oolrt,oalrt,oaelrt,oscore,ascore,aescore
  scal ti2,ci,cii,Uw,i22
  for [ntimes=1; index=i]
    calc ci=Py$[i]
    calc cii = P$[i;i]
    calc ti2=(ci*ci)/(es2*cii)
    calc tsqrd=ti2
    if ti2.gt.1
      calc di=unit.eq.i
      calc Uw=cii*(ti2-1)/2
      calc tdiPZ=(T(di)*+P)*+Z
" a) Full LRT "
      vcomp [cad=no] random=di+trt; con=pos
      reml [p=*; rmeth=all] y
      vkeep [dev=dev_aom] di; comp=vc
      calc lrt=dev_0-dev_aom
" b) Partial LRT "
      calc plrt=(n-2)*log((n-2)/(n-1-ti2))-log(ti2)
" c) Full score test "
      if eg.gt.1e-5: calc I12$[1;1]=tdiPZ*+T(tdiPZ)/2: endif
      calc I12$[2;1]=cii/(2*es2)
      calc I22=(cii*cii)/2
      calc ie22 = 1/(I22-T(I12)*+inv(I11)*+I12)
      calc score=Uw*ie22*Uw
" d) Partial score test "
      calc pscore=(n-1)*(ti2-1)*(ti2-1)/(2*(n-2))
" e) one-step LRT expected "
      calc o1 = i1-ie22*inv(I11)*+I12*Uw
      calc ow = ie22*Uw
      calc ogam,os2 = o1$[1,2;1]
      " if w or s2 < 0, update has failed "
      if ow.lt.0 .or. os2.lt.0
        calc olrt=c(’*’)
      " if gamma<0, fix at zero and recalculate update "
      elsif ogam.le.0
```

```
        calc aci=PXpy$[i]
        calc acii = PXp$[i;i]
        calc ati2=(aci*aci)/(es2*acii)
        calc aUw=acii*(ati2-1)/2
        calc II11=(n-1)/(2*es2*es2)
        calc II12=(acii)/(2*es2)
        calc II22=(acii*acii)/2
        calc i22 = 1/(II22-II12*II12/II11)
        calc os2 = es2-i22*II12*Uw/II11
        calc ow = i22*Uw
        if ow.lt.0 .or. os2.lt.0
          calc olrt=c('*')
        else
          vcomp [cad=no] random=di; con=fix; init=ow,1*os2
          reml [p=*] y
          vkeep [dev=dev_aom]
          calc olrt=dev_0-dev_aom
        endif
      else
        vcomp [cad=no] random=di+trt; con=fix; init=ow,ogam,1*os2
        reml [p=*] y
        vkeep [dev=dev_aom]
        calc olrt=dev_0-dev_aom
      endif
" f) one-step LRT observed "
      if eg.gt.1e-5: calc Io12$[1;1]=ci*tdiPZ*+ZtPy/es2-tdiPZ*+T(tdiPZ)/2: endif
      calc Io12$[2;1]=ci*ci/(2*es2*es2)
      calc Io22=(cii*cii)*(2*ti2-1)/2
      calc io22 = 1/(Io22-T(Io12)*+inv(Io11)*+Io12)
      calc io12 = -io22*inv(Io11)*+Io12
      calc o1 = i1+io12*Uw
      calc ow = io22*Uw
     calc ogam,os2 = o1$[1,2;1]
      " if w or s2 < 0, update has failed "
      if ow.lt.0 .or. os2.lt.0
        calc oolrt=c('*')
      " if gamma<0, fix at zero and recalculate update "
      elsif ogam.le.0
        calc aci=PXpy$[i]
        calc acii = PXp$[i;i]
        calc ati2=(aci*aci)/(es2*acii)
        calc aUw=acii*(ati2-1)/2
        calc II11=t(y)*+PXpy/(es2**3)-(n-1)/(2*es2*es2)
        calc II12=aci*aci/(2*es2*es2)
        calc II22=(acii*acii)*(2*ati2-1)/2
        calc i22 = 1/(II22-II12*II12/II11)
        calc os2 = es2-i22*II12*Uw/II11
        calc ow = i22*Uw
        if ow.lt.0 .or. os2.lt.0
          calc oolrt=c('*')
        else
          vcomp [cad=no] random=di; con=fix; init=ow,1*os2
          reml [p=*] y
          vkeep [dev=dev_aom]
          calc oolrt=dev_0-dev_aom
        endif
      else
        vcomp [cad=no] random=di+trt; con=fix; init=ow,ogam,1*os2
        reml [p=*] y
        vkeep [dev=dev_aom]
        calc oolrt=dev_0-dev_aom
      endif
```

```
" g) one-step LRT approx average "
      if eg.gt.1e-5: calc Ia12$[1;1]=ci*tdiPZ*+ZtPy/(2*es2): endif
      calc Ia12$[2;1]=ci*ci/(2*es2*es2)
      calc Ia22=(cii*cii)*ti2/2
      calc ia22 = 1/(Ia22-T(Ia12)*+inv(Ia11)*+Ia12)
      calc ia12 = -ia22*inv(Ia11)*+Ia12
      calc o1 = i1+ia12*Uw
      calc ow = ia22*Uw
      calc ogam,os2 = o1$[1,2;1]
      " if w or s2 < 0, update has failed "
      if ow.lt.0 .or. os2.lt.0
        calc oalrt=c('*')
      " if gamma<0, fix at zero and recalculate update "
      elsif ogam.le.0
        calc aci=PXpy$[i]
        calc acii = PXp$[i;i]
        calc ati2=(aci*aci)/(es2*acii)
        calc aUw=acii*(ati2-1)/2
        calc II11=t(y)*+PXpy/(2*es2**3)
        calc II12=aci*aci/(2*es2*es2)
        calc II22=(acii*acii)*ati2/2
        calc i22 = 1/(II22-II12*II12/II11)
        calc os2 = es2-i22*II12*Uw/II11
        calc ow = i22*Uw
        if ow.lt.0 .or. os2.lt.0
          calc oalrt=c('*')
        else
          vcomp [cad=no] random=di; con=fix; init=ow,1*os2
          reml [p=*] y
          vkeep [dev=dev_aom]
          calc oalrt=dev_0-dev_aom
        endif
      else
        vcomp [cad=no] random=di+trt; con=fix; init=ow,ogam,1*os2
        reml [p=*] y
        vkeep [dev=dev_aom]
        calc oalrt=dev_0-dev_aom
      endif
" h) one-step LRT exact average "
      calc Iae12=0.5*(I12+Io12)
      calc Iae22=0.5*(I22+Io22)
      calc iae22 = 1/(Iae22-T(Iae12)*+inv(Iae11)*+Iae12)
      calc iae12 = -iae22*inv(Iae11)*+Iae12
      calc o1 = i1+iae12*Uw
      calc ow = iae22*Uw
      calc ogam,os2 = o1$[1,2;1]
      " if w or s2 < 0, update has failed "
      if ow.lt.0 .or. os2.lt.0
        calc olrt=c('*')
      " if gamma<0, fix at zero and recalculate update "
      elsif ogam.le.0
        calc aci=PXpy$[i]
        calc acii = PXp$[i;i]
        calc ati2=(aci*aci)/(es2*acii)
        calc aUw=acii*(ati2-1)/2
        calc II11=t(y)*+PXpy/(2*es2**3)
        calc II12=(acii)/(4*es2)+aci*aci/(4*es2*es2)
        calc II22=(acii*acii)/4+(acii*acii)*(2*ati2-1)/4
        calc i22 = 1/(II22-II12*II12/II11)
        calc os2 = es2-i22*II12*Uw/II11
        calc ow = i22*Uw
        if ow.lt.0 .or. os2.lt.0
```

```
          calc oaelrt=c('*')
        else
          vcomp [cad=no] random=di; con=fix; init=ow,1*os2
          reml [p=*] y
          vkeep [dev=dev_aom]
          calc oaelrt=dev_0-dev_aom
        endif
      else
        vcomp [cad=no] random=di+trt; con=fix; init=ow,ogam,1*os2
        reml [p=*] y
        vkeep [dev=dev_aom]
        calc oaelrt=dev_0-dev_aom
      endif
" i) score test using observed information "
      calc oscore=Uw*io22*Uw
" j) score test using observed information "
      calc ascore=Uw*ia22*Uw
" k) score test using observed information "
      calc aescore=Uw*iae22*Uw
    else
      calc lrt=0
      calc plrt=0
      calc score=0
      calc pscore=0
      calc olrt=0
      calc oolrt=0
      calc oalrt=0
      calc oaelrt=0
      calc oscore=0
      calc ascore=0
      calc aescore=0
    endif
  endfor
  prin [ch=2; ip=*; sq=yes] jj,ti2,lrt,plrt,score,pscore,olrt,oolrt,oalrt,oaelrt,oscore,ascore,aescore
  calc (vlrt,vplrt,vscore,vpscore,volrt,voolrt,voalrt,voaelrt,voscore,vascore,vaescore)$[jj]=lrt,plrt,score,pscore,olrt,oolrt,oalrt,\
oaelrt,oscore,ascore,aescore
" Do simulations to get empirical distribution of order statistics "
" Set up save structures  "
  vari [nval=nit] _lrt,_plrt,_score,_pscore,_olrt,_oolrt,_oalrt,_oaelrt,_oscore,_ascore,_aescore
  scal _tsqrd
  vari [nval=n] _y,_e,_Py,_PXpy
  vari [nval=p] _etrt
"
  flrv P; lrv=lP
  matrix [r=n; c=n] L
  calc L = lP[1]*+sqrt(lP[2])
  vari [nv=n] diagP; val=diag(P)
"
" Generate and analyse nit simulated data sets "
  for [ntimes=nit; index=kk]

" Where are we? every 100 iterations "
  if kk.eq.500*int(kk/500)
    calc now=now(0)
    prin [sq=yes; ip=*] '***** nit =',kk,now; fie=0,6,6; dec=0,0,*; skip=0,1,1; \
      just=left; drep=0,0,34
  endif
" Generate full data based on original fit "
    calc _etrt = grnormal(p;0;es2*eg)
    calc _e = grnormal(n;0;es2)
    calc _y = ec + _etrt$[trt] + _e
    calc _PXpy = PXp*+_y
```

```
" Fit model to simulated data "
    vcomp trt; con=pos
    reml [p=*; meth=ai] _y
" Keep results of analysis "
    vkeep [sigma2=_es2; res=_Py; dev=dev_0] trt; comp=vct
    scal _eg; val=vct/_es2
" Vector of initial values: gamma, s2 "
    vari i1; val=!(_eg,_es2)
" Get new H & P "
    calc H = _eg*Z*+T(Z) + ident(n)
    calc iH = inv(H)
    calc XtiHX = T(X)*+iH*+X
    calc iXtiHX = inv(XtiHX)
    calc iHX = iH*+X
    symm [r=n] _P
    calc _P = iH - (iHX*+iXtiHX)*+T(iHX)
" Get information matrices "
    calc ZtPZ=(T(Z)*+_P)*+Z
    calc ZtPy = T(Z)*+_Py
    calc TrZtPZ = trace( ZtPZ )
    calc TrZtPZZtPZ = trace( ZtPZ*+ZtPZ )
" initialise to zero "
    calc I11,Io11,Ia11,Iae11=0
" expected information matrix "
    if _eg.gt.1e-5
      calc I11$[1;1]=TrZtPZZtPZ/2
      calc I11$[2;1]=TrZtPZ/(2*_es2)
    endif
    calc I11$[2;2]=(n-1)/(2*_es2*_es2)
" observed information matrix "
    if _eg.gt.1e-5
      calc Io11$[1;1]=T(ZtPy)*+(ZtPZ*+ZtPy)/_es2 - TrZtPZZtPZ/2
      calc Io11$[2;1]=T(ZtPy)*+ZtPy/(2*_es2*_es2)
    endif
    calc Io11$[2;2]=t(_y)*+_Py/(_es2**3)-(n-1)/(2*_es2*_es2)
" approx average information matrix "
    if _eg.gt.1e-5
      calc Ia11$[1;1]=T(ZtPy)*+(ZtPZ*+ZtPy)/(2*_es2)
      calc Ia11$[2;1]=T(ZtPy)*+ZtPy/(2*_es2*_es2)
    endif
    calc Ia11$[2;2]=t(_y)*+_Py/(2*_es2**3)
" exact average information matrix "
    calc Iae11=0.5*(I11+Io11)
" initialise I12 matrices "
    calc I12,Io12,Ia12,Iae12=0
" Do AOM LRT and score test for each unit "
    for [ntimes=1; index=i]
      calc ci=_Py$[i]
      calc cii = _P$[i;i]
      calc ti2=(ci*ci)/(_es2*cii)
      calc _tsqrd=ti2
      if ti2.gt.1
        calc di=unit.eq.i
        calc Uw=cii*(ti2-1)/2
        calc tdiPZ=(T(di)*+_P)*+Z
" a) Full LRT "
        vcomp [cad=no] random=di+trt; con=pos
        reml [p=*; rmeth=all] _y
        vkeep [dev=dev_aom] di; comp=vc
        calc _lrt$[kk]=dev_0-dev_aom
" b) Partial LRT "
        calc _plrt$[kk]=(n-2)*log((n-2)/(n-1-ti2))-log(ti2)
```

```
" c) Full score test "
        if (_eg.gt.1e-5): calc I12$[1;1]=tdiPZ*+T(tdiPZ)/2: endif
        calc I12$[2;1]=cii/(2*_es2)
        calc I22=(cii*cii)/2
        calc ie22 = 1/(I22-T(I12)*+inv(I11)*+I12)
        calc _score$[kk]=Uw*ie22*Uw
" d) Partial score test "
        calc _pscore$[kk]=(n-1)*(ti2-1)*(ti2-1)/(2*(n-2))
" e) one-step LRT expected "
        calc o1 = i1-ie22*inv(I11)*+I12*Uw
        calc ow = ie22*Uw
        calc ogam,os2 = o1$[1,2;1]
      " if w or s2 < 0, update has failed "
        if ow.lt.0 .or. os2.lt.0
          calc _olrt$[kk]=c('*')
      " if gamma<0, fix at zero and recalculate update "
        elsif ogam.le.0
          calc aci=_PXpy$[i]
          calc acii = PXp$[i;i]
          calc ati2=(aci*aci)/(_es2*acii)
          calc aUw=acii*(ati2-1)/2
          calc II11=(n-1)/(2*_es2*_es2)
          calc II12=(acii)/(2*_es2)
          calc II22=(acii*acii)/2
          calc i22 = 1/(II22-II12*II12/II11)
          calc os2 = _es2-i22*II12*Uw/II11
          calc ow = i22*Uw
          if ow.lt.0 .or. os2.lt.0
            calc _olrt$[kk]=c('*')
          else
            vcomp [cad=no] random=di; con=fix; init=ow,1*os2
            reml [p=*] _y
            vkeep [dev=dev_aom]
            calc _olrt$[kk]=dev_0-dev_aom
          endif
        else
          vcomp [cad=no] random=di+trt; con=fix; init=ow,ogam,1*os2
          reml [p=*] _y
          vkeep [dev=dev_aom]
          calc _olrt$[kk]=dev_0-dev_aom
        endif
" f) one-step LRT observed "
        if (_eg.gt.1e-5): calc Io12$[1;1]=ci*tdiPZ*+ZtPy/_es2-tdiPZ*+T(tdiPZ)/2: endif
        calc Io12$[2;1]=ci*ci/(2*_es2*_es2)
        calc Io22=(cii*cii)*(2*ti2-1)/2
        calc io22 = 1/(Io22-T(Io12)*+ginv(Io11)*+Io12)
        calc io12 = -io22*ginv(Io11)*+Io12
        calc o1 = i1+io12*Uw
        calc ow = io22*Uw
        calc ogam,os2 = o1$[1,2;1]
      " if w or s2 < 0, update has failed "
        if ow.lt.0 .or. os2.lt.0
          calc _oolrt$[kk]=c('*')
      " if gamma<0, fix at zero and recalculate update "
        elsif ogam.le.0
          calc aci=_PXpy$[i]
          calc acii = PXp$[i;i]
          calc ati2=(aci*aci)/(_es2*acii)
          calc aUw=acii*(ati2-1)/2
          calc II11=t(_y)*+_PXpy/(_es2**3)-(n-1)/(2*_es2*_es2)
          calc II12=aci*aci/(2*_es2*_es2)
          calc II22=(acii*acii)*(2*ati2-1)/2
```

```
          calc i22 = 1/(II22-II12*II12/II11)
          calc os2 = _es2-i22*II12*Uw/II11
          calc ow = i22*Uw
          if ow.lt.0 .or. os2.lt.0
            calc _oolrt$[kk]=c('*')
          else
            vcomp [cad=no] random=di; con=fix; init=ow,1*os2
            reml [p=*] _y
            vkeep [dev=dev_aom]
            calc _oolrt$[kk]=dev_0-dev_aom
          endif
        else
          vcomp [cad=no] random=di+trt; con=fix; init=ow,ogam,1*os2
          reml [p=*] _y
          vkeep [dev=dev_aom]
          calc _oolrt$[kk]=dev_0-dev_aom
        endif
" g) one-step LRT approx average "
        if (_eg.gt.1e-5): calc Ia12$[1;1]=ci*tdiPZ*+ZtPy/(2*_es2): endif
        calc Ia12$[2;1]=ci*ci/(2*_es2*_es2)
        calc Ia22=(cii*cii)*ti2/2
        calc ia22 = 1/(Ia22-T(Ia12)*+inv(Ia11)*+Ia12)
        calc ia12 = -ia22*inv(Ia11)*+Ia12
        calc o1 = i1+ia12*Uw
        calc ow = ia22*Uw
        calc ogam,os2 = o1$[1,2;1]
      " if w or s2 < 0, update has failed "
        if ow.lt.0 .or. os2.lt.0
          calc _oalrt$[kk]=c('*')
      " if gamma<0, fix at zero and recalculate update "
        elsif ogam.le.0
          calc aci=_PXpy$[i]
          calc acii = PXp$[i;i]
          calc ati2=(aci*aci)/(_es2*acii)
          calc aUw=acii*(ati2-1)/2
          calc II11=t(_y)*+_PXpy/(2*_es2**3)
          calc II12=aci*aci/(2*_es2*_es2)
          calc II22=(acii*acii)*ati2/2
          calc i22 = 1/(II22-II12*II12/II11)
          calc os2 = _es2-i22*II12*Uw/II11
          calc ow = i22*Uw
          if ow.lt.0 .or. os2.lt.0
            calc _oalrt$[kk]=c('*')
          else
            vcomp [cad=no] random=di; con=fix; init=ow,1*os2
            reml [p=*] _y
            vkeep [dev=dev_aom]
            calc _oalrt$[kk]=dev_0-dev_aom
          endif
        else
          vcomp [cad=no] random=di+trt; con=fix; init=ow,ogam,1*os2
          reml [p=*] _y
          vkeep [dev=dev_aom]
          calc _oalrt$[kk]=dev_0-dev_aom
        endif
" h) one-step LRT exact average "
        calc Iae12=0.5*(I12+Io12)
        calc Iae22=0.5*(I22+Io22)
        calc iae22 = 1/(Iae22-T(Iae12)*+inv(Iae11)*+Iae12)
        calc iae12 = -iae22*inv(Iae11)*+Iae12
        calc o1 = i1+iae12*Uw
        calc ow = iae22*Uw
```

```
          calc ogam,os2 = o1$[1,2;1]
        " if w or s2 < 0, update has failed "
          if ow.lt.0 .or. os2.lt.0
            calc _oaelrt$[kk]=c('*')
        " if gamma<0, fix at zero and recalculate update "
          elsif ogam.le.0
            calc aci=_PXpy$[i]
            calc acii = PXp$[i;i]
            calc ati2=(aci*aci)/(_es2*acii)
            calc aUw=acii*(ati2-1)/2
            calc II11=t(_y)*+_PXpy/(2*_es2**3)
            calc II12=(acii)/(4*_es2)+aci*aci/(4*_es2*_es2)
            calc II22=(acii*acii)/4+(acii*acii)*(2*ati2-1)/4
            calc i22 = 1/(II22-II12*II12/II11)
            calc os2 = _es2-i22*II12*Uw/II11
            calc ow = i22*Uw
            if ow.lt.0 .or. os2.lt.0
              calc _oaelrt$[kk]=c('*')
            else
              vcomp [cad=no] random=di; con=fix; init=ow,1*os2
              reml [p=*] _y
              vkeep [dev=dev_aom]
              calc _oaelrt$[kk]=dev_0-dev_aom
            endif
          else
            vcomp [cad=no] random=di+trt; con=fix; init=ow,ogam,1*os2
            reml [p=*] _y
            vkeep [dev=dev_aom]
            calc _oaelrt$[kk]=dev_0-dev_aom
          endif
  " i) score test using observed information "
        calc _oscore$[kk]=Uw*io22*Uw
  " j) score test using observed information "
        calc _ascore$[kk]=Uw*ia22*Uw
  " k) score test using observed information "
        calc _aescore$[kk]=Uw*iae22*Uw
        else
          calc _lrt$[kk]=0
          calc _plrt$[kk]=0
          calc _score$[kk]=0
          calc _pscore$[kk]=0
          calc _olrt$[kk]=0
          calc _oolrt$[kk]=0
          calc _oalrt$[kk]=0
          calc _oaelrt$[kk]=0
          calc _oscore$[kk]=0
          calc _ascore$[kk]=0
          calc _aescore$[kk]=0
        endif
      endfor

  " end of kkth simulation data set "
    endfor
  " are the aoms larger than the percentiles? "
    calc pc_lrt[]$[jj] = percentile(_lrt;90,95,97.5,99)
    calc pc_plrt[]$[jj] = percentile(_plrt;90,95,97.5,99)
    calc pc_score[]$[jj] = percentile(_score;90,95,97.5,99)
    calc pc_pscore[]$[jj] = percentile(_pscore;90,95,97.5,99)
    calc pc_olrt[]$[jj] = percentile(_olrt;90,95,97.5,99)
    calc pc_oolrt[]$[jj] = percentile(_oolrt;90,95,97.5,99)
    calc pc_oalrt[]$[jj] = percentile(_oalrt;90,95,97.5,99)
    calc pc_oaelrt[]$[jj] = percentile(_oaelrt;90,95,97.5,99)
```

```
  calc pc_oscore[]$[jj] = percentile(_oscore;90,95,97.5,99)
  calc pc_ascore[]$[jj] = percentile(_ascore;90,95,97.5,99)
  calc pc_aescore[]$[jj] = percentile(_aescore;90,95,97.5,99)
  calc ind_lrt[]$[jj] = (lrt)4.gt.percentile(_lrt;90,95,97.5,99)
  calc ind_plrt[]$[jj] = (plrt)4.gt.percentile(_plrt;90,95,97.5,99)
  calc ind_score[]$[jj] = (score)4.gt.percentile(_score;90,95,97.5,99)
  calc ind_pscore[]$[jj] = (pscore)4.gt.percentile(_pscore;90,95,97.5,99)
  calc ind_olrt[]$[jj] = (olrt)4.gt.percentile(_olrt;90,95,97.5,99)
  calc ind_oolrt[]$[jj] = (oolrt)4.gt.percentile(_oolrt;90,95,97.5,99)
  calc ind_oalrt[]$[jj] = (oalrt)4.gt.percentile(_oalrt;90,95,97.5,99)
  calc ind_oaelrt[]$[jj] = (oaelrt)4.gt.percentile(_oaelrt;90,95,97.5,99)
  calc ind_oscore[]$[jj] = (oscore)4.gt.percentile(_oscore;90,95,97.5,99)
  calc ind_ascore[]$[jj] = (ascore)4.gt.percentile(_ascore;90,95,97.5,99)
  calc ind_aescore[]$[jj] = (aescore)4.gt.percentile(_aescore;90,95,97.5,99)
prin [ch=3; ip=*; sq=yes] jj; dec=0
prin [ch=3; ip=*; sq=yes] (pc_lrt[],pc_plrt[],pc_score[],pc_pscore[],pc_olrt[])$[jj]
prin [ch=3; ip=*; sq=yes] \
       (pc_oolrt[],pc_oalrt[],pc_oaelrt[],pc_oscore[],pc_ascore[],pc_aescore[])$[jj]
" end of run for jjth data set "
endfor
describe ind_lrt[]
vari [nval=11] ind[1,2,3,4],mean[1,2,3,4],sd[1,2,3,4],%cv[1,2,3,4]
calc ind[]$[1] = mean(ind_lrt[])
calc ind[]$[2] = mean(ind_plrt[])
calc ind[]$[3] = mean(ind_score[])
calc ind[]$[4] = mean(ind_pscore[])
calc ind[]$[5] = mean(ind_olrt[])
calc ind[]$[6] = mean(ind_oolrt[])
calc ind[]$[7] = mean(ind_oalrt[])
calc ind[]$[8] = mean(ind_oaelrt[])
calc ind[]$[9] = mean(ind_oscore[])
calc ind[]$[10] = mean(ind_ascore[])
calc ind[]$[11] = mean(ind_aescore[])
calc mean[]$[1] = mean(pc_lrt[])
calc mean[]$[2] = mean(pc_plrt[])
calc mean[]$[3] = mean(pc_score[])
calc mean[]$[4] = mean(pc_pscore[])
calc mean[]$[5] = mean(pc_olrt[])
calc mean[]$[6] = mean(pc_oolrt[])
calc mean[]$[7] = mean(pc_oalrt[])
calc mean[]$[8] = mean(pc_oaelrt[])
calc mean[]$[9] = mean(pc_oscore[])
calc mean[]$[10] = mean(pc_ascore[])
calc mean[]$[11] = mean(pc_aescore[])
calc sd[]$[1] = sd(pc_lrt[])
calc sd[]$[2] = sd(pc_plrt[])
calc sd[]$[3] = sd(pc_score[])
calc sd[]$[4] = sd(pc_pscore[])
calc sd[]$[5] = sd(pc_olrt[])
calc sd[]$[6] = sd(pc_oolrt[])
calc sd[]$[7] = sd(pc_oalrt[])
calc sd[]$[8] = sd(pc_oaelrt[])
calc sd[]$[9] = sd(pc_oscore[])
calc sd[]$[10] = sd(pc_ascore[])
calc sd[]$[11] = sd(pc_aescore[])
calc %cv[] = sd[]/mean[]
for [ntimes=4; index=i]
prin ind[i],mean[i],sd[i],%cv[i]
endfor
calc end=now(0)
calc duration=end-start
print [ip=*] 'Duration of run =',duration; fie=0,12; dec=0,*; j=1; drep=0,38
```

```
" make results filename "
prin [ch=tn; ip=*; sq=yes] n; dec=0
prin [ch=tr; ip=*; sq=yes] r; dec=0
prin [ch=tp; ip=*; sq=yes] p; dec=0
prin [ch=tg; ip=*; sq=yes] g; dec=1
prin [ch=tnrep; ip=*; sq=yes] nrep; dec=0
prin [ch=tnit; ip=*; sq=yes] nit; dec=0
prin [ch=tout; ip=*; sq=yes] out; dec=0
concat [new=file] 'results_',tn,'_',tr,'_',tp,'_',tg,'_',tnrep,'_',tnit,'_',tout,'.xls'; skip=-1
" export results "
export [out=#file; method=add; sheet='Percentiles'] pc_lrt[],pc_plrt[],pc_score[],pc_pscore[],pc_olrt[],\
    pc_oolrt[],pc_oalrt[],pc_oaelrt[],pc_oscore[],pc_ascore[],pc_aescore[]
export [out=#file; method=add; sheet='Indicators'] ind_lrt[],ind_plrt[],ind_score[],ind_pscore[],\
  ind_olrt[],ind_oolrt[],ind_oalrt[],ind_oaelrt[],ind_oscore[],ind_ascore[],ind_aescore[]
export [out=#file; method=add; sheet='Tests'] vlrt,vplrt,vscore,vpscore,volrt,voolrt,voalrt,voaelrt,voscore,vascore,vaescore
close channel=2,3; filetype=out
endfor
```

D.2 Additional power graphics

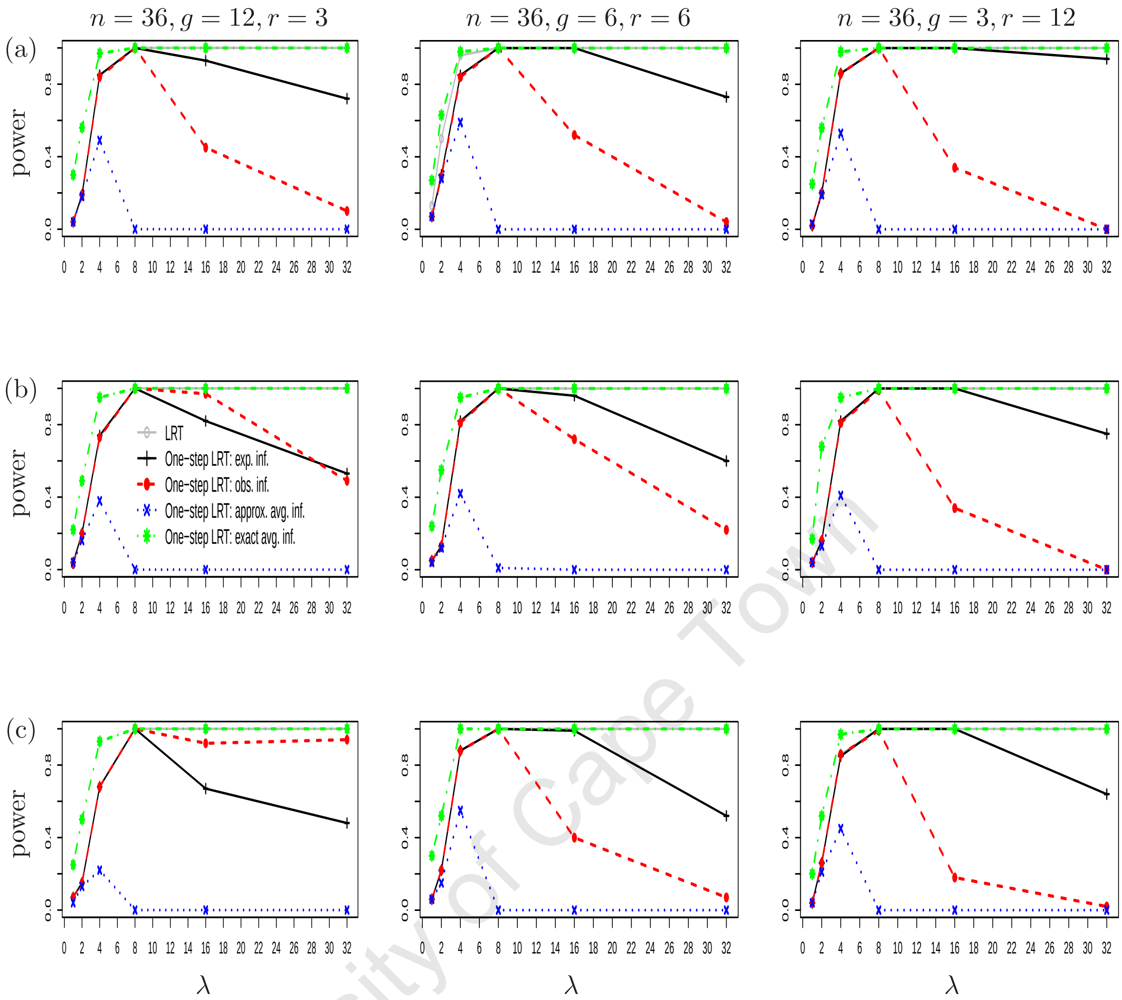

Figure D.1: *Empirical power of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets with 36 units each for (a)* $\gamma = 0.1$*, (b)* $\gamma = 1$ *and (c)* $\gamma = 10$*:* $\alpha = 0.01$.

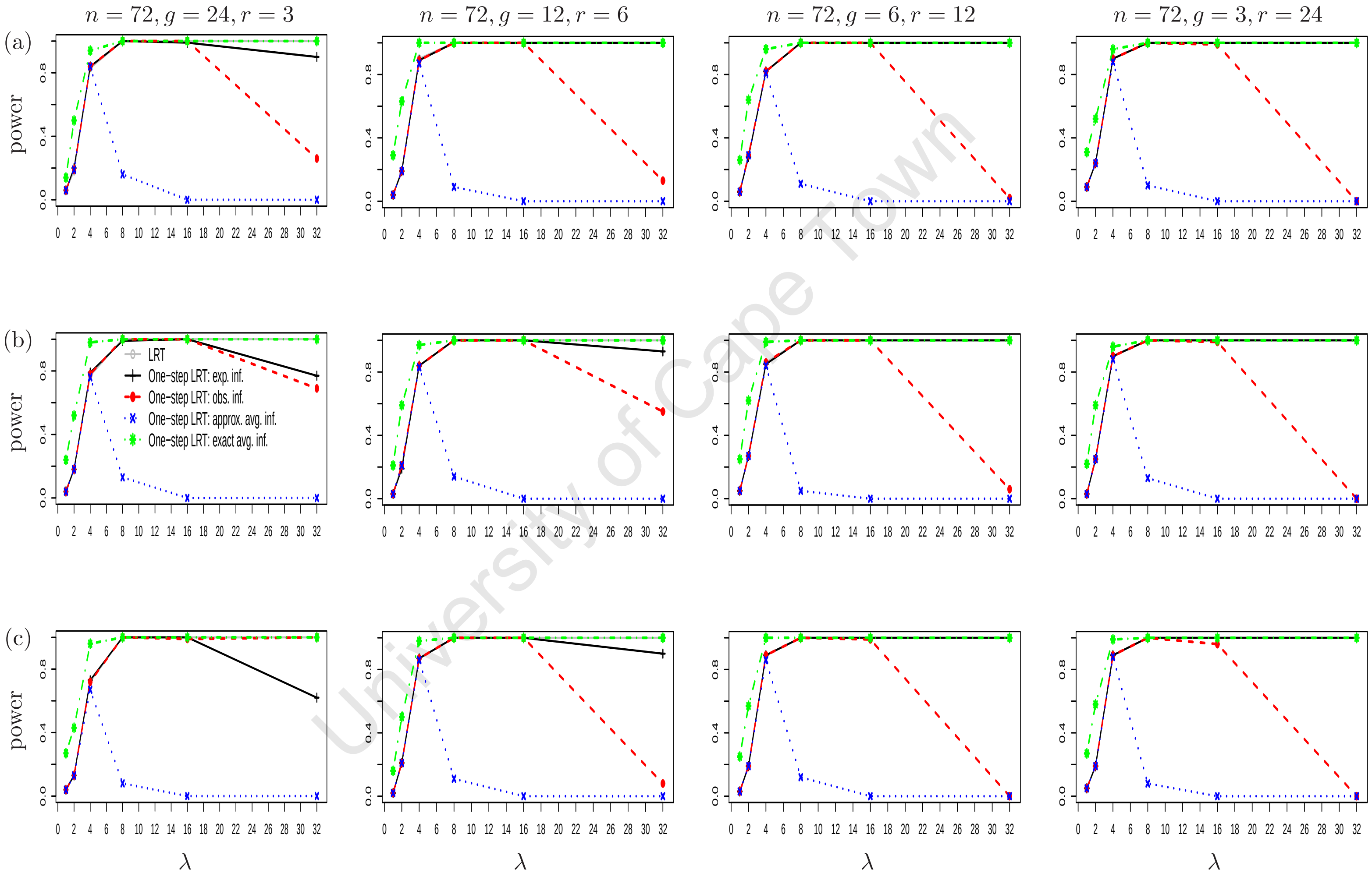

Figure D.2: *Empirical power of the likelihood ratio test statistics for the first unit based on 2500 simulations of 500 data sets with 72 units each for (a)* $\gamma = 0.1$*, (b)* $\gamma = 1$ *and (c)* $\gamma = 10$*:* $\alpha = 0.01$.