www.uhasselt.be

Universiteit Hasselt | Campus Diepenbeek
Agoralaan | Gebouw D | BE-3590 Diepenbeek | België
Tel.: +32(0)11 26 81 11

The impact of a misspecified random-effects distribution on estimation in generalized linear mixed models

Saskia LITIERE

DOCTORAATSPROEFSCHRIFT

2007 | Faculteit Wetenschappen

The impact of a misspecified random-effects distribution on estimation in generalized linear mixed models

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting wiskunde, te verdedigen door:

Saskia LITIERE

Promotor: Prof. dr. G. Molenberghs
Copromotor: Prof. dr. A. Alonso

universiteit hasselt

DOCTORAATSPROEFSCHRIFT

2007 | Faculteit Wetenschappen

The impact of a misspecified random-effects distribution on estimation in generalized linear mixed models

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting wiskunde, te verdedigen door:

Saskia LITIERE

Promotor: Prof. dr. G. Molenberghs
Copromotor: Prof. dr. A. Alonso

D/2007/2451/57

universiteit hasselt

VOOR MIJN OUDERS

VOORKANT OMSLAG:
De Terugkeer (detail), door René Magritte, 1940
(Le Retour)
Olieverf op doek, 50 x 65 cm
Koninklijke Musea voor Schone Kunsten van België, Brussel

Dankwoord

Heel veel mensen hebben de afgelopen jaren een belangrijke rol in mijn leven gespeeld, zowel op persoonlijk als op professioneel vlak. Zonder hen zou ik niet in staat geweest zijn dit avontuur tot een goed einde te brengen. Dit lijkt me dan ook een ideaal moment om aan al die mensen een woord van dank te richten.

Ik acht mijzelf gelukkig dat ik hier, aan de universiteit van Hasselt, op de steun en het vertrouwen van maar liefst drie mentoren kon rekenen. Mijn eerste mentor, Prof. dr. Herman Callaert zette mij meer dan zes jaar geleden aan het werk in een onderwijsproject om de toenmalige Master in Applied Statistics in een nieuw kleedje te steken. Herman, bedankt dat je me een kans gaf en bedankt voor de mooie herinneringen die ik aan onze samenwerking overhou!

Een paar jaar later besloot mijn tweede mentor, Prof. dr. Geert Molenberghs, me een kans te geven om binnen CenStat aan een doctoraat te beginnen. Geert, ik voel me bijzonder gelukkig en geprivilegieerd dat dit werk tot stand gekomen is mede dankzij jouw talrijke adviezen en suggesties. Bedankt hiervoor!

Tenslotte heb ik ook veel te danken aan mijn derde mentor, Prof. dr. Ariel Alonso, zonder wie ik nooit zover geraakt zou zijn. Zijn nieuwsgierigheid en toewijding zijn voor mij een bron van motivatie en inspiratie. Muchas gracias Ariel, que me has ayudado a terminar este doctorado, a pesar de mi misma. Gracias por tu paciencia, por tu amistad, por todo!

I would also like to use this opportunity to thank my colleagues at CenStat. Thanks for the stimulating multicultural atmosphere during the work hours, the lunches and the many activities. Dames, ik heb genoten van onze ladies-nights, en kijk dan ook volop uit naar het vervolg.

Uiteraard wil ik ook de inbreng van familie en vrienden niet onbesproken laten. In het bijzonder, Nadine, Kristien en Lien, bedankt voor de steun, voor de lach en de traan, voor het luisterend oor! Ariel en Annouschka, jullie waren en zijn voor mij een stukje familie ver van huis! Mama en papa, jullie steun, onvoorwaardelijke liefde en vertrouwen hebben mij aangemoedigd om de keuzes te maken die me zover gebracht hebben. Hopelijk zijn jullie even trots op mij als ik ben op jullie!

Saskia Litière

Diepenbeek, December 2007

Contents

Chapter 1

Introduction

Correlated data often occur in health-related research, like clinical trials or epidemiological studies, and consist of responses grouped according to matched sets, such as subjects or clusters. Examples include measurements on children within classes, within schools, within regions,..., or a subject characteristic measured repeatedly over time (the so-called *longitudinal data*). Naturally, observations within a cluster tend to be more alike than observations from different clusters. Therefore, classical modeling techniques such as linear regression, analysis of variance, or generalized linear models, may not be valid. Instead, these data require specific methodology which take into account the multiple sources of variation.

In the case of longitudinal data several model families are available. In so-called *marginal models*, focus is on population averages while the joint dependence structure is treated as nuisance. However, when interest lies in the association structure or specific effects for each unit, subject-level terms are added to the model. These unobserved parameters take the same value for each observation within a subject, but different values among subjects. As such, the so-called *random-effects model* contains a vector $\boldsymbol{b}$ of unobserved subject-specific effects, conditional upon which it is then often assumed that the observed responses within each subject are independent.

Random-effects models for normal or Gaussian responses are well established. In this setting, the *linear mixed model* (Verbeke and Molenberghs, 2000) offers a unifying framework which can handle a wide variety of correlated data, including repeated measurements and longitudinal data as well as clustered, hierarchical and

spatial data. Thanks to the choice of the normal distribution for the random effects, the model marginalizes to a multivariate normal with easily interpretable mean and variance components.

When dealing with non-Gaussian data, the most commonly used subject-specific model is the *generalized linear mixed model* (Diggle *et al.*, 2002; Molenberghs and Verbeke, 2005). It is a straightforward extension of the generalized linear model (McCullagh and Nelder, 1989), and accounts for correlation among clustered observations. Model fit is based on classical likelihood techniques, and requires maximization of the marginal likelihood, which is obtained by integrating out the random effects over their assumed distribution. The commonly used normal distribution for these random effects generally leads to integrals that cannot be calculated in a closed form. However, several numerical approximations have been implemented and are available in standard software tools.

Obviously, estimation and inferences, based on such models, depend on the assumption that the model and therefore, the random-effects distribution is correctly specified. Since the random effects are unobserved, the validity of this assumption can be difficult to verify. Therefore, the question naturally arising is concerned with the impact of a misspecified random-effects distribution on the maximum likelihood estimators and the inferential procedures in generalized linear mixed models. In the present work, we will address this question via an extensive simulation study. Additionally, we will propose a set of diagnostic tools which can detect this, and even more general types of model misspecification. Finally, we will conclude with some guidelines on how to proceed when facing the consequences of possible random-effects misspecification.

This thesis can be structurally divided into three parts. The first part presents a concise introduction to mental health, and to the key motivating study on schizophrenia, which is used throughout this work (Chapter 2). Chapter 3 provides a brief review of the generalized linear mixed model and a short discussion on some of the challenges involved in its application. The introductory part ends with Chapter 4, which describes an initial analysis of the case study using a logistic-normal model.

The second part, consisting of 4 chapters, presents an overall picture of the effect of random-effects misspecification on maximum likelihood estimation. First, some important contributions by White (1982) on likelihood inferences under general model misspecification are summarized in Chapter 5. Chapter 6 focuses on random-effects misspecification in the special case of linear mixed models, whereas Chapter 7 un-

dertakes the study of the impact of this type of misspecification on the maximum likelihood estimators in generalized linear mixed models. Guidelines are supplied to distinguish those situations in which the misspecification has a negligible impact, from those in which the misspecification can have serious consequences. Finally, Chapter 8 provides a theoretical result which states that whenever a subset of fixed-effects parameters, not included in the random-effects structure, equals zero, the corresponding maximum likelihood estimator will consistently estimate zero. This implies that under certain conditions a significant effect could be considered as a reliable result, even if the random-effects distribution is misspecified.

The third part of this work comprises 5 additional chapters, which focus on remedial measures for random-effects misspecification. Following some ideas by White (1982), Chapters 9 and 10 introduce a set of diagnostic tools to detect misspecification. The availability of such a toolbox then naturally raises the issue of how to proceed in the presence of misspecification. When the number of subjects and the number of repeated measurements per subject are sufficiently large, it will be shown in Chapter 11 that the maximum likelihood estimators of the mean structure remain asymptotically robust, irrespective of the distribution of the random effects. However, when the available information is not sufficiently large to rely on asymptotic results, alternative approaches need to be considered. Since at the moment there does not seem to exist a model family which is generally robust against this type of misspecification, Chapter 12 proposes a sensitivity analysis, where different distributions are considered for the random effects. In some specific situations, robust alternative models can be found. For instance, the linear mixed model for normal responses is known to be robust against random-effects misspecification. In Chapter 13, it will be shown that another example is given by the Poisson-gamma model for repeated counts. Finally, Chapter 14 recapitulates some concluding remarks and offers a perspective on possible future research.

Chapter 2

A Case Study in Mental Health

2.1 Introduction

The World Health Organization (WHO) defines mental health as "not merely the absence of disease" but rather as "a state of complete physical, mental and social well-being". Together with physical health, mental health contributes to the overall well-being of individuals, societies and communities. Recent advances in neuroscience and behavioral medicine have shown that, like many physical illnesses, mental and behavioral disorders are the result of a complex interaction between biological, psychological and social factors. They are known to have a basis in the brain and they can affect people of all ages in all countries.

Today, around 450 million people are estimated to be suffering from a mental disorder, and mental disorders are estimated to account for 12% of the global disease burden. Fortunately, in most cases, the presence of mental disorders can be diagnosed and treated cost-effectively, so that people affected by them can have a better chance of living a full and productive life within their own community. However, only few of those affected receive even the most basic treatment. In most parts of the world, mental disorders are not regarded with the same importance as physical disorders. Today, more than 40% of the countries have no mental health policy. Those that do foresee a

mental health budget, spend less than 1% of their total health expenditures on mental and behavioral disorders. In developing countries, most individuals diagnosed with severe mental disorders do not receive any help to cope with depression, dementia, schizophrenia and substance dependence. Instead, they become targets of stigma and discrimination (The World Health Report 2001, available at http://www.who.int).

The economic cost of mental disorders on society should not be underestimated. Indeed, it is long lasting and huge. Studies estimating the aggregate economic burden on social service needs, lost employment and reduced productivity obtained that mental disorders account for about 2.5% of the gross national product in the United States. In Europe, some studies have estimated the expenses of mental disorders as a proportion of all health costs. In the Netherlands, this amounted to 23.2%, and in the United Kingdom, for patients residing in hospitals only, the costs were up to 22% (Burzykowski, Molenberghs and Buyse, 2005).

The occurrence of mental disorders is truly universal, affecting people of all countries and societies, all ages, both women and men, rich and poor, from urban and rural environments. Mental disorders are said to be present in about 10% of the adult population at any given point in time. One in four families is likely to have at least one member diagnosed with a mental disorder. And the problem will only grow as experts expect further increases in the number of diagnoses in view of the ageing population, worsening social problems and civil unrest. Already, mental disorders represent four of the 10 leading causes of disability worldwide. Most common disorders, with usually severe disabilities, include depressive disorders, substance use disorders, epilepsy, Alzheimer, mental retardation and schizophrenia (The World Health Report 2001, available at http://www.who.int).

The topic of our case study, schizophrenia, is one of the most disabling and emotionally devastating illnesses known to man. This disease has been misunderstood for a very long time, and its victims were frequently undeservingly stigmatized. Although accounts relating to symptoms of schizophrenia go back as far as 2000 BC, it was not until 1908 that the term was first used by Eugene Bleuler, a Swiss psychiatrist, to refer to the lack of interaction between thought processes and perception. The term is actually derived from two Greek words and literally means “split” or “shattered mind”. This perhaps confusing term has led to the common misconception that affected people suffer from a split personality. Nevertheless, although patients may suffer from hallucinations or delusions, it generally does not involve changing among distinct personalities.

Schizophrenia is most commonly diagnosed during late adolescence or early adulthood, and approximately equally in men and women. Symptoms are often described in terms of "positive" and "negative" manifestations. Positive symptoms include delusions, auditory hallucinations and thought disorder and are typically regarded as manifestations of psychosis. Negative symptoms include the loss or absence of normal traits or abilities, and include features such as a lack of emotional expression, poverty of speech and lack of motivation. Additionally, deficits in the form of reduced or impaired psychological functions such as memory, attention, problem-solving, executive function or social cognition may be present.

By the first half of the twentieth century, schizophrenia was considered by many a hereditary disease, and individuals affected by schizophrenia were often removed from the evolutionary cycle through sterilization. During the regime of Adolf Hitler, between 75,000 to 250,000 people diagnosed with schizophrenia, or labeled as "mentally unfit" were murdered in the context of the Nazi "cleansing" program.

Nowadays, it is suggested that schizophrenia can affect anyone at any point in life. Although no common cause of schizophrenia has been identified in all individuals diagnosed with the condition, it is believed that genetic vulnerability and environmental stressors can act together to result in its diagnosis. The extent to which these factors influence the likelihood of being diagnosed with schizophrenia is debated widely, and currently controversial. Although it does have a strong heritable component (some estimates are as high as 80%), research is showing that also stressful life events can cause or trigger the disease (Day *et al.*, 1987; Harrison and Owen, 2003; Corcoran *et al.*, 2003).

With modern advances in science, including drug therapy and psychological care, almost half of the individuals developing schizophrenia can expect a full and lasting recovery. Of the remainder, only about one-fifth continue to face serious limitations in their day-to-day activities. Part of the treatment of schizophrenia focuses on the treatment of the psychotic symptoms using so-called *antipsychotic* drugs. Risperidone is one such antipsychotic, which is generally accepted as a first-line treatment for newly-diagnosed patients. Although not effective for everyone, its side effects are usually minimal at regular maintenance dosages. In the case study that motivated the present work, the effect of risperidone was compared to that of conventional antipsychotic agents for the treatment of *chronic* schizophrenia (Alonso *et al.*, 2004). The data obtained from this randomized clinical trial will be described in the next section.

Table 2.1: *The Clinical Global Impression (CGI) questionnaire.*

Considering your total clinical experience with this particular population, how severe is the patient's schizophrenia at this time?
☐ 1. Normal (not at all ill)
☐ 2. Borderline mentally ill
☐ 3. Mildly ill
☐ 4. Moderately ill
☐ 5. Markedly ill
☐ 6. Severely ill
☐ 7. Extremely ill

2.2 Data From Clinical Trials in Schizophrenia

Like many mental illnesses, the diagnosis of schizophrenia is based upon the behavior of the person being assessed. Several measures can be used to evaluate a patient's global condition. The *Clinical Global Impression* (CGI) is generally accepted as a subjective but useful clinical measure of change. It is a 7-grade scale used to characterize a subject's mental condition (see the questionnaire in Table 2.1). In the case study, the binary response variable Y_i for the ith patient is a dichotomous version of the CGI which equals 1 for those subjects classified as normal to mildly ill, and 0 for those classified as moderately to extremely ill. Since it is recommended that risperidone is most effective at doses ranging from 4 to 6 mg/day, we considered only those patients receiving either these doses of risperidone, i.e., the treatment group $z_i = 1$, or the active control, i.e., the control group $z_i = 0$. Treatment was administered for 2 months and the outcome was measured at weeks 0, 1, 2, 4, 6, and 8. In total, 128 patients were included in the trial, from which 64 were assigned to the treatment group. These data are shown in Appendix A.

Figure 2.1 summarizes the probability of being classified as normal to mildly ill ($P(Y = 1)$), by time point and treatment group. Further, Table 2.2 summarizes the dropout per time point and per treatment group. From this table we can see that the dropout is slightly higher in the control group. Additionally, in both groups around 50% of the participants are missing near the end of the study. A full discussion and analysis of this missing data problem goes beyond the scope of the present work. Therefore, in what follows, we will assume that the underlying missing data generating mechanism was *missing at random* (MAR), i.e., conditional on the observed data the missingness is independent of the unobserved measurements. We

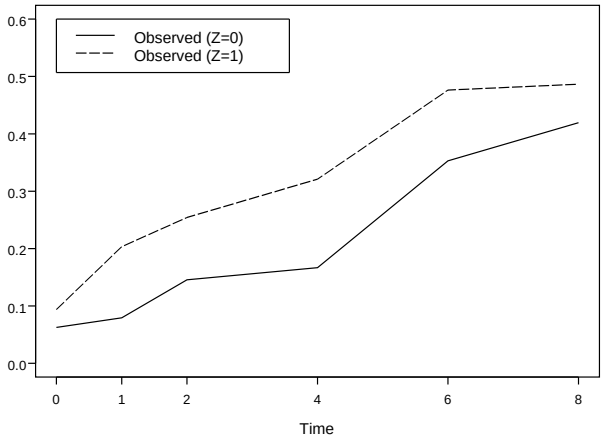


Figure 2.1: *Schizophrenia data. Evolution over time of the observed probabilities of being classified as normal to mildly ill, by treatment group ($Z = 0$ for the control group and $Z = 1$ for the treatment group).*

Table 2.2: *Schizophrenia data. Dropout (in %) per time point and treatment group, where $z = 0$ ($z = 1$) represents the control (treatment) group.*

Time	$z = 0$	$z = 1$
0	0	0
1	2	0
2	14	8
4	25	17
6	47	34
8	52	42

refer to Molenberghs and Verbeke (2005) and Kenward and Molenberghs (2007) for more details on this topic in the context of discrete longitudinal data.

As mentioned in Chapter 1, the analysis of these data requires a model which takes into account the correlation between observations coming from the same subject. One such approach, given by the generalized linear mixed model, will be introduced in the next chapter along with a discussion of some of the challenges involved in its application.

Chapter 3

Generalized Linear Mixed Models

The generalized linear mixed model has become a powerful parametric tool for the analysis of non-Gaussian longitudinal data with multiple sources of variation. Its implementation in popular statistical packages, such as the SAS procedures MIXED, NLMIXED, and GLIMMIX, or the R functions `lme` and `glmm`, has substantially contributed to its wide use in different areas like, for example, toxicology (Molenberghs and Verbeke, 2005), epidemiology (Kleinman, Lazarus and Platt, 2004), dairy science (Tempelman, 1998), etcetera. In this chapter, the model is briefly introduced, and some general issues in its estimation are discussed.

3.1 Model Formulation

Let y_{ij} be the jth response of subject i, $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Conditional on a vector of individual random effects $\boldsymbol{b}_i$, the outcome variables are assumed to be independent, with density functions belonging to the exponential family

$$f(y_{ij}|\theta_{ij}, \varphi) = \exp[\varphi^{-1}\{y_{ij}\theta_{ij} - \psi(\theta_{ij})\} + c(y_{ij}, \varphi)], \tag{3.1}$$

where φ is a scale parameter, $c(.)$ is a function only depending on y_{ij} and φ, and $\psi(.)$ is a function satisfying $\mathrm{E}(y_{ij}|\boldsymbol{b}_i) = \psi'(\theta_{ij})$ and $\mathrm{Var}(y_{ij}|\boldsymbol{b}_i) = \varphi\psi''(\theta_{ij})$. Further,

$\mu_{ij} = \mathrm{E}(y_{ij}|\boldsymbol{b}_i) = v(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)$, where $v(.)$ denotes a known link function, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are vectors of covariates, and $\boldsymbol{\beta}$ is a vector of unknown fixed regression coefficients. The subject-specific effects $\boldsymbol{b}_i$ are often assumed to be normally distributed with mean zero and variance-covariance matrix D. Examples of generalized linear mixed models include the linear mixed model for continuous data (see Section 3.3), the logistic-normal model for binary data, and the Poisson-normal model for count data. Fitting these models requires maximization of the marginal likelihood, obtained by integrating out the random effects. Let the contribution to the likelihood of subject i be given by

$$f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, D, \varphi) = \int \prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, \varphi) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i, \tag{3.2}$$

then we can derive the marginal likelihood as

$$L(\boldsymbol{\beta}, D, \varphi) = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, \varphi) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i. \tag{3.3}$$

The commonly used normal distribution for the random effects generally leads to an intractable likelihood function. Only in a few special cases can (3.2) be worked out analytically. For instance, in linear mixed models, (3.2) corresponds to the density of a multivariate normal distribution. The extension of the linear mixed model to non-Gaussian data does not exhibit the same behavior. One of the problems is that there is no analogue to the multivariate normal distribution for generalized linear mixed models. This, together with the nonlinearity introduced by the link function v, implies that the parameters of the random-effects model and the induced marginal model have different interpretations.

Even though (3.2) cannot be calculated in closed form, several numerical approximations to the likelihood have been implemented in the available software tools. These include approximations of the integrand, approximations of the data and approximations of the integral itself. When interest is in approximating the integrand, Laplace-type approximations can be used to obtain a closed-form expression for the likelihood. The second class of approximations is based on the decomposition of the data into a Taylor series expansion of the mean and an appropriate error term. This approach includes, for example, penalized and marginal quasi-likelihood. An extensive discussion of these numerical approximation techniques is given in Molenberghs and Verbeke (2005). In this manuscript, we will mainly focus on approximations of the integral via Gaussian quadrature, as implemented in the SAS procedure NLMIXED.

3.1.1 Gaussian Quadrature

Gaussian quadrature is a technique designed to approximate integrals of the form

$$\int f(z)\phi(z)dz, \tag{3.4}$$

where $f(z)$ is a known function and $\phi(z)$ represents the density of the (multivariate) standard normal distribution. If we define a new set of random effects $\boldsymbol{a}_i = D^{-1/2}\boldsymbol{b}_i$, then $\boldsymbol{a}_i$ follows a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix I. Hence, the likelihood contribution (3.2) for every subject i can be rewritten as

$$\begin{aligned} f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, D, \varphi) &= \int \prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, \varphi) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i \\ &= \int \prod_{j=1}^{n_i} f[y_{ij}|\theta_{ij}(\boldsymbol{a}_i), \varphi, D]\phi(\boldsymbol{a}_i) d\boldsymbol{a}_i. \end{aligned} \tag{3.5}$$

Expression (3.5) is of the form (3.4), and can easily be evaluated via Gauss-Hermite polynomials, i.e.,

$$\int_{-\infty}^{\infty} f(z)\phi(z)dz \approx \sum_{q=1}^{Q} P(z_q)w_q. \tag{3.6}$$

In this formula, the integral is approximated by a weighted sum evaluated at Q values z_q, called quadrature points. The weights w_q depend only on Q and the normal density. In the simple setting of univariate integration, the approximation consists of subdividing the integration region in intervals, and approximating the area under the curve by the sum of the areas of the so-obtained rectangles. In general, the higher Q, the smaller the intervals and the better the approximation. This technique is called Gaussian quadrature and is illustrated in the left panel of Figure 3.1, for $Q = 10$.

Note that the quadrature points z_q are independent of the function $f(z)$, so depending on the support of $f(z)$, the z_q will or will not lie in the region of interest. For example, in the left panel of Figure 3.1, the mode $\tilde{z}$ of $f(z)$ lies remote from 0. Then, for a small value of Q, the quadrature points z_q will be inappropriate. In that case, it might be useful to rescale and shift the quadrature points such that more points lie in the region of interest: centered at $\tilde{z}$ and with spread depending on the shape of f (as illustrated in the right panel of Figure 3.1). This goes by the name of *adaptive* Gaussian quadrature, where the quadrature points are centered and scaled as if $f(z)\phi(z)$ would follow a normal distribution. The mean of this normal distribution

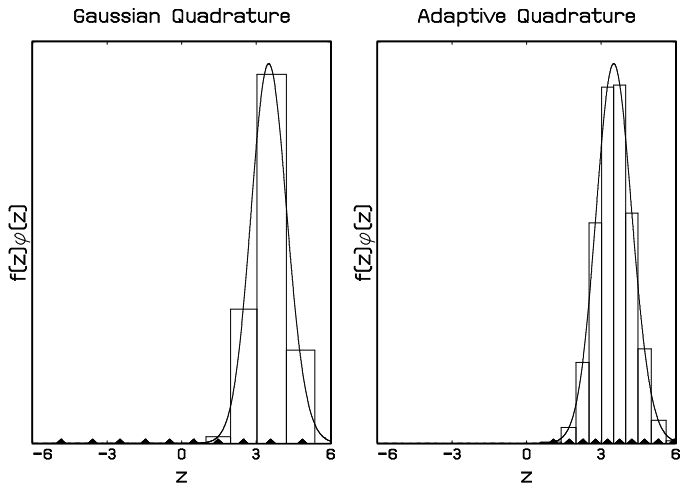


Figure 3.1: *Gaussian and adaptive Gaussian quadrature obtained from 10 quadrature points. The black triangles indicate the position of the quadrature points, while the rectangles indicate the contribution of each point to the integral (ref. Molenberghs and Verbeke, 2005).*

corresponds to the mode $\tilde{z}$ of $\ln[f(z)\phi(z)]$; the variance of this distribution equals

$$\hat{\tau} = \left[-\frac{\partial^2}{\partial z^2} \ln[f(z)\phi(z)] \Big|_{z=\tilde{z}} \right]^{-1}. \tag{3.7}$$

Hence, the new quadrature points are given by $z_q^* = \tilde{z} + \sqrt{\hat{\tau}} z_q$, the corresponding new weights by $w_q^* = \frac{\phi(z_q^*)}{\phi(z_q)} \sqrt{\hat{\tau}} w_q$, and, as a result, the integral in (3.5) is approximated by

$$\int_{-\infty}^{\infty} f(z)\phi(z)dz \approx \sum_{q=1}^{Q} P(z_q^*) w_q^*. \tag{3.8}$$

Maximizing this approximate likelihood still involves first and second order derivatives. Additionally, an approximation of the likelihood contribution for each one of the n subjects is needed to determine the mode of $\ln[f(z)\phi(z)]$. So, even though typically, with adaptive Gaussian quadrature (much) less quadrature points are needed than with non-adaptive Gaussian quadrature, this does not always imply fewer function evaluations, and therefore, adaptive Gaussian quadrature can become time consuming.

3.2 Inferential Procedures

Once an approximation to the marginal likelihood is available, inferences for the model parameters are obtained from classical maximum likelihood theory. Indeed, assuming that the model is correctly specified, the maximum likelihood estimators are asymptotically normally distributed with the correct values as means, and the inverse Fisher information matrix as covariance matrix. Hence, Wald-type tests for the mean structure parameters can easily be constructed, as well as likelihood ratio and score tests.

However, when inferences for some variance components in D are of interest, classical tests may not be valid as the hypotheses to be tested tend to be on the boundary of the parameter space. For instance, when testing for the presence of a random intercept, one generally tests the hypothesis that the variance σ_b^2 of the random effect equals zero. Since variances cannot be negative, the null-hypothesis $H_0 : \sigma_b^2 = 0$ is clearly on the boundary of the parameter space. As a result, none of the classical Wald, likelihood ratio or score tests are valid. In this setting, appropriate alternatives can be constructed for which the asymptotic null distribution is a mixture of chi-squared distributions (Verbeke and Molenberghs, 2003; Silvapulle and Silvapulle, 1995; and Hall and Praestgaard, 2001).

It should be pointed out that inferences in generalized linear mixed models are based on the marginal model (3.2) rather than on the original hierarchical model given by (3.1). In practice this could lead to negative estimates of the variance components. Indeed, the representation of the marginal model does not explicitly assume the presence of random effects representing the natural heterogeneity between the subjects, but merely a specific structure for the marginal covariance matrix. As far as this covariance matrix is positive definite, a valid marginal model is obtained. However, in this case the resulting model does not allow any hierarchical interpretation since no random-effects structure could ever induce such a model. Although the marginal model follows naturally from the random-effects model, both models are not equivalent. Different random-effects models can produce the same marginal model and some marginal models cannot be obtained from any hierarchical counterpart.

Finally, note that the estimates of the random effects $\boldsymbol{b}_i$ can be useful for predicting cluster-specific evolutions. They reflect between-subject variability and can therefore be valuable to detect outlying profiles or groups of individuals evolving different over time. Since random effects are assumed to be random variables, a natural method to

obtain estimates follows from Bayesian methodology. Indeed, the posterior density of the random effects is given by

$$f_i(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, D, \phi) = \frac{\prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, \phi) f(\boldsymbol{b}_i|D)}{\int \prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, \phi) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i}. \tag{3.9}$$

Therefore, the so-called empirical Bayes (EB) estimates of $\boldsymbol{b}_i$ are readily given by the posterior mode of this density, i.e., the value of $\boldsymbol{b}_i$ that maximizes the posterior density, in which the unknown parameters have been replaced by their maximum likelihood estimates.

3.3 Special Case: the Linear Mixed Model

An important special case of generalized linear mixed models is obtained when the response vector $\boldsymbol{y}_i$ is continuous and assumed to follow a normal distribution. In this setting, the following linear mixed model was proposed by Laird and Ware (1982). Conditional on a vector of subject-specific random effects $\boldsymbol{b}_i$, the outcomes $\boldsymbol{y}_i$ are modeled as

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon_i}, \tag{3.10}$$

where X_i and Z_i now represent $n_i \times p$ and $n_i \times q$ matrices of known covariates, and the random effects $\boldsymbol{b}_i$ are assumed to be sampled from a multivariate normal distribution with mean zero and covariance matrix D. Finally, the residual vector $\boldsymbol{\varepsilon_i}$ is assumed to be independent from $\boldsymbol{b}_i$ and to be normally distributed with zero mean and some covariance matrix Σ_i.

As before, estimation is based on maximization of the marginal likelihood of $\boldsymbol{y}_i$, i.e., $f_i(\boldsymbol{y}_i) = \int f_i(\boldsymbol{y}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i$. In this expression, the normal random-effects distribution is conjugate to the normal distribution of the outcome, conditional on the random effects. As a result, the linear mixed model (3.10) implies a multivariate normal marginal model for $\boldsymbol{y}_i$ with mean $X_i\boldsymbol{\beta}$ and covariance matrix $V_i = Z_i D Z_i^T + \Sigma_i$. This connection between the hierarchical and the marginal specification of the model allows both a marginal and a hierarchical interpretation for the fixed-effects parameters $\boldsymbol{\beta}$. The fitting of a linear mixed model is usually based on the marginal model for the response vector $\boldsymbol{y}_i$. However, an extensive description of estimation and inferences related to this model would be outside the scope of this manuscript. Instead, we refer to Verbeke and Molenberghs (2000) for an elaborate discussion.

Chapter 4

Initial Analysis of the Case Study

Let us recall the case study introduced in Section 2.2. Given the discrete nature of the repeated outcomes, the generalized linear mixed model can be considered as an appropriate choice for the analysis of these data. Therefore, in this chapter we will use the techniques previously introduced in Chapter 3 to study the effect of risperidone on the evolution of patients suffering from chronic schizophrenia.

We analyzed the data using a random-intercept model by considering different link functions and linear predictors. In the model building exercise, a total of nine models were fitted. These models were constructed as combinations of three link functions, i.e., the logit, log-log and probit link, and three different linear predictors (LP) which can be summarized as

$$\bullet \text{ LP1: } \beta_0 \quad +\beta_1 z_i \quad +\beta_2 t_j + \beta_3 z_i t_j \tag{4.1}$$

$$\bullet \text{ LP2: } \beta_0 \quad +\beta_1 z_i \quad +\beta_2 t_j \tag{4.2}$$

$$\bullet \text{ LP3: } \beta_0 \quad\quad\quad +\beta_2 t_j + \beta_3 z_i t_j \tag{4.3}$$

where $z_i = 1$ (0) denotes the treatment (control) group and t_j denotes the occasion of measurement. The random intercept b_i was always assumed to follow a normal distribution with mean zero and variance σ_b^2. All the previous models were fitted using the SAS procedure NLMIXED with adaptive Gaussian quadrature and 20 quadrature

Table 4.1: *Schizophrenia data. Maximum likelihood estimates (standard errors) and AIC values from the different random-intercept models obtained as combinations of three link functions and the three linear predictors defined in (4.1)-(4.3).*

Link		β_0	β_1	β_2	β_3	σ_b^2	AIC
Logit	LP1	-7.36 (1.23)	2.12 (1.25)	0.65 (0.14)	0.006 (0.170)	21.01 (6.81)	393.9
	LP2	-7.37 (1.18)	2.14 (1.08)	0.65 (0.09)		21.01 (6.81)	391.9
	LP3	-6.20 (0.90)		0.56 (0.12)	0.16 (0.14)	21.17 (6.65)	394.9
Log-log	LP1	-6.14 (0.94)	1.80 (0.95)	0.51 (0.11)	-0.059 (0.127)	11.74 (3.85)	395.1
	LP2	-5.99 (0.87)	1.57 (0.81)	0.47 (0.07)		11.76 (3.86)	393.3
	LP3	-5.15 (0.67)		0.43 (0.09)	0.068 (0.106)	12.08 (3.90)	396.9
Probit	LP1	-4.07 (0.67)	1.18 (0.69)	0.36 (0.07)	0.002 (0.091)	6.47 (2.07)	394.2
	LP2	-4.07 (0.64)	1.18 (0.60)	0.36 (0.05)		6.47 (2.07)	392.9
	LP3	-3.42 (0.48)		0.31 (0.06)	0.081 (0.077)	6.48 (2.02)	395.3

points. The corresponding maximum likelihood estimates obtained from all combinations of link functions and linear predictors are displayed in Table 4.1. We used Akaike's Information Criterion (AIC) to select the best fitting model (the smaller the value of the criterion, the better the model). Given the AIC values shown in the last column of Table 4.1, it follows that the model with the best fit is given by

$$\text{logit}\{P(y_{ij} = 1|b_i)\} = \beta_0 + \beta_1 z_i + \beta_2 t_j + b_i. \tag{4.4}$$

Figure 4.1 displays the fitted probabilities obtained from this model against the observed probability of being classified as normal to mildly ill (i.e., $P(Y = 1)$) by time point and treatment group. The fitted probabilities are calculated by numerically integrating out the random effect for each subject. Until week 4 there seems to be a reasonable agreement between the fitted and the observed values. Nevertheless, some discrepancy is observed in the last two measurement occasions. As stated before, the proportion of dropouts is significantly high for these two measurements, specially in week 8 for the control group (see Table 2.2). In the presence of missing data such a discrepancy is not necessarily an evidence of lack of fit (Molenberghs and Verbeke, 2005). However, a full discussion of the missing data problem would be outside the scope of this work. Instead, as stated in Chapter 2, we will assume that the missing data generating mechanism is MAR, making our likelihood approach a valid option.

Note that, even though the model given by (4.4) emerged as the best fitting model

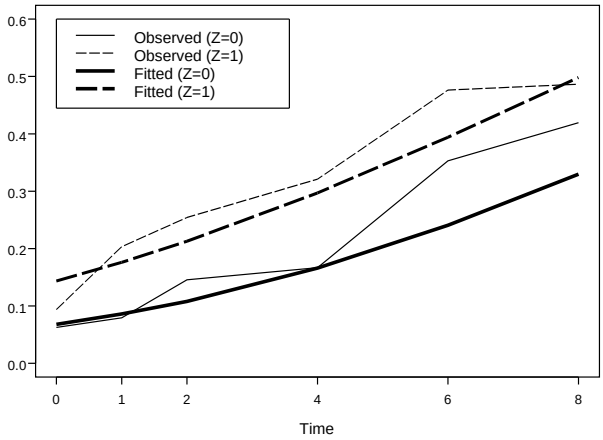


Figure 4.1: *Schizophrenia data. Evolution of the observed and estimated (using model (4.4)) probabilities of being classified as normal to mildly ill, by treatment group.*

among all the ones considered in the model building exercise, it produces relatively extreme estimates for the intercept and the variance component. We believe this is the result of some extreme response pattern in the data. For example, in the control group a high proportion of the patients (75%) have a response pattern of nothing but zeros, whereas in the treatment group a more variable pattern of responses is observed. There, only 56% of the patients have a response pattern consisting solely of zeros. Hence, the large estimate for the variance of the random component could be explained by the high intra-subject correlation that these data seem to suggest. Allowing σ_b^2 to vary among the treatment groups did not improve the fit. Indeed, this analysis resulted in a random-effect variance of 20.00 (s.e. 7.93) for the treatment group and 22.61 (s.e. 10.69) for the control group. Clearly, these high variances hint on a very strong within-subject correlation within both treatment groups.

Arguably, these circumstances could render the assumption of a normal distribution for the random effects questionable. However, this situation should not be considered exceptional or infrequent. Indeed, in a typical placebo controlled clinical trial such an extreme pattern of all zeros could be expected in the placebo control group, whereas a more variable pattern should be expected in the responses of the

treated group. This naturally leads to concerns about the impact of a misspecified random-effects distribution on our estimates and related inference procedures. Therefore, in what follows, we will study how such misspecification can affect maximum likelihood estimation, starting with a review in Chapter 5 of some available theory on likelihood inference under general model misspecification.

Chapter 5

Maximum Likelihood Estimation in Misspecified Models

In a landmark paper, White (1982) analyzed in detail the whole problem of likelihood inferences under general model misspecification. Some of his results will play a central role in the present work, and therefore this chapter is devoted to summarize his findings.

Let us consider a random variable $\boldsymbol{y}$ with density function h, and a parametric family of density functions $\mathfrak{F} = \{f(\boldsymbol{y}, \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Upsilon\}$. If there exists a $\boldsymbol{\xi}_0 \in \Upsilon$ such that $\mathfrak{F}$ contains the true distribution (i.e., $h(\boldsymbol{y})$ can be written as $f(\boldsymbol{y}, \boldsymbol{\xi}_0)$), then the maximum likelihood estimator $\widehat{\boldsymbol{\xi}}_n$ of $\boldsymbol{\xi}_0$ is consistent and asymptotically normal. However, since in practice h is unknown, it can be difficult to check whether h belongs to $\mathfrak{F}$ or not.

In general, when h does not belong to $\mathfrak{F}$, White (1982) has shown that the maximum likelihood estimator $\widehat{\boldsymbol{\xi}}_n$ will (strongly) converge to the value of $\boldsymbol{\xi}$, denoted by $\boldsymbol{\xi}^*$, which minimizes the so-called Kullback-Leibler Information Criterion (KLIC)

$$I(h : f, \boldsymbol{\xi}) = \mathrm{E}\left[\log \frac{h(\boldsymbol{y})}{f(\boldsymbol{y}, \boldsymbol{\xi})}\right]. \tag{5.1}$$

Here and in what follows, the expectations are taken with respect to the true distri-

bution h. The following assumptions provide the necessary conditions to prove this result.

Assumption 5.1 *The independent random vectors $\boldsymbol{Y}_i$ $(i = 1, \ldots, n)$ have common joint distribution function H on Ω, a measurable Euclidean space, with measurable Radon-Nikodým density $h = \dfrac{dH}{dv}$.*

Assumption 5.2 *The family of distribution functions $F(\boldsymbol{y}, \boldsymbol{\xi})$ has Radon-Nikodým densities $f(\boldsymbol{y}, \boldsymbol{\xi}) = \dfrac{dF(\boldsymbol{y}; \boldsymbol{\xi})}{dv}$, such that*

- *f is measurable in $\boldsymbol{y}$ for every $\boldsymbol{\xi} \in \Upsilon$, where Υ is a compact subset of a p-dimensional Euclidean space, and*
- *f is continuous in $\boldsymbol{\xi}$, for every $\boldsymbol{y} \in \Omega$.*

Assumption 5.3 *The following properties hold:*

- *$E[\log h(\boldsymbol{y})]$ exists and $|\log f(\boldsymbol{y}, \boldsymbol{\xi})| \leq m(\boldsymbol{y})$ for all $\boldsymbol{\xi} \in \Upsilon$, where m is integrable with respect to H, and*
- *$I(h : f; \boldsymbol{\xi})$ has a unique minimum at $\boldsymbol{\xi}^* \in \Upsilon$.*

Note that Assumption 5.3 ensures that the KLIC is well-defined, and together with Assumptions 5.1 and 5.2 it specifies the regularity conditions needed for the following consistency theorem.

Theorem 5.1 (Consistency) *Given Assumptions 5.1-5.3, $\widehat{\boldsymbol{\xi}}_n \xrightarrow{a.s.} \boldsymbol{\xi}^*$.*

Next, White (1982) studied the asymptotic normality of the maximum likelihood estimators under model misspecification. To this end, we need to introduce the following additional notation

$$A(\boldsymbol{\xi}) = \mathrm{E}\left\{ \frac{\partial^2 \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k \partial \xi_\ell} \right\}, \tag{5.2}$$

$$B(\boldsymbol{\xi}) = \mathrm{E}\left\{ \frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k} \cdot \frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_\ell} \right\}, \tag{5.3}$$

$$A_n(\boldsymbol{\xi}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k \partial \xi_\ell} \right\}, \tag{5.4}$$

$$B_n(\boldsymbol{\xi}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k} \cdot \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_\ell} \right\}, \tag{5.5}$$

for $k, \ell = 1, \ldots, p$, where p refers to the number of parameters in the model. Further, let $V(\boldsymbol{\xi}) = A^{-1}(\boldsymbol{\xi})B(\boldsymbol{\xi})A^{-1}(\boldsymbol{\xi})$, and $V_n(\boldsymbol{\xi}) = A_n^{-1}(\boldsymbol{\xi})B_n(\boldsymbol{\xi})A_n^{-1}(\boldsymbol{\xi})$. Finally, we need to consider the following additional assumptions.

Assumption 5.4 *The functions*

$$\frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k}$$

are measurable with respect to $\boldsymbol{y}$ *for each* $\boldsymbol{\xi} \in \Upsilon$*, and continuously differentiable of* $\boldsymbol{\xi}$ *for each* $\boldsymbol{y} \in \Omega$*.*

Assumption 5.5 *The functions*

$$\left| \frac{\partial^2 \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k \partial \xi_\ell} \right| \quad \textit{and} \quad \left| \frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k} \cdot \frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_\ell} \right|$$

are dominated by functions integrable with respect to H *for all* $\boldsymbol{y} \in \Omega$ *and all* $\boldsymbol{\xi} \in \Upsilon$*.*

Assumption 5.6 *The following properties hold:*

- $\boldsymbol{\xi}^*$ *is an interior point of* Υ*,*
- $B(\boldsymbol{\xi}^*)$ *is nonsingular, and*
- $\boldsymbol{\xi}^*$ *is a regular point of* $A(\boldsymbol{\xi})$ *(i.e.,* $A(\boldsymbol{\xi})$ *has constant rank in some open neighborhood of* $\boldsymbol{\xi}$*)*

Using these elements, White (1982) showed that

Theorem 5.2 (Asymptotic normality) *Given Assumptions 5.1-5.6,*

$$\sqrt{n}(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}^*) \sim N(\mathbf{0}, V(\boldsymbol{\xi}^*)). \tag{5.6}$$

Moreover, $V_n(\widehat{\boldsymbol{\xi}}_n) \xrightarrow{a.s.} V(\boldsymbol{\xi}^*)$*, element by element.*

Under a correctly specified model, classical asymptotic normality of the maximum likelihood estimators can be recovered from Theorem 5.2. However, before this idea can be formalized, an extra assumption is needed.

Assumption 5.7 *The functions*

$$\left| \frac{\partial}{\partial \xi_\ell} \left[\frac{\partial f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k} \cdot f(\boldsymbol{y}, \boldsymbol{\xi}) \right] \right|$$

are dominated by functions integrable with respect to v*, for all* $\boldsymbol{\xi} \in \Upsilon$*, and the minimal support of* $f(\boldsymbol{y}, \boldsymbol{\xi})$ *does not depend on* $\boldsymbol{\xi}$*.*

Theorem 5.3 (Information Matrix Equivalence) *Given Assumptions 5.1-5.7, and given that* $h(\boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{\xi}_0)$ *belongs to* $\mathfrak{F}$*, then*

$$1. \quad \boldsymbol{\xi}^* = \boldsymbol{\xi}_0, \text{ and} \tag{5.7}$$

$$2. \quad A(\boldsymbol{\xi}_0) = -B(\boldsymbol{\xi}_0). \tag{5.8}$$

This theorem states that, under a correctly specified model, the information criterion (5.1) attains its unique minimum at $\boldsymbol{\xi}^* = \boldsymbol{\xi}_0$. Hence, $\widehat{\boldsymbol{\xi}}_n$ is a consistent estimator for $\boldsymbol{\xi}_0$, and

$$V(\boldsymbol{\xi}_0) = -A^{-1}(\boldsymbol{\xi}_0). \tag{5.9}$$

This theorem also implies that, under misspecification, the information matrix equivalence given by (5.8) does not necessarily hold. As a result, classical tests such as the Wald test may no longer be valid. Nevertheless, an appropriate correction of the Wald test can be obtained from Theorem 5.2. To this end, suppose that the null and alternative hypotheses are of the following form

$$\begin{aligned} H_0 : s(\boldsymbol{\xi}^*) &= 0, \\ H_A : s(\boldsymbol{\xi}^*) &\neq 0, \end{aligned}$$

where $s : \Re^p \to \Re^r$ is a continuous function such that its Jacobian $\nabla s(\boldsymbol{\xi}^*)$ is finite and of full rank r. The appropriate form of the Wald statistic is then given by

Theorem 5.4 (Wald Test)

$$n s(\widehat{\boldsymbol{\xi}}_n)^T [\nabla s(\widehat{\boldsymbol{\xi}}_n) V_n(\widehat{\boldsymbol{\xi}}_n) \nabla s(\widehat{\boldsymbol{\xi}}_n)^T]^{-1} s(\widehat{\boldsymbol{\xi}}_n) \sim \chi^2_r. \tag{5.10}$$

Additionally, from Theorem 5.3, it follows that, under a correctly specified model, $A(\boldsymbol{\xi}_0) + B(\boldsymbol{\xi}_0) = 0$. As a consequence, deviations from the model assumptions are expected to distort this equality. Therefore $A(\boldsymbol{\xi}^*) + B(\boldsymbol{\xi}^*)$ could be used as a potential indicator of misspecification. Note that, although these two matrices are unobservable, they can be consistently estimated using $A_n(\widehat{\boldsymbol{\xi}}_n)$ and $B_n(\widehat{\boldsymbol{\xi}}_n)$.

As discussed by White (1982), it may be prohibitive to base a test for misspecification on all elements of $A_n(\widehat{\boldsymbol{\xi}}_n) + B_n(\widehat{\boldsymbol{\xi}}_n)$. So, for simplicity reasons, we will focus only on its diagonal elements. Let $\boldsymbol{d}(\boldsymbol{y}, \boldsymbol{\xi})$ represent the $p \times 1$ vector with elements

$$d_k(\boldsymbol{y}, \boldsymbol{\xi}) = \left\{ \frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_k} \right\}^2 + \frac{\partial^2 \log f(\boldsymbol{y}, \boldsymbol{\xi})}{(\partial \xi_k)^2}. \tag{5.11}$$

Then,

$$\boldsymbol{D}_n(\widehat{\boldsymbol{\xi}}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}(\boldsymbol{y}_i, \widehat{\boldsymbol{\xi}}_n) \tag{5.12}$$

represents the $p \times 1$ vector containing the diagonal elements of $A_n(\widehat{\boldsymbol{\xi}}_n) + B_n(\widehat{\boldsymbol{\xi}}_n)$. Further, let $\boldsymbol{D}(\boldsymbol{\xi}) = \mathrm{E}[\boldsymbol{d}(\boldsymbol{y}, \boldsymbol{\xi})]$ and define the $p \times p$ Jacobian matrices

$$\begin{aligned} \nabla \boldsymbol{D}(\boldsymbol{\xi}) &= \mathrm{E}\left\{ \frac{\partial d_k(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_\ell} \right\}, \qquad \text{and} \\ \nabla \boldsymbol{D}_n(\boldsymbol{\xi}) &= \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial d_k(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_\ell} \right\}. \end{aligned} \tag{5.13}$$

Finally, let us consider

$$\begin{aligned} C(\boldsymbol{\xi}) &= \mathrm{E}\left\{ [\boldsymbol{d}(\boldsymbol{y}, \boldsymbol{\xi}) - \nabla \boldsymbol{D}(\boldsymbol{\xi}) A^{-1}(\boldsymbol{\xi}) \nabla \log f(\boldsymbol{y}, \boldsymbol{\xi})] \right. \\ &\quad \left. \times [\boldsymbol{d}(\boldsymbol{y}, \boldsymbol{\xi}) - \nabla \boldsymbol{D}(\boldsymbol{\xi}) A^{-1}(\boldsymbol{\xi}) \nabla \log f(\boldsymbol{y}, \boldsymbol{\xi})]^T \right\}. \end{aligned}$$

The following assumptions guarantee that $C(\boldsymbol{\xi})$ is well-defined.

Assumption 5.8 *The functions*

$$\frac{\partial d_k(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_\ell}$$

exist and are continuous functions of $\boldsymbol{\xi}$ *for each* $\boldsymbol{y}$.

Assumption 5.9 *The functions*

$$|d_k(\boldsymbol{y}, \boldsymbol{\xi}) d_m(\boldsymbol{y}, \boldsymbol{\xi})|, \qquad \left| \frac{\partial d_k(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_\ell} \right|, \qquad \textit{and} \qquad \left| d_k(\boldsymbol{y}, \boldsymbol{\xi}) \frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \xi_\ell} \right|$$

are dominated by functions integrable with respect to H *for all* $\boldsymbol{y}$ *and* $\boldsymbol{\xi}$ *in* Υ.

Assumption 5.10 $C(\boldsymbol{\xi}^*)$ *is nonsingular.*

White (1982) showed that $C(\boldsymbol{\xi}^*)$ is the asymptotic covariance matrix of $\sqrt{n}\boldsymbol{D}_n(\widehat{\boldsymbol{\xi}}_n)$, and a consistent estimator for $C(\boldsymbol{\xi}^*)$ is given by

$$\begin{aligned} C_n(\widehat{\boldsymbol{\xi}}_n) &= \frac{1}{n}\sum_{i=1}^{n} [\boldsymbol{d}(\boldsymbol{y}_i, \widehat{\boldsymbol{\xi}}_n) - \nabla \boldsymbol{D}_n(\widehat{\boldsymbol{\xi}}_n) A_n^{-1}(\widehat{\boldsymbol{\xi}}_n) \nabla \log f(\boldsymbol{y}_i, \widehat{\boldsymbol{\xi}}_n)] \\ &\quad \times [\boldsymbol{d}(\boldsymbol{y}_i, \widehat{\boldsymbol{\xi}}_n) - \nabla \boldsymbol{D}_n(\widehat{\boldsymbol{\xi}}_n) A_n^{-1}(\widehat{\boldsymbol{\xi}}_n) \nabla \log f(\boldsymbol{y}_i, \widehat{\boldsymbol{\xi}}_n)]^T. \end{aligned} \tag{5.14}$$

Using these elements, White (1982) proposed the following general test for model misspecification.

Theorem 5.5 (Information Matrix Test) *Given the Assumptions 5.1-5.10, and if the model is correctly specified, then*

1. $\sqrt{n}\boldsymbol{D}_n(\widehat{\boldsymbol{\xi}}_n) \sim N(0, C(\boldsymbol{\xi}_0))$,
2. $C_n(\widehat{\boldsymbol{\xi}}_n) \xrightarrow{a.s.} C(\boldsymbol{\xi}_0)$, *and* $C_n(\widehat{\boldsymbol{\xi}}_n)$ *is nonsingular almost surely for all* n *sufficiently large,*
3. *the Information Matrix Test (IMT) statistic*

$$\Im(n) = n\boldsymbol{D}_n^T(\widehat{\boldsymbol{\xi}}_n)C_n^{-1}(\widehat{\boldsymbol{\xi}}_n)\boldsymbol{D}_n(\widehat{\boldsymbol{\xi}}_n) \tag{5.15}$$

is distributed asymptotically as χ_p^2.

The IMT provides a unified framework for testing misspecification in a wide variety of settings. One could expect the test to be consistent against any alternative which renders the usual maximum likelihood inference invalid. If model misspecification is detected, this may imply inconsistency of some maximum likelihood estimators. In this case, proper inferences can be drawn for $\boldsymbol{\xi}^*$ using, for example, the corrected Wald test as described in Theorem 5.4, but inferences for the parameters of interest may be difficult to obtain. Clearly, in such a situation it is of the utmost importance to know how much bias may be introduced in the estimation procedures. If the maximum likelihood estimators are severely affected, then alternative approaches would need to be considered.

In the present work, we are mainly concerned with the impact of random-effects misspecification in generalized linear mixed models. Verbeke and Lesaffre (1997) studied in detail how estimation and inferential procedures associated with linear mixed models are affected by this type of misspecification. In the next chapter, we summarize their findings for this relevant special case.

Chapter 6

Random-effects Misspecification in Linear Mixed Models

Classical likelihood theory has shown that, if the assumed model is correctly specified, the maximum likelihood estimators are consistent and asymptotically normally distributed with the inverse of the Fisher information matrix as asymptotic covariance matrix. Similar results can be obtained in linear mixed models, when the random-effects distribution is misspecified as normal (Verbeke and Lesaffre, 1997). Let us start by introducing some model notation.

6.1 Model Notation

First, it is important to explicitly distinguish the correct model from the model used for parameter estimation. Let the correct model for the continuous outcomes $\boldsymbol{y}_i$, conditional on a vector of random effects $\boldsymbol{b}_i$, be specified as

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta}_0 + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon_i}. \tag{6.1}$$

Under this correct model, the vector $\boldsymbol{\beta}_0$ describes the true population mean, while the random effects $\boldsymbol{b}_i$ have zero mean and density function $f_0(\boldsymbol{b}_i|\boldsymbol{\psi}_0)$ with $\boldsymbol{\psi}_0$ a vector

of unknown parameters. Further, the error terms $\boldsymbol{\varepsilon_i}$ are assumed to be independent from the $\boldsymbol{b}_i$, and to follow a normal distribution with mean zero and covariance matrix $\sigma_0^2 I_{n_i}$. Marginally, this model induces a multivariate normal distribution for $\boldsymbol{y}_i$, with mean $X_i\boldsymbol{\beta}_0$ and covariance matrix $V_{i0} = Z_i D_0 Z_i^T + \sigma_0^2 I_{n_i}$, where $D_0 = D_0(\boldsymbol{\psi}_0) = \mathrm{Var}(\boldsymbol{b}_i)$.

On the other hand, for parameter estimation, we will assume that the responses $\boldsymbol{y}_i$, conditionally on the random effects $\boldsymbol{b}_i$ can by modeled by (3.10), where $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, D)$ and $\boldsymbol{\varepsilon_i} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 I_{n_i})$. In this case, the marginal distribution for $\boldsymbol{y}_i$ is a normal distribution with mean $X_i\boldsymbol{\beta}$ and covariance matrix $V_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$.

6.2 Consistency and Asymptotic Normality

Let $\boldsymbol{\xi}$ denote the vector of all parameters in model (3.10), including the fixed effects $\boldsymbol{\beta}$, all variance components in D, and σ^2; let $\boldsymbol{\xi}_0$ represent the vector of true parameter values $\boldsymbol{\beta}_0$, D_0 and σ_0^2. Maximum likelihood estimates $\widehat{\boldsymbol{\beta}}_n$, $\widehat{D}_n$ and $\widehat{\sigma}_n^2$ for the parameters in $\boldsymbol{\xi}$ are obtained by maximizing the marginal likelihood function for $\boldsymbol{y}_i$. Without assuming any specific structure for D, Verbeke and Lesaffre (1994, 1997) proved that

Theorem 6.1 (Consistency) *Under general regularity conditions, $\widehat{\boldsymbol{\beta}}_n$, $\widehat{D}_n$, and $\widehat{\sigma}_n^2$ are strongly consistent estimators for $\boldsymbol{\beta}_0$, D_0, and σ_0^2, as $n \to \infty$.*

This theorem implies that, under general regularity conditions, the vector $\boldsymbol{\xi}^*$, which minimizes the KLIC in (5.1), and to which the maximum likelihood estimators converge, equals the vector of true parameters $\boldsymbol{\xi}_0$. Hence, the maximum likelihood estimators for all parameters in the model, including the variance components, are consistent, even when the distribution of $\boldsymbol{b}_i$ is misspecified. Simulations with different distributions for the random effects have confirmed these results. However, these simulations also suggested that the rate of convergence heavily depends on the shape of the correct random-effects distribution, especially for the components in D.

Further, note that, although the inverse Fisher information matrix does not yield correct standard errors when the model is misspecified, valid asymptotic estimates for the standard errors of the maximum likelihood estimators can still be obtained using Theorem 5.2. Indeed, using the notation introduced by the Expressions (5.2)-(5.5), we have that

Theorem 6.2 (Asymptotic Normality) *Under general regularity conditions,* $\widehat{\boldsymbol{\xi}}_n$ *is asymptotically normally distributed with mean* $\boldsymbol{\xi}_0$ *and with the covariance matrix given by* $\frac{1}{n}A^{-1}(\boldsymbol{\xi}_0)B(\boldsymbol{\xi}_0)A^{-1}(\boldsymbol{\xi}_0)$*, as* $n \to \infty$*.*

Although calculation of $A(\boldsymbol{\xi}_0)$ and $B(\boldsymbol{\xi}_0)$ in Theorem 6.2 requires knowledge of the correct model, correct standard errors can still be obtained from the following theorem.

Theorem 6.3 (Correction of Standard Errors) *The results in Theorem 6.2 remain valid when* $A(\boldsymbol{\xi}_0)$ *and* $B(\boldsymbol{\xi}_0)$ *are replaced by* $A_n(\widehat{\boldsymbol{\xi}}_n)$ *and* $B_n(\widehat{\boldsymbol{\xi}}_n)$ *respectively.*

Note that the corrected asymptotic covariance matrix suggested by Theorem 6.3 corrects for possible misspecification of the random-effects distribution. To study the effect of this last correction, we can compare the corrected covariance with the naive uncorrected one obtained from classical likelihood theory. To this end, we can compare the corrected variance of any linear combination $\boldsymbol{v}^T\widehat{\boldsymbol{\xi}}_n$, given by $\frac{1}{n}\boldsymbol{v}^T A_n^{-1}(\widehat{\boldsymbol{\xi}}_n)B_n(\widehat{\boldsymbol{\xi}}_n)A_n^{-1}(\widehat{\boldsymbol{\xi}}_n)\boldsymbol{v}$, with the uncorrected variance $\frac{1}{n}\boldsymbol{v}^T A_n^{-1}(\widehat{\boldsymbol{\xi}}_n)\boldsymbol{v}$ by studying the following ratio

$$\lambda_{\min} \leq \frac{\boldsymbol{v}^T A_n^{-1}(\widehat{\boldsymbol{\xi}}_n)B_n(\widehat{\boldsymbol{\xi}}_n)A_n^{-1}(\widehat{\boldsymbol{\xi}}_n)\boldsymbol{v}}{\boldsymbol{v}^T A_n^{-1}(\widehat{\boldsymbol{\xi}}_n)\boldsymbol{v}} \leq \lambda_{\max}, \tag{6.2}$$

where $\lambda_{\min}$ ($\lambda_{\max}$) is the smallest (largest) eigenvalue of $-B_n(\widehat{\boldsymbol{\xi}}_n)A_n^{-1}(\widehat{\boldsymbol{\xi}}_n)$. Note that the left hand (right hand) inequality in (6.2) becomes an equality for $\boldsymbol{v}$ equal to an eigenvector associated with $\lambda_{\min}$ ($\lambda_{\max}$). Obviously, $\lambda_{\min} = \lambda_{\max} = 1$ would indicate that both inferences yield similar results. Therefore, $\lambda_{\min} \approx \lambda_{\max} \approx 1$ may be an indicator of the random effects being approximately normally distributed.

Through extensive simulations, Verbeke and Lesaffre (1997) showed that the corrected and uncorrected standard errors for the parameters in the mean structure are very similar, even when the random effects are not normally distributed. This was not observed for the parameters in D. The simulations showed that, in this case, the corrected standard errors clearly outperformed the uncorrected ones. However, it should be noted that occasionally, even the corrected standard errors for these parameters were not performing adequately. This was the case for instance when the random effects were generated from a lognormal distribution. Further, although the corrected standard errors are generally good estimates for the variability of the parameter estimators, it should be noted that they may still yield incorrect confidence intervals for small samples, due to the bias present in the parameter estimates.

6.3 Discussion

The implications of these results are clearly very important. Indeed, even if the random-effects distribution is misspecified, linear mixed models still yield reliable conclusions, as far as the mean and the covariance structure are correctly specified. The maximum likelihood estimators of the fixed effects remain consistent and, even though we need to correct the standard errors, this correction seems to have a minor impact on the standard errors of the mean structure. Thus, the misspecification of the random-effects distribution does not dramatically affect the type I error and the power for testing the parameters of the mean structure. In the next chapter, we will study via simulations whether this is also true for the generalized linear mixed model.

Chapter 7

Random-effects Misspecification in Generalized Linear Mixed Models: A Simulation Study

The conventional belief among data analysts seems to be that the choice of the random-effects distribution is not crucial for the quality of the inferences related to the regression coefficients. This believe appears to be reinforced by the results for the linear mixed model acquired by Verbeke and Lesaffre (1997) and previously described in Chapter 6. However, results obtained in recent years show that moving away from the realm of normality leads to qualitative differences. For instance, Neuhaus, Hauck, and Kalbfleisch (1992) examined the performance of a random-intercept logistic regression model with misspecified random-effect distribution. They showed that the maximum likelihood estimators of the model parameters are inconsistent but that the magnitude of the bias is typically small. Simulations by Chen, Zhang, and Davidian (2002) with a comparable model also indicate that the estimation of the regression coefficients may be subject to negligible bias only. According to Agresti, Caffo, and Ohman-Strickland (2004), the choice of the random-effects distribution seems to have,

in most situations, little effect on the maximum likelihood estimators. However, when there is a severe polarization of subjects, e.g., by omitting an influential binary covariate, this can affect the predictive qualities of characteristics involving the random effects as well as the fixed effects. Similarly, Heagerty and Kurland (2001) found substantial bias while using a logistic-normal model, when the variance of the random effects depends on measured covariates.

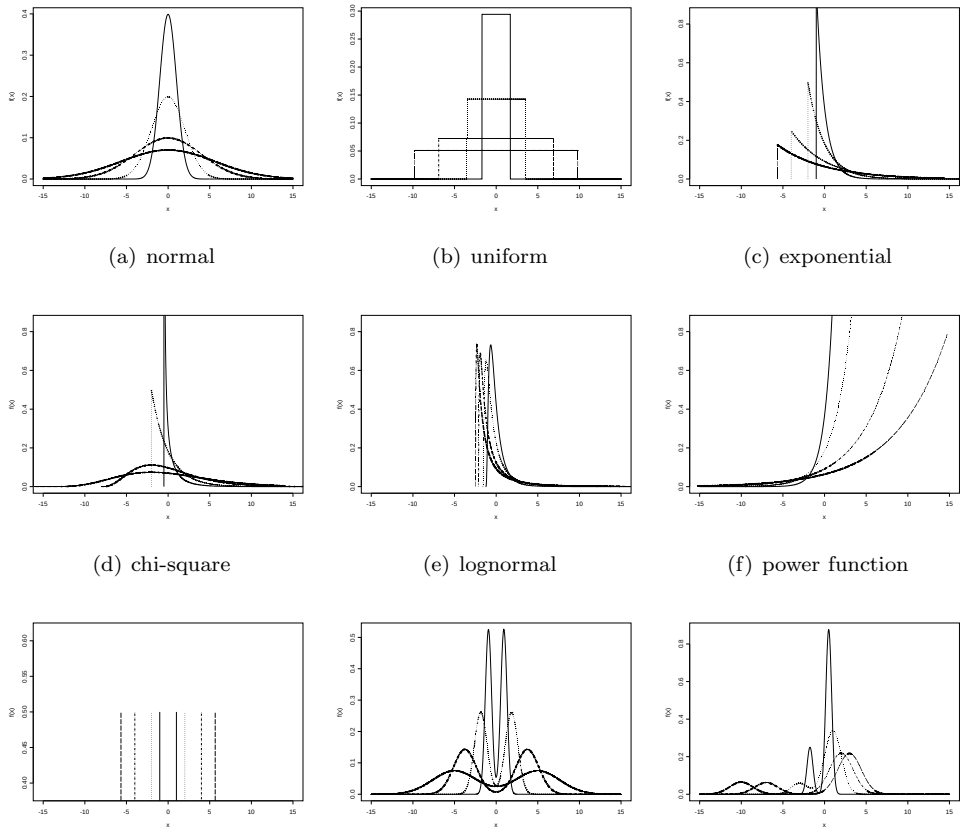
These results clearly illustrate the wide range of opinions which exist in the literature regarding the impact of misspecifying the random-effects distribution on the maximum likelihood estimators in generalized linear mixed models. In general, there seems to be a consensus about the presence of bias due to the misspecification. However, most of the research till now seems to indicate that this bias is typically small. It is important to note that each of these simulation studies was performed using a limited number of distributions for the random effects, and in most of them, only small variances for the random effects were considered. As we will show in the subsequent sections, the magnitude of this variance can have an important effect on the bias induced by the misspecification, where larger biases associate with larger variances. Moreover, as was seen from the analysis of the case study in Chapter 4, small variances may not always be realistic in some important practical settings. Another issue which has not received so much attention in the previous studies concerns the impact of the misspecification on commonly used inferential procedures such as the Wald test. Therefore, the main objective of this chapter is to use a wide set of simulations to study the impact of random-effects misspecification on the quality of the maximum likelihood estimators, as well as on the power and type I error rate of frequently used tests for the linear predictor parameters. The results presented in this chapter are based on Litière, Alonso, and Molenberghs (2007b).

7.1 Simulation Settings

In this simulation study, binary response data were generated using the logistic random-intercept model obtained from the analysis of the case study, i.e.,

$$\text{logit}\{P(y_{ij} = 1|b_i)\} = \beta_0 + \beta_1 z_i + \beta_2 t_j + b_i. \tag{7.1}$$

Recall that this model includes a binary covariate z_i and a within-cluster covariate t_j, reminiscent of time, with values 0, 1, 2, 4, 6, and 8. For the linear predictor parameters, values close to the estimates obtained from the analysis of the case study



(a) normal (b) uniform (c) exponential

(d) chi-square (e) lognormal (f) power function

(g) discrete (h) symmetric mixture (i) asymmetric mixture

Figure 7.1: *Graphical representation of the random-effects distributions with variance* σ^2_{0b} *used in the simulation study:* $\sigma^2_{b0} = 1$ *(solid line),* $\sigma^2_{b0} = 4$ *(dotted line),* $\sigma^2_{b0} = 16$ *(dash-dotted line) and* $\sigma^2_{b0} = 32$ *(dashed line).*

were chosen: $\beta_0^0 = -8$, $\beta_1^0 = 2$, and $\beta_2^0 = 1$. Further, 9 different distributions for the random intercept b_i, each with variance $\sigma^2_{0b} = 1$, 4, 16, and 32, were included in the study. These were a mean zero normal density, a uniform, an exponential, a chi-square, a lognormal, a power function distribution, a discrete distribution with equal probability at two support points, and finally both a symmetric and an asymmetric mixture of two normal densities (see Figure 7.1). Observe that this selection covers

a wide range of densities varying from very symmetric to very skewed, and with potentially very heavy tails. Further, the settings where $\sigma^2_{0b} = 16$ and 32 will help us to investigate scenarios with variances in the same order of magnitude as the one observed in the case study. On the other hand, the smaller values considered for σ^2_{0b} should allow us to study the performance of the maximum likelihood estimators in less extreme settings. In this way, we cover a wide range of practically relevant situations.

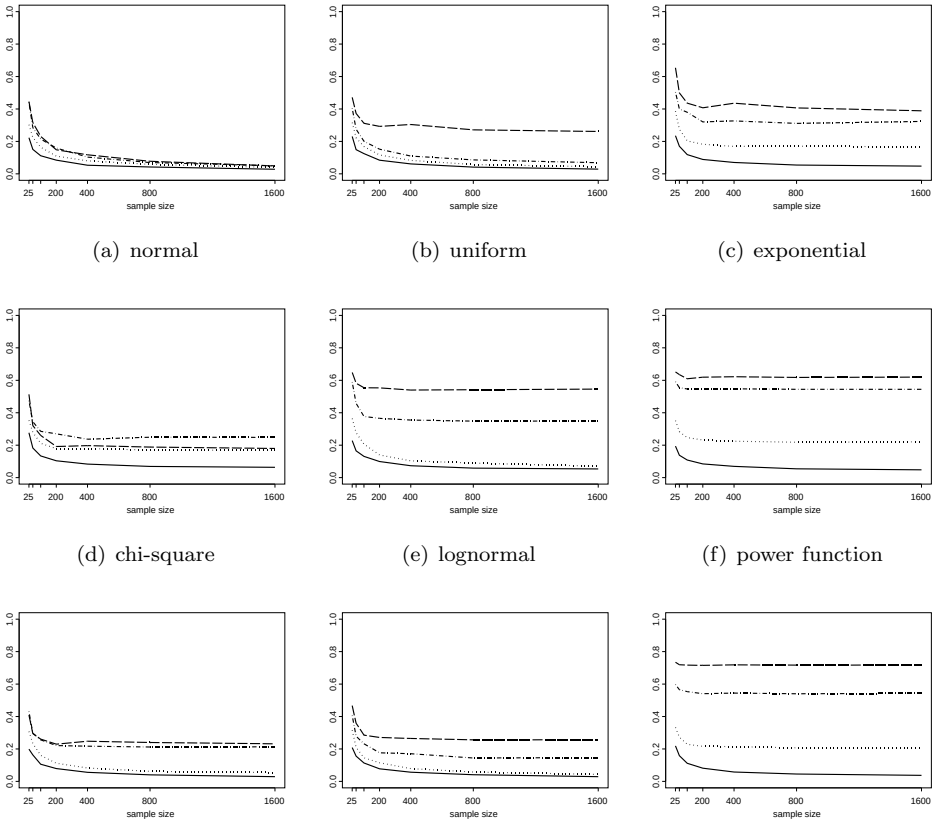
The simulations were performed with 7 different sample sizes, including 25, 50, 100, 200, 400, 800, and 1600 subjects. For each setting, 500 data sets were generated, and the model given by (7.1) was then fitted to the generated data, assuming normally distributed random effects. All analyses were carried out using the SAS procedure NLMIXED with adaptive Gaussian quadrature and 50 quadrature points to approximate the likelihood function.

7.2 Consistency

Consistency was studied through the evolution of the relative distance between the estimates and the real values, over increasing sample size. Let $\boldsymbol{\xi}_0 = (\beta^0_0, \beta^0_1, \beta^0_2, \sigma^2_{0b})^T$ represent the vector of true parameter values and $\widehat{\boldsymbol{\xi}}_n = (\widehat{\beta}_{0n}, \widehat{\beta}_{1n}, \widehat{\beta}_{2n}, \widehat{\sigma}^2_{bn})^T$ the corresponding vector of maximum likelihood estimates, then the relative distance between $\boldsymbol{\xi}_0$ and $\widehat{\boldsymbol{\xi}}_n$ is defined by

$$d_n(\boldsymbol{\xi}_0) = \frac{||\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0||}{||\boldsymbol{\xi}_0||}, \tag{7.2}$$

where $||.||$ denotes Euclidean distance. If the estimators remain consistent after misspecification of the random-effects distribution, then $d_n(\boldsymbol{\xi}_0)$ should converge to zero as the sample size increases. Figure 7.2 displays the evolution of the median relative distance $d_n(\boldsymbol{\xi}_0)$ over increasing sample size and for the different values of σ^2_{0b}. Although the bias observed for small values of σ^2_{0b} is generally negligible, it should be noted that for the exponential density, the power function, and the asymmetric mixture, the overall bias exceeds 20%, even for $\sigma^2_{0b} = 4$. On the other hand, substantial bias is always observed for larger values of σ^2_{0b}, especially for skewed densities such as the exponential, the lognormal function, the power function, and the asymmetric mixture. Therefore, skewness of the underlying random-effects distribution could be considered an indicator for strong inconsistency of the maximum likelihood estimators.



(a) normal (b) uniform (c) exponential

(d) chi-square (e) lognormal (f) power function

(g) discrete (h) symmetric mixture (i) asymmetric mixture

Figure 7.2: *Consistency of the parameter estimates - evolution of the median relative distance* $d_n(\boldsymbol{\xi}_0)$ *for each random-effects distribution, over increasing sample size and for the different values of the random-effects variance:* $\sigma^2_{b0} = 1$ *(solid line),* $\sigma^2_{b0} = 4$ *(dotted line),* $\sigma^2_{b0} = 16$ *(dash-dotted line) and* $\sigma^2_{b0} = 32$ *(dashed line).*

By definition, $d_n(\boldsymbol{\xi}_0)$ is an indicator of the relative bias for all parameters jointly. However, it is also of interest to know which parameters are most affected. To this end, we have displayed in Table 7.1 the median maximum likelihood estimates obtained for the variance component σ^2_b. Note that Table 7.1, as well as all other tables discussed in this thesis, displays only the results from the converging analyses. A lack of convergence occurred mainly for small σ^2_{0b}, in combination with small sample sizes.

Table 7.1: *Median of the maximum likelihood estimates $\widehat{\sigma}^2_{bn}$ of σ^2_{0b}, obtained from fitting the logistic-normal model given by (7.1) to the binary data generated using model (7.1), considering different sample sizes (n) and different random-intercept distributions with variance σ^2_{0b}.*

n	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
	Normal distribution				Uniform distribution				Exponential distribution			
25	1.335	3.427	14.702	30.378	1.380	3.864	16.509	39.841	1.769	5.188	20.359	46.930
50	1.046	3.845	15.681	31.986	1.022	4.117	16.787	41.612	1.215	5.119	21.200	46.754
100	1.013	3.740	15.645	32.524	0.977	4.355	17.160	41.663	1.326	5.276	21.784	45.450
200	0.977	3.980	15.723	31.445	0.913	4.248	17.289	41.374	1.211	5.392	21.534	45.081
400	0.997	3.888	15.832	32.382	0.938	4.218	17.122	42.058	1.288	5.440	21.687	46.251
800	1.001	4.030	15.761	32.106	0.975	4.210	17.186	40.952	1.321	5.504	21.437	45.247
1600	0.990	3.977	15.927	32.406	0.962	4.222	17.052	40.619	1.302	5.470	21.673	44.619
	Chi-square distribution				Lognormal distribution				Power function distribution			
25	1.847	5.009	18.180	37.477	1.392	3.802	9.336	14.111	1.034	2.327	6.272	10.960
50	1.420	5.182	19.014	37.874	1.305	4.222	9.358	13.854	0.747	2.038	6.149	11.207
100	1.355	5.259	20.117	38.540	1.286	4.074	10.004	13.906	0.680	1.904	6.221	11.928
200	1.358	5.386	20.597	37.271	1.275	4.296	9.560	13.854	0.685	1.967	6.184	11.577
400	1.445	5.532	20.197	38.390	1.302	4.248	9.691	14.225	0.661	1.979	6.183	11.499
800	1.456	5.500	20.376	38.165	1.320	4.397	9.827	14.147	0.672	2.009	6.217	11.610
1600	1.460	5.472	20.429	37.900	1.338	4.354	9.811	14.036	0.656	1.994	6.214	11.546
	Discrete distribution				Symmetric mixture of two normals				Asymmetric mixture of two normals			
25	1.101	4.173	19.939	37.158	1.225	3.885	17.655	41.795	1.138	2.167	5.992	7.998
50	1.029	4.272	19.404	38.611	0.963	4.345	18.222	41.058	1.003	2.155	5.932	8.358
100	1.035	4.309	20.077	39.341	0.922	3.995	18.856	40.093	0.806	2.103	6.095	8.388
200	0.976	4.332	19.788	39.225	0.987	4.218	18.443	40.845	0.799	2.109	6.306	8.432
400	0.981	4.330	19.873	40.156	0.999	4.227	18.911	40.714	0.788	2.116	6.241	8.324
800	1.003	4.396	19.812	39.884	0.980	4.220	18.497	40.462	0.805	2.163	6.315	8.354
1600	1.009	4.386	19.819	39.613	1.006	4.194	18.616	40.460	0.792	2.156	6.258	8.359

The proportion of non-converging analyses could be as high as 30%, when the data were generated using a power function distribution with $\sigma^2_{0b} = 1$ and only 25 subjects. However, this rate quickly drops to 7% and less, as of 100 subjects, or when σ^2_{0b} was increased to 4.

The results displayed in Table 7.1 clearly show that the estimates of the variance component are severely affected by the misspecification in most settings. Note that substantial bias can occur, even for small variance of the random intercept. This is clear in the results for the lognormal distribution, the power function distribution and the asymmetric mixture of two normals. Additionally, the direction of the bias can change depending on the true underlying distribution. For most of the distributions considered here, the variance component is overestimated. However, in the case of the lognormal distribution, the power function distribution and the asymmetric mixture, we observe serious underestimation of the variance of the random intercept.

On the other hand, the maximum likelihood estimates of the linear predictor parameters seem to be less affected by the misspecification (see Tables 7.2-7.4). The observed bias is generally small when the variance of the random intercept is small. However, more substantial biases associate with larger variances. For instance, when the random intercept was generated from an exponential or a lognormal distribution, with $\sigma^2_{0b} = 16$, bias of 15% and more for the intercept β_0, and 25% and more for the treatment effect β_1 can occur, even for relatively big sample sizes of 400 subjects. Although less dramatic, similar results can be observed for the power function distribution and the asymmetric mixture of two normals. Given that the estimate of the variance component is the only tool to study the variability of the true random-effects distribution, this highly biased estimate makes it difficult to evaluate whether or not problems can occur in the linear predictor as well.

Note that the relative bias of the time effect remained under 5% in all scenarios, even for moderate sample sizes of 100 subjects. This concurs with results obtained by Heagerty and Kurland (2001) and Chen *et al.* (2002). The latter argue that, since the estimation of the treatment effect and the intercept is subject to between-individual variation, we could expect misspecification of the random-effects distribution to affect the quality of these estimates. However, a covariate which changes within subjects, would be roughly orthogonal to between-individual effects and therefore less affected by the misspecification.

Table 7.2: *Median of the maximum likelihood estimates $\widehat{\beta}_{0n}$ of β_0^0, obtained from fitting the logistic-normal model given by (7.1) to the binary data generated using model (7.1), considering different sample sizes (n) and different random-intercept distributions with variance σ_{0b}^2 (note that $\beta_0^0 = -8$).*

n	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$
	Normal distribution				Uniform distribution				Exponential distribution			
25	-8.714	-8.089	-8.153	-8.185	-8.949	-8.410	-8.271	-7.321	-8.807	-8.697	-9.269	-11.347
50	-8.258	-8.181	-8.115	-8.111	-8.209	-8.069	-7.892	-8.257	-8.201	-8.223	-9.108	-10.773
100	-8.115	-7.924	-7.936	-8.150	-8.053	-8.042	-7.797	-7.991	-8.096	-8.253	-9.063	-10.107
200	-8.040	-8.099	-8.052	-8.027	-7.955	-8.098	-7.989	-8.051	-8.073	-8.295	-9.092	-9.816
400	-7.993	-7.956	-8.008	-8.032	-8.029	-8.006	-7.797	-7.861	-8.108	-8.302	-9.214	-10.269
800	-8.022	-8.016	-7.993	-8.027	-8.009	-8.011	-7.831	-7.785	-8.091	-8.349	-9.115	-10.130
1600	-7.978	-7.989	-8.008	-8.056	-7.997	-8.012	-7.807	-7.740	-8.079	-8.320	-9.181	-10.054
	Chi-square distribution				Lognormal distribution				Power function distribution			
25	-8.849	-8.486	-8.738	-8.763	-8.716	-8.566	-9.393	-9.638	-8.447	-8.173	-7.425	-6.454
50	-8.434	-8.407	-8.597	-9.149	-8.221	-8.665	-9.010	-9.782	-8.150	-7.888	-7.410	-6.767
100	-8.182	-8.430	-8.624	-8.738	-8.212	-8.432	-9.099	-9.485	-8.062	-7.855	-7.330	-6.851
200	-8.197	-8.357	-8.653	-8.606	-8.104	-8.390	-8.934	-9.566	-8.016	-7.776	-7.318	-6.924
400	-8.144	-8.372	-8.577	-8.697	-8.108	-8.360	-8.956	-9.395	-7.951	-7.765	-7.240	-6.670
800	-8.148	-8.369	-8.597	-8.699	-8.104	-8.442	-8.934	-9.463	-7.922	-7.748	-7.276	-6.749
1600	-8.160	-8.354	-8.599	-8.618	-8.084	-8.386	-8.951	-9.434	-7.934	-7.760	-7.262	-6.732
	Discrete distribution				Symmetric mixture of two normals				Asymmetric mixture of two normals			
25	-8.554	-8.514	-8.321	-7.817	-8.557	-8.234	-8.129	-8.594	-8.614	-7.882	-7.229	-6.450
50	-8.301	-8.204	-8.019	-7.184	-8.148	-8.172	-8.009	-7.741	-8.427	-7.555	-7.143	-6.343
100	-8.151	-8.145	-8.044	-7.354	-7.963	-8.013	-8.002	-7.972	-8.033	-7.577	-7.091	-6.350
200	-8.014	-8.115	-7.942	-7.069	-8.042	-8.049	-7.876	-7.874	-7.993	-7.493	-7.162	-6.314
400	-7.999	-8.149	-8.142	-7.410	-8.021	-8.048	-7.910	-7.847	-7.976	-7.476	-7.155	-6.272
800	-8.029	-8.117	-7.982	-7.348	-7.985	-8.056	-7.910	-7.819	-7.952	-7.472	-7.133	-6.330
1600	-8.040	-8.141	-8.039	-7.178	-8.005	-8.026	-7.937	-7.805	-7.959	-7.489	-7.122	-6.298

Table 7.3: *Median of the maximum likelihood estimates $\widehat{\beta}_{1n}$ of β_1^0 obtained from fitting the logistic-normal model given by (7.1) to the binary data generated using model (7.1), considering different sample sizes (n) and different random-intercept distributions with variance σ_{0b}^2 (note that $\beta_1^0 = 2$).*

n	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$
	Normal distribution				Uniform distribution				Exponential distribution			
25	2.154	2.038	2.102	2.110	2.432	2.323	2.916	0.794	2.330	2.472	2.650	4.356
50	2.116	2.059	2.097	2.109	2.047	1.905	1.960	2.570	2.052	2.069	1.974	2.955
100	2.012	2.028	2.039	2.038	1.929	1.794	1.759	2.330	2.043	2.139	2.290	2.642
200	2.021	2.048	2.029	1.979	1.984	1.895	1.954	2.375	2.055	2.200	2.416	2.147
400	2.006	1.969	1.977	2.023	1.986	1.898	1.873	1.858	2.116	2.233	2.525	2.891
800	2.009	2.008	1.983	2.022	1.993	1.971	1.924	1.845	2.063	2.202	2.364	2.609
1600	1.989	1.992	1.994	2.023	1.982	1.931	1.858	1.833	2.054	2.191	2.473	2.601
	Chi-square distribution				Lognormal distribution				Power function distribution			
25	2.401	2.656	1.959	2.311	2.236	2.338	2.478	2.573	2.115	1.889	1.582	1.339
50	2.166	2.260	2.165	2.427	2.074	2.257	2.460	2.636	2.044	1.946	1.768	1.926
100	2.129	2.213	2.080	2.090	2.095	2.202	2.462	2.602	1.998	1.919	1.736	1.807
200	2.095	2.211	2.161	2.113	2.047	2.199	2.374	2.599	1.973	1.896	1.749	1.969
400	2.085	2.240	2.121	2.046	2.060	2.202	2.375	2.562	1.952	1.836	1.638	1.634
800	2.092	2.217	2.098	2.088	2.076	2.227	2.423	2.552	1.951	1.858	1.728	1.706
1600	2.098	2.247	2.126	2.081	2.080	2.211	2.387	2.553	1.939	1.839	1.682	1.702
	Discrete distribution				Symmetric mixture of two normals				Asymmetric mixture of two normals			
25	2.254	2.180	1.809	2.883	2.193	1.861	1.981	1.931	2.086	2.107	1.778	1.670
50	2.088	1.939	1.553	1.356	2.048	2.004	1.822	1.904	2.104	1.920	1.794	1.781
100	2.070	1.982	1.642	1.688	2.015	1.989	1.731	2.013	1.969	1.944	1.717	1.782
200	2.031	2.033	1.751	1.340	2.033	1.960	1.723	1.909	1.969	1.924	1.796	1.703
400	2.066	2.117	1.993	2.041	2.020	1.952	1.720	1.928	1.974	1.924	1.798	1.703
800	2.036	2.040	1.793	1.895	1.984	1.956	1.786	1.886	1.970	1.915	1.760	1.715
1600	2.072	2.069	1.890	1.517	1.996	1.955	1.761	1.919	1.964	1.925	1.760	1.702

Table 7.4: *Median of the maximum likelihood estimates $\widehat{\beta}_{2n}$ of β_2^0, obtained from fitting the logistic-normal model given by (7.1) to the binary data generated using model (7.1), considering different sample sizes (n) and different random-intercept distributions with variance σ_{0b}^2 (note that $\beta_2^0 = 1$).*

n	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$	$\sigma_{0b}^2=1$	$\sigma_{0b}^2=4$	$\sigma_{0b}^2=16$	$\sigma_{0b}^2=32$
	Normal distribution				Uniform distribution				Exponential distribution			
25	1.089	1.015	1.011	1.000	1.099	1.038	0.985	0.984	1.091	1.046	0.994	1.065
50	1.037	1.012	1.017	1.022	1.023	1.017	0.996	1.000	1.021	1.000	1.037	1.051
100	1.013	0.988	0.998	1.018	1.013	1.021	0.981	0.974	1.016	0.997	1.005	1.006
200	1.005	1.013	0.999	1.012	0.995	1.011	0.988	0.973	1.004	0.994	0.998	1.001
400	0.999	0.998	1.000	1.008	1.006	1.006	0.981	0.978	1.002	1.000	1.004	1.008
800	1.001	1.002	0.997	1.002	1.000	1.005	0.980	0.972	1.000	1.004	1.002	1.006
1600	0.997	0.999	0.999	1.007	1.001	1.004	0.980	0.971	1.001	1.000	0.999	1.009
	Chi-square distribution				Lognormal distribution				Power function distribution			
25	1.093	0.996	1.023	1.003	1.081	1.027	1.045	1.036	1.055	1.062	1.029	0.990
50	1.041	1.013	1.000	1.038	1.015	1.048	1.006	1.039	1.036	1.021	1.023	0.987
100	1.013	1.009	1.018	1.022	1.009	1.014	1.024	1.003	1.018	1.012	1.012	0.992
200	1.004	1.002	1.009	1.001	1.003	1.006	0.995	1.005	1.012	1.007	1.008	0.995
400	1.007	1.006	1.011	1.009	1.004	1.004	1.001	0.993	1.003	1.008	1.008	0.982
800	1.006	1.005	1.003	1.002	1.002	1.008	1.002	1.000	0.999	1.004	1.007	0.985
1600	1.006	1.002	1.005	1.005	1.002	1.004	0.999	0.995	1.002	1.004	1.009	0.984
	Discrete distribution				Symmetric mixture of two normals				Asymmetric mixture of two normals			
25	1.070	1.069	1.000	0.880	1.062	1.043	0.986	1.007	1.098	1.044	1.046	1.028
50	1.044	1.030	0.997	0.878	1.025	1.035	0.991	0.989	1.058	1.016	1.022	1.016
100	1.016	1.018	1.003	0.883	1.003	1.004	1.006	0.976	1.010	1.011	0.997	1.006
200	1.001	1.014	0.987	0.874	1.002	1.016	0.984	0.965	1.006	1.001	1.019	1.003
400	0.996	1.012	0.989	0.871	1.004	1.011	0.987	0.965	1.003	1.002	1.011	0.997
800	0.999	1.016	0.986	0.871	0.998	1.007	0.985	0.967	0.998	1.003	1.013	1.005
1600	1.001	1.011	0.988	0.874	1.003	1.005	0.989	0.961	1.001	1.002	1.009	1.000

Further, to study the extent to which the results obtained from a logistic-normal model generalize to scenarios with more than one random effect, we generated binary responses using the model given by

$$\text{logit}\{P(y_{ij}=1|\boldsymbol{b}_i)\} = \beta_0 + \beta_1 z_i + \beta_2 t_j + b_{0i} + b_{1i}t_j, \tag{7.3}$$

which now also includes a random slope b_{1i} for time. For the linear predictor parameters, we considered $\beta_0^0 = -6$, $\beta_1^0 = 2$, and $\beta_2^0 = 1$. The random effects were generated from two multivariate distributions, including a multivariate normal $\boldsymbol{b}_i \sim \text{N}(\boldsymbol{0}, V)$ and a symmetric mixture of two multivariate normals $\boldsymbol{b}_i \sim \frac{1}{2}\text{N}(\boldsymbol{\mu}, D) + \frac{1}{2}\text{N}(-\boldsymbol{\mu}, D)$, where

$$D = \begin{pmatrix} d & d_{12} \\ d_{12} & d \end{pmatrix}.$$

In the case of the mixture, $\boldsymbol{\mu} = (4,4)^T$, $d = 1, 4$, and d_{12} was chosen such that $\rho = \text{corr}(b_{0i}, b_{1i}) = 0.5, 0.9$. Note that these values lead to the following overall covariance matrices $V = \text{Var}(\boldsymbol{b}_i) = \{\sigma_{k\ell}\}_{k,\ell=1,2}$,

$$\begin{aligned} V_1 &= \begin{pmatrix} 5 & 4.5 \\ 4.5 & 5 \end{pmatrix}, \quad V_2 = \begin{pmatrix} 5 & 4.9 \\ 4.9 & 5 \end{pmatrix}, \\ V_3 &= \begin{pmatrix} 8 & 6 \\ 6 & 8 \end{pmatrix}, \quad V_4 = \begin{pmatrix} 8 & 7.6 \\ 7.6 & 8 \end{pmatrix}. \end{aligned} \tag{7.4}$$

These same covariance matrices were then also used to generate the multivariate normal random effects $\boldsymbol{b}_i \sim \text{N}(\boldsymbol{0}, V)$.

The simulations were performed with 50 and 100 subjects. For each setting, 500 data sets were generated, and the model given by (7.3) was fitted to the generated data, assuming normally distributed random effects. The medians of the corresponding maximum likelihood estimates are shown in Table 7.5. Clearly, including an additional random effect increased the impact of the misspecification. Even though the variances used to generate the random effects were moderate, considerable bias is now also observed for all parameters in the linear predictor. Interestingly, and unlike the results previously obtained from the logistic-normal model, now the time effect is also severely underestimated. Likely, adding a random time effect induced a bias in the corresponding fixed effect estimate under misspecification. Finally, observe that the estimates of the variance components were again most affected in this scenario, where a large bias was observed for all elements in the covariance matrix.

Table 7.5: *Median of the maximum likelihood estimates obtained from fitting model (7.3) assuming normally distributed random effects, to binary data generated using model (7.3), considering different sample sizes (n) and random effects generated from a multivariate normal, i.e. $\boldsymbol{b}_i \sim N(\mathbf{0}, V)$, as well as a symmetric mixture of two multivariate normal densities, i.e., $\boldsymbol{b}_i \sim \frac{1}{2}N(\boldsymbol{\mu}, D) + \frac{1}{2}N(-\boldsymbol{\mu}, D)$, such that $Var(\boldsymbol{b}_i) = V$.*

	Real		Normal		Mixture	
	value	V	$n = 50$	$n = 100$	$n = 50$	$n = 100$
Fixed effects						
β_0	-6	V_1	-6.45	-6.14	-10.34	-10.52
		V_2	-6.34	-6.33	-9.76	-9.28
		V_3	-6.46	-6.19	-9.65	-9.37
		V_4	-6.71	-6.37	-10.06	-9.49
β_1	2	V_1	2.09	2.04	3.27	2.52
		V_2	2.10	2.09	3.19	2.43
		V_3	2.03	2.10	3.17	2.04
		V_4	2.11	2.01	3.69	2.73
β_2	1	V_1	1.04	1.04	-0.63	-0.13
		V_2	0.99	1.02	-0.14	-0.25
		V_3	0.98	0.88	-0.09	0.15
		V_4	0.99	0.94	-0.23	0.03
Variance structure						
σ_{11}	5	V_1	4.98	5.00	48.60	57.38
		V_2	5.32	5.89	45.43	49.18
	8	V_3	8.63	8.12	41.97	49.32
		V_4	8.87	8.32	49.16	59.76
σ_{22}	5	V_1	6.84	5.84	967.80	500.26
		V_2	6.41	6.37	969.68	597.18
	8	V_3	9.23	8.21	912.06	415.04
		V_4	11.17	10.20	962.09	453.80
σ_{12}	4.5	V_1	4.20	3.64	189.05	125.19
	4.9	V_2	4.10	4.82	174.16	134.14
	6	V_3	6.15	5.60	167.51	115.42
	7.6	V_4	7.76	7.36	176.39	131.02

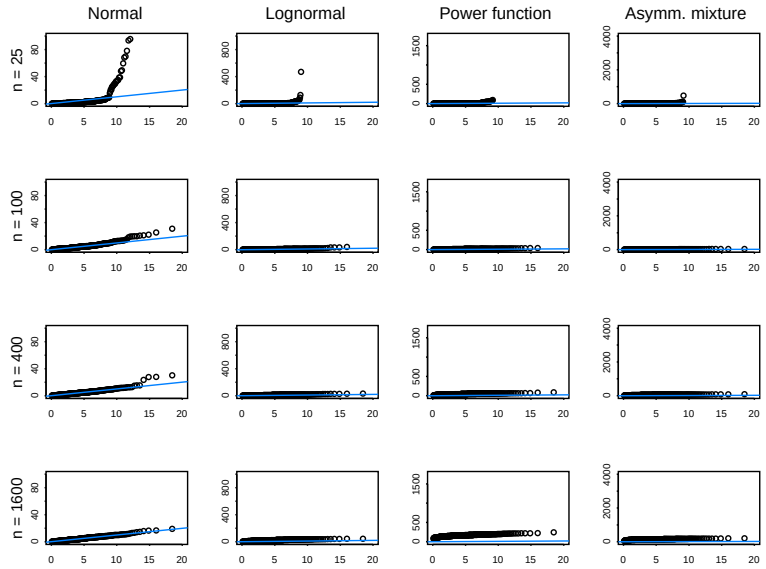
We should note that, when the random effects were generated from a mixture, we observed a high proportion of non-converging analyses (ranging between 57% and 72% of the total of 500 runs). We also observed that, as a result of the misspecification, in some of the simulations the procedure to maximize the likelihood had converged to an ill-conditioned maximum, leading to some extreme estimates. In any case, these results clearly illustrate that the impact of the random-effects misspecification is even worse in the presence of complicated covariance structures.

7.3 Asymptotic Normality

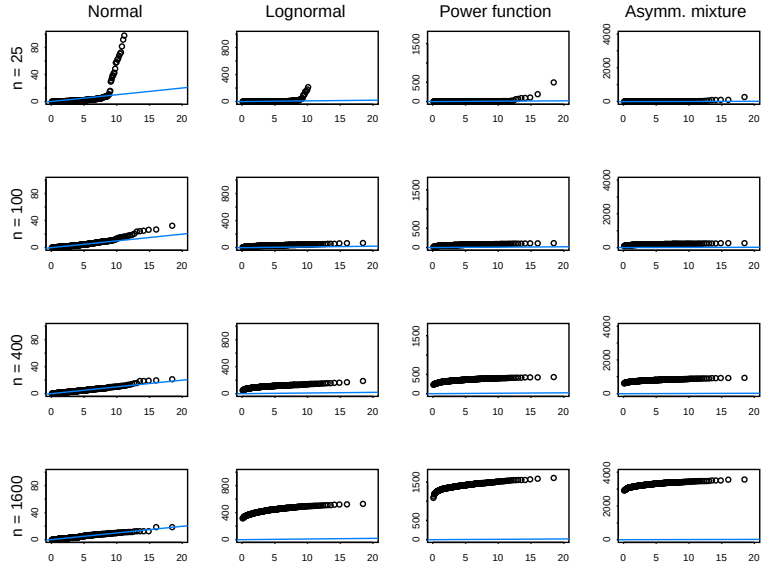
To check the extent to which asymptotic normality holds under misspecification, we constructed gamma plots of the maximum likelihood estimates of $\boldsymbol{\xi}$. Note that these plots are based on the ordered (from smallest to largest) squared general distances

$$d_k^2 = (\widehat{\boldsymbol{\xi}}_{kn} - \boldsymbol{\xi}_0)^T S^{-1} (\widehat{\boldsymbol{\xi}}_{kn} - \boldsymbol{\xi}_0),$$

where $\widehat{\boldsymbol{\xi}}_{kn}$ refers to the vector of maximum likelihood estimates obtained from the kth simulation, and S denotes the corresponding estimated sample covariance matrix, obtained from the simulated replicas. When the maximum likelihood estimates follow a multivariate normal distribution, then each of the squared distances d_k^2 should behave like a chi-square random variable (Johnson and Wichern, 1998). To verify this result, we have displayed in Figure 7.3 the pairs $(q_{\chi^2,4}[(k-1/2)/N], d_k^2)$, where N denotes the number of simulations and $q_{\chi^2,4}[(k-1/2)/N]$ is the $100(k-1/2)/N$ quantile of the chi-square distribution with 4 degrees of freedom. The graphical display of these pairs in Figure 7.3 is restricted to $\sigma_{0b}^2 = 4$ and 32, and to those settings for which consistency was most problematic, i.e. the lognormal, the power function and the asymmetric mixture distribution, for a selection of sample sizes. Further, gamma plots for the normal distribution were included to illustrate the asymptotic behavior of the maximum likelihood estimates under a correctly specified model. In this figure, a clear deviation from multivariate asymptotic normality can be observed when the random-effects distribution is misspecified, especially for large variance of the random effects. Interestingly, this deviation becomes more pronounced with increasing sample size.



(a) $\sigma^2_{b0} = 4$



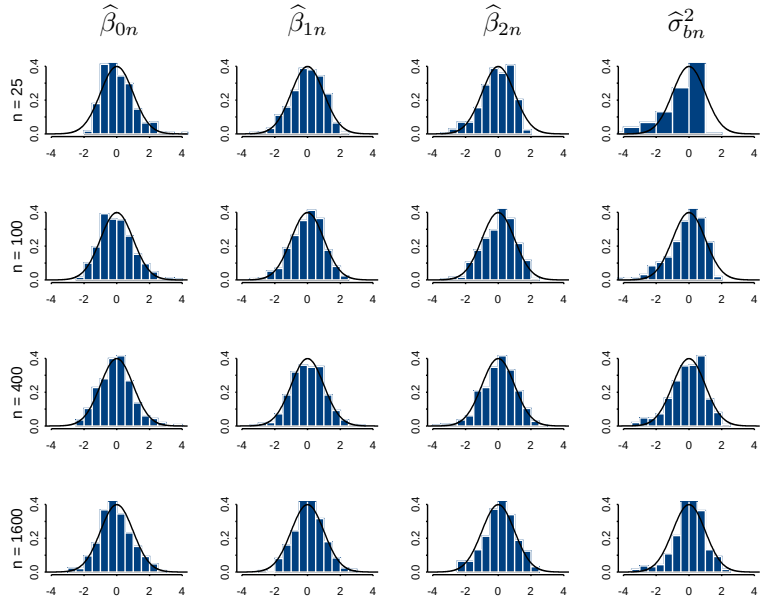
(b) $\sigma^2_{b0} = 32$

Figure 7.3: *Asymptotic normality of* $\widehat{\boldsymbol{\xi}}_n$ *- Gamma plots of the pairs* $(q_{\chi^2,4}[(k-1/2)/N], d_k^2)$*, where* d_k^2 *is obtained from fitting the logistic-normal model given by (7.1) to the binary data generated using model (7.1), with a random intercept sampled from a normal, lognormal, power function and asymmetric mixture distribution with variance* σ^2_{0b}*; each row represents a different sample size* (n)*.*

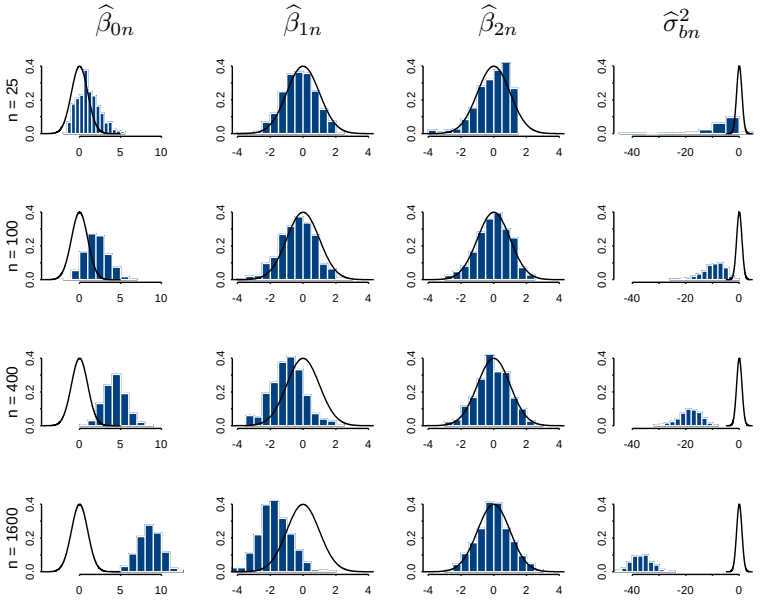
Further, to study asymptotic normality on the level of each parameter separately, histograms were constructed of the standardized (again w.r.t. $\boldsymbol{\xi}_0$) parameter estimates obtained from fitting the logistic-normal model given by (7.1). For instance, Figures 7.4(a) and (b) show the distributions of the standardized estimates, when the true random-effects distribution corresponds to a normal and an asymmetric mixture of two normal distributions, respectively. In comparison, the continuous curve represents the standard normal distribution. Note that we have limited the display to the rather extreme scenario of $\sigma^2_{0b} = 32$ for illustrative purposes.

Contrary to the gamma plots, the histograms in Figures 7.4(a) and (b) reveal approximately normally distributed maximum likelihood estimates. However, under the misspecified model, the histograms are shifted such that the estimates are not centered on the real parameter values given by $\boldsymbol{\xi}_0$. Further, note that this shift did not only occur for the highly biased variance component. When focusing on fixed effects such as the intercept and the treatment effect, although less pronounced, we observe the same behavior. They are over- and underestimated, respectively, when the true random-effects distribution is an asymmetric mixture of normal distributions with variance $\sigma^2_{0b} = 32$. Finally, our previous statements on how the time effects seems to be unaffected by the misspecification, are again confirmed in the histograms of $\widehat{\beta}_{2n}$ in Figure 7.4(b).

These findings concur with the results described in Chapter 5. It is clear from these simulations that, under misspecification of the random-effects distribution, the maximum likelihood estimators $\widehat{\boldsymbol{\xi}}_n$ are no longer consistent or asymptotically normal with respect to $\boldsymbol{\xi}_0$. The higher the variance and the skewness of the underlying random-effects distribution, the bigger the discrepancy between the two vectors. Further, it is important to realize that the estimates of the variance components are always subject to considerable bias when the random-effects distribution is misspecified. Although variance components are generally treated as nuisance parameters, this highly biased component can have an important impact in studies where they are of primary interest. This is the case, for instance, in fields like surrogate marker validation, the evaluation of the reliability of rating scales, or studies of the criterion and predictive validity of psychiatric scales.



(a) Normal distribution



(b) Asymmetric mixture of two normal distributions

Figure 7.4: *Asymptotic normality of* $\widehat{\boldsymbol{\xi}}_n$ *- Histograms of the standardized maximum likelihood estimates obtained from fitting the logistic-normal model given by (7.1) to binary response data generated using model (7.1), with the random intercept sampled from (a) a normal and (b) an asymmetric mixture, each with variance* $\sigma^2_{0b} = 32$*. Each row represents a different sample size* (n)*. The continuous curve represents the standard normal distribution.*

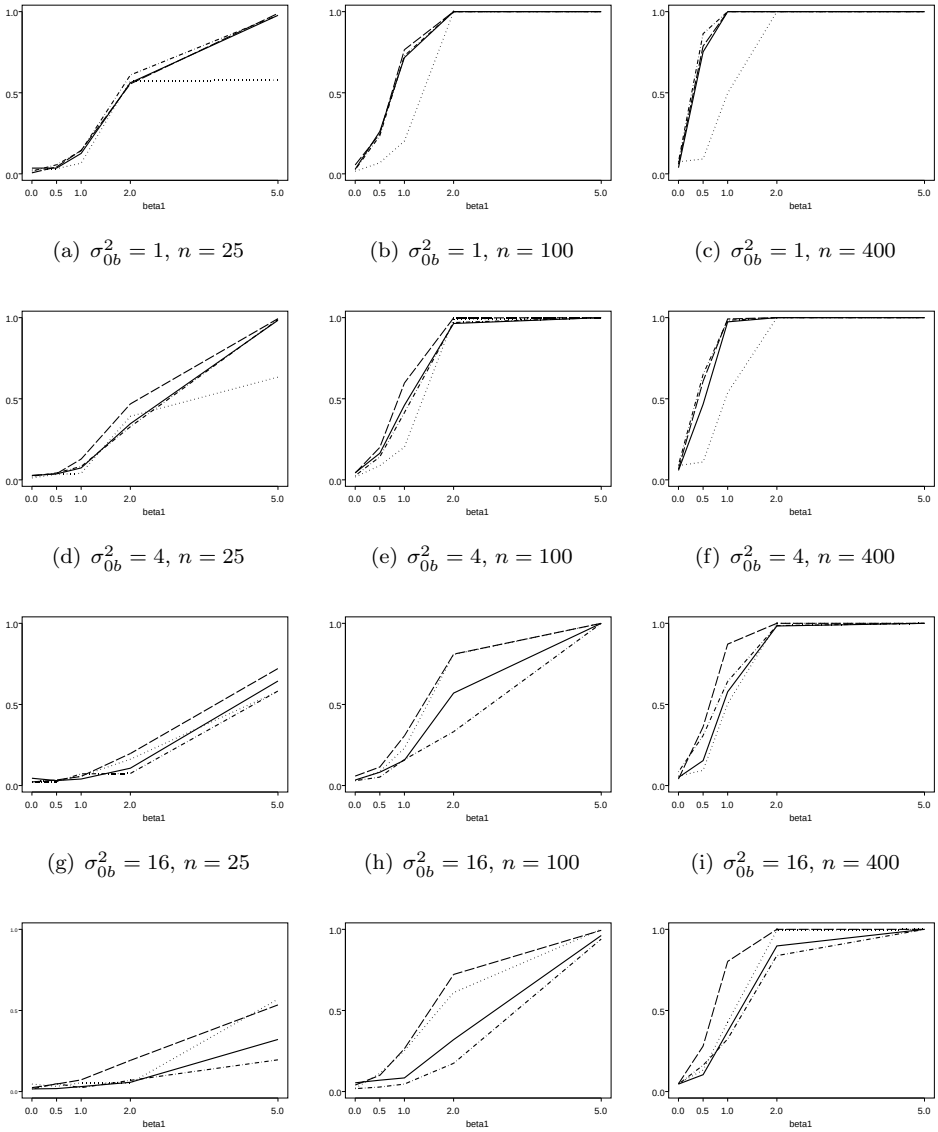
As stated before, the bias induced in the estimates of the linear predictor parameters appears to depend on the magnitude of the variance components, whereby large bias is associated with large random-effects variances. Clearly, in any practical situation, the bias present in the estimators for the variance components under misspecification, will make it hard to distinguish between the two scenarios, that is, small or larger variance components. As a consequence, it can also be difficult to determine how severe the impact on the parameter estimates can be.

7.4 Hypothesis Testing

In many situations, data analysts consider test statistics and corresponding p-values to evaluate, for example, whether or not a drug has a significant influence. Therefore, the impact of misspecifying the random-effects distribution on the type I and the type II error is very important from a practical point of view. Even though consistency has been studied to some extent in the literature, there does not seem to be too much research done on the behaviour of the test statistics.

To explore this effect, additional simulations were carried out for different values of the treatment effect β_1^0. Again, the binary responses were generated using the logistic random-intercept model given by (7.1) with $\beta_0^0 = -8$ and $\beta_2^0 = 1$. However, five different values for the treatment effect β_1^0 were considered: 0, 0.5, 1, 2 and 5. The simulations were performed for three different sample sizes, namely 25, 100, and 400 subjects, and considering four random-effects distributions with variance $\sigma_{0b}^2 = 1$, 4, 16 and 32, including the normal, the power function, the discrete, and the asymmetric mixture of two normals. For each setting, 500 data sets were generated, and the model given by (7.1) was used to analyze these generated data, assuming normally distributed random effects. We then determined the proportion of cases in which a treatment effect different from zero (at a 5% significance level) was detected. When $\beta_1^0 = 0$, this proportion corresponds to the type I error; otherwise, it represents the power of the test. The results of these simulations are displayed in Figure 7.5.

It is clear from these graphs that misspecification can severely affect the power of the analysis, depending on the shape and the variance of the real random-effects distribution. Actually, the power can be seriously affected even in settings where the random intercept accounts for a small variability. For example, let us consider in Figure 7.5 the graphs corresponding to a sample of 100 patients, when $\beta_1^0 = 1$. Even with $\sigma_{0b}^2 = 1$, the power to detect a significant treatment effect can drop as low as



(a) $\sigma^2_{0b} = 1$, $n = 25$ (b) $\sigma^2_{0b} = 1$, $n = 100$ (c) $\sigma^2_{0b} = 1$, $n = 400$

(d) $\sigma^2_{0b} = 4$, $n = 25$ (e) $\sigma^2_{0b} = 4$, $n = 100$ (f) $\sigma^2_{0b} = 4$, $n = 400$

(g) $\sigma^2_{0b} = 16$, $n = 25$ (h) $\sigma^2_{0b} = 16$, $n = 100$ (i) $\sigma^2_{0b} = 16$, $n = 400$

(j) $\sigma^2_{0b} = 32$, $n = 25$ (k) $\sigma^2_{0b} = 32$, $n = 100$ (l) $\sigma^2_{0b} = 32$, $n = 400$

Figure 7.5: *Power of the analysis with the logistic-normal model (7.1) to detect a significant treatment effect in binary response data generated using model (7.1), over a range of possible β_1^0 values, sample sizes n, and for 4 random-effects distributions with variance σ^2_{0b}: normal (solid line), power function (dotted line), discrete (dash-dotted line) and asymmetric mixture (dashed line).*

Table 7.6: *Type I error for detecting a significant treatment effect when $\beta_1^0 = 0$, when the logistic-normal model given by (7.1) is fitted to binary response data generated using model (7.1), considering different sample sizes (n) and a random intercept sampled from a normal (No), a power function (PF), a discrete (D) or an asymmetric mixture of two normal distributions (AM), each distribution with variance σ_{0b}^2. Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.*

	n	$\sigma_{0b}^2 = 1$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$
No	25	0.012	0.025	0.029	0.025
	100	0.041	0.052	0.050	0.026
	400	0.050	0.046	0.052	0.058
PF	25	0.008	0.023	0.036	0.016
	100	0.041	0.040	0.050	0.028
	400	0.046	0.064	**0.076**	0.050
D	25	0.023	0.012	0.014	0.004
	100	0.032	0.016	**0.084**	0.018
	400	0.048	**0.080**	0.024	**0.088**
AM	25	0.014	0.014	0.018	0.038
	100	0.053	0.066	0.036	0.038
	400	0.053	0.057	0.036	0.032

20% for the power function distribution, whereas for the correctly specified model, we observed a value around 70%. This makes it difficult to interpret negative results, i.e., do we fail to reject the null hypothesis because there is no real treatment effect or because of a lack of power due to misspecification?

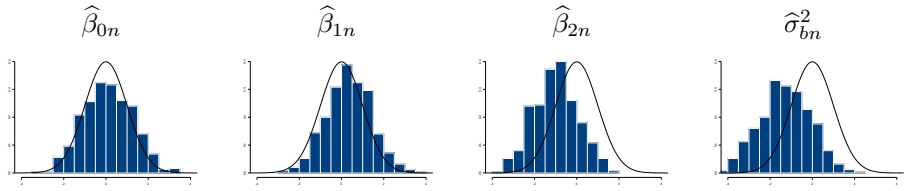
Interestingly, the corresponding type I error rate (presented in Table 7.6) rarely exceeded the specified 5% level of significance in all the scenarios displayed in Figure 7.5. These findings concur with results obtained by Neuhaus *et al.* (1992). Indeed, these authors showed for a similar logistic random-intercept model that when $\beta_1^0 = 0$, the corresponding maximum likelihood estimator consistently estimates zero. It is possible to prove that in this situation the type I error rate will be asymptotically preserved and therefore, a significant effect could be considered as a reliable result, even though caution may be needed in the interpretation of the point estimates. Whether this

statement holds also in more general settings than the one presented here, will be studied in Chapter 8.

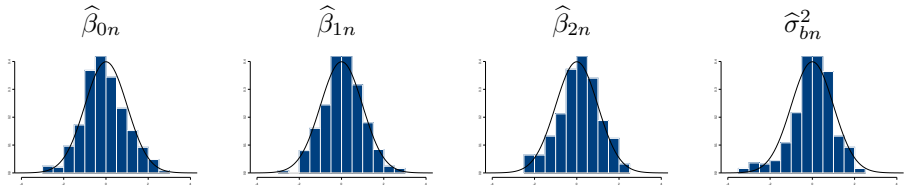
7.5 Numerical Precision

In general, lack of consistency can also be due to issues other than those related to the statistical procedures. Calculation of the maximum likelihood estimates is computationally intensive, and rounding errors and inadequate precision can interfere with the statistical results. In this section, we will show that inconsistency and lack of normality are not only due to model misspecification; arguably, they can also be ascribed to numerical approximations. Let us illustrate this with an example. Consider binary responses generated using the model given by (7.1) with $\beta_0^0 = -8$, $\beta_1^0 = 2$, and $\beta_2^0 = 1$, and with a random intercept sampled from a normal distribution with variance $\sigma_{0b}^2 = 32$. In total, 500 data sets were generated using this model, each with information on 1600 subjects, and the logistic-normal model (7.1) was used to analyze these generated data. During the first run, the SAS procedure NLMIXED was allowed to automatically determine the number of quadrature points. By default, the number of quadrature points is selected adaptively by evaluating the log-likelihood function until the relative distance between two successive likelihood calculations is sufficiently small. Figure 7.6(a) shows the results of these first analyses. Surprisingly, some lack of normality with respect to the real values already occurs in the estimation of the parameters, even when the random-effects distribution is correctly specified and the sample size is very large. When we compare this to Figure 7.6(b), where the estimates are obtained by fixing the number of quadrature points to 50, it is clear that the lack of normality was introduced by imprecise numerical approximation.

Further, let us reconsider the motivating case study. In order to investigate the accuracy of the numerical integration method for the estimation of the treatment effect in this example, the data were analyzed using the logistic-normal model given by (7.1), with varying numbers of quadrature points. The results are summarized in Table 7.7. With the default setting, considering in this case 3 quadrature points, the analysis resulted in a non-significant treatment effect. However, increasing the number of quadrature points leads to an increase of the estimated treatment effect, and as a result, we can also observe a change from a non-significant to a significant effect. In a similar example, Lesaffre and Spiessens (2001) showed that increasing the number of quadrature points could even change a treatment effect from being



(a) default number of quadrature points



(b) 50 quadrature points

Figure 7.6: *Histograms of standardized maximum likelihood estimates, obtained from analyzing simulated binary data with a correctly specified model, using NLMIXED with (a) the SAS default and (b) 50 quadrature points.*

Table 7.7: *Schizophrenia data. Effect of the number of quadrature points on the estimates of the treatment effect* β_1 *obtained from fitting the logistic-normal model (7.1) with NLMIXED.*

Q	$\widehat{\beta}_{1n}$	s.e.	p-value	-2loglik
3 (default)	1.89	1.21	0.1225	388.8
4	2.09	1.06	0.0508	379.1
5	2.09	0.99	0.0367	389.3
10	2.17	1.14	0.0594	384.4
20	2.14	1.08	0.0490	383.9
50	2.15	1.09	0.0499	384.0
100	2.15	1.09	0.0498	384.0

highly significant to not significant at all. In practice, this could lead to accepting a potentially inferior drug over a standard treatment or, as was the case for the schizophrenia data, not being able to detect the potential (borderline) benefit of risperidone over the conventional antipsychotic agents. It is therefore important to

remember that, even for an efficient method like adaptive Gaussian quadrature, it is not always safe to rely on the default settings of standard software.

Note that from Table 7.7 it follows that both the estimates and standard errors seem to stabilize when 20 or more quadrature points are used. The same can be observed for the other parameter estimates obtained from model (7.1). This result indicates that in this particular setting using 50 quadrature points would be sufficient to obtain adequate approximations to the likelihood function and its derivatives. Therefore, in all our simulations, analyses were performed using adaptive Gaussian quadrature with 50 quadrature points.

7.6 Summary

A commonly encountered perception among data analysts is that the choice of the random-effects distribution is not crucial for the quality of the inferences related with model parameters in generalized linear mixed models. However, this is not a generally valid conclusion. We found that the induced bias in the linear predictor estimators is negligible only when the variance of the underlying random-effects distribution is small. This was the case for $\sigma^2_{0b} = 1$ and 4 in our simulations. However, caution is necessary when the variance of the random effects is 16 or higher. Note that large random-effects variances are not exceptional in clinical trials, like our case study, when little variability in the response is expected in one of the two groups. In such a scenario, the linear predictor parameters, including the treatment effect, could be subject to considerable bias under misspecification.

On the other hand, the estimates of the variance components are always severely affected, even for small variances of the underlying random-effects distribution. Given that these estimates are the only available tool to study the variability of the true distribution, the bias induced by the misspecification can make it difficult to evaluate whether problems in the linear predictor will also occur. Additionally, as stated before, this bias in the variance components can have severe consequences in applications in which the main interest is in the association structure.

Finally, note that the power can also be affected in important ways by such misspecification, regardless of the variance of the random effects. Interestingly, the type I error rate seems to be maintained for the treatment effect. In the next chapter we will study whether this robustness of the type I error rate holds as well in more general scenarios.

Chapter 8

Type I Error under Misspecification of the Random-effects Distribution

In the previous chapter we studied how misspecification of the random-effects distribution can affect the properties of the maximum likelihood estimators. Interestingly, we found that the type I error associated with the treatment effect was maintained around its pre-specified level, irrespectively of the underlying random-effects distribution. In this chapter, we will further explore whether this finding is valid only for the settings considered in our simulations, or whether the type I error also remains unaffected under more general conditions for certain covariates in the model. The results in this chapter are based on Litière, Alonso and Molenberghs (2007a).

8.1 To Have or Not To Have an Associated Random Effect

First, let us note that a distinctive characteristic of the treatment covariate in the model given by (7.1) is that it does not have an associated random effect. It then seems plausible to presume that the type I error related to covariates, that are also

Table 8.1: *Type I error for detecting a significant intercept when $\beta_0^0 = 0$, when the logistic-normal model given by (7.1) is fitted to binary response data generated using model (7.1), considering different sample sizes (n) and a random intercept sampled from a normal (No), a power function (PF), a discrete (D) or an asymmetric mixture of two normal distributions (AM), each distribution with variance σ_{0b}^2. Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.*

	n	$\sigma_{0b}^2 = 1$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$
No	25	0.014	0.035	0.016	0.023
	100	0.042	0.048	0.040	0.034
	400	0.060	0.046	0.054	0.050
PF	25	0.019	0.031	0.028	0.022
	100	0.043	**0.164**	**0.320**	**0.370**
	400	**0.158**	**0.682**	**0.946**	**0.962**
D	25	0.021	0.046	**0.087**	0.073
	100	0.040	0.060	**0.136**	**0.156**
	400	**0.080**	**0.252**	**0.594**	**0.604**
AM	25	0.015	0.025	0.011	0.045
	100	0.030	**0.328**	**0.408**	**0.886**
	400	**0.076**	**0.924**	**0.986**	**1.000**

included in the random-effects structure, may be more sensitive to misspecification of the random-effects distribution. To clarify this issue, we designed a new simulation study. We generated binary responses using the model given by (7.1), considering now $\beta_0^0 = 0$, $\beta_1^0 = 2$, and $\beta_2^0 = 1$. Similar to the simulations described in the previous chapter, the random intercept was sampled from a normal distribution, a power function, a discrete and an asymmetric mixture of two normal distributions, each distribution with variance $\sigma_{0b}^2 = 1$, 4, 16, and 32. The sample sizes were set at 25, 100, and 400 subjects. For each setting, 500 data sets were generated, and the model given by (7.1) was fitted to the generated data, assuming normally distributed random effects. The performance of the Wald test associated with the intercept parameter β_0 can be seen in Table 8.1. Unlike for the treatment effect, the results presented in this table clearly illustrate that the type I error associated with β_0 can be severely

affected by the misspecification. Even when the variance of the random intercept is small, e.g., when $\sigma^2_{0b} = 1$, the type I error rate can be inflated up to 16%. With $\sigma^2_{0b} = 4$, we observed a type I error as high as 68% when the random intercept was sampled from a power function distribution, or even up to 92% using an asymmetric mixture.

Remarkably, the situation seems to worsen as the sample size increases. This could be explained using the results introduced in previous chapters. Indeed, in Chapters 5 and 7 we saw that under misspecification, the maximum likelihood estimator $\widehat{\beta}_{0n}$ is consistent, not with respect to the real value of the parameter, i.e., β^0_0, but with respect to β^*_0, the value of β_0 which minimizes the KLIC (5.1). Likely, in the settings considered in our simulations $\beta^*_0 \neq 0$ even though $\beta^0_0 = 0$. As a consequence, the Wald test implemented in SAS is not testing the hypothesis $H_0 : \beta^0_0 = 0$ as we would expect, but rather the new hypothesis $H_0 : \beta^*_0 = 0$. Obviously, larger sample sizes would increase the power to detect any deviation of β^*_0 from zero. This results in the observed inflation of the type I error associated with the hypothesis of interest, i.e., $H_0 : \beta^0_0 = 0$.

The previous results suggest that the type I error rate could be robust for parameters which do not have an associated random effect, like for instance the treatment effect β_1 in (7.1). On the other hand, the type I error could be severely affected for parameters which do have a random counterpart, like for the intercept β_0. The question remains whether this is true only for the specific model used in these simulations, or whether the type I error also remains unaffected in more general situations. The following theorem will help us to answer this question.

Theorem 8.1 *Let y_{ij} denote the jth measurement for the ith subject, with $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Conditional on a vector $\boldsymbol{b}_i$ of individual random effects for subject i, it is assumed that all responses y_{ij} are independent with density belonging to the exponential family (3.1), where θ_{ij} is modeled as*

$$\theta_{ij} = \eta(\beta_0 + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i), \tag{8.1}$$

and $\eta(.)$ denotes a known function, β_0 is an intercept, $\boldsymbol{x}_{ij} = (\boldsymbol{x}^M_{ij}, \boldsymbol{x}^R_{ij})$ denotes a p-dimensional vector of covariates with $\boldsymbol{x}^M_{ij} \bigcap \boldsymbol{x}^R_{ij} = \emptyset$, $\boldsymbol{z}_{ij} = \boldsymbol{x}^R_{ij}$ is a q-dimensional vector, $\boldsymbol{\beta} = (\boldsymbol{\beta}^M, \boldsymbol{\beta}^R)$ is a vector of fixed parameters and $\boldsymbol{b}_i$ is a vector of random effects assumed to follow a density $f(\boldsymbol{b}_i, D)$ with $E(\boldsymbol{b}_i) = \boldsymbol{0}$. Without loss of generality, the covariates are assumed to be centered around zero, i.e. $E(\boldsymbol{x}_{ij}) = \boldsymbol{0}$.

If $h(\boldsymbol{b}_i)$ represents the true random-effects distribution ($h(\boldsymbol{b}_i) \neq f(\boldsymbol{b}_i, D)$) and if for certain subset $\boldsymbol{x}_{Sij}^M$ of $\boldsymbol{x}_{ij}^M$, the vector of true parameter values $\boldsymbol{\beta}_S^{M0} = \mathbf{0}$, then under Assumptions 5.1-5.3, $\boldsymbol{\beta}_S^{M*}$, which minimizes the KLIC, is also zero. Therefore, the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_{Sn}^M$, based on a model with a misspecified random-effects distribution, satisfies

$$\widehat{\boldsymbol{\beta}}_{Sn}^M \xrightarrow{P}_n \mathbf{0}. \tag{8.2}$$

The general idea of the proof of Theorem 8.1 is as follows (the full proof can be found in Appendix B). For simplicity of notation we will work out the proof for $\boldsymbol{x}_{Sij}^M = \boldsymbol{x}_{ij}^M$. The proof for any other subset $\boldsymbol{x}_{Sij}^M$ can be obtained in a similar way.

First note that there always exists a lower triangular matrix U, so that $\boldsymbol{b}_i = U\boldsymbol{a}_i$ with $E(\boldsymbol{a}_i) = \mathbf{0}$ and $V(\boldsymbol{a}_i) = I$. This allows to write (8.1) as $\theta_{ij} = \eta(\beta_0 + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T U\boldsymbol{a}_i)$. Let us further denote by H and F the true and the assumed distribution functions of the random effects. According to White (1982), the maximum likelihood estimator of $\boldsymbol{\xi} = (\beta_0, \boldsymbol{\beta}, U)$ converges to the unique value $\boldsymbol{\xi}^* = (\beta_0^*, \boldsymbol{\beta}^*, U^*)$ which minimizes the KLIC (5.1), i.e. $\boldsymbol{\xi}^*$ minimizes

$$I(H : F, \boldsymbol{\xi}) = E_{\boldsymbol{x}} E_{\boldsymbol{y}|\boldsymbol{x}} \log\left\{ \frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0, \boldsymbol{x}, \boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{x}, \boldsymbol{z})} \right\}, \tag{8.3}$$

where the expectation is taken with respect to the true model. In the previous expression,

$$\begin{aligned} f_H(\boldsymbol{y}|\boldsymbol{\xi}_0, \boldsymbol{x}, \boldsymbol{z}) &= \int \prod_j \exp[\varphi^{-1}\{y_j\theta_j^0 - \psi(\theta_j^0)\} + c(y_j, \varphi)] dH(\boldsymbol{a}), \\ f_F(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{x}, \boldsymbol{z}) &= \int \prod_j \exp[\varphi^{-1}\{y_j\theta_j - \psi(\theta_j)\} + c(y_j, \varphi)] dF(\boldsymbol{a}), \end{aligned}$$

with

$$\begin{aligned} \theta_j^0 &= \eta(\beta_0^0 + \boldsymbol{x}_j^T\boldsymbol{\beta}^0 + \boldsymbol{z}_j^T U^0\boldsymbol{a}), \\ \theta_j &= \eta(\beta_0 + \boldsymbol{x}_j^T\boldsymbol{\beta} + \boldsymbol{z}_j^T U\boldsymbol{a}). \end{aligned}$$

For simplicity of notation, the subject index i has been omitted from the previous equations. To find $\boldsymbol{\xi}^*$ we have to differentiate (8.3) with respect to β_0, $\boldsymbol{\beta}$ and U. This

leads to the following system of simultaneous equations

$$E_{\boldsymbol{x}}\left[\int \lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})\left\{\int \sum_i k_i(\boldsymbol{x},\boldsymbol{a})dF(\boldsymbol{a})\right\}dy\right] = 0, \tag{8.4}$$

$$E_{\boldsymbol{x}}\left[\int \lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})\left\{\int \sum_i \boldsymbol{x}_i^M k_i(\boldsymbol{x},\boldsymbol{a})dF(\boldsymbol{a})\right\}dy\right] = 0, \tag{8.5}$$

$$E_{\boldsymbol{x}}\left[\int \lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})\left\{\int \sum_i \boldsymbol{x}_i^R k_i(\boldsymbol{x},\boldsymbol{a})dF(\boldsymbol{a})\right\}dy\right] = 0, \tag{8.6}$$

$$E_{\boldsymbol{x}}\left[\int \lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})\left\{\int \sum_i Q_i k_i(\boldsymbol{x},\boldsymbol{a})dF(a)\right\}dy\right] = 0, \tag{8.7}$$

where $\lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z}) = f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})/f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})$ and

$$k_i(\boldsymbol{x},\boldsymbol{a}) = \eta'(\boldsymbol{\xi})\varphi^{-1}\{y_i - \psi'(\theta_i)\}\prod_j(\exp[\varphi^{-1}\{y_j\theta_j - \psi(\theta_j)\} + c(y_j,\varphi)]).$$

If $\boldsymbol{\beta}^{M0} = \boldsymbol{\beta}^{M*} = \mathbf{0}$, then $\lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})$, $\eta(\beta_0^* + \boldsymbol{x}_i^T\boldsymbol{\beta}^* + \boldsymbol{z}_i^T U^*\boldsymbol{a})$, $\psi(\theta_i^*)$, $\psi'(\theta_j^*)$ and $k_i(\boldsymbol{x},\boldsymbol{a})$ are functions which are independent of $\boldsymbol{x}_i^M$. Hence, the left-hand side of (8.5) is zero for all β_0^0, $\boldsymbol{\beta}^{R0}$, U^0, β_0^*, $\boldsymbol{\beta}^{R*}$ and U^*. Note now that equations (8.4), (8.6), and (8.7) determine β_0^*, $\boldsymbol{\beta}^{R*}$, and U^* in terms of β_0^0, $\boldsymbol{\beta}^{R0}$, and U^0. Thus, when $\boldsymbol{\beta}^{M0} = \boldsymbol{\beta}^{M*} = \mathbf{0}$, we have found the unique solution for $\boldsymbol{\xi}^*$ given by $(\beta_0^*, \mathbf{0}, \boldsymbol{\beta}^{R*}, U^*)$. Therefore, if $\boldsymbol{\beta}^{M0} = \mathbf{0}$, then $\boldsymbol{\beta}^{M*} = \mathbf{0}$, and the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_n^M$ consistently estimates zero.

It is important to note that this result, as well as the ones that will follow, should be interpreted in an asymptotic way, i.e., when n tends to infinity. From a practical perspective these results can be considered approximately correct when n is sufficiently large. How large is sufficiently large, is difficult to establish theoretically and will be studied later on using simulations.

Theorem 8.1 implies that if the parameters associated with a subset of variables, which are not included in the random-effect structure, equal zero then the corresponding maximum likelihood estimators will consistently estimate zero. The main implication of this theorem is stated in the following corollary.

Corollary 8.1 *Consider the hypothesis testing problem*

$$H_0 : \boldsymbol{\beta}_S^{M0} = \mathbf{0} \qquad vs \qquad H_1 : \boldsymbol{\beta}_S^{M0} \neq \mathbf{0}, \tag{8.8}$$

and the corresponding Wald test statistic $W = (\widehat{\boldsymbol{\beta}}_{Sn}^M)^T \widehat{V}_n^{-1}(\widehat{\boldsymbol{\beta}}_{Sn}^M)$*, where* $\widehat{V}_n$ *is the sandwich estimator of the asymptotic covariance matrix corresponding to the maximum likelihood estimate* $\widehat{\boldsymbol{\beta}}_{Sn}^M$*, calculated using a misspecified model. Let* p_0 *denote the*

dimension of $\boldsymbol{\beta}_S^{M0}$*. Then, given the assumptions stated in Theorem 8.1, the type I error rate associated with the critical region* $W > \chi^2_{p_0,1-\frac{\alpha}{2}}$ *is asymptotically preserved, even under misspecification of the random-effects distribution, i.e.,*

$$P(W > \chi^2_{p_0,1-\frac{\alpha}{2}} \mid \boldsymbol{\beta}_S^{M0} = \mathbf{0}) \leq \alpha \tag{8.9}$$

Proof

White (1982), studying the hypothesis testing problem

$$H_0 : \boldsymbol{\beta}_S^{M*} = \mathbf{0} \qquad \text{vs} \qquad H_1 : \boldsymbol{\beta}_S^{M*} \neq \mathbf{0}, \tag{8.10}$$

showed that the Wald statistic $W = (\widehat{\boldsymbol{\beta}}_{Sn}^{M})^T \widehat{V}_n^{-1} (\widehat{\boldsymbol{\beta}}_{Sn}^{M})$ under the null hypothesis follows a χ^2 distribution with p_0 degrees of freedom (see Theorem 5.4). It then follows that asymptotically

$$\alpha = P(W > \chi^2_{p_0,1-\frac{\alpha}{2}} \mid \boldsymbol{\beta}_S^{M*} = \mathbf{0}). \tag{8.11}$$

However, in our context, we are not directly interested in the hypothesis defined in (8.10) but in testing (8.8). We propose to test (8.8) using the same testing statistic and critical region previously defined for (8.10), and we will find an upper bound for $P(W > \chi^2_{p_0,1-\frac{\alpha}{2}} \mid \boldsymbol{\beta}_S^{M0} = \mathbf{0})$.

Note that, if $\boldsymbol{\beta}_S^{M0} = \mathbf{0}$ then, according to Theorem 8.1, also $\boldsymbol{\beta}_S^{M*} = \mathbf{0}$, and as a consequence

$$P(W > \chi^2_{p_0,1-\frac{\alpha}{2}} \mid \boldsymbol{\beta}_S^{M0} = \mathbf{0}) \leq P(W > \chi^2_{p_0,1-\frac{\alpha}{2}} \mid \boldsymbol{\beta}_S^{M*} = \mathbf{0}) = \alpha. \tag{8.12}$$

Therefore, even under misspecification of the random effects, the type I error associated with (8.8) and W will be upper-bounded by α. □

This corollary implies that the type I error will be maintained, even under a misspecified random-effects distribution, provided that the corresponding subset of covariates is not included in the random-effects structure. Therefore, the corollary fully explains our previous findings. Indeed, the treatment variable in the logistic random-intercept model given by (7.1) is not included in the random-effects structure. Therefore, in this case, the type I error shown in Table 7.6 was maintained in almost all settings, even with relatively small sample sizes. Unlike the treatment variable, the intercept does have an associated random effect. This means that β_0^* may be different from zero, and as a result the type I error, shown in Table 8.1, is severely affected.

Table 8.2: *Type I error, based on corrected standard errors, for detecting a significant treatment effect when $\beta_1^0 = 0$, and a significant intercept when $\beta_0^0 = 0$, when the logistic-normal model given by (7.1) is fitted to binary response data generated using model (7.1), considering different sample sizes (n) and a random intercept sampled from a normal (No), a power function (PF), a discrete (D) or an asymmetric mixture of two normal distributions (AM), each distribution with variance σ^2_{0b} =1, 4, 16 and 32. Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.*

		$\beta_1^0 = 0$				$\beta_0^0 = 0$			
	n	$\sigma^2_{0b} = 1$	$\sigma^2_{0b} = 4$	$\sigma^2_{0b} = 16$	$\sigma^2_{0b} = 32$	$\sigma^2_{0b} = 1$	$\sigma^2_{0b} = 4$	$\sigma^2_{0b} = 16$	$\sigma^2_{0b} = 32$
No	25	0.015	0.029	0.047	0.041	0.020	0.043	0.022	0.021
	100	0.036	0.052	0.050	0.028	0.044	0.052	0.038	0.036
	400	0.052	0.044	0.054	0.058	0.060	0.046	0.054	0.050
PF	25	0.031	0.030	0.056	0.028	0.024	0.046	0.049	0.065
	100	0.038	0.038	0.050	0.026	0.045	**0.174**	**0.308**	**0.366**
	400	0.046	0.060	**0.074**	0.050	**0.174**	**0.708**	**0.952**	**0.966**
D	25	0.051	0.032	0.026	0.022	0.029	0.054	0.064	**0.082**
	100	0.034	0.018	**0.100**	0.018	0.040	0.052	**0.098**	**0.098**
	400	0.050	**0.078**	0.024	**0.088**	**0.074**	**0.212**	**0.494**	**0.510**
AM	25	0.027	0.032	0.034	0.040	0.025	0.042	0.071	**0.249**
	100	0.048	0.064	0.032	0.036	0.028	**0.354**	**0.428**	**0.892**
	400	0.053	0.057	0.036	0.032	**0.080**	**0.938**	**0.992**	**0.998**

Clearly, both theorems can play a relevant role in studies where randomization is used. For example, in a clinical trial, where patients are randomized, the treatment variable will usually not be included in the random-effects structure and therefore, if a significant treatment effect is observed, one could generally be confident about this result.

8.2 The Sandwich Correction

So far, only the inverse of the Fisher information matrix has been used in all our analyses and simulations to obtain standard errors for the estimates of the model parameters. However, as discussed in Chapter 5, this matrix only yields valid results

when the model is correctly specified. In the presence of model misspecification, appropriate standard errors can be obtained by replacing the estimate of the asymptotic covariance matrix by $\frac{1}{n}A_n^{-1}(\boldsymbol{\xi})B_n(\boldsymbol{\xi})A_n^{-1}(\boldsymbol{\xi})$ (see Theorem 5.2). The results presented in Table 8.2 now reflect the impact of using such corrected standard errors on the type I error rates from the previously presented simulations. Clearly, in these settings, the use of the sandwich correction had only a mild impact on the results. Therefore, it seems unnecessary to carry out some extra programming in order to obtain this correction. In principle, the results offered in the standard SAS output could be directly used to test the hypothesis of interest.

8.3 Implications for the Schizophrenia Data

Let us now recall the case study on schizophrenia, analyzed in Chapter 4. Assuming that the linear predictor is correctly specified by (4.4), the results obtained in Theorem 8.1 and Corollary 8.1 allow us to feel relatively confident about the presence of a treatment effect of risperidone on the CGI scores of patients suffering from chronic schizophrenia. Although care is necessary when interpreting the estimated size of the effect, this is undoubtedly a valuable clinical finding.

Chapter 9

A Family of Diagnostic Tools

In the previous chapters we have shown that the estimates of variance components are always subject to considerable bias when the random-effects distribution is misspecified. This can have an important impact in studies where these parameters are of main interest. Further, we found that bias can also be present in the estimates of the linear predictor parameters, especially when complex random-effects structures are considered. Finally, we observed that the misspecification can also affect both the power and the type I error rate associated with the most commonly used inferential procedures.

Evidently, in these circumstances, the development of diagnostic tools is of great importance. The problem of studying the impact of random-effects misspecification is complex due to the latent nature of the random effects. This renders difficult the evaluation of the associated distributional assumptions. Diagnostic tools to analyze the random-effects distribution are therefore not straightforward. For instance, one should be careful in using empirical Bayes estimates of the random effects to detect departures from normality. Indeed, unlike in the linear mixed model, the posterior density of the empirical Bayes estimates in generalized linear mixed models is, in general, not normal, even when the random-effects distribution is correctly specified as normal (Molenberghs and Verbeke, 2005).

Waagepetersen (2006) proposed a simulation-based test to evaluate the appropriateness of the choice of the random-effects distribution, by generating random effects while conditioning on the observations. The intuition behind this test is that if the

joint model for the observations and the random effects is correctly specified, then the marginal distribution of the simulated random effects should coincide with the assumed distribution. Although simulations with Poisson responses showed a reasonable power, this test required very large cluster and sample sizes to produce similar results with binary outcomes. Tchetgen and Coull (2006) introduced a diagnostic test to verify the validity of the choice of the random-effects distribution by comparing marginal and conditional maximum likelihood estimators of a subset of fixed effects in the model. They argue that the conditional estimators are robust to the choice of the random-effects distribution, whereas the estimators from the marginal model will be affected if this distribution is misspecified. Therefore, they propose a test statistic based on the difference between these estimates, focusing on the covariates which vary within each cluster. Clearly, the test is restricted to those applications which involve at least one within-cluster covariate. This would make it inapplicable, for instance, to study the appropriateness of the normal distribution to describe the heterogeneity of the latent trait involved in the Rasch model and other item response models (Agresti, 2002).

White (1982) proposed a general test for model misspecification. However, this Information Matrix Test (IMT) requires third-order partial derivatives of the likelihood function. Even though the calculation of higher order derivatives might not be an issue in cases where the likelihood is available in a closed form, it can become an important problem when working with complicated likelihood functions, like in generalized linear mixed models. Hence, one has to resort to numerical approximations, which can be burdensome and less than straightforward to carry out using conventional statistical packages. In what follows, we propose three new tests for misspecification, following the general idea introduced in Chapter 5. Given that under a correctly specified model, and using the notation of Chapter 5, $A(\boldsymbol{\xi}_0) = -B(\boldsymbol{\xi}_0)$, deviations from the model assumptions are expected to distort this equality. Our proposals focus on the matrix $B_n(\boldsymbol{\xi})[-A^{-1}(\boldsymbol{\xi})]$ and on its properties under a correctly specified model, as a potential indicator of misspecification. All proposals share the desirable property of avoiding the use of third order derivatives, and can be easily implemented using standard software like the SAS procedures NLMIXED and IML (Appendix C shows some exemplary SAS code). The results in this chapter are based partly on Alonso, Litière and Molenberghs (2007).

9.1 The Determinant Tests

Let us start by considering the score statistic for each subject $i = 1, \ldots, n$, given by

$$S(\boldsymbol{y}_i, \boldsymbol{\xi}) = \left\{ \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \right\}, \tag{9.1}$$

where f denotes the marginal model given by (3.2), derived from the hierarchical model defined in (3.1). If the model is correctly specified then, following the notation in Chapter 5, there exists a $\boldsymbol{\xi}_0 \in \Upsilon$ such that $h(\boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{\xi}_0)$. Further assuming that $S(\boldsymbol{y}_i, \boldsymbol{\xi}_0) \sim \mathrm{N}_p[\mathbf{0}, -A(\boldsymbol{\xi}_0)]$, where p denotes the dimension of $\boldsymbol{\xi}$, and given the results established by Anderson (1963) and Girschick (1939) on the large sample distribution of the eigenvalues of a covariance matrix, it is easy to show that if $\gamma_1, \ldots, \gamma_p$ represent the eigenvalues of $-A(\boldsymbol{\xi}_0)$ and $\widehat{\gamma}_{1n}, \ldots, \widehat{\gamma}_{pn}$ the eigenvalues of $B_n(\boldsymbol{\xi}_0)$, then under a correctly specified model, asymptotically,

$$\sqrt{n}(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}) \sim \mathrm{N}_p(\mathbf{0}, 2\Gamma^2), \tag{9.2}$$

where Γ is the diagonal matrix of eigenvalues $\gamma_1, \ldots, \gamma_p$, or after applying the delta method, $\sqrt{n}(\log \widehat{\boldsymbol{\gamma}}_n - \log \boldsymbol{\gamma}) \sim \mathrm{N}_p(\mathbf{0}, 2I)$, where I is the $p \times p$ identity matrix. Using these results, we can obtain the following theorem.

Theorem 9.1 (Determinant Tests) *Let us define* $\delta_{d1}(n) = \log |B_n(\boldsymbol{\xi}_0)[-A^{-1}(\boldsymbol{\xi}_0)]|$ *and* $\delta_{d2}(n) = |B_n(\boldsymbol{\xi}_0)||-A^{-1}(\boldsymbol{\xi}_0)|$. *Then, if the model is correctly specified,*

1. $$\frac{n}{2p}[\delta_{d1}(n)]^2 \sim \chi^2_1. \tag{9.3}$$
2. $$\frac{n}{2p}[\delta_{d2}(n) - 1]^2 \sim \chi^2_1. \tag{9.4}$$

Proof

To simplify the notation, we will omit the index n that denotes the functional dependence of $\widehat{\boldsymbol{\gamma}}_n$ on the sample size.

1. The first test statistic can be rewritten as

$$\delta_{d1}(n) = \log\left(\prod_{k=1}^{p} \widehat{\gamma}_k\right) - \log\left(\prod_{k=1}^{p} \gamma_k\right) = \sum_{k=1}^{p} (\log \widehat{\gamma}_k - \log \gamma_k). \tag{9.5}$$

   Since for each $k = 1, \ldots, p$, the distribution of $\log \widehat{\gamma}_k$ is given by $(\log \widehat{\gamma}_k - \log \gamma_k) \sim \mathrm{N}(0, 2/n)$, it follows that

$$\delta_{d1}(n) = \sum_{k=1}^{p} (\log \widehat{\gamma}_k - \log \gamma_k) \sim \mathrm{N}\left(0, \frac{2p}{n}\right), \tag{9.6}$$

which is equivalent to $\frac{n}{2p}[\delta_{d1}(n)]^2 \sim \chi^2_1$.

2. The second test statistic can also be rewritten as a function of the eigenvalues of the matrices $B_n(\boldsymbol{\xi}_0)$ and $-A(\boldsymbol{\xi}_0)$, such that

$$\delta_{d2}(n) = |B_n(\boldsymbol{\xi}_0)||-A^{-1}(\boldsymbol{\xi}_0)| = \prod_{k=1}^{p} \frac{\widehat{\gamma}_k}{\gamma_k}. \tag{9.7}$$

The variance of $\delta_{d2}(n)$ follows from applying the delta method. First, we determine the gradient of $\delta_{d2}(n)$ evaluated in $\widehat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$:

$$\left.\frac{\partial \delta_{d2}(n)}{\partial \widehat{\gamma}_k}\right|_{\widehat{\boldsymbol{\gamma}}=\boldsymbol{\gamma}} = \left(\prod_{\ell \neq k} \frac{\widehat{\gamma}_\ell}{\gamma_\ell}\right) \frac{1}{\gamma_k}\Bigg|_{\widehat{\boldsymbol{\gamma}}=\boldsymbol{\gamma}} = \frac{1}{\gamma_k}. \tag{9.8}$$

Next, using the distribution of $\widehat{\gamma}_k$ given in (9.2) and applying the delta method, leads to $\delta_{d2}(n) \sim \mathrm{N}(1, \frac{2}{n}\Delta^T\Gamma^2\Delta)$, where $\Delta^T = \left(\frac{1}{\gamma_1} \ldots \frac{1}{\gamma_p}\right)$. Finally, note that $\Delta^T\Gamma^2\Delta = p$. Hence, under a correctly specified model $\delta_{d2}(n) \sim \mathrm{N}(1, 2p/n)$, or equivalently, $\frac{n}{2p}[\delta_{d2}(n) - 1]^2 \sim \chi^2_1$. □

When Theorem 9.1 is applied in a practical situation, $A^{-1}(\boldsymbol{\xi}_0)$ in (9.3) and (9.4) can be substituted by its consistent estimator under the null, given by $A_n^{-1}(\widehat{\boldsymbol{\xi}}_n)$. Each subject's contribution to $A_n(\widehat{\boldsymbol{\xi}}_n)$ and $B_n(\widehat{\boldsymbol{\xi}}_n)$ can be readily obtained from NLMIXED, by fitting the final model by subject, keeping all parameters fixed (`maxiter = 0`) and saving the corresponding first and second order derivatives (see Appendix C).

Note further that, $\delta_{d1}(n)$ and $\delta_{d2}(n)$ are merely two variations to the same theme. However, whether or not the logarithm is used can play an important role in the asymptotic behavior of the tests as well as in their small sample performance.

Essentially, (9.3) and (9.4) try to detect departures from the equality $B(\boldsymbol{\xi}_0) = -A(\boldsymbol{\xi}_0)$ using the determinant of the matrix $B_n(\boldsymbol{\xi}_0)[-A^{-1}(\boldsymbol{\xi}_0)]$. The use of the determinant in this setting is a plausible and sensible choice to quantify the "distance" between $B_n(\boldsymbol{\xi}_0)$ and $-A^{-1}(\boldsymbol{\xi}_0)$. Another intuitive and appealing possibility consists in combining the determinant and the trace into a test statistic to quantify the "distance" between the two matrices of interest. We will explore this approach in the next section.

9.2 The Determinant-Trace Test

Consider the test statistic, which incorporates both the trace and the determinant of $A(\boldsymbol{\xi}_0)$ and $B_n(\boldsymbol{\xi}_0)$, given by

$$\delta_{dt}(n) = \frac{\text{tr}[B_n(\boldsymbol{\xi}_0)]}{\text{tr}[-A(\boldsymbol{\xi}_0)]} - \frac{|B_n(\boldsymbol{\xi}_0)|}{|-A(\boldsymbol{\xi}_0)|}, \tag{9.9}$$

or equivalently by

$$\delta_{dt}(n) = \sum_{k=1}^{p} \left(\frac{\widehat{\gamma}_{kn}}{\sum_\ell \gamma_\ell} \right) - \prod_{k=1}^{p} \frac{\widehat{\gamma}_{kn}}{\gamma_k}. \tag{9.10}$$

If we further denote

$$\sigma_{\delta_n} = \sum_{k=1}^{p} \left(\frac{\gamma_k}{\sum_\ell \gamma_\ell} - 1 \right)^2,$$

then the following result allows us to establish the distribution under the null of $\delta_{dt}(n)$.

Theorem 9.2 (Determinant-Trace Test) *If the model is correctly specified, then*

$$\frac{n[\delta_{dt}(n)]^2}{2\sigma_{\delta_n}} \sim \chi^2_1. \tag{9.11}$$

Proof

Like before, to simplify the notation, we will omit the index n that denotes the functional dependence of $\widehat{\boldsymbol{\gamma}}_n$ on the sample size. Here again, we can apply the delta method to obtain an expression for the variance of $\delta_{dt}(n)$. First, we determine the gradient of $\delta_{dt}(n)$, evaluated in $\widehat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$:

$$\left. \frac{\partial \delta_{dt}(n)}{\partial \widehat{\gamma}_k} \right|_{\widehat{\boldsymbol{\gamma}}=\boldsymbol{\gamma}} = \frac{1}{\sum_\ell \gamma_\ell} - \left(\prod_{\ell \neq k} \frac{\widehat{\gamma}_\ell}{\gamma_\ell} \right) \frac{1}{\gamma_k} \Bigg|_{\widehat{\boldsymbol{\gamma}}=\boldsymbol{\gamma}} = \frac{1}{\sum_\ell \gamma_\ell} - \frac{1}{\gamma_k}. \tag{9.12}$$

If we let $\Delta^T = \begin{pmatrix} \frac{1}{\sum_\ell \gamma_\ell} - \frac{1}{\gamma_1} & \cdots & \frac{1}{\sum_\ell \gamma_\ell} - \frac{1}{\gamma_p} \end{pmatrix}$, then under a correctly specified model $\delta_{dt}(n) \sim \text{N}\left(0, \frac{2}{n}\sigma_{\delta_n}\right)$, where

$$\sigma_{\delta_n} = \Delta^T \Gamma^2 \Delta = \sum_{k=1}^{p} \left(\frac{\gamma_k}{\sum_\ell \gamma_\ell} - 1 \right)^2. \tag{9.13}$$

This is equivalent to the distribution specified in (9.11). □

In principle, other alternatives using these eigenvalues could be considered as well. We have chosen three that are intuitively appealing and mathematically tractable when calculating their null distribution. Obviously, these tests are based on some assumptions. For instance, the individual contributions to the score are assumed to be normally distributed. Departures from this assumption may affect the distributional results in (9.2) and the performance of the proposed tests. In this case, Waternaux (1976) showed that the estimators of the eigenvalues obtained from the observed covariance matrix will still be normally distributed and centered around their population values, however, the covariance matrix of these estimators may require a correction depending on the shape of the real distribution. Essentially, this assumption is the price to pay in order to gain simplicity and avoid the use of high order derivatives.

In what follows, we will empirically study the performance of these tests via simulations. This study will help us to evaluate the impact of this assumption and the behavior of these asymptotic results in finite sample sizes.

9.3 Simulation Study

To explore the behavior of the aforementioned diagnostic tools, we designed a simulation study comparing the performance of these tests to detect random-effects misspecification in both linear and generalized linear mixed models. We considered a number of practically relevant settings, including small sample sizes, and small and large variances for the random effect.

9.3.1 Linear Mixed Models

When dealing with mixed models for normal responses, the marginal likelihood can be written down in a closed form (see Section 3.3). Therefore, all derivatives involved in the calculation of the IMT can be solved analytically. An overview of these derivatives is presented in Appendix D. Hence, in this setting, we can easily compare the performance of the proposed diagnostic tools with this general test for model misspecification. For this purpose, normal responses were generated using the linear random-intercept model given by

$$y_{ij} = \beta_0 + b_i + \beta_1 z_i + \beta_2 t_j + \varepsilon_{ij}, \tag{9.14}$$

including an intercept, a binary covariate z_i, a within-cluster covariate t_j taking values 0, 1, 2, 4, 6, and 8, measurement error terms ε_{ij} drawn from $\mathrm{N}(0, 1)$, and a random intercept b_i sampled from the 5 following mean-zero distributions, each with variance $\sigma^2_{0b} = 4$ and 32: a normal, a power function, and a lognormal distribution, as well as a discrete distribution with equal probability at two support points, and an asymmetric mixture of two normal densities.

The parameters in the mean structure were fixed at $\beta_0^0 = -8$, $\beta_1^0 = 2$, and $\beta_2^0 = 1$. Six different sample sizes were considered, including 50, 100, 200, 350, 500, and 1000 subjects. For each setting, 500 data sets were generated and the model given by (9.14) was fitted to these data, assuming normally distributed random effects. We then determined the proportion of cases in which each test produced a significant result at the 5% significance level. When the random effects were generated from a normal distribution, this proportion corresponds to the type I error; otherwise, it represents the power of the test to detect random-effects misspecification. The results of these simulations are shown in Table 9.1.

In general, White's IMT shows a good power and type I error rate for reasonable sample sizes ($n = 350$). The only exception to the previous behavior is observed when the true distribution of the random effect is an asymmetric mixture with small variance. Nevertheless, it should be noted that for small variances this asymmetric mixture can be reasonably well approximated by a normal density.

The newly proposed tests exhibit an excellent type I error rate, especially for small sample sizes. However, in general they are less powerful than the IMT to detect misspecification. Only when the true random-effects distribution is a very heterogenous asymmetric mixture are they able to outperform the IMT. In contrast, they fail to detect the asymmetric mixture with small variance. Additionally, these tests may require up to 500 subjects to achieve an acceptable power of 60% and more when the random intercept was generated from a power function distribution.

As a conclusion we could state that the three tests showed a similar general performance in the linear case. Further, the IMT showed a very good power to detect misspecifications of the random-effects distribution in this scenario. In the following section we will approach this problem in the more challenging generalized linear mixed model setting.

Table 9.1: *Power and type I error for detecting a misspecified random-effects distribution, using* $\Im(n)$, $\delta_{d1}(n)$, $\delta_{d2}(n)$ *and* $\delta_{dt}(n)$ *in the setting of linear mixed models: a normal random intercept is assumed, whereas the random effects are generated from a normal (No), a power function (PF), a discrete (D), an asymmetric mixture of two normals (AM) or a lognormal distribution (LN), each with variance* $\sigma^2_{0b} = 4$ *or* 32. *Those settings for which the new proposals outperform the IMT are highlighted.*

		$\sigma^2_{0b} = 4$				$\sigma^2_{0b} = 32$			
	N	$\Im(n)$	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$	$\Im(n)$	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$
No	50	0.266	**0.164**	**0.000**	**0.022**	0.232	**0.172**	**0.000**	**0.012**
	100	0.208	**0.110**	**0.006**	**0.038**	0.190	**0.098**	**0.012**	**0.028**
	200	0.106	**0.084**	**0.024**	**0.038**	0.116	**0.080**	**0.034**	**0.036**
	350	0.084	**0.066**	**0.046**	**0.056**	0.096	**0.074**	**0.036**	**0.042**
	500	0.048	0.058	**0.042**	0.054	0.078	0.080	**0.056**	**0.058**
	1000	0.042	0.042	**0.036**	0.052	0.054	0.060	**0.046**	0.056
PF	50	0.372	0.160	0.004	0.020	0.368	0.242	0.004	0.034
	100	0.432	0.122	0.110	0.144	0.460	0.166	0.120	0.144
	200	0.640	0.256	0.288	0.322	0.676	0.264	0.310	0.336
	350	0.778	0.508	0.550	0.576	0.848	0.488	0.534	0.574
	500	0.876	0.650	0.692	0.720	0.888	0.614	0.648	0.678
	1000	0.958	0.898	0.904	0.918	0.974	0.870	0.878	0.904
D	50	1.000	1.000	0.968	0.824	1.000	1.000	1.000	0.916
	100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	350	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AM	50	0.136	**0.226**	0.008	0.040	0.562	**0.990**	0.420	**0.586**
	100	0.156	**0.178**	0.024	0.052	0.562	**0.998**	**0.996**	**0.992**
	200	0.220	0.170	0.056	0.112	0.610	**1.000**	**1.000**	**1.000**
	350	0.504	0.208	0.124	0.148	0.642	**1.000**	**1.000**	**1.000**
	500	0.666	0.202	0.134	0.162	0.702	**1.000**	**1.000**	**1.000**
	1000	0.960	0.318	0.256	0.324	0.784	**1.000**	**1.000**	**1.000**
LN	50	0.694	0.138	0.014	0.042	0.850	0.344	0.053	0.115
	100	0.840	0.202	0.266	0.320	0.941	0.239	0.306	0.361
	200	0.960	0.734	0.778	0.796	0.982	0.806	0.843	0.857
	350	0.970	0.958	0.968	**0.982**	0.988	0.976	0.982	0.984
	500	0.970	**0.992**	**0.992**	**0.992**	0.992	**0.996**	**0.998**	**1.000**
	1000	0.966	**1.000**	**1.000**	**1.000**	0.990	**1.000**	**1.000**	**1.000**

9.3.2 Generalized Linear Mixed Models

To study the performance of the new proposals with generalized linear mixed models, binary responses were generated using the logistic random-intercept model given by (7.1). Recall that the linear predictor of this model included an intercept, a binary covariate z_i taking at random values 0 and 1, and a within-cluster covariate t_j taking values 0, 1, 2, 4, 6, and 8. The random intercept b_i, sampled from the same distributions considered in the previous simulation study, formed the random part. The parameters in the linear predictor were fixed at $\beta_0^0 = -8$, $\beta_1^0 = 2$, and $\beta_2^0 = 1$, in accordance with the values in Table 4.1, estimated from the case study. As in the previous section, we considered 50, 100, 200, 350, 500, and 1000 subjects. For each setting, 500 data sets were generated and the model given by (7.1) was fitted to these data under the assumption of normally distributed random effects. The type I error and the power of the proposed diagnostic tools are reported in Table 9.2. The first part of the table displays the results for the small variance setting ($\sigma_{0b}^2 = 4$). In this scenario, most of the tests exhibit a good type I error rate for reasonable samples of 200 or 350 subjects. The only exception seems to be the determinant-trace test $\delta_{dt}(n)$ with a considerably large type I error rate of up to 13%, even with 1000 subjects.

The determinant tests now present the best global behavior when both power and type I error are taken into account. Especially the test based on $\delta_{d1}(n)$ successfully detects the discrete and lognormal distribution for sample sizes of 350 or larger. The determinant-trace test $\delta_{dt}(n)$ had a high power to detect many of the considered misspecifications. However, its inflated type I error would make its results unclear in a real situation where the true distribution is unknown. Finally, note that for small variances, the asymmetric mixture and power densities can be reasonably well approximated by a normal distribution. As a consequence, all tests need a large sample size to achieve a moderate power in detecting these misspecifications.

The second part of the table summarizes the results for the large variance scenario ($\sigma_{0b}^2 = 32$). Unlike in the previous setting, all the tests now exhibit a very good type I error rate, even for small samples of 50 subjects. The best overall behavior is now observed for the determinant-trace test. Even when the random intercept was drawn from the asymmetric mixture, the determinant-trace test displayed a remarkable power to detect this misspecification. Regarding power, all tests perform quite well already with sample sizes of 200 subjects.

Table 9.2: *Power and type I error for detecting a misspecified random-effects distribution, using $\delta_{d1}(n)$, $\delta_{d2}(n)$ and $\delta_{dt}(n)$ in the setting of generalized linear mixed models: a normal random intercept is assumed, whereas the random effects are generated from a normal (No), a power function (PF), a discrete (D), an asymmetric mixture of two normals (AM) or a lognormal distribution (LN), each with variance $\sigma^2_{0b} = 4$ or 32. The test with the best performance for each setting is highlighted.*

		$\sigma^2_{0b} = 4$			$\sigma^2_{0b} = 32$		
	n	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$
No	50	0.126	**0.010**	0.048	0.074	0.042	**0.032**
	100	0.106	**0.030**	0.098	0.052	0.026	**0.020**
	200	0.096	**0.054**	0.126	0.048	0.044	**0.032**
	350	0.098	**0.072**	0.152	0.062	0.050	**0.026**
	500	0.100	**0.084**	0.164	0.054	0.040	**0.018**
	1000	0.074	**0.068**	0.126	0.052	0.038	**0.012**
PF	50	**0.167**	0.004	0.094	0.422	0.014	**0.448**
	100	0.208	0.034	**0.240**	0.542	0.214	**0.754**
	200	0.262	0.092	**0.422**	0.734	0.558	**0.968**
	350	0.334	0.200	**0.570**	0.926	0.846	**0.996**
	500	0.392	0.278	**0.648**	0.990	0.980	**1.000**
	1000	0.640	0.580	**0.894**	**1.000**	**1.000**	**1.000**
D	50	0.412	0.010	**0.266**	**0.914**	0.458	0.840
	100	0.502	0.176	**0.600**	0.986	0.942	**0.992**
	200	0.716	0.550	**0.858**	**1.000**	0.998	**1.000**
	350	0.894	0.782	**0.972**	**1.000**	**1.000**	**1.000**
	500	0.954	0.926	**0.994**	**1.000**	**1.000**	**1.000**
	1000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
AM	50	**0.145**	0.006	0.095	0.266	0.014	**0.500**
	100	0.116	0.016	**0.148**	0.322	0.094	**0.722**
	200	0.152	0.054	**0.288**	0.456	0.264	**0.904**
	350	0.164	0.086	**0.328**	0.602	0.474	**0.998**
	500	0.170	0.108	**0.404**	0.704	0.602	**1.000**
	1000	0.312	0.260	**0.620**	0.934	0.918	**1.000**
LN	50	0.076	0.088	**0.169**	0.091	0.113	**0.178**
	100	0.267	0.369	**0.583**	0.172	0.260	**0.474**
	200	0.676	0.766	**0.936**	0.476	0.578	**0.868**
	350	0.914	0.954	**0.996**	0.690	0.786	**0.982**
	500	0.994	0.996	**1.000**	0.874	0.924	**1.000**
	1000	**1.000**	**1.000**	**1.000**	0.998	**1.000**	**1.000**

Clearly, the power of the tests considerably improved when the variance of the random effects was large. This is a desirable behavior given the results obtained in Chapter 7. Indeed, we found that for small values of the random effects variance, the bias introduced by the misspecification on the linear predictor parameters was negligible. Nevertheless, considerable bias could appear under misspecification when the variance was more substantial. Precisely, the setting where the proposed tests showed the larger power and best general performance.

9.4 Application: The Schizophrenia Data

In this section, we will apply the different tests to assess the suitability of the logistic-normal model given by (4.4) for the analysis of the case study in Chapter 4. The determinant tests lead to $\delta_{d1}(n) = 0.075$ and $\delta_{d2}(n) = 1.078$ with corresponding p-values of 0.763 and 0.754, respectively, while the determinant-trace test delivers the following output: $\delta_{dt}(n) = 0.001$ accompanied by $p = 0.996$.

These results imply that, with the data at hand, we do not have evidence of a departure from the assumption of normally distributed random effects or any other misspecification in the model. While the sample size is likely not large enough to detect very small deviations, $n = 128$ might be considered sufficient to detect gross departures. Therefore, the clearly non-significant p-values certainly allow us to entertain a comfortable level of confidence in the final model.

Even though overall the proposed tests show good performance, none of them consistently detects random-effects misspecification in all the settings considered in this chapter. Clearly there is still some room for improvement. For instance, the previous tests impose some distributional assumptions on the subject's contribution to the score function that do not need to hold in a practical situation. In the next chapter we will introduce and study two diagnostic tools that try to overcome some of the limitations and problems of the tests analyzed here.

Chapter 10

Alternative Information Matrix Tests

In the previous chapter, we introduced a family of tests for detecting model misspecifications based on the eigenvalues of the matrix $B_n(\boldsymbol{\xi}_0)[-A^{-1}(\boldsymbol{\xi}_0)]$. Even though in the simulation studies, these tests showed a good general performance, the new proposals are clearly not powerful enough in some settings. Therefore, in this chapter we will continue the search for more efficient diagnostic tools. We propose two new alternatives, along the ideas of the IMT, and study the power of these tests to detect random-effects misspecification, as well as violations of other important model assumptions. This chapter is based on results obtained in Litière, Alonso and Molenberghs (2007c) and Alonso *et al.* (2007).

10.1 The Sandwich Estimator Test

In this section, we focus on the difference between $V(\boldsymbol{\xi}_0)$ and $-A^{-1}(\boldsymbol{\xi}_0)$ as a potential indicator of misspecification. More specifically, we will focus on the diagonal elements of $V(\boldsymbol{\xi}_0) + A^{-1}(\boldsymbol{\xi}_0)$.

Let us first define $\tilde{V}_n(\boldsymbol{\xi}) = A^{-1}(\boldsymbol{\xi})B_n(\boldsymbol{\xi})A^{-1}(\boldsymbol{\xi})$ and $\boldsymbol{v}_n(\boldsymbol{\xi}) = \text{diag}[\tilde{V}_n(\boldsymbol{\xi}) + A^{-1}(\boldsymbol{\xi})]$. Note that $\boldsymbol{v}_n(\boldsymbol{\xi})$ can also be written as $\Delta\text{vec}[\tilde{V}_n(\boldsymbol{\xi}) + A^{-1}(\boldsymbol{\xi})]$, where Δ is the $p \times p^2$

matrix specified as

$$\Delta = \begin{cases} 1 & \text{for } k = 1, \ldots, p \text{ and } \ell = (k-1)p + k \\ 0 & \text{otherwise,} \end{cases} \tag{10.1}$$

and p refers to the number of parameters in the model. Further, let us consider

$$\boldsymbol{b}_i(\boldsymbol{\xi}) = \text{vec}\left\{ \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k} \cdot \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_\ell} \right\}, \tag{10.2}$$

and let $\boldsymbol{\mu}_b(\boldsymbol{\xi})$ and $V_b(\boldsymbol{\xi})$ represent the mean and the covariance matrix of $\boldsymbol{b}_i(\boldsymbol{\xi})$. An unbiased estimator of $\boldsymbol{\mu}_b(\boldsymbol{\xi})$ is given by $\widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{b}_i(\boldsymbol{\xi}) = \text{vec}[B_n(\boldsymbol{\xi})]$, whereas $V_b(\boldsymbol{\xi})$ can be estimated through

$$\widehat{V}_b(\boldsymbol{\xi}) = \frac{1}{n-1}\sum_{i=1}^n [\boldsymbol{b}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi})][\boldsymbol{b}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi})]^T. \tag{10.3}$$

Additionally, let us denote by

$$C_v(\boldsymbol{\xi}) = \frac{1}{n}\Delta[A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]V_b(\boldsymbol{\xi})[A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\Delta^T. \tag{10.4}$$

Using all of these elements, we can now establish the following result.

Theorem 10.1 (Sandwich Estimator Test) *Under general regularity conditions, if the model is correctly specified, then as* $n \to \infty$

$$\boldsymbol{v}_n(\boldsymbol{\xi}_0) \sim N_p(\boldsymbol{0}, C_v(\boldsymbol{\xi}_0)),$$

and therefore

$$\delta_s(n) = \boldsymbol{v}_n^T(\boldsymbol{\xi}_0)[C_v(\boldsymbol{\xi}_0)]^{-1}\boldsymbol{v}_n(\boldsymbol{\xi}_0) \sim \chi^2_p. \tag{10.5}$$

Proof

It is possible to show, using some properties of the Kronecker product and the vec operator, that

$$\begin{aligned} \text{vec}[\tilde{V}_n(\boldsymbol{\xi})] &= [A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\text{vec}[B_n(\boldsymbol{\xi})]. \\ &= [A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{b}_i(\boldsymbol{\xi})\right] \end{aligned} \tag{10.6}$$

Further, we have that

$$\begin{aligned} \mathrm{E}\{\mathrm{vec}[\tilde{V}_n(\boldsymbol{\xi})]\} &= [A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\boldsymbol{b}_i(\boldsymbol{\xi})] \\ &= [A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\boldsymbol{\mu}_b(\boldsymbol{\xi}), \end{aligned}$$

and

$$\mathrm{cov}\{\mathrm{vec}[\tilde{V}_n(\boldsymbol{\xi})]\} = [A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\mathrm{cov}\{\mathrm{vec}[B_n(\boldsymbol{\xi})]\}[A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]^T. \quad (10.7)$$

Since

$$\begin{aligned} \mathrm{cov}\{\mathrm{vec}[B_n(\boldsymbol{\xi})]\} &= \mathrm{cov}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{b}_i(\boldsymbol{\xi})\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{cov}[\boldsymbol{b}_i(\boldsymbol{\xi})] \\ &= \frac{1}{n}V_b(\boldsymbol{\xi}), \end{aligned} \quad (10.8)$$

we can rewrite (10.7) as

$$\mathrm{cov}\{\mathrm{vec}[\tilde{V}_n(\boldsymbol{\xi})]\} = \frac{1}{n}[A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]V_b(\boldsymbol{\xi})[A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]. \quad (10.9)$$

Under general regularity conditions, (10.6) and the central limit theorem imply that, asymptotically,

$$\mathrm{vec}[\tilde{V}_n(\boldsymbol{\xi})] \sim \mathrm{N}_{p^2}\left([A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\boldsymbol{\mu}_b(\boldsymbol{\xi}), \mathrm{cov}\{\mathrm{vec}[\tilde{V}_n(\boldsymbol{\xi})]\}\right). \quad (10.10)$$

Since

$$\begin{aligned} \boldsymbol{v}_n(\boldsymbol{\xi}) &= \Delta\mathrm{vec}[\tilde{V}_n(\boldsymbol{\xi}) + A^{-1}(\boldsymbol{\xi})] \\ &= \Delta\mathrm{vec}[\tilde{V}_n(\boldsymbol{\xi})] + \Delta\mathrm{vec}[A^{-1}(\boldsymbol{\xi})], \end{aligned}$$

it easily follows that

$$\boldsymbol{v}_n(\boldsymbol{\xi}) \sim \mathrm{N}_p\left(\Delta[A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\boldsymbol{\mu}_b(\boldsymbol{\xi}) + \Delta\mathrm{vec}[A^{-1}(\boldsymbol{\xi})], \Delta\mathrm{cov}\{\mathrm{vec}[\tilde{V}_n(\boldsymbol{\xi})]\}\Delta^T\right),$$

or

$$\boldsymbol{v}_n(\boldsymbol{\xi}) \sim \mathrm{N}_p\left(\Delta[A^{-1}(\boldsymbol{\xi}) \otimes A^{-1}(\boldsymbol{\xi})]\boldsymbol{\mu}_b(\boldsymbol{\xi}) + \Delta\mathrm{vec}[A^{-1}(\boldsymbol{\xi})], C_v(\boldsymbol{\xi})\right). \quad (10.11)$$

Under a correctly specified model, $B(\boldsymbol{\xi}_0) = -A(\boldsymbol{\xi}_0)$. In this case,

$$\begin{aligned}
\boldsymbol{\mu}_b(\boldsymbol{\xi}_0) &= \mathrm{E}\left[\mathrm{vec}\left\{\frac{\partial \log f(\boldsymbol{y},\boldsymbol{\xi}_0)}{\partial \xi_k}\cdot\frac{\partial \log f(\boldsymbol{y},\boldsymbol{\xi}_0)}{\partial \xi_\ell}\right\}\right] \\
&= \mathrm{vec}\left[\mathrm{E}\left\{\frac{\partial \log f(\boldsymbol{y},\boldsymbol{\xi}_0)}{\partial \xi_k}\cdot\frac{\partial \log f(\boldsymbol{y},\boldsymbol{\xi}_0)}{\partial \xi_\ell}\right\}\right] \\
&= \mathrm{vec}[B(\boldsymbol{\xi}_0))] \\
&= \mathrm{vec}[-A(\boldsymbol{\xi}_0)].
\end{aligned}$$

It can now be easily seen that

$$\begin{aligned}
\mathrm{E}[\boldsymbol{v}_n(\boldsymbol{\xi}_0)] &= \Delta[A^{-1}(\boldsymbol{\xi}_0)\otimes A^{-1}(\boldsymbol{\xi}_0)]\boldsymbol{\mu}_b(\boldsymbol{\xi}_0) + \Delta\mathrm{vec}[A^{-1}(\boldsymbol{\xi}_0)] \\
&= \Delta[A^{-1}(\boldsymbol{\xi}_0)\otimes A^{-1}(\boldsymbol{\xi}_0)]\mathrm{vec}[-A(\boldsymbol{\xi}_0)] + \Delta\mathrm{vec}[A^{-1}(\boldsymbol{\xi}_0)].
\end{aligned}$$

The first term in this expression can also be written as

$$[A^{-1}(\boldsymbol{\xi}_0)\otimes A^{-1}(\boldsymbol{\xi}_0)]\mathrm{vec}[-A(\boldsymbol{\xi}_0)] = -\mathrm{vec}[A^{-1}(\boldsymbol{\xi}_0)A(\boldsymbol{\xi}_0)A^{-1}(\boldsymbol{\xi}_0)] = -\mathrm{vec}[A^{-1}(\boldsymbol{\xi}_0)].$$

Therefore,

$$\mathrm{E}[\boldsymbol{v}_n(\boldsymbol{\xi}_0)] = -\Delta[A^{-1}(\boldsymbol{\xi}_0)] + \Delta\mathrm{vec}[A^{-1}(\boldsymbol{\xi}_0)] = \mathbf{0} \tag{10.12}$$

and as a consequence

$$\boldsymbol{v}_n(\boldsymbol{\xi}_0) \sim \mathrm{N}_p(\mathbf{0}, C_v(\boldsymbol{\xi}_0)).$$

□

When Theorem 10.1 is applied in a practical situation, $A^{-1}(\boldsymbol{\xi}_0)$ in (10.5) is substituted by its consistent estimator under the null, given by $A^{-1}(\widehat{\boldsymbol{\xi}}_n)$, and $V_b(\boldsymbol{\xi}_0)$ by its consistent estimator $\widehat{V}_b(\widehat{\boldsymbol{\xi}}_n)$ given by (10.3). All the necessary calculations are illustrated with some exemplary SAS code in Appendix C.

Unlike the previously defined tests, the SET uses the available information in a more sophisticated fashion and does not impose distributional assumptions on the $\boldsymbol{b}_i$'s. This could have a positive impact on its asymptotic and small sample behavior. To explore this issue further, we have applied the test to the data generated in Chapter 9. In the first panel of Table 10.1, we compare the performance of the SET $\delta_s(n)$ with the IMT, when the data are generated from the linear mixed model given by (9.14), considering different random-effects distributions, and analyzed using model (9.14), assuming a normally distributed random intercept.

Table 10.1: *Power and type I error over different sample sizes* (n)*, for detecting a misspecified random-effects distribution, using the IMT* $\Im(n)$*, the SET* $\delta_s(n)$ *and the MIMT* $\Im_m(n)$ *in linear and/or generalized linear mixed models: a normal random intercept is assumed, whereas the random effects are generated from a normal (No), a power function (PF), a discrete (D), an asymmetric mixture of two normals (AM) or a lognormal distribution (LN), each with variance* $\sigma^2_{0b} = 4$ *or 32.*

		LMM						GLMM			
		$\sigma^2_{0b} = 4$			$\sigma^2_{0b} = 32$			$\sigma^2_{0b} = 4$		$\sigma^2_{0b} = 32$	
	n	$\Im(n)$	$\delta_s(n)$	$\Im_m(n)$	$\Im(n)$	$\delta_s(n)$	$\Im_m(n)$	$\delta_s(n)$	$\Im_m(n)$	$\delta_s(n)$	$\Im_m(n)$
No	50	0.266	0.298	0.344	0.232	0.284	0.300	0.278	0.259	0.234	0.154
	100	0.208	0.204	0.238	0.190	0.172	0.222	0.192	0.242	0.122	0.102
	200	0.106	0.088	0.122	0.116	0.090	0.126	0.098	0.180	0.072	0.072
	350	0.084	0.058	0.090	0.096	0.060	0.104	0.046	0.108	0.038	0.056
	500	0.048	0.034	0.054	0.078	0.064	0.082	0.044	0.080	0.040	0.048
	1000	0.042	0.026	0.044	0.054	0.040	0.056	0.026	0.054	0.016	0.032
PF	50	0.372	0.458	0.474	0.368	0.482	0.524	0.242	0.589	0.264	0.762
	100	0.432	0.490	0.510	0.460	0.516	0.498	0.172	0.515	0.288	0.922
	200	0.640	0.636	0.658	0.676	0.686	0.688	0.092	0.620	0.406	0.996
	350	0.778	0.780	0.793	0.848	0.848	0.858	0.086	0.710	0.650	1.000
	500	0.876	0.864	0.882	0.888	0.886	0.892	0.096	0.820	0.812	1.000
	1000	0.958	0.952	0.958	0.974	0.972	0.974	0.164	0.952	0.998	1.000
D	50	1.000	1.000	1.000	1.000	1.000	1.000	0.304	0.782	0.710	0.982
	100	1.000	1.000	1.000	1.000	1.000	1.000	0.306	0.914	0.830	0.998
	200	1.000	1.000	1.000	1.000	1.000	1.000	0.430	0.968	0.948	1.000
	350	1.000	1.000	1.000	1.000	1.000	1.000	0.688	0.988	0.996	1.000
	500	1.000	1.000	1.000	1.000	1.000	1.000	0.848	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000	0.986	1.000	1.000	1.000
AM	50	0.136	0.228	0.222	0.562	0.660	0.634	0.230	0.471	0.242	0.788
	100	0.156	0.182	0.188	0.562	0.592	0.592	0.146	0.374	0.172	0.926
	200	0.220	0.226	0.264	0.610	0.612	0.628	0.092	0.458	0.114	0.996
	350	0.504	0.500	0.526	0.642	0.618	0.650	0.070	0.514	0.136	1.000
	500	0.666	0.646	0.674	0.702	0.692	0.704	0.064	0.588	0.098	1.000
	1000	0.960	0.948	0.960	0.784	0.782	0.786	0.058	0.784	0.166	1.000
LN	50	0.694	0.736	0.750	0.850	0.915	0.904	0.355	0.446	0.480	0.547
	100	0.840	0.856	0.868	0.941	0.972	0.972	0.251	0.545	0.338	0.772
	200	0.960	0.958	0.966	0.982	0.984	0.990	0.198	0.792	0.430	0.984
	350	0.970	0.968	0.978	0.988	0.990	0.988	0.224	0.964	0.724	1.000
	500	0.970	0.972	0.976	0.992	0.990	0.992	0.332	0.996	0.892	1.000
	1000	0.966	0.964	0.972	0.990	0.988	0.992	0.780	1.000	1.000	1.000

Clearly, the IMT and the SET have a very similar performance in all the settings considered. Combined with its computational simplicity, this is a strong point for the SET. Further, when comparing the results in Tables 9.1 and 10.1, we observed that its performance is stabler than that of the diagnostic tools introduced in Chapter 9. Indeed, although the tests based on the eigenvalues of the matrix $-B_n(\boldsymbol{\xi}_0)A^{-1}(\boldsymbol{\xi}_0)$ may show a good power in some settings, they encounter serious problems to detect the misspecification in others, even when a large sample size is available. However, the IMT and the SET always achieve a power of at least 67%, even when the true distribution of the random effect is an asymmetric mixture with small variance.

The power of the SET to detect random-effects misspecification, when the data are generated using the logistic random-intercept model given by (7.1), considering different random-effects distributions, and analyzed using the same model, but assuming a normal random intercept, is shown in the second panel of Table 10.1. Compared to the results in Table 9.2, we now observe a similar behavior of the SET and the diagnostic tools in Chapter 9, especially when the variance of the random intercept is large. In this case, a good power and type I error rate is generally observed for the SET with reasonable samples as of 350 subjects. As discussed in the previous chapter this is a desirable behavior given that, under misspecification, considerable bias can appear in the fixed-effects estimates when the variance of the random effects is large. Still, like for the linear mixed model, the asymmetric mixture appears to be a misspecification difficult to detect. In the next section, we will introduce another diagnostic tool that will exhibit a greater power to detect this misspecification.

10.2 The Modified Information Matrix Test

While developing the SET, the variability of $\text{vec}[B_n(\boldsymbol{\xi})]$ was estimated using the empirical covariance estimator (10.3). In this section, we will use the same approach to obtain an empirical estimate of the variability of $\text{vec}[A_n(\boldsymbol{\xi})]$. This will allow us to approximate the variability of the IMT test statistic, without the need for third order derivatives. Consider again $\boldsymbol{D}_n(\boldsymbol{\xi})$, as defined in (5.12), which corresponds to the vector of diagonal elements of $A_n(\boldsymbol{\xi}) + B_n(\boldsymbol{\xi})$. Note that $\boldsymbol{D}_n(\boldsymbol{\xi}) = \Delta\text{vec}[A_n(\boldsymbol{\xi}) + B_n(\boldsymbol{\xi})]$, where Δ is given by (10.1). From this expression it follows that

$$\begin{aligned} C_D(\boldsymbol{\xi}) &= \text{cov}[\boldsymbol{D}_n(\boldsymbol{\xi})] = \text{cov}\{\Delta\text{vec}[A_n(\boldsymbol{\xi}) + B_n(\boldsymbol{\xi})]\} \\ &= \Delta\text{cov}\{\text{vec}[A_n(\boldsymbol{\xi})] + \text{vec}[B_n(\boldsymbol{\xi})]\}\Delta^T, \end{aligned}$$

where

$$\begin{aligned}\text{cov}\{\text{vec}[A_n(\boldsymbol{\xi})] + \text{vec}[B_n(\boldsymbol{\xi})]\} &= \text{cov}\{\text{vec}[A_n(\boldsymbol{\xi})]\} + \text{cov}\{\text{vec}[B_n(\boldsymbol{\xi})]\} \\ &+ \text{cov}\{\text{vec}[A_n(\boldsymbol{\xi})], \text{vec}[B_n(\boldsymbol{\xi})]\} + \text{cov}\{\text{vec}[B_n(\boldsymbol{\xi})], \text{vec}[A_n(\boldsymbol{\xi})]\}. \quad (10.13)\end{aligned}$$

Note that, from (10.8), we have that $\text{cov}\{\text{vec}[B_n(\boldsymbol{\xi})]\} = n^{-1}V_b(\boldsymbol{\xi})$, and $V_b(\boldsymbol{\xi})$ can be consistently estimated using (10.3). Similarly to (10.2), we can now define

$$\boldsymbol{a}_i(\boldsymbol{\xi}) = \text{vec}\left(\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k \partial \xi_\ell}\right). \quad (10.14)$$

Let $\boldsymbol{\mu}_a(\boldsymbol{\xi})$ and $V_a(\boldsymbol{\xi})$ represent the mean and the covariance of $\boldsymbol{a}_i(\boldsymbol{\xi})$. It is then easy to show that $\text{cov}\{\text{vec}[A_n(\boldsymbol{\xi})]\} = n^{-1}V_a(\boldsymbol{\xi})$, and an unbiased estimator of $V_a(\boldsymbol{\xi})$ is given by

$$\widehat{V}_a(\boldsymbol{\xi}) = \frac{1}{n-1}\sum_{i=1}^{n}[\boldsymbol{a}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi})][\boldsymbol{a}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi})]^T, \quad (10.15)$$

where $\widehat{\mu}_a(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i(\boldsymbol{\xi})$. Finally, let $C_{ab}(\boldsymbol{\xi})$ and $C_{ba}(\boldsymbol{\xi})$ denote the covariance between $\boldsymbol{a}_i(\boldsymbol{\xi})$ and $\boldsymbol{b}_i(\boldsymbol{\xi})$, and $\boldsymbol{b}_i(\boldsymbol{\xi})$ and $\boldsymbol{a}_i(\boldsymbol{\xi})$ respectively. Then,

$$\begin{aligned}\text{cov}\{\text{vec}[A_n(\boldsymbol{\xi})], \text{vec}[B_n(\boldsymbol{\xi})]\} &= \text{cov}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i(\boldsymbol{\xi}), \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{b}_i(\boldsymbol{\xi})\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{cov}[\boldsymbol{a}_i(\boldsymbol{\xi}), \boldsymbol{b}_i(\boldsymbol{\xi})] \\ &= \frac{1}{n}C_{ab}(\boldsymbol{\xi}), \quad (10.16)\end{aligned}$$

and similarly, $\text{cov}\{\text{vec}[B_n(\boldsymbol{\xi})], \text{vec}[A_n(\boldsymbol{\xi})]\} = \frac{1}{n}C_{ba}(\boldsymbol{\xi})$. Unbiased estimators for these quantities are given by

$$\begin{aligned}\widehat{C}_{ab}(\boldsymbol{\xi}) &= \frac{1}{n-1}\sum_{i=1}^{n}[\boldsymbol{a}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi})][\boldsymbol{b}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi})]^T \\ \widehat{C}_{ba}(\boldsymbol{\xi}) &= \frac{1}{n-1}\sum_{i=1}^{n}[\boldsymbol{b}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_b(\boldsymbol{\xi})][\boldsymbol{a}_i(\boldsymbol{\xi}) - \widehat{\boldsymbol{\mu}}_a(\boldsymbol{\xi})]^T.\end{aligned}$$

Using all these elements, we can now formulate the following theorem.

Theorem 10.2 (Modified Information Matrix Test) *Under general regularity conditions, and if the model is correctly specified, then, as* $n \to \infty$*,*

$$\boldsymbol{D}_n(\boldsymbol{\xi}_0) \sim N_p(\boldsymbol{0}, C_D(\boldsymbol{\xi}_0)), \quad (10.17)$$

and therefore

$$\Im_m(n) = \boldsymbol{D}_n^T(\boldsymbol{\xi}_0)[C_D(\boldsymbol{\xi}_0)]^{-1}\boldsymbol{D}_n(\boldsymbol{\xi}_0) \sim \chi_p^2. \tag{10.18}$$

Proof

First, let us recall that

$$\text{vec}[A_n(\boldsymbol{\xi}) + B_n(\boldsymbol{\xi})] = \frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{a}_i(\boldsymbol{\xi}) + \boldsymbol{b}_i(\boldsymbol{\xi})], \tag{10.19}$$

and therefore

$$\text{E}\{\text{vec}[A_n(\boldsymbol{\xi}) + B_n(\boldsymbol{\xi})]\} = \boldsymbol{\mu}_a(\boldsymbol{\xi}) + \boldsymbol{\mu}_b(\boldsymbol{\xi}).$$

Using (10.19) and the central limit theorem, under general regularity conditions, it follows that, asymptotically,

$$\text{vec}[A_n(\boldsymbol{\xi}) + B_n(\boldsymbol{\xi})] \sim \text{N}_{p^2}\left(\boldsymbol{\mu}_a(\boldsymbol{\xi}) + \boldsymbol{\mu}_b(\boldsymbol{\xi}), C^*(\boldsymbol{\xi})\right),$$

where $C^*(\boldsymbol{\xi})$ is given by (10.13). Hence,

$$\boldsymbol{D}_n(\boldsymbol{\xi}) \sim \text{N}_p(\Delta[\boldsymbol{\mu}_a(\boldsymbol{\xi}) + \boldsymbol{\mu}_b(\boldsymbol{\xi})], C_D(\boldsymbol{\xi})).$$

Finally, note that under a correctly specified model,

$$\begin{aligned}\boldsymbol{\mu}_a(\boldsymbol{\xi}_0) + \boldsymbol{\mu}_b(\boldsymbol{\xi}_0) &= \text{vec}[A(\boldsymbol{\xi}_0)] + \text{vec}[B(\boldsymbol{\xi}_0)] \\ &= \text{vec}[A(\boldsymbol{\xi}_0) + B(\boldsymbol{\xi}_0)] = \boldsymbol{0},\end{aligned}$$

such that (10.17) holds. □

To study how the Modified Information Matrix Test (MIMT) $\Im_m(n)$ performs, relative to the SET and the IMT in the setting of linear mixed models, we have applied the test to the normal response data generated in Chapter 9. The power obtained from the MIMT is displayed next to the results from the IMT and the SET in the first panel of Table 10.1. From this table we can see that the MIMT performs very similar to both the IMT and the SET.

Further, the second panel of Table 10.1 displays the type I error and the power of the MIMT to detect random-effects misspecification in generalized linear mixed models. Recall that, in this setting, the SET encountered problems to detect misspecification when the random intercept was generated from a power function or an

asymmetric mixture of two normal distributions, especially when $\sigma^2_{0b} = 4$. The MIMT clearly outperforms the SET, as well as the determinant and the determinant-trace tests, in the settings considered. Indeed, a good power can generally be observed with samples of 350 subjects or larger. Additionally, the MIMT has a very high power to detect the misspecification when $\sigma^2_{0b} = 32$, irrespective of the shape of the real distribution. In most settings displayed here, we observed a power of 70%, even when the data contained information on only 50 subjects. This clearly shows the potential of the MIMT to detect misspecification of the random-effects distribution, especially in those settings in which the misspecification can have a substantial negative impact.

Note that even though the SET and the MIMT were initially developed as tools to detect misspecification of the random-effects distribution, they are also suitable to detect other types of misspecifications. In the following section, we will further explore this possibility.

10.3 Misspecification of the Random-effects Structure

Heagerty and Kurland (2001) studied the impact of some random-effects misspecifications in a number of different settings. They estimated the bias induced by different violations of the random effects assumptions, including the fitting of a logistic-normal model, (i) when the random effect is generated from a non-normal distribution, (ii) when the variance of the random effect depends on a covariate in the linear predictor, (iii) when the random structure includes both a random intercept and slope, and (iv) when the random effects are auto-correlated.

Up to now we have mainly focused on the first setting, i.e., the random effects are generated from a non-normal distribution, but assumed to be normal in the model fitting. However, these authors found that the other misspecifications above could also induce an important bias in the estimates of the fixed effects parameters. Therefore, in this section, we will move away from the misspecification that constitutes the core of the present work and we will evaluate the performance of the previous tests to detect the other violations of the model assumptions studied by Heagerty and Kurland (2001).

10.3.1 Random Intercept Variance Depending on a Binary Covariate

In this setting, binary responses were generated using the model given by

$$\text{logit}\{P(y_{ij} = 1|b_{ij})\} = \beta_0 + \beta_1 z_i + \beta_2 t_j + \beta_3 z_i t_j + b_{ij}, \tag{10.20}$$

where z_i is a binary covariate (randomly assigned $z_i = 0$ or $z_i = 1$ with equal probabilities), t_j is a within-cluster covariate representing a linear trend, with $t_j = (j-1)/(n_i - 1)$, the variance of the random intercept $b_{ij} = b_{i0}$ is sampled from a distribution given by

$$b_{i0} \sim \begin{cases} \text{N}(0, \sigma_0^2) & \text{when} \quad z_i = 0 \\ \text{N}(0, \sigma_1^2) & \text{when} \quad z_i = 1, \end{cases} \tag{10.21}$$

and $n_i = 6$. The parameters in the linear predictor were fixed at $\beta_0^0 = -2$, $\beta_1^0 = 1$, $\beta_2^0 = 0.5$ and $\beta_3^0 = -0.25$. Further, three sample sizes were considered, including 100, 350 and 500 subjects. In total, 500 data sets were generated using the previous specifications, and model (10.20) was fitted to these generated data, assuming that $b_{ij} = b_{i0} \sim \text{N}(0, \sigma_b^2)$.

Heagerty and Kurland (2001) found that substantial bias can occur for all coefficients in the model, when σ_0 and σ_1 are very different. For example, they reported 38% and 31% of relative bias in the estimation of β_1 and β_3 respectively, when $\sigma_0 = 1$ and $\sigma_1 = 2$. Additionally, they observed that as the discrepancy between the two parameters increases, so does the bias in the parameter estimates.

To study the performance of our proposals in this particular setting, we applied the tests introduced in Chapter 9, the SET and the MIMT to the generated data sets and determined the proportion out of the 500 repetitions in which the tests were able to detect the misspecification (at a 5% significance level). The corresponding powers, for $n = 500$ are displayed in Table 10.2 as a function of σ_0 and σ_1 (the results for the other sample sizes are shown in Table E.1). Note that, when $\sigma_0 = \sigma_1$, these values correspond to the type I error rate.

Remarkably, all the tests introduced in Chapter 9 have, in general, a poor performance in this setting. The observed type I error largely exceeds the pre-specified value in some scenarios and the power is usually very small. They failed to detect the misspecification, even when the difference between σ_0 and σ_1 was largest. For instance, the determinant-trace test $\delta_{dt}(n)$ shows an excellent power when $\sigma_1 = 3.0$

Table 10.2: *Power of the determinant tests $\delta_{d1}(n)$ and $\delta_{d2}(n)$, the determinant-trace test $\delta_{dt}(n)$, the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ to detect model misspecification, when a logistic-normal model is assumed, but the variance of the random intercept depends on a binary cluster-level covariate, $[b_{i0}|z_i = 0] \sim N(0, \sigma_0^2)$ and $[b_{i0}|z_i = 1] \sim N(0, \sigma_1^2)$. (sample size $n = 500$).*

σ_1	σ_0	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$	$\delta_s(n)$	$\Im_m(n)$
0.5	0.5	0.100	0.080	0.106	0.040	0.064
	1.0	0.056	0.060	0.090	0.102	0.514
	2.0	0.176	0.108	0.026	0.984	1.000
	3.0	0.594	0.478	0.248	1.000	1.000
1.0	0.5	0.134	0.072	0.184	0.078	0.696
	1.0	0.050	0.042	0.070	0.018	0.038
	2.0	0.072	0.048	0.044	0.620	0.980
	3.0	0.227	0.165	0.107	0.994	1.000
2.0	0.5	0.364	0.236	0.638	0.770	1.000
	1.0	0.154	0.116	0.242	0.452	0.980
	2.0	0.062	0.046	0.064	0.020	0.020
	3.0	0.062	0.044	0.028	0.244	0.608
3.0	0.5	0.740	0.584	0.926	0.992	1.000
	1.0	0.444	0.328	0.654	0.974	1.000
	2.0	0.100	0.068	0.086	0.184	0.630
	3.0	0.066	0.064	0.046	0.016	0.014

and $\sigma_0 = 0.5$, but fails to detect the reverse situation, when $\sigma_1 = 0.5$ combined with $\sigma_0 = 3.0$, in 75% of the cases. Conversely, the SET showed a good power if differences between the two variance parameters were larger than 1.0. Nevertheless, the test with the best overall performance is the MIMT. This test is clearly able to detect problems in most of the considered settings, and especially in those for which the maximum likelihood estimators of the linear predictor are most affected. It is important to point out that both the SET and the MIMT can produce inflated type I error rates when small sample sizes are used (see Table E.1). However, this problem disappears when the sample size is increased.

10.3.2 Ignoring a Random Effect

Another type of misspecification in the random structure occurs when a random slope is incorrectly ignored. To study the performance of our proposals in this scenario, we have generated binary responses from the model given by (10.20), with $b_{ij} = b_{i0} + b_{i1}t_j$, and σ_0^2 and σ_1^2 representing the variance of the random intercept b_{i0} and slope b_{i1}, respectively.

Simulations by Heagerty and Kurland (2001) showed that when these data are analyzed wrongly assuming that $b_{ij} = b_{i0}$, moderate bias can appear in the estimation of the regression coefficients. For instance, they observed asymptotic relative biases as large as 30-50% in the estimates of β_2 and β_3 when σ_0 is small and σ_1 is large. On the other hand, the bias for the estimators of the intercept β_0 and the cluster-level covariate effect β_1 remained below 15% for all considered pairs of (σ_0, σ_1).

Table 10.3 shows the power of the diagnostic tools, for $n = 500$ subjects, to detect this type of misspecification, as a function of σ_0 and σ_1 (the results for the other sample sizes are displayed in Table E.2). As one would expect, all tests fail to detect the misspecification when σ_1 is small. However, the bias calculations by Heagerty and Kurland (2001) showed that little bias is present in this case. The SET and the MIMT increase their power when also σ_1 is increased, relative to σ_0. Nevertheless, when $\sigma_1 = 1$ and $\sigma_0 = 0.5$, precisely the setting in which bias as large as 52% was obtained for β_2 and β_3, we only observed a power of 55% with the SET to detect the misspecification, and 61% with the MIMT.

10.3.3 Autoregressive Random Effects

In the analysis of longitudinal data one often observes that the dependence between repeated measurements within a subject seems to decay as the time separation between the measurements increases. This could be accounted for with a generalized linear mixed model including autocorrelated random effects b_{ij} for which $\text{cov}(b_{ij}, b_{ik}) = \sigma^2\rho^{|t_{ij}-t_{ik}|}$. Simulations by Heagerty and Kurland (2001) with this type of misspecification have shown that substantial negative bias can arise in the estimated fixed effects, with increasing bias as σ increases and especially when ρ is small. Note that the random intercept model follows as a special case of the autoregressive model when $\rho = 1$. For models with $\rho < 1$, a potentially large negative bias can be observed in $\widehat{\sigma}_n$, given that it estimates the common variance and therefore approximates the true covariances $\sigma^2\rho^{|t_{ij}-t_{ik}|}$. These authors observed that as ρ de-

Table 10.3: *Power of the determinant tests $\delta_{d1}(n)$ and $\delta_{d2}(n)$, the determinant-trace test $\delta_{dt}(n)$, the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ to detect model misspecification, when a logistic-normal model is assumed, but the data are generated using both a random intercept and slope ($b_{ij} = b_{i0} + b_{i1}t_j$), with variance σ_0^2 and σ_1^2, respectively. (sample size $n = 500$).*

σ_1	σ_0	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$	$\delta_s(n)$	$\Im_m(n)$
0.2	0.5	0.078	0.070	0.110	0.030	0.066
	1.0	0.056	0.040	0.076	0.012	0.032
	2.0	0.050	0.048	0.048	0.008	0.018
	3.0	0.070	0.060	0.042	0.022	0.036
0.5	0.5	0.070	0.068	0.094	0.052	0.098
	1.0	0.066	0.068	0.102	0.026	0.056
	2.0	0.062	0.090	0.086	0.006	0.024
	3.0	0.058	0.067	0.053	0.013	0.032
0.8	0.5	0.068	0.116	0.120	0.230	0.282
	1.0	0.120	0.166	0.172	0.070	0.156
	2.0	0.178	0.242	0.224	0.022	0.076
	3.0	0.190	0.234	0.198	0.014	0.034
1.0	0.5	0.138	0.178	0.174	0.546	0.610
	1.0	0.210	0.300	0.260	0.234	0.394
	2.0	0.356	0.434	0.380	0.054	0.166
	3.0	0.368	0.468	0.394	0.044	0.012

creases, the negative bias in $\widehat{\sigma}_n$ increases, ranging between -30% and -50% when $\rho = 0.7$, and between -47% and -70% when $\rho = 0.5$. As a result, substantial negative bias can also arise in the estimated regression coefficients, with increasing bias as σ increases. For instance, when $(\rho, \sigma) = (0.5, 3.0)$ negative bias as high as -45% occurred in each of the linear predictor parameter estimates.

Table 10.4 shows the power of the diagnostic tools to detect this type of misspecification, as a function of σ and ρ. From the table it follows that, unlike in the previous misspecification settings, the SET and especially the MIMT are not as powerful as $\delta_{d1}(n)$, $\delta_{d2}(n)$ and $\delta_{dt}(n)$. When $\sigma \geq 2$, these tests are able to detect the misspecification in over 80% of the data. Further, all the diagnostic tools exhibit

Table 10.4: *Power of the determinant tests $\delta_{d1}(n)$ and $\delta_{d2}(n)$, the determinant-trace test $\delta_{dt}(n)$, the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ to detect model misspecification, when a logistic-normal model is assumed, but the data are generated using autocorrelated random effects b_{ij} such that $cov(b_{ij}, b_{ik}) = \sigma^2\rho^{|t_{ij}-t_{ik}|}$. (sample size $n = 500$).*

ρ	σ	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$	$\delta_s(n)$	$\Im_m(n)$
0.5	0.5	0.067	0.080	0.103	0.013	0.054
	1.0	0.182	0.267	0.333	0.046	0.064
	2.0	0.852	0.910	0.938	0.696	0.264
	3.0	0.992	0.998	0.998	0.980	0.594
0.7	0.5	0.095	0.099	0.128	0.019	0.071
	1.0	0.232	0.306	0.366	0.044	0.056
	2.0	0.954	0.970	0.972	0.888	0.544
	3.0	1.000	1.000	1.000	1.000	0.958
0.9	0.5	0.076	0.062	0.103	0.024	0.064
	1.0	0.092	0.148	0.190	0.018	0.034
	2.0	0.768	0.864	0.834	0.368	0.306
	3.0	0.998	0.998	0.998	0.922	0.924

a very good performance for $\sigma = 3$. Given that the bias in the estimation of the linear predictor parameters was seen to be more substantial as of $\sigma \geq 2$ (Heagerty and Kurland, 2001), this is a very desirable property.

10.4 Application: The Schizophrenia Data

In this section, we will apply the SET and the MIMT to assess the suitability of model (4.4) with normal random effects for the analysis of the case study. It follows that $\delta_s(n) = 0.324$ and compared to a χ^2 distribution with 4 degrees of freedom, this leads to $p = 0.988$. Additionally, $\Im_m(n) = 0.695$ with corresponding $p = 0.952$. These results imply that, with the data at hand, we do not have evidence of a departure from the assumption of normally distributed random effects or, in fact, of any other misspecification in the model. Given the power exhibited by the MIMT, even for small sample sizes, $n = 128$ might be considered sufficient to detect some gross departures. Therefore, these clearly non-significant p-values again allow us a comfortable level of

confidence in the final model.

Obviously, the availability of easily accessible diagnostic tools raises questions on how to proceed when the tests do produce significant results. Frequently, in statistics, estimators and inferential procedures show asymptotic robustness against some departures from the assumptions they are based on. In Chapter 11, we will explore this issue further. If the available sample size does not allow to rely on asymptotic arguments, alternative approaches are of the utmost importance. In Chapter 12, we will discuss some of these alternatives, including the use of the heterogeneity model and its implementation within a more general sensitivity analysis framework.

Chapter 11

Asymptotic Robustness

In the context of shared parameter models, where interest lies in the association structure between a longitudinal and a survival process, Rizopoulos, Verbeke and Molenberghs (2007) studied the effect of misspecifying the random-effects distribution on the parameter estimates. They argued that the impact of the misspecification on the estimation of the mean structure in the longitudinal process diminishes as the number of repeated observations per subject increases. They also claimed that the maximum likelihood estimator of the variance components remains biased, even when both the number of subjects and the number of observations per subject go to infinity. In this chapter, we bring this result into the generalized linear mixed models framework including any number of random effects, and analyze its practical implications in the light of all our previous findings.

11.1 Asymptotic Robustness: Consistency

In the context of linear mixed models, Verbeke and Lesaffre (1997) have shown that the maximum likelihood estimators are consistent, even when the random-effects distribution is misspecified, as far as the mean and the covariance structure are correctly specified (see Chapter 6). These authors analyzed consistency when the number of subjects increases, and the number of observations per subject is kept constant. As was illustrated in Chapter 7, such a result does not necessarily hold for generalized linear mixed models. We observed that severe bias is usually present in the esti-

mates of the variance components, and the fixed effects estimates can also be severely affected, especially when the underlying population is very heterogeneous.

Regularly, consistency for the model parameters is stated as the number of clusters (n) increases (Agresti, 2002). In the following theorem, we will argue that, in order to obtain maximum likelihood estimates of the fixed effects parameters close to the real values, both the number of clusters n and the number of observations per cluster n_i should be sufficiently large.

Theorem 11.1 (Asymptotic Robustness) *Let y_{ij} denote the jth measurement for the ith subject, with $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Conditional on a vector $\boldsymbol{b}_i$ of individual random effects for subject i, it is assumed that all responses y_{ij} are independent with density belonging to the exponential family (3.1), where*

$$\mu_{ij} = E(y_{ij}|\boldsymbol{b}_i) = v(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i), \tag{11.1}$$

$v(.)$ denotes a known link function, $\boldsymbol{x}_{ij}$ denotes a p-dimensional vector of covariates, $\boldsymbol{z}_{ij}$ is a q-dimensional vector, $\boldsymbol{\beta}$ is a vector of fixed parameters and $\boldsymbol{b}_i$ is a vector of random effects assumed to follow a density $f(\boldsymbol{b}_i|\boldsymbol{\delta})$, with $E(\boldsymbol{b}_i) = \boldsymbol{0}$, which differs from the true random-effects distribution denoted by $h(\boldsymbol{b}_i)$. Further, let $\widehat{\boldsymbol{\beta}}_F(n, m)$ denote the maximum likelihood estimator of $\boldsymbol{\beta}$, under the assumed model, where $m = min(n_i)$, and let $\boldsymbol{\beta}_0$ represent the true value for $\boldsymbol{\beta}$.

Then, under general regularity conditions, as $n \to \infty$ and $m \to \infty$,

$$\widehat{\boldsymbol{\beta}}_F(n, m) \xrightarrow{P} \boldsymbol{\beta}_0. \tag{11.2}$$

Proof

Under the assumed model, the likelihood contribution for every subject i can be written as

$$f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\delta}) = \int f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i|\delta) d\boldsymbol{b}_i, \tag{11.3}$$

and the marginal loglikelihood has the form

$$\ell(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \log \int f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{\delta}) d\boldsymbol{b}_i. \tag{11.4}$$

Further, the posterior distribution of $\boldsymbol{b}_i$ can be written as

$$f_i(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{\delta})}{\int f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{\delta}) d\boldsymbol{b}_i}, \tag{11.5}$$

therefore,

$$\int f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i)f(\boldsymbol{b}_i|\boldsymbol{\delta})d\boldsymbol{b}_i = \frac{f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i)f(\boldsymbol{b}_i|\boldsymbol{\delta})}{f_i(\boldsymbol{b}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\delta})}. \tag{11.6}$$

This implies that

$$\log \int f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i)f(\boldsymbol{b}_i|\boldsymbol{\delta})d\boldsymbol{b}_i = \log f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i) - \log f_i(\boldsymbol{b}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\delta}) + \log f(\boldsymbol{b}_i|\boldsymbol{\delta}).$$

As a result, the marginal loglikelihood can be written as

$$\ell(\boldsymbol{\beta},\boldsymbol{\delta}) = \sum_{i=1}^{n}[\log f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i) - \log f_i(\boldsymbol{b}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\delta}) + \log f(\boldsymbol{b}_i|\boldsymbol{\delta})]. \tag{11.7}$$

Since $f(\boldsymbol{b}_i|\boldsymbol{\delta})$ does not depend on $\boldsymbol{\beta}$, maximizing (11.7) with respect to $\boldsymbol{\beta}$ is equivalent to maximizing

$$\ell_1(\boldsymbol{\beta},\boldsymbol{\delta}) = \sum_{i=1}^{n}[\log f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i) - \log f_i(\boldsymbol{b}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\delta})]. \tag{11.8}$$

Now, if $\boldsymbol{b}_{0i}$ denotes the value of the random effects for subject i, then under some regularity conditions, the Bayesian central limit theorem (see Appendix F) guarantees that, as $n_i \to \infty$,

$$f_i(\boldsymbol{b}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\delta}) \xrightarrow{P} \mathrm{N}(\boldsymbol{b}_{0i}, H^{-1}(\boldsymbol{b}_{0i})),$$

where

$$H(\boldsymbol{b}_{0i}) = E\left(-\frac{\partial^2 \log f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i)}{\partial b_{ij}\partial b_{ik}}\right)\Bigg|_{\boldsymbol{b}_i=\boldsymbol{b}_{0i}}. \tag{11.9}$$

Note that (11.9) depends only on $f_i(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i)$, i.e., the conditional distribution of the responses, which is assumed to be correctly specified. Therefore, when $m \to \infty$, we have that, irrespective of the prior distribution used to model the random effects,

$$\ell_1(\boldsymbol{\beta},\boldsymbol{\delta}) \xrightarrow{P} Q(\boldsymbol{\beta},\boldsymbol{\delta}), \tag{11.10}$$

where

$$Q(\boldsymbol{\beta},\boldsymbol{\delta}) = \sum_{i=1}^{n}[\log f(\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{b}_i) - \log \mathrm{N}(\boldsymbol{b}_{0i}, H^{-1}(\boldsymbol{b}_{0i}))]. \tag{11.11}$$

Now, let $\ell_{1F}(\boldsymbol{\beta},\boldsymbol{\delta})$ and $\ell_{1H}(\boldsymbol{\beta},\boldsymbol{\delta})$ correspond to (11.8) under the assumed and the correct distribution for the random effects. Then, (11.10) implies that, for all $\varepsilon > 0$ when $m \to \infty$,

$$P[|\ell_{1F}(\boldsymbol{\beta},\boldsymbol{\delta}) - Q(\boldsymbol{\beta},\boldsymbol{\delta})| < \varepsilon] \quad \to \quad 1 \tag{11.12}$$

$$P[|\ell_{1H}(\boldsymbol{\beta},\boldsymbol{\delta}) - Q(\boldsymbol{\beta},\boldsymbol{\delta})| < \varepsilon] \quad \to \quad 1. \tag{11.13}$$

Further, the inequality

$$|\ell_{1F}(\boldsymbol{\beta},\boldsymbol{\delta}) - \ell_{1H}(\boldsymbol{\beta},\boldsymbol{\delta})| \leq |\ell_{1F}(\boldsymbol{\beta},\boldsymbol{\delta}) - Q(\boldsymbol{\beta},\boldsymbol{\delta})| + |\ell_{1H}(\boldsymbol{\beta},\boldsymbol{\delta}) - Q(\boldsymbol{\beta},\boldsymbol{\delta})|,$$

implies that

$$\begin{aligned} &P[|\ell_{1F}(\boldsymbol{\beta},\boldsymbol{\delta}) - \ell_{1H}(\boldsymbol{\beta},\boldsymbol{\delta})| < \varepsilon] \\ &\qquad \geq \; P\left[|\ell_{1F}(\boldsymbol{\beta},\boldsymbol{\delta}) - Q(\boldsymbol{\beta},\boldsymbol{\delta})| < \frac{\varepsilon}{2} \text{ and } |\ell_{1H}(\boldsymbol{\beta},\boldsymbol{\delta}) - Q(\boldsymbol{\beta},\boldsymbol{\delta})| < \frac{\varepsilon}{2}\right]. \end{aligned}$$

This last inequality, together with (11.12) and (11.13) implies that, when $m \to \infty$,

$$P[|\ell_{1F}(\boldsymbol{\beta},\boldsymbol{\delta}) - \ell_{1H}(\boldsymbol{\beta},\boldsymbol{\delta})| < \varepsilon] \to 1. \tag{11.14}$$

If we denote by $\widehat{\boldsymbol{\beta}}_F(n,m)$ and $\widehat{\boldsymbol{\beta}}_H(n,m)$ the maximas associated with $\ell_{1F}(\boldsymbol{\beta},\boldsymbol{\delta})$ and $\ell_{1H}(\boldsymbol{\beta},\boldsymbol{\delta})$ respectively, then under certain regularity conditions (11.14) implies that, for all $\varepsilon > 0$ as $m \to \infty$,

$$P[|\widehat{\boldsymbol{\beta}}_F(n,m) - \widehat{\boldsymbol{\beta}}_H(n,m)| < \varepsilon] \to 1. \tag{11.15}$$

Further, $\widehat{\boldsymbol{\beta}}_H(n,m)$ is a consistent estimator for $\boldsymbol{\beta}_0$, so as $n \to \infty$,

$$P[|\widehat{\boldsymbol{\beta}}_H(n,m) - \boldsymbol{\beta}_0| < \varepsilon] \to 1. \tag{11.16}$$

Finally, the following inequality

$$|\widehat{\boldsymbol{\beta}}_F(n,m) - \boldsymbol{\beta}_0| \leq |\widehat{\boldsymbol{\beta}}_F(n,m) - \widehat{\boldsymbol{\beta}}_H(n,m)| + |\widehat{\boldsymbol{\beta}}_H(n,m) - \boldsymbol{\beta}_0|$$

implies that

$$\begin{aligned} &P[|\widehat{\boldsymbol{\beta}}_F(n,m) - \boldsymbol{\beta}_0| < \varepsilon] \\ &\qquad \geq \; P\left[|\widehat{\boldsymbol{\beta}}_F(n,m) - \widehat{\boldsymbol{\beta}}_H(n,m)| < \frac{\varepsilon}{2} \text{ and } |\widehat{\boldsymbol{\beta}}_H(n,m) - \boldsymbol{\beta}_0| < \frac{\varepsilon}{2}\right]. \end{aligned}$$

Therefore, this last inequality, together with (11.15) and (11.16), implies that, as $n \to \infty$ and $m \to \infty$, we have

$$P[|\widehat{\boldsymbol{\beta}}_F(n,m) - \boldsymbol{\beta}_0| < \varepsilon] \to 1. \tag{11.17}$$

□

In practice, this suggests that we need both the number of clusters as well as the number of observations per cluster to go to infinity, to guarantee consistency of the

maximum likelihood estimators for the fixed effects in the model. The same result cannot be reproduced for the variance components associated with $\boldsymbol{b}_i$. In this case, the prior distribution of the random effects cannot be discarded from the likelihood and its impact does not fade away. This essentially implies that the maximum likelihood estimator of the variance components will be inconsistent under misspecification, even when both the number of clusters as well as the number of observations per cluster go to infinity.

Note that the previous result is intuitively appealing, especially from a Bayesian perspective. Indeed, it is known from asymptotic Bayesian analysis that the impact of the prior distribution on our inferences fades away when a lot of data are available. Precisely the situation we have when both m and n go to infinity. Even though this theorem represents an elegant asymptotic result, its use can be limited if unrealistic sample sizes are required. Hence, it is of interest to know how many repeated measurements per subject are required for this result to hold. In the next section, we will study the practical implications of Theorem 11.1 via simulations with different numbers of subjects and observations per subject.

11.2 Impact of Increasing the Number of Observations per Cluster

The practical scope of Theorem 11.1 basically depends on two important factors: i) the magnitude of the sample size required to achieve a reasonable level of precision, and ii) the practical plausibility of the assumptions it relies on. In this section, we will explore the first factor via simulations, and in what follows we will try to analyze the second issue in some detail. Essentially, Theorem 11.1 is based on the asymptotic normality of the posterior distribution. There is a large literature available on the regularity conditions required to justify mathematically this important result. Those who have contributed to the field include, Chen (1985), Sweeting and Adekola (1987), Fu and Kass (1988), Fraser and McDunnough (1984), Sweeting (1992), and Gosh *et al.* (1994).

Gelman *et al.* (1995) offered some counterexamples to the Bayesian central limit theorem. These counterexamples generally correspond to situations in which the prior distribution has an impact on the posterior, even in the limit of infinite sample sizes. They claim that problems can arise, for example, when the model is underi-

dentified, in the presence of aliasing or when the posterior has thick tails. Bernardo and Smith (1994, Section 5.3) proposed three basic conditions which are sufficient to ensure a valid normal approximation for the posterior. He called these assumptions steepness, smoothness and concentration. A detailed description of all of them can be found in Appendix F. A closer look to the concentration assumption shows that it requires the probability outside any neighborhood of the posterior mode to become negligible when the sample size increases. This may not be the case for multimodal posteriors, to mention one example. In general, multimodal priors could in principle lead to multimodal posteriors and therefore, Theorem 11.1 could fail if the real unknown distribution has several modes. In any practical application, the real random-effects distribution will be totally unknown and previous asymptotic results should be considered with care. Further, when the variance components are of interest, misspecification of the random-effects distribution may continue to be a problem.

In this section, we will study the implications of Theorem 11.1 via simulations with a logistic random-intercept model. Let binary responses be generated using model (7.1), including an intercept, a binary covariate z_i, a within-cluster covariate t_j, and a random intercept b_i. We will consider longitudinal sequences of 3 different lengths, including

- 4 repeated measurements, taking values 0, 1, 2, and 4,
- 6 repeated measurements, taking values 0, 1, 2, 4, 6, and 8, and
- 8 repeated measurements, taking values 0, 1, 2, 4, 6, 8, 10, and 12.

The random intercept b_i was sampled from 5 distinct mean-zero distributions, including a normal, power function, and lognormal distribution, as well as a discrete distribution with equal probability at two support points, and an asymmetric mixture of two normal densities, and each with variances $\sigma^2_{0b} = 1$, 4, 16, and 32. Further, the parameters in the linear predictor were fixed at $\beta^0_0 = -8$, $\beta^0_1 = 2$, and $\beta^0_2 = 1$. Six different sample sizes were considered, namely 50, 100, 200, 400, 800, and 1600 subjects. For each setting, 500 data sets were generated, and the model given by (7.1) was used to analyze these generated data, assuming normally distributed random effects.
First, we would like to point out that a high number of non-converging analyses were obtained, especially for those settings with few repeated measurements and subjects (see Table 11.1 for the convergence rates of the analyses based on 4 repeated measure-

Table 11.1: *Percentage of converged analyses with model (7.1) assuming a normal random intercept, when the data were generated using model (7.1), with 4 repeated measurements, n subjects and 5 different random-intercept distributions incl. a normal (No), a lognormal (LN), a power function (PF) a discrete (D), and an asymmetric mixture (AM) of two normal distributions, each with variance* σ^2_{0b}.

n	No	LN	PF	D	AM	n	No	LN	PF	D	AM
	$\sigma^2_{0b}=1$						$\sigma^2_{0b}=4$				
50	42	54	33	35	32	50	77	80	49	67	63
100	56	72	38	44	34	100	94	93	61	82	77
200	70	88	47	59	50	200	99	100	69	91	90
400	82	99	59	68	60	800	100	100	90	99	99
800	85	100	63	76	62	800	100	100	90	99	99
1600	95	100	68	82	65	1600	100	100	95	100	100
	$\sigma^2_{0b}=16$						$\sigma^2_{0b}=32$				
50	99	86	95	100	95	50	98	90	100	100	100
100	100	97	100	100	99	100	100	97	100	100	100
200	100	100	100	100	100	200	100	98	100	100	100
400	100	100	100	100	100	400	100	100	100	100	100
800	100	100	100	100	100	800	100	100	100	100	100
1600	100	100	100	100	100	1600	100	100	100	100	100

ments per subject). For instance, more than half of the analyses with 50 subjects failed to converge when $\sigma^2_{0b}=1$. This low convergence rate can most likely be attributed to the limited amount of available information. Likely, with only 50 subjects and 4 time points, there is not enough information to distinguish the variability introduced by a random intercept with variance $\sigma^2_{0b}=1$, from the overall variability. Additionally, we observed that in many simulations, the procedure NLMIXED converged to an ill-conditioned maxima, resulting in extreme values for both the parameter estimates and their standard errors.

The median of the maximum likelihood estimates of β_0, β_1, β_2, and σ^2_b obtained from the converged analyses are displayed in Tables 11.2 to 11.5. In general, the results of this new simulation study fully resemble our previous findings.

Table 11.2: *Median of the maximum likelihood estimates $\widehat{\beta}_{0n}$ obtained from fitting model (7.1) to the data generated using different numbers of time points, sample sizes n and different random-effects distributions including a normal (No), a lognormal (LN), a power function (PF) a discrete (D) distribution, as well as an asymmetric mixture (AM) of two normal distributions, each with variance σ^2_{0b} (note that $\beta^0_0 = -8$ was used to generate the data).*

		4 time points				6 time points				8 time points			
	n	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
No	50	-23.92	-8.95	-8.02	-8.12	-8.26	-8.18	-8.12	-8.11	-8.18	-8.00	-8.06	-8.19
	100	-9.25	-8.12	-8.00	-7.83	-8.12	-7.92	-7.94	-8.15	-8.06	-8.03	-8.05	-7.94
	200	-8.64	-8.13	-8.17	-8.00	-8.04	-8.10	-8.05	-8.03	-8.05	-8.03	-8.09	-7.95
	400	-8.14	-8.01	-8.00	-8.05	-7.99	-7.96	-8.01	-8.03	-7.98	-8.01	-8.03	-8.00
	800	-8.20	-8.02	-8.04	-8.02	-8.02	-8.02	-7.99	-8.03	-8.00	-7.99	-7.98	-8.03
	1600	-8.06	-8.00	-7.98	-8.01	-7.98	-7.99	-8.01	-8.06	-8.01	-8.00	-8.00	-8.00
LN	50	-20.50	-12.31	-13.46	-16.84	-8.22	-8.67	-9.01	-9.78	-8.19	-8.22	-8.53	-8.63
	100	-10.31	-10.70	-13.62	-15.58	-8.21	-8.43	-9.10	-9.48	-8.06	-8.15	-8.39	-8.70
	200	-9.25	-10.63	-12.91	-14.56	-8.10	-8.39	-8.93	-9.57	-7.99	-8.20	-8.45	-8.63
	400	-9.06	-10.91	-12.68	-14.69	-8.11	-8.36	-8.96	-9.40	-8.02	-8.18	-8.41	-8.59
	800	-8.95	-10.76	-12.74	-14.38	-8.10	-8.44	-8.93	-9.46	-8.05	-8.19	-8.40	-8.61
	1600	-8.95	-10.65	-12.72	-14.06	-8.08	-8.39	-8.95	-9.43	-8.04	-8.18	-8.41	-8.58
PF	50	-34.93	-11.00	-6.85	-6.01	-8.15	-7.89	-7.41	-6.77	-8.25	-8.03	-7.93	-7.51
	100	-10.81	-8.28	-6.62	-5.89	-8.06	-7.85	-7.33	-6.85	-8.10	-8.08	-7.78	-7.29
	200	-8.72	-7.74	-6.60	-5.93	-8.02	-7.78	-7.32	-6.92	-8.01	-8.06	-7.73	-7.19
	400	-8.51	-7.71	-6.62	-5.94	-7.95	-7.76	-7.24	-6.67	-8.05	-8.00	-7.64	-7.21
	800	-8.13	-7.56	-6.60	-5.97	-7.92	-7.75	-7.28	-6.75	-8.03	-8.00	-7.74	-7.26
	1600	-8.01	-7.48	-6.66	-5.94	-7.93	-7.76	-7.26	-6.73	-8.00	-8.00	-7.70	-7.20
D	50	-32.94	-7.84	-5.74	-4.98	-8.30	-8.20	-8.02	-7.18	-8.04	-7.92	-7.54	-7.29
	100	-9.58	-7.60	-5.77	-4.88	-8.15	-8.14	-8.04	-7.35	-8.03	-7.99	-7.65	-7.63
	200	-8.39	-7.43	-5.80	-4.92	-8.01	-8.11	-7.94	-7.07	-8.07	-8.00	-7.86	-7.71
	400	-8.29	-7.40	-5.79	-4.96	-8.00	-8.15	-8.14	-7.41	-8.05	-8.07	-7.92	-7.96
	800	-8.07	-7.28	-5.76	-4.90	-8.03	-8.12	-7.98	-7.35	-8.02	-8.01	-7.81	-7.80
	1600	-7.96	-7.25	-5.75	-4.83	-8.04	-8.14	-8.04	-7.18	-8.03	-8.01	-7.85	-7.77
AM	50	-34.33	-8.21	-6.83	-5.89	-8.43	-7.55	-7.14	-6.34	-8.14	-7.64	-8.03	-7.41
	100	-10.76	-7.84	-6.67	-5.72	-8.03	-7.58	-7.09	-6.35	-8.06	-7.66	-8.07	-7.41
	200	-8.65	-7.51	-6.69	-5.70	-7.99	-7.49	-7.16	-6.31	-8.03	-7.63	-8.02	-7.37
	400	-8.28	-7.41	-6.62	-5.74	-7.98	-7.48	-7.15	-6.27	-8.03	-7.65	-7.95	-7.44
	800	-8.09	-7.31	-6.64	-5.68	-7.95	-7.47	-7.13	-6.33	-8.02	-7.64	-7.95	-7.38
	1600	-7.95	-7.31	-6.60	-5.72	-7.96	-7.49	-7.12	-6.30	-8.01	-7.61	-7.99	-7.38

Table 11.3: *Median of the maximum likelihood estimates $\widehat{\beta}_{1n}$ obtained from fitting model (7.1) to the data generated using different numbers of time points, sample sizes n and different random-effects distributions including a normal (No), a lognormal (LN), a power function (PF) a discrete (D) distribution, as well as an asymmetric mixture (AM) of two normal distributions, each with variance σ^2_{0b} (note that $\beta^0_1 = 2$ was used to generate the data).*

		4 time points				6 time points				8 time points			
	n	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
No	50	2.48	2.47	2.03	2.06	2.12	2.06	2.10	2.11	2.07	1.99	2.00	1.87
	100	2.27	2.12	2.12	1.85	2.01	2.03	2.04	2.04	2.00	2.04	2.04	1.95
	200	2.28	1.99	2.11	2.05	2.02	2.05	2.03	1.98	2.01	2.02	2.03	1.94
	400	2.07	2.00	1.99	2.00	2.01	1.97	1.98	2.02	1.99	1.99	2.03	1.95
	800	2.07	2.00	2.02	2.01	2.01	2.01	1.98	2.02	2.01	2.00	1.97	2.01
	1600	2.02	2.00	2.01	1.99	1.99	1.99	1.99	2.02	2.00	1.99	1.99	1.99
LN	50	2.24	2.67	3.12	3.24	2.07	2.26	2.46	2.64	2.05	2.04	2.02	2.03
	100	2.50	2.21	2.53	2.52	2.10	2.20	2.46	2.60	2.04	1.99	2.02	2.02
	200	2.33	2.28	2.47	2.41	2.05	2.20	2.37	2.60	2.01	2.02	2.04	2.03
	400	2.11	2.39	2.44	2.37	2.06	2.20	2.38	2.56	2.01	2.01	2.01	2.02
	800	2.11	2.34	2.40	2.28	2.08	2.23	2.42	2.55	2.02	2.02	1.99	2.02
	1600	2.09	2.32	2.36	2.23	2.08	2.21	2.39	2.55	2.02	2.01	2.00	1.99
PF	50	2.11	2.36	1.99	1.76	2.04	1.95	1.77	1.93	2.09	2.05	2.16	2.29
	100	2.43	2.12	1.86	1.81	2.00	1.92	1.74	1.81	2.03	2.02	2.04	1.99
	200	2.19	2.00	1.93	1.89	1.97	1.90	1.75	1.97	2.00	1.96	1.85	1.86
	400	2.13	1.99	1.93	1.85	1.95	1.84	1.64	1.63	1.97	1.93	1.79	1.69
	800	2.05	1.98	1.92	1.84	1.95	1.86	1.73	1.71	1.99	1.96	1.84	1.80
	1600	2.02	1.95	1.95	1.87	1.94	1.84	1.68	1.70	1.98	1.92	1.80	1.75
D	50	2.06	2.28	1.64	1.79	2.09	1.94	1.55	1.36	1.86	1.76	1.59	2.01
	100	2.48	2.22	1.81	1.76	2.07	1.98	1.64	1.69	2.00	1.97	2.05	2.65
	200	2.28	2.07	1.83	1.77	2.03	2.03	1.75	1.34	1.99	2.06	2.32	2.90
	400	2.17	2.10	1.90	1.91	2.07	2.12	1.99	2.04	2.07	2.14	2.46	3.27
	800	2.08	2.04	1.81	1.77	2.04	2.04	1.79	1.89	2.03	2.06	2.29	2.96
	1600	2.04	2.00	1.77	1.60	2.07	2.07	1.89	1.52	2.04	2.08	2.33	3.01
AM	50	2.03	2.28	2.06	2.04	2.10	1.92	1.79	1.78	2.04	1.97	1.87	1.61
	100	2.45	2.26	1.97	1.97	1.97	1.94	1.72	1.78	2.00	1.99	1.88	1.73
	200	2.31	2.10	1.98	1.96	1.97	1.92	1.80	1.70	1.98	1.99	1.89	1.74
	400	2.11	2.03	1.98	1.95	1.97	1.92	1.80	1.70	2.02	1.97	1.86	1.75
	800	2.10	2.01	1.97	1.93	1.97	1.92	1.76	1.71	2.00	2.01	1.85	1.69
	1600	2.05	2.00	1.97	1.94	1.96	1.92	1.76	1.70	2.00	1.99	1.87	1.69

Table 11.4: *Median of the maximum likelihood estimates $\widehat{\beta}_{2n}$ obtained from fitting model (7.1) to the data generated using different numbers of time points, sample sizes n and different random-effects distributions including a normal (No), a lognormal (LN), a power function (PF) a discrete (D) distribution, as well as an asymmetric mixture (AM) of two normal distributions, each with variance σ^2_{0b} (note that $\beta^0_2 = 1$ was used to generate the data).*

		4 time points				6 time points				8 time points			
	n	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
No	50	1.55	1.06	1.00	1.02	1.04	1.01	1.02	1.02	1.01	1.00	1.01	1.02
	100	1.06	1.01	1.02	1.01	1.01	0.99	1.00	1.02	1.01	1.00	1.01	1.01
	200	1.03	1.02	1.02	0.99	1.00	1.01	1.00	1.01	1.00	1.00	1.01	1.00
	400	1.01	0.99	1.00	1.01	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.01
	800	1.02	1.00	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1600	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.00
LN	50	1.14	1.08	1.02	1.14	1.02	1.05	1.01	1.04	1.03	1.03	1.04	1.01
	100	1.07	1.02	1.08	1.07	1.01	1.01	1.02	1.00	1.01	1.02	1.02	1.04
	200	1.03	1.03	1.05	1.06	1.00	1.01	0.99	1.00	1.00	1.02	1.03	1.02
	400	1.03	1.05	1.02	1.07	1.00	1.00	1.00	0.99	1.00	1.02	1.03	1.02
	800	1.02	1.02	1.03	1.06	1.00	1.01	1.00	1.00	1.00	1.02	1.02	1.02
	1600	1.01	1.03	1.04	1.05	1.00	1.00	1.00	0.99	1.00	1.02	1.02	1.02
PF	50	7.66	1.12	1.02	0.99	1.04	1.02	1.02	0.99	1.04	1.01	1.01	0.98
	100	1.17	1.07	1.00	0.98	1.02	1.01	1.01	0.99	1.01	1.02	1.01	0.98
	200	1.02	1.03	0.98	0.98	1.01	1.01	1.01	1.00	1.00	1.02	1.01	0.97
	400	1.05	1.02	0.98	0.98	1.00	1.01	1.01	0.98	1.01	1.01	1.01	0.99
	800	1.02	1.02	0.99	0.98	1.00	1.00	1.01	0.98	1.01	1.01	1.01	0.98
	1600	1.01	1.00	0.99	0.97	1.00	1.00	1.01	0.98	1.00	1.01	1.01	0.98
D	50	3.08	1.06	0.95	0.93	1.04	1.03	1.00	0.88	1.01	1.00	0.94	0.83
	100	1.06	1.03	0.96	0.92	1.02	1.02	1.00	0.88	1.01	1.00	0.93	0.81
	200	1.04	1.02	0.96	0.92	1.00	1.01	0.99	0.87	1.01	1.00	0.93	0.81
	400	1.04	1.00	0.96	0.92	1.00	1.01	0.99	0.87	1.00	1.00	0.93	0.81
	800	1.03	1.00	0.95	0.92	1.00	1.02	0.99	0.87	1.00	1.00	0.93	0.81
	1600	1.01	0.99	0.95	0.92	1.00	1.01	0.99	0.87	1.00	0.99	0.93	0.80
AM	50	5.64	1.10	1.03	0.99	1.06	1.02	1.02	1.02	1.03	1.01	1.03	1.01
	100	1.12	1.06	1.01	1.00	1.01	1.01	1.00	1.01	1.01	1.01	1.05	1.01
	200	1.07	1.02	1.01	0.99	1.01	1.00	1.02	1.00	1.01	1.01	1.03	1.01
	400	1.03	1.02	1.00	0.99	1.00	1.00	1.01	1.00	1.00	1.01	1.02	1.01
	800	1.02	1.00	1.00	0.98	1.00	1.00	1.01	1.00	1.00	1.01	1.02	1.01
	1600	1.01	1.00	0.99	0.99	1.00	1.00	1.01	1.00	1.00	1.00	1.02	1.01

Table 11.5: *Median of the maximum likelihood estimates $\widehat{\sigma}^2_{bn}$ obtained from fitting model (7.1) to the data generated using different numbers of time points, sample sizes n and different random-effects distributions including a normal (No), a lognormal (LN), a power function (PF) a discrete (D) distribution, as well as an asymmetric mixture (AM) of two normal distributions, each with variance σ^2_{0b}.*

		4 time points				6 time points				8 time points			
	n	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
No	50	2.26	4.13	14.98	30.80	1.05	3.84	15.68	31.99	0.97	3.61	15.43	30.29
	100	1.99	3.99	15.55	31.55	1.01	3.74	15.64	32.52	0.96	3.83	16.08	31.07
	200	1.50	3.82	16.31	31.44	0.98	3.98	15.72	31.45	0.96	3.96	16.24	31.80
	400	1.09	3.92	15.92	31.96	1.00	3.89	15.83	32.38	0.98	4.00	15.95	32.25
	800	1.11	4.06	16.03	32.25	1.00	4.03	15.76	32.11	0.99	3.97	15.82	32.08
	1600	0.99	4.00	15.89	31.97	0.99	3.98	15.93	32.41	0.99	4.00	16.00	32.04
LN	50	3.26	16.34	35.33	71.68	1.30	4.22	9.36	13.85	1.08	3.00	6.12	7.47
	100	4.05	14.33	39.67	68.44	1.29	4.07	10.00	13.91	1.00	3.12	6.02	8.10
	200	3.87	14.19	38.62	59.57	1.27	4.30	9.56	13.85	0.99	3.04	6.33	7.84
	400	3.96	15.40	36.52	61.19	1.30	4.25	9.69	14.22	1.05	3.17	6.22	8.09
	800	3.87	15.24	38.10	59.92	1.32	4.40	9.83	14.15	1.06	3.18	6.27	8.08
	1600	3.94	14.78	37.90	57.86	1.34	4.35	9.81	14.04	1.07	3.22	6.25	8.07
PF	50	2.01	2.05	2.43	5.30	0.75	2.04	6.15	11.21	0.99	3.16	10.11	17.55
	100	1.79	1.66	2.37	5.14	0.68	1.90	6.22	11.93	0.97	3.22	9.94	17.22
	200	1.28	1.02	2.42	5.40	0.68	1.97	6.18	11.58	0.92	3.31	9.97	16.87
	400	0.87	1.00	2.54	5.44	0.66	1.98	6.18	11.50	0.92	3.26	10.00	17.52
	800	0.58	0.88	2.48	5.51	0.67	2.01	6.22	11.61	0.94	3.32	10.05	17.41
	1600	0.48	0.84	2.55	5.49	0.66	1.99	6.21	11.55	0.92	3.26	10.02	17.27
D	50	2.24	1.80	3.61	9.92	1.03	4.27	19.40	38.61	0.93	3.84	15.32	36.05
	100	1.75	1.54	3.67	10.03	1.04	4.31	20.08	39.34	0.95	3.90	15.61	37.49
	200	1.08	1.33	3.72	10.19	0.98	4.33	19.79	39.22	0.99	3.91	16.00	38.02
	400	0.96	1.25	3.60	10.09	0.98	4.33	19.87	40.16	0.95	3.92	16.01	38.52
	800	0.71	1.25	3.67	10.14	1.00	4.40	19.81	39.88	0.98	3.92	15.95	38.08
	1600	0.67	1.26	3.73	10.19	1.01	4.39	19.82	39.61	1.00	3.96	15.94	37.81
AM	50	2.13	1.92	3.12	3.39	1.00	2.16	5.93	8.36	1.00	2.79	14.05	22.06
	100	1.72	1.51	3.03	3.41	0.81	2.10	6.09	8.39	0.92	2.86	14.71	22.22
	200	1.26	1.38	2.89	3.45	0.80	2.11	6.31	8.43	0.95	2.89	14.59	22.26
	400	0.79	1.31	2.95	3.52	0.79	2.12	6.24	8.32	0.99	2.93	14.55	22.84
	800	0.65	1.26	2.99	3.41	0.80	2.16	6.31	8.35	0.97	2.96	14.49	22.74
	1600	0.46	1.33	2.90	3.45	0.79	2.16	6.26	8.36	0.98	2.92	14.52	22.57

Table 11.6: *Median maximum likelihood estimates obtained from fitting the logistic-normal model (7.1) to binary response data generated using 20 time points,* $\sigma^2_{0b} = 32$*, and a random intercept sampled from an asymmetric mixture of two normals, a power function and a discrete distribution.*

	Asymmetric				Power function				Discrete			
n	$\widehat{\beta}_{0n}$	$\widehat{\beta}_{1n}$	$\widehat{\beta}_{2n}$	$\widehat{\sigma}^2_{bn}$	$\widehat{\beta}_{0n}$	$\widehat{\beta}_{1n}$	$\widehat{\beta}_{2n}$	$\widehat{\sigma}^2_{bn}$	$\widehat{\beta}_{0n}$	$\widehat{\beta}_{1n}$	$\widehat{\beta}_{2n}$	$\widehat{\sigma}^2_{bn}$
50	-7.78	2.15	0.98	30.46	-7.65	1.96	0.99	24.61	-7.39	1.73	0.97	31.45
100	-7.86	2.07	0.97	31.54	-7.76	2.08	0.99	25.21	-7.78	2.41	0.97	32.73
200	-7.83	2.05	0.98	31.96	-7.74	2.03	0.99	25.67	-7.65	2.17	0.97	32.73
400	-7.85	2.09	0.97	31.85	-7.78	2.09	0.99	26.20	-7.85	2.58	0.96	33.06
800	-7.84	2.07	0.97	32.11	-7.73	2.02	0.99	25.94	-7.81	2.42	0.97	32.99
1600	-7.81	2.05	0.97	32.04	-7.69	1.99	0.99	25.78	-7.71	2.27	0.96	33.01

Indeed, even though the bias decreases when the number of observations per subject increases, the rate of convergence to the true values is smaller when the variance of the random effects is large. Similarly, the estimators of parameters which are included in the random-effects structure seem to have a slower convergence rate to the true value. So here, like before, large variances are associated with poorer performance of our estimators, and parameters included in the random-effects structure are more difficult to estimate.

Overall, increasing the number of time points seems to reduce the observed bias in the estimation of the linear predictor parameters. Still, some misspecifications seem to require more repeated measurements for asymptotic robustness to hold. For instance, in the case of the asymmetric mixture, the maximum likelihood estimators tend to underestimate the treatment effect, even when the data contain 8 time points if $\sigma^2_{0b} = 16$ or 32. Actually, increasing the number of time points in these settings seems to increase the magnitude of the bias. However, as can be seen from the first panel in Table 11.6, increasing the number of time points to 20 reduces the bias of the treatment effect to a magnitude similar to the one observed with 4 time points.

Unlike in the previous case, when the random effects were generated from a power function distribution, increasing the number of time points seems to decrease the

bias in a more consistent way (see the second panel in Table 11.6). Note that in both cases the intercept was more poorly estimated than the treatment and the time effect. Actually, here like in Chapter 7, the time effect was estimated with a negligible bias, even when only a small sample of 50 subjects was used. Nevertheless, we have also shown in that chapter that the situation immediately worsened when a random slope was also included in the model.

Remarkably, increasing the number of time points does not seem to reduce significantly the bias in the parameter estimates, when the random effects are generated from a two-point discrete distribution. For instance, even though for the other random-effects distributions displayed here, the bias related to the estimation of the time effect remained below 10%, in the case of the discrete distribution, increasing the number of repeated measurements seems to increase the bias for the time effect up to 20%. Additionally, we also observed problems when estimating β_1, especially for very heterogenous random-effects distributions. Increasing the number of time points to 20, in this case, does not considerably improve these results (see the third panel in Table 11.6).

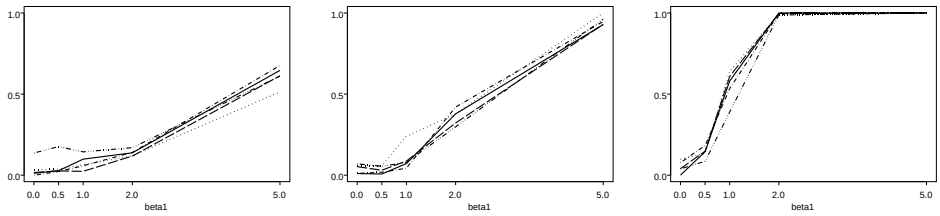
This may be an illustration of the impact of violating the conditions underlying the Bayesian central limit theorem. Indeed, it can be seen in Appendix F that this theorem is based on a Taylor series expansion of both the likelihood and the prior distribution. This requires these functions to be continuous and differentiable, and even though this holds for the assumed normal distribution, this condition is violated for the true discrete distribution. Therefore, the estimates from the assumed model may no longer converge to the true effects.

In Section 7.4 we have shown that the power and the type I error can also be affected by random-effects misspecification. We will now repeat these simulations, varying the number of repeated measurements, to study the impact on the corresponding Wald test. As in the aforementioned section, binary responses were generated using the logistic random-intercept model given by (7.1) with $\beta_0^0 = -8$, $\beta_2^0 = 1$, and five different values for the treatment effect β_1^0: 0, 0.5, 1, 2, and 5. Further, we considered again 4, 6, and 8 time points. The simulations were performed for three different sample sizes, namely 25, 100, and 400 subjects. The random intercept was drawn from a normal, a power function, a discrete, and an asymmetric mixture of two normals, each with variance $\sigma_{0b}^2 = 1$, 4, 16, and 32. For each setting, 500 data sets were generated and the model given by (7.1) was fitted to these data under the assumption of normally distributed random effects. We determined the proportion

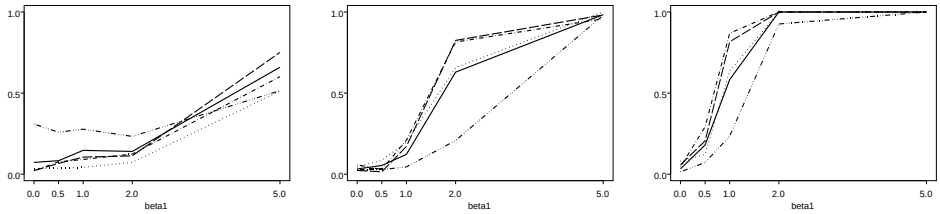
of cases in which a treatment effect different from zero (at a 5% significance level) was detected. When $\beta_1^0 = 0$, this proportion corresponds to the type I error; otherwise, it represents the power of the test. The results of these simulations are displayed in Figures 11.1 and 11.2 for 4 and 8 repeated measurements, and in Figure 7.5 for 6 repeated measurements. As can be seen from these graphs, increasing the number of repeated measurements improves the power of the analysis to detect a significant treatment effect. This is an expectable result, given that more information is now available.

Further, the corresponding type I error rates displayed in Table 11.7 again confirm the results in Theorem 8.1. Indeed, since β_1 does not have an associated random effect, its type I error seems to be maintained around 5% in most settings. On the other hand, the type I error rate related to β_0, obtained from similar simulations with parameter values $\beta_0^0 = 0$, $\beta_1^0 = 2$ and $\beta_2^0 = 1$, and displayed in Table 11.8, is seriously affected by the misspecification. Increasing the cluster size only induces a minor improvement in most of the cases considered here, while increasing the sample size further inflates the type I error rate. Indeed, in Chapters 7 and 8 we saw that, under misspecification, the maximum likelihood estimator $\widehat{\beta}_{0n}$ is consistent with respect to β_0^*, the value of β_0 which minimizes the KLIC (5.1). Likely, in the settings considered in our simulations, $\beta_0^* \neq 0$. Therefore, the Wald test, implemented in SAS, is not testing the hypothesis $H_0 : \beta_0^0 = 0$ as one would expect, but rather the hypothesis $H_0 : \beta_0^* = 0$. In this context, larger sample sizes will increase the power to detect any deviation of β_0^* from zero. This results in the observed inflation of the type I error associated with the hypothesis of interest, i.e., $H_0 : \beta_0^0 = 0$. Even though increasing the number of observations per subject would decrease the bias in the maximum likelihood estimates, it will not remove it completely. When additionally the number of subjects grows, so does the power to detect this bias, even if it is now smaller.

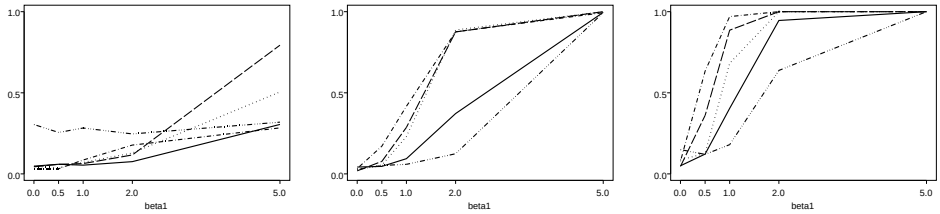
These simulations illustrate, that the previous asymptotic results should be considered in a careful way, especially in finite samples with only few number of repeated observations per subject. Note that in some scenarios a large number of time points may be needed before reasonable results can be obtained. In the following chapter we will discuss alternative strategies for the analysis when the normality assumption is questionable and the sample size at hand does not allow us to resort to asymptotic arguments.



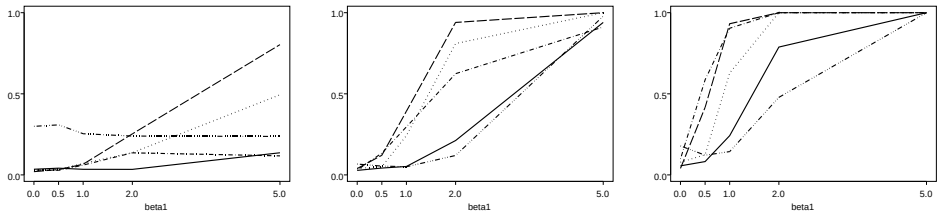
(a) $\sigma^2_{0b} = 1$, $n = 25$ (b) $\sigma^2_{0b} = 1$, $n = 100$ (c) $\sigma^2_{0b} = 1$, $n = 400$



(d) $\sigma^2_{0b} = 4$, $n = 25$ (e) $\sigma^2_{0b} = 4$, $n = 100$ (f) $\sigma^2_{0b} = 4$, $n = 400$

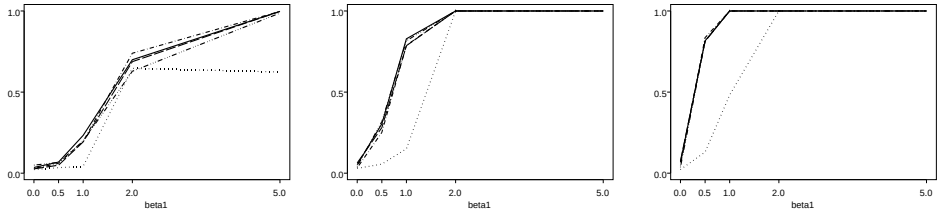


(g) $\sigma^2_{0b} = 16$, $n = 25$ (h) $\sigma^2_{0b} = 16$, $n = 100$ (i) $\sigma^2_{0b} = 16$, $n = 400$

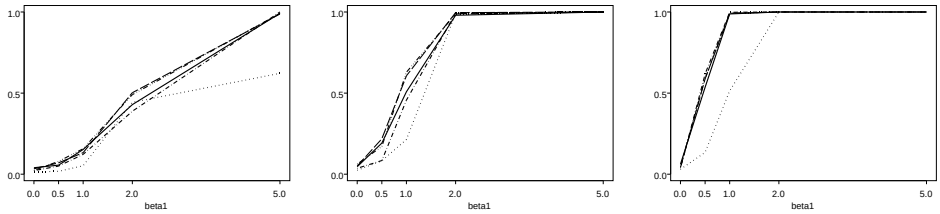


(j) $\sigma^2_{0b} = 32$, $n = 25$ (k) $\sigma^2_{0b} = 32$, $n = 100$ (l) $\sigma^2_{0b} = 32$, $n = 400$

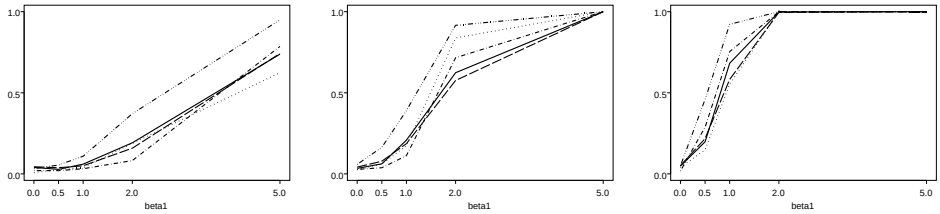
Figure 11.1: *Power of the analysis of the logistic-normal model (7.1) to detect a significant treatment effect in binary response data generated using model (7.1), over a range of possible β_1^0 values, sample sizes (n), considering 4 time points and 5 random-effects distributions with variance σ^2_{0b}: normal (solid line), power function (dotted line), discrete (dash-dotted line), lognormal (dash-triple dot) and asymmetric mixture (dashed line).*



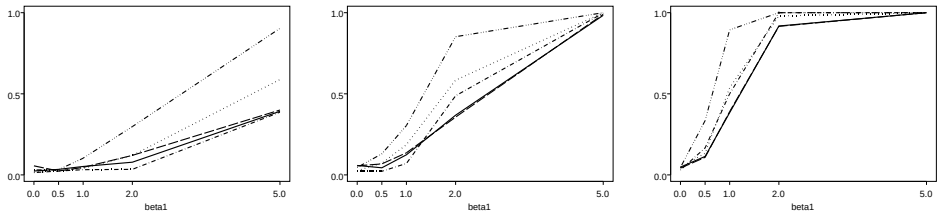
(a) $\sigma^2_{0b} = 1$, $n = 25$ (b) $\sigma^2_{0b} = 1$, $n = 100$ (c) $\sigma^2_{0b} = 1$, $n = 400$



(d) $\sigma^2_{0b} = 4$, $n = 25$ (e) $\sigma^2_{0b} = 4$, $n = 100$ (f) $\sigma^2_{0b} = 4$, $n = 400$



(g) $\sigma^2_{0b} = 16$, $n = 25$ (h) $\sigma^2_{0b} = 16$, $n = 100$ (i) $\sigma^2_{0b} = 16$, $n = 400$



(j) $\sigma^2_{0b} = 32$, $n = 25$ (k) $\sigma^2_{0b} = 32$, $n = 100$ (l) $\sigma^2_{0b} = 32$, $n = 400$

Figure 11.2: *Power of the analysis of the logistic-normal model (7.1) to detect a significant treatment effect in binary response data generated using model (7.1), over a range of possible β_1^0 values, sample sizes (n), considering 8 time points and 5 random-effects distributions with variance σ^2_{0b}: normal (solid line), power function (dotted line), discrete (dash-dotted line), lognormal (dash-triple dot) and asymmetric mixture (dashed line).*

Table 11.7: *Type I error for detecting a significant treatment effect when $\beta_1^0 = 0$, when the logistic-normal model given by (7.1) is fitted to binary response data generated using model (7.1), considering different sample sizes (n) and a random intercept sampled from a normal (No), a power function (PF), a discrete (D) or an asymmetric mixture of two normal distributions (AM), each distribution with variance σ^2_{0b}. Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.*

		4 time points				6 time points				8 time points			
	n	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
No	25	0.013	0.073	0.048	0.035	0.012	0.025	0.029	0.025	0.035	0.040	0.040	0.024
	100	0.010	0.032	0.042	0.028	0.041	0.052	0.050	0.026	0.052	0.046	0.034	0.058
	400	0.000	0.033	0.050	0.056	0.050	0.046	0.052	0.058	0.068	0.062	0.048	0.048
LN	25	0.136	**0.309**	**0.304**	**0.299**	0.008	0.020	0.017	0.017	0.050	0.023	0.032	0.030
	100	0.070	0.058	0.026	0.067	0.035	0.048	0.032	0.048	0.042	0.062	0.058	0.046
	400	0.039	0.014	**0.149**	**0.178**	0.046	0.060	0.034	0.042	0.036	0.044	0.050	0.046
PF	25	0.033	0.034	0.047	0.031	0.008	0.023	0.036	0.016	0.022	0.014	0.008	0.018
	100	0.054	0.044	0.030	0.040	0.041	0.040	0.050	0.028	0.030	0.024	0.026	0.026
	400	**0.098**	**0.080**	**0.086**	**0.078**	0.046	0.064	**0.076**	0.050	0.022	0.030	0.040	0.046
D	25	0.000	0.019	0.032	0.030	0.023	0.012	0.014	0.004	0.029	0.022	0.020	0.014
	100	0.010	0.028	0.033	0.030	0.032	0.016	**0.084**	0.018	0.033	0.036	0.028	0.024
	400	0.077	0.045	**0.078**	**0.098**	0.048	**0.080**	0.024	**0.088**	0.048	0.038	0.018	0.030
AM	25	0.016	0.023	0.044	0.020	0.014	0.014	0.018	0.038	0.027	0.033	0.044	0.056
	100	0.053	0.023	0.019	0.039	0.053	0.066	0.036	0.038	0.063	0.050	0.042	0.056
	400	0.040	0.057	0.046	0.040	0.053	0.050	0.036	0.032	**0.074**	0.042	0.052	0.042

Table 11.8: *Type I error for detecting a significant intercept when $\beta_0^0 = 0$, when the logistic-normal model given by (7.1) is fitted to binary response data generated using model (7.1), considering different sample sizes (n) and a random intercept sampled from a normal (No), a power function (PF), a discrete (D) or an asymmetric mixture of two normal distributions (AM), each distribution with variance σ^2_{0b}. Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.*

		4 time points				6 time points				8 time points			
	n	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=1$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
No	25	0.019	0.023	0.018	0.016	0.014	0.035	0.016	0.023	0.022	0.023	0.032	0.018
	100	0.042	0.046	0.050	0.052	0.042	0.048	0.040	0.034	0.028	0.036	0.046	0.032
	400	0.053	0.052	0.046	0.044	0.060	0.046	0.054	0.050	0.052	0.044	0.046	0.050
LN	25	0.014	0.051	**0.154**	**0.309**	0.014	0.042	**0.177**	**0.284**	0.023	0.051	**0.175**	**0.296**
	100	0.040	**0.215**	**0.634**	**0.847**	0.056	**0.159**	**0.667**	**0.880**	0.051	**0.206**	**0.673**	**0.882**
	400	**0.116**	**0.597**	**0.994**	**1.000**	**0.114**	**0.600**	**0.996**	**1.000**	**0.109**	**0.618**	**0.992**	**1.000**
PF	25	0.016	0.011	0.004	0.007	0.019	0.031	0.028	0.022	0.003	0.021	0.021	0.034
	100	0.068	**0.110**	**0.196**	**0.275**	0.043	**0.164**	**0.320**	**0.370**	0.039	**0.146**	**0.260**	**0.368**
	400	**0.110**	**0.510**	**0.912**	**0.934**	**0.158**	**0.682**	**0.946**	**0.962**	**0.112**	**0.674**	**0.944**	**0.972**
D	25	0.020	0.066	**0.076**	0.073	0.021	0.046	**0.087**	0.073	0.009	0.045	**0.092**	**0.107**
	100	0.054	**0.106**	**0.180**	**0.134**	0.040	0.060	**0.136**	**0.156**	0.043	**0.082**	**0.190**	**0.154**
	400	0.060	**0.278**	**0.526**	**0.448**	**0.080**	**0.252**	**0.594**	**0.604**	0.034	**0.160**	**0.418**	**0.438**
AM	25	0.013	0.034	0.027	0.021	0.015	0.025	0.011	0.045	0.010	0.039	0.020	0.036
	100	0.037	**0.352**	**0.461**	**0.946**	0.030	**0.328**	**0.408**	**0.886**	0.039	**0.306**	**0.418**	**0.739**
	400	**0.074**	**0.936**	**0.980**	**1.000**	**0.076**	**0.924**	**0.986**	**1.000**	0.072	**0.956**	**0.954**	**1.000**

Chapter 12

Alternative Approaches

In the previous chapter we have illustrated that under certain conditions, consistency of the maximum likelihood estimators of the linear predictor parameters can be achieved under random-effects misspecification, if the number of clusters as well as the number of repeated observations per cluster go to infinitive. Although this robustness is a nice asymptotic result, its use in practice may, however, be limited. For instance, when the measurement of the response of interest in a clinical trial is invasive for the patient and/or very expensive, it may be unethical and/or costly to subject the patient to 6 or more repeated measurements.

This obviously raises the question of how to proceed with only few repeated observations, when the diagnostic tools introduced in the Chapters 9 and 10 hint on the presence of misspecification. In this case, one could consider a few of the alternative approaches which have been presented in recent literature. For instance, a non-parametric approach could guard against possible misspecification. In this case, one replaces the normal random-effects distribution by a non-parametric distribution, which is estimated through a set of mass points and their corresponding probabilities (Aitkin, 1999). Although it is an appealing approach in many settings, there can be some loss of efficiency when using a non-parametric method, compared to parametric assumptions close to the true distribution (Agresti, 2004). Additionally, model comparison can be difficult as standard asymptotic theory does not apply. Finally, a non-parametric approach is definitely not appropriate when the distribution of the random effects is of primary interest like, for example, in surrogate marker evalu-

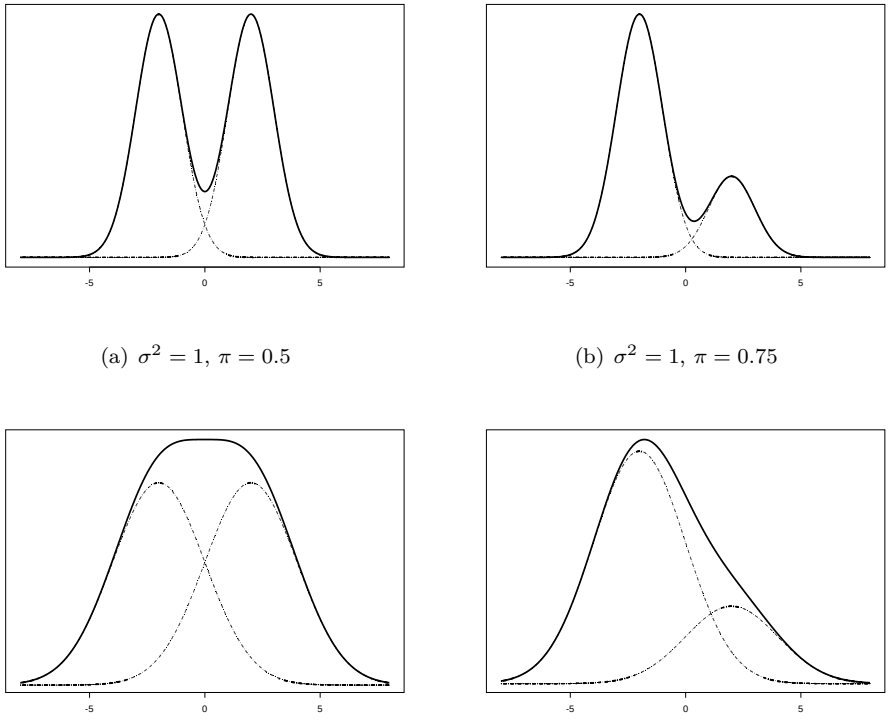
ation, the evaluation of the psychometric properties of rating scales, or when one wants to predict individual profiles or evolutions. Chen *et al.* (2002) suggested a semi-parametric random-effects distribution, allowing the random-effects density to be skewed, multi-modal, fat- or thin-tailed, and including the normal as a special case. Lee and Thompson (2007) used Markov Chain Monte Carlo methods to fit models with random effects following a t distribution, and skew extensions to the normal and to the t distribution. In the present chapter, we will study another approach which consists in replacing the normal random-effects distribution by a finite mixture of normals, the so-called heterogeneity model (Fieuws, Spiessens and Draney, 2004; Molenberghs and Verbeke, 2005). This allows one to cover a wide range of shapes for the random-effects density, including unimodal as well as multimodal, and symmetric as well as very skewed distributions.

Further, we will show that, although using more flexible families of distributions can be a valid strategy in some settings, these more general families are not fully robust either. Therefore, we will propose to incorporate them into a more general sensitivity analysis framework. In this scenario, different distributions are considered for the random effects. If the estimates of the parameters of interest and the associated inferential procedures are similar, irrespective of the distribution used to obtain them, the analyst can feel relatively confident about his/her results. On the other hand, if the results vary considerably, then they are obviously sensitive to the distributional assumptions for the random effects, and caution is needed. The results presented in this chapter are based partly on Litière *et al.* (2007b).

12.1 The Heterogeneity Model

The heterogeneity model is an extension of the generalized linear mixed model, obtained by sampling the random effects $\boldsymbol{b}_i$ from a mixture of k normal distributions with mean vectors $\boldsymbol{\mu}_r$ and covariance matrix D, i.e., $\boldsymbol{b}_i \sim \sum_{r=1}^{k} \pi_r \mathrm{N}(\boldsymbol{\mu}_r, D)$. In principle, many distributions can be approximated with high precision by finite mixtures of normal densities, making this approach theoretically very appealing. For instance, Figure 12.1 shows a few realizations of the two-component mixture $\pi \mathrm{N}(\mu_1, \sigma^2) + (1-\pi)\mathrm{N}(\mu_2, \sigma^2)$.

Observe that, with this distribution for the random-effects, the probability for a subject to belong to component r is π_r, and $\sum_{r=1}^{k} \pi_r = 1$. Note further that each component has the same covariance matrix D. This constraint is necessary to avoid



(a) $\sigma^2 = 1$, $\pi = 0.5$

(b) $\sigma^2 = 1$, $\pi = 0.75$

(c) $\sigma^2 = 4$, $\pi = 0.5$

(d) $\sigma^2 = 4$, $\pi = 0.75$

Figure 12.1: *Density functions of the mixture* $\pi N(-2, \sigma^2)+(1-\pi)N(2, \sigma^2)$*, for varying values of* π *and* σ^2*. The dashed lines represents the densities of the normal components; the solid line represents the density of the mixture.*

unbounded likelihoods (Böhning, 2000). Let us now define $\boldsymbol{\pi}^T = (\pi_1, ..., \pi_k)$ and let $\boldsymbol{\xi}$ be the vector containing the remaining parameters, i.e., the vector $\boldsymbol{\beta}$ of unknown parameters common to all subjects, as well as all parameters in $\boldsymbol{\mu}_r$ and D. The joint density function of $\boldsymbol{y}_i$ can then be written as $f_i(\boldsymbol{y}_i) = \sum_{r=1}^{k} \pi_r f_{ir}(\boldsymbol{y}_i|\boldsymbol{\xi})$ where

$$f_{ir}(\boldsymbol{y}_i|\boldsymbol{\xi}) = \int f_i(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i)\phi_r(\boldsymbol{b}_i)d\boldsymbol{b}_i,$$

and $\phi_r(\boldsymbol{b}_i)$ denotes the multivariate normal with mean $\boldsymbol{\mu}_r$ and covariance matrix D.

Estimation is now based on the maximization of

$$\ell(\boldsymbol{\gamma}|\boldsymbol{y}) = \sum_{i=1}^{n} \ln \left\{ \sum_{r=1}^{k} \pi_r f_{ir}(\boldsymbol{y}_i|\boldsymbol{\xi}) \right\},$$

where $\boldsymbol{\gamma}^T = (\boldsymbol{\xi}^T, \boldsymbol{\pi}^T)$, using the Expectation-Maximization (EM) algorithm described in Dempster, Laird and Rubin (1977).

Initially, the EM algorithm was developed for missing data problems. In our context, the algorithm is very useful if we treat the component membership indicator z_{ir}, defined as

$$z_{ir} = \begin{cases} 1 & \text{if } \boldsymbol{b}_i \text{ is sampled from the } r\text{th component in the mixture} \\ 0 & \text{otherwise,} \end{cases} \tag{12.1}$$

as missing observations. Using these indicators, the log-likelihood function can be rewritten as

$$\ell(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{r=1}^{k} z_{ir}[\ln \pi_r + \ln f_{ir}(\boldsymbol{y}_i|\boldsymbol{\xi})], \tag{12.2}$$

where $\boldsymbol{z}$ is the vector of all unobserved z_{ir}. This function is easier to maximize, however maximizing $\ell(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{z})$ with respect to $\boldsymbol{\gamma}$ will lead to estimates of $\boldsymbol{\gamma}$ which depend on the unobserved indicators z_{ir}. To avoid this, it has been suggested to use the EM algorithm so that the expected value of (12.2) rather than $\ell(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{z})$ itself, will be maximized with respect to $\boldsymbol{\gamma}$ (with the expectation taken over all unobserved z_{ir}). More specifically, in the E step (Expectation) the conditional expectation of $\ell(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{z})$, given the observed data $\boldsymbol{y}$, is determined. In the M step (Maximization), the so-obtained expected log-likelihood function is maximized with respect to $\boldsymbol{\gamma}$, providing an updated estimate for $\boldsymbol{\gamma}$. The algorithm is repeated until the difference between two successive loglikelihood evaluations is small enough.

In practice, the heterogeneity model can easily be fitted using a SAS macro based on the SAS procedures NLMIXED and IML (described in Fieuws *et al.*, 2004). This macro allows the fitting of nonlinear and generalized linear mixed models with finite normal mixtures as random-effects distributions.

To explore the actual performance of this model, we applied it to the binary response data, generated for the simulation study described in Section 7.1. Recall that these data were generated using the logistic random-intercept model given by (7.1). We considered only those data for which the random intercept was drawn from a mean zero normal density, a uniform distribution, a lognormal distribution, a power

function distribution and an asymmetric mixture of two normal densities, each with variance $\sigma^2_{0b} = 4$, 16, and 32. Further, we limited this study to 50, 100, and 200 subjects, and to 100 repetitions per setting. Finally, model (7.1) was fitted to the generated data, assuming that the random intercept followed a mixture of two normal distributions. The median maximum likelihood estimates obtained from this model are shown in Table 12.1.

When comparing these results to the estimates displayed in Tables 7.1 to 7.4, obtained from fitting a logistic-normal model, we can see that the difference between the performance of the two models is negligible for the linear predictor parameters, when the variance of the random effect is small. However, as the variance increases, the heterogeneity model seems to perform better in the estimation of the intercept and the treatment effect, especially when the sample size is small. Additionally, like in the generalized linear mixed model, from Table 12.1 we can observe that the heterogeneity model seems to be robust to the random-effects misspecification when estimating the time effect. The relative bias remained under 5% in all scenarios considered, even for $\sigma^2_{0b} = 32$.

However, we still observed substantial bias when estimating the variance of the random effects, especially for the lognormal and the power function distribution. Expectedly, using the heterogeneity model considerably improved the bias in the case of the asymmetric mixture of normals. Clearly, the heterogeneity model is a correctly specified model in this case. Nevertheless, the overall variance of the random intercept was still considerably underestimated.

It should be noted that, when the random intercept was drawn from a normal distribution, the heterogeneity model loses efficiency compared to the classical generalized linear mixed model. However, in a practical model building exercise, the comparison of the one- and two-component mixture would most likely lead to the simpler and more efficient generalized linear mixed model.

Further, recall that the simulations in Section 7.4 showed that the type I error and the power of the analysis with a generalized linear mixed model can be seriously affected by random-effects misspecification. To study whether the heterogeneity model offers a solution to this problem, additional simulations were carried out for different values of the treatment effect β_1. These simulations were performed for 3 different sample sizes (namely, 50, 100 and 200 subjects) and a total of 5 different β_1^0 values (including 0, 0.5, 1, 2 and 5).

Table 12.1: *Median of the maximum likelihood estimates $\widehat{\beta}_{0n}$, $\widehat{\beta}_{1n}$, $\widehat{\beta}_{2n}$ and $\widehat{\sigma}^2_{bn}$ obtained from fitting model (7.1) with the random intercept assumed to follow a mixture of two normal distributions, to the binary response data generated using different random-effects distributions with variance σ^2_{0b} and considering different sample sizes (n). (Note that $\beta^0_0 = -8$, $\beta^0_1 = 2$ and $\beta^0_2 = 1$.)*

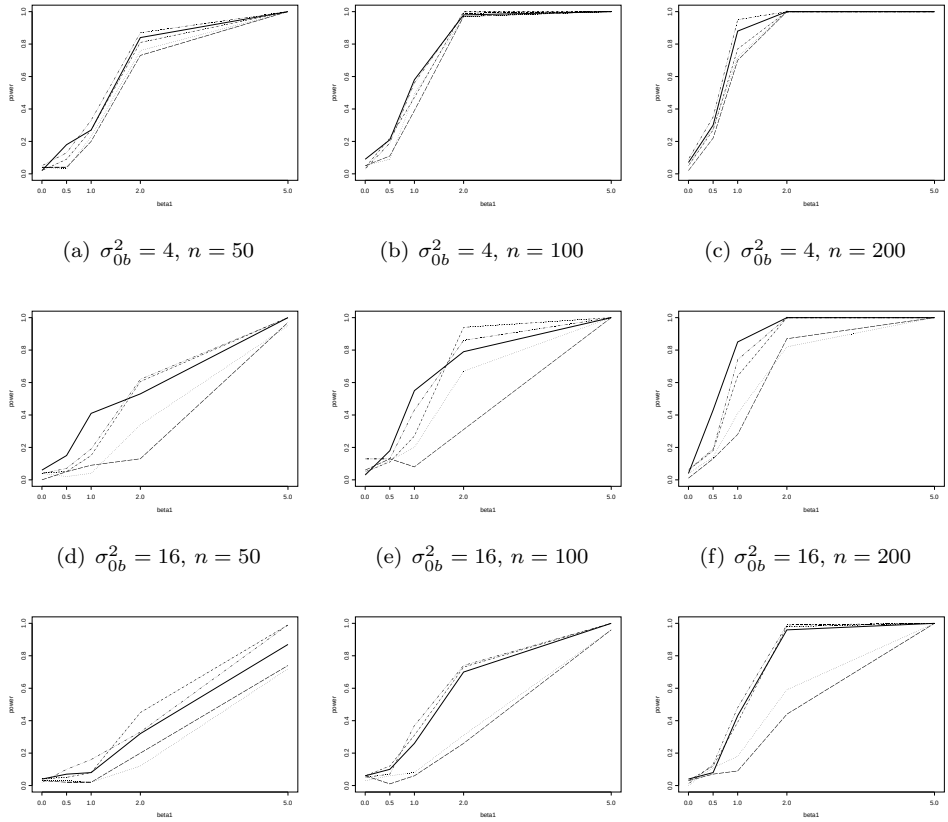
			$\widehat{\beta}_{0n}$			$\widehat{\beta}_{1n}$			$\widehat{\beta}_{2n}$			$\widehat{\sigma}^2_{bn}$		
	n	σ^2_{0b}:	4	16	32	4	16	32	4	16	32	4	16	32
Normal	50		-7.86	-8.37	-8.34	2.02	2.06	1.56	1.02	1.03	1.01	3.92	14.93	24.50
	100		-8.00	-7.86	-8.14	1.97	1.94	1.72	0.97	0.99	1.03	3.80	12.96	25.27
	200		-7.87	-8.08	-7.87	2.02	1.86	1.95	1.00	1.01	1.00	3.68	14.32	27.17
Uniform	50		-8.48	-8.13	-9.09	2.20	1.99	2.64	1.06	1.04	0.99	4.98	15.65	38.26
	100		-8.25	-8.05	-8.34	2.21	2.16	2.36	1.01	0.99	1.00	4.02	15.12	33.92
	200		-8.20	-7.88	-8.26	2.15	1.99	2.38	1.02	0.99	0.97	4.50	14.53	35.53
Lognormal	50		-8.51	-8.65	-8.86	2.19	2.29	2.03	1.07	1.05	1.06	3.53	7.15	9.87
	100		-8.41	-8.47	-8.49	2.15	2.16	2.11	1.00	1.00	1.00	3.50	6.89	9.21
	200		-8.31	-8.27	-8.73	2.13	1.98	2.13	1.02	0.99	1.02	3.91	6.70	9.45
Power	50		-8.43	-7.81	-7.64	2.21	2.33	2.55	1.07	1.02	1.03	3.33	8.49	13.91
function	100		-8.13	-7.53	-7.04	2.13	2.08	2.17	1.01	0.98	0.98	3.16	6.92	12.07
	200		-7.97	-7.43	-7.23	1.99	2.00	2.17	1.02	1.00	1.00	2.79	7.46	13.60
Asymmetric	50		-7.63	-7.92	-7.33	2.16	2.09	1.92	1.01	1.04	1.01	3.50	11.51	14.96
mixture	100		-7.99	-8.01	-7.67	2.15	2.05	2.05	1.04	1.01	1.02	3.04	12.36	21.08
	200		-7.50	-7.81	-7.43	1.99	1.94	2.05	1.00	0.99	1.01	2.64	11.92	21.24

For each setting, 100 data sets were generated, and the model given by (7.1) was fitted to these data, assuming that the random intercept follows a normal and a mixture of two normal distributions. The proportion of the cases in which these models detected a treatment effect different from zero (on a 5% significance level) was recorded. Like before, when there is no treatment effect, this proportion corresponds to the type I error; for the other values of β_1^0, this proportion represents the power of the analysis. The results of these simulations are summarized in Figures 12.2 and 12.3.

Again we observe very little difference between the performance of both models, when σ_{0b}^2 is small. However, as the variance increases, we can clearly see an increased power of the heterogeneity model to detect a treatment effect. For example, let us consider in Figures 12.2(e) and 12.3(e) the graphs corresponding to a sample of 100 patients, when $\beta_1^0 = 2$ and the random intercept was drawn from a uniform distribution. By increasing the number of components k in the assumed random-effects distribution, the power to detect a significant treatment effect increases from 31% for $k = 1$ to 73% for $k = 2$. Similarly, when the random effects are generated from an asymmetric mixture, the use of two components increases the power from 79% to 97%.

Additionally, the graphs also support the results implied by Theorem 8.1, i.e., that the type I error rate associated with the test for the presence of a covariate effect will not be affected by the (possibly incorrect) choice of the random-effects distribution, as far as this covariate is not included in the random-effects structure. Indeed, the type I error rates corresponding to the treatment effect (presented in Table 12.2) rarely exceeded the specified 5% level of significance in all the scenarios displayed in Figures 12.2 and 12.3. To study whether the type I error associated with β_0 improves using the heterogeneity model, we repeated the simulations described in Section 7.4, with $\beta_0^0 = 0$ (and $\beta_1^0 = 2$, $\beta_2^0 = 1$), and with random effects generated from a mean zero normal density, a uniform distribution, a lognormal distribution, a power function distribution and an asymmetric mixture of two normal densities. We considered again 50, 100 and 200 subjects, and for each of these settings 100 data sets were generated. Model (7.1) was then used to analyze these data, assuming that the random intercept followed a normal and a mixture of two normal distributions. The corresponding type I error rates associated with the intercept parameter β_0 are shown in Table 12.3.

The results clearly illustrate that here, although the type I error is severely affected by the random-effects misspecification when a normal random intercept is assumed,



(a) $\sigma^2_{0b} = 4$, $n = 50$ (b) $\sigma^2_{0b} = 4$, $n = 100$ (c) $\sigma^2_{0b} = 4$, $n = 200$

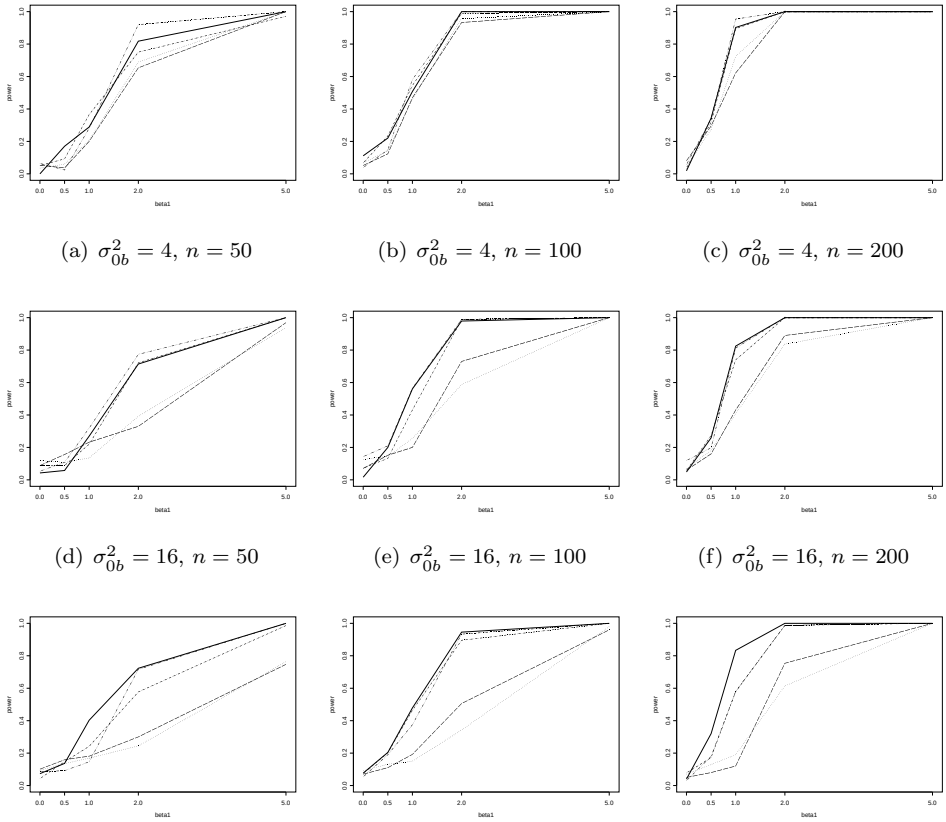
(d) $\sigma^2_{0b} = 16$, $n = 50$ (e) $\sigma^2_{0b} = 16$, $n = 100$ (f) $\sigma^2_{0b} = 16$, $n = 200$

(g) $\sigma^2_{0b} = 32$, $n = 50$ (h) $\sigma^2_{0b} = 32$, $n = 100$ (i) $\sigma^2_{0b} = 32$, $n = 200$

Figure 12.2: *Power of the logistic-normal model given by (7.1) under the assumption of normal random effects, to detect a significant treatment effect in binary response data generated using model (7.1), over a range of possible* β_1^0 *values, sample sizes* (n)*, and for 5 random-effects distributions with variance* σ^2_{0b}*: asymmetric mixture (solid line), normal (dotted line), lognormal (dash-dotted line), uniform (dashed line), and power function (dash-triple dotted line).*

it now remains below the specified significance level when the data are analyzed using the two-component heterogeneity model.

Therefore, although we cannot provide a clear indication that the model is fully



(a) $\sigma^2_{0b} = 4$, $n = 50$ (b) $\sigma^2_{0b} = 4$, $n = 100$ (c) $\sigma^2_{0b} = 4$, $n = 200$

(d) $\sigma^2_{0b} = 16$, $n = 50$ (e) $\sigma^2_{0b} = 16$, $n = 100$ (f) $\sigma^2_{0b} = 16$, $n = 200$

(g) $\sigma^2_{0b} = 32$, $n = 50$ (h) $\sigma^2_{0b} = 32$, $n = 100$ (i) $\sigma^2_{0b} = 32$, $n = 200$

Figure 12.3: *Power of the heterogeneity model given by (7.1), when the random effects are assumed to be drawn from a mixture of two normals, to detect a significant treatment effect in binary response data generated using model (7.1), over a range of possible* β_1^0 *values, sample sizes* (n)*, and for 5 random-effects distributions with variance* σ^2_{0b}*: asymmetric mixture (solid line), normal (dotted line), lognormal (dash-dotted line), uniform (dashed line), and power function (dash-triple dotted line).*

robust against random-effects misspecification, we have seen from the simulations that the heterogeneity model tends to perform slightly or considerably better than the classical generalized linear mixed model, especially for small sample sizes. It is difficult

Table 12.2: *Type I error of the heterogeneity model and the logistic-normal model (GLMM) for detecting a significant treatment effect when $\beta_1^0 = 0$ in binary response data generated using model (7.1), considering different samples sizes (n) and 5 distributions with variance σ^2_{0b} for the random intercept. Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.*

		Heterogeneity model			GLMM		
Distribution	n	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$	$\sigma^2_{0b}=4$	$\sigma^2_{0b}=16$	$\sigma^2_{0b}=32$
Normal	50	0.035	**0.120**	0.094	0.040	0.040	0.030
	100	0.068	**0.126**	0.091	0.050	0.040	0.030
	200	0.043	0.070	0.070	0.050	0.040	0.000
Uniform	50	0.000	0.043	0.073	0.020	0.060	0.040
	100	0.111	0.016	0.077	0.090	0.030	0.060
	200	0.019	0.048	0.040	0.070	0.040	0.040
Lognormal	50	0.051	0.090	0.042	0.020	0.042	0.041
	100	0.067	0.072	0.074	0.040	0.050	0.060
	200	0.033	0.057	0.078	0.050	0.060	0.030
Power	50	0.053	0.088	0.101	0.040	0.000	0.030
function	100	0.052	0.071	0.070	0.050	0.060	0.060
	200	0.081	0.060	0.050	0.020	0.010	0.030
Asymmetric	50	0.065	0.054	0.084	0.050	0.040	0.020
mixture	100	0.038	**0.143**	0.053	0.030	**0.130**	0.050
	200	0.057	0.119	0.030	0.090	0.060	0.010

to assess the full power of the heterogeneity model using simulations, mainly due to computational and time constraints. For instance, we have only considered models with two components having different means but equal variances. One could wonder if considering two or more components with equal means and different variances for the random effects could significantly improve the performance in some cases. This, together with the promising results obtained from the inferential procedures, leads us to believe that the heterogeneity model is a tool worthy of consideration.

Table 12.3: *Type I error of the heterogeneity model and the logistic-normal model (GLMM) for detecting a significant intercept when $\beta_0^0 = 0$, in binary response data generated using model (7.1), considering different samples sizes (n) and 5 distributions with variance σ_{0b}^2 for the random intercept. Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.*

		Heterogeneity model			GLMM		
Distribution	n	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$	$\sigma_{0b}^2 = 4$	$\sigma_{0b}^2 = 16$	$\sigma_{0b}^2 = 32$
Normal	50	0.000	0.021	0.095	0.040	0.000	0.070
	100	0.000	0.020	0.030	0.080	0.040	0.040
	200	0.000	0.000	0.010	0.060	0.030	0.040
Uniform	50	0.000	0.011	0.079	0.000	0.030	0.020
	100	0.000	0.010	0.020	0.060	0.030	0.080
	200	0.000	0.030	0.040	0.050	0.040	0.060
Lognormal	50	0.000	0.000	0.019	**0.160**	**0.440**	**0.620**
	100	0.000	0.000	0.028	**0.180**	**0.540**	**0.880**
	200	0.000	0.000	0.039	**0.360**	**0.930**	**1.000**
Power	50	0.000	0.000	0.084	0.050	**0.160**	**0.120**
function	100	0.033	0.010	0.010	0.060	**0.250**	**0.320**
	200	0.010	0.000	0.000	**0.290**	**0.580**	**0.760**
Asymmetric	50	0.000	0.000	0.000	**0.160**	**0.113**	**0.429**
mixture	100	0.026	0.010	0.025	**0.370**	**0.370**	**0.910**
	200	0.000	0.040	0.011	**0.660**	**0.790**	**1.000**

12.2 The Schizophrenia Data: A Sensitivity Analysis

The simulation studies discussed in this and previous chapters, seem to lead to two relevant remarks: i) misspecification of the random effects distribution can induce a severe bias in the estimation of the variance components as well as a large bias in the estimation of the linear predictor parameters, ii) more robust alternatives like the heterogeneity model can represent a significant improvement in some scenarios but can still suffer from severe bias in others. In this section, we propose to incorporate these alternatives within a sensitivity analysis framework, considering different distributions for the random effects. This approach will be illustrated with the schizophrenia data.

Table 12.4: *Schizophrenia data. Parameter estimates and standard errors using a logistic random-intercept model with the random effect (RE) assumed to follow a normal distribution (GLMM), a chi-square (χ^2), an exponential, a uniform, a lognormal, a mixture of 2 or 3 normals, and finally a non-parametric (NP) distribution with 2 or 3 support points.*

Model	$\widehat{\beta}_{0n}$ (s.e.)	$\widehat{\beta}_{1n}$ (s.e.)	$\widehat{\beta}_{2n}$ (s.e.)	$\widehat{\sigma}^2_{bn}$ (s.e.)	AIC
GLMM	-7.37 (1.18)	2.14 (1.08)	0.65 (0.096)	21.01 (6.81)	391.9
RE, χ^2	-6.99 (1.18)	1.92 (0.88)	0.66 (0.096)	18.20 (6.07)	392.5
RE, exp.	-6.35 (1.00)	1.70 (0.88)	0.64 (0.098)	10.71 (2.76)	397.3
RE, uni.	-5.47 (0.75)	1.38 (0.64)	0.56 (0.082)	9.54 (1.93)	408.3
RE, logn.	-5.17 (0.78)	1.20 (0.71)	0.57 (0.086)	19.34 (7.52)	409.6
Mixture, $k = 2$	-7.88 (1.23)	1.99 (0.94)	0.67 (0.096)	28.03 (8.66)	395.1
Mixture, $k = 3$	-7.77 (4.28)	2.70 (0.85)	0.68 (0.094)	20.20 (31.6)	396.0
NP, $k = 2$	-4.84 (0.59)	1.28 (0.33)	0.52 (0.085)	-	-
NP, $k = 3$	-5.67 (0.72)	2.06 (0.54)	0.59 (0.097)	-	-

First, we will analyze these data using a heterogeneity model with a mixture of two and three normal distributions, and a non-parametric model with two and three mass points. As discussed in the previous sections, these mixture models can easily be fitted using the SAS macro for fitting nonlinear and generalized linear mixed models with finite normal mixtures as random-effects distributions. Additionally, we will also consider some non-normal distributions for the random intercept, including a chi-square, an exponential, a uniform and a lognormal density. Recent research has shown that such analysis can easily be carried out with standard statistical software packages like the SAS procedure NLMIXED using probability integral transformations (Nelson *et al*, 2006). The parameter estimates obtained from fitting model 7.1, with these random-effects distributions are displayed in Table 12.4.

Observe that the point estimates are all similar, especially if we exclude the ones coming from the models that used the uniform and lognormal distributions (and incidently produced the highest AIC values). Moreover, the inferential results were similar in all the cases as well. For instance, the treatment effect was significant in all

the models except for the ones with the lognormal and exponential random effects, which produced borderline p-values of 0.054 and 0.094, respectively. The variance of the random effect was rather large in all scenarios considered, with a median value around 19. Therefore, all the models consistently hint on a very strong within-subject association. One could now apply some known model selection criteria to select the distribution that fits the data best. For instance, using the AIC displayed in the last column of Table 12.4, the logistic-normal model emerged as the most appropriate one.

Further, the task of choosing the number of components k in the heterogeneity model is not an easy one. One approach consists in fitting models with increasing numbers of components, and comparing them using likelihood ratio tests. For instance, we can consider testing $H_0 : k = 1$ versus $H_A : k = 2$. However, this null hypothesis can also be expressed as $H_0 : \pi_2 = 0$, which is clearly on the boundary of the parameter space. In this case, bootstrap simulations are required to derive the distribution of the likelihood ratio test statistic under the null (Böhning, 2000). An alternative ad-hoc approach consists in increasing the number of components k until some of the subpopulations reflect very small weights π_r, or until some subpopulations coincide (Verbeke and Molenberghs, 2000). Here we will use the AIC to select the best fitting model. This results, again, in a preference for the one-component logistic-normal model.

Finally, as stated before, when the random-effects distribution is not primarily of interest, one can resort to a nonparametric approach. In this case the random-effects distribution can be approximated by a discrete distribution with a finite number of support points. The results of this analysis are shown in the final part of Table 12.4. The estimates stabilized with 3 support points, and with this model we obtained values very close to the ones from the logistic-normal model. For instance, the estimate of the treatment effect was close to the one obtained with the logistic-normal model and it was significantly different from zero. After carrying out this analysis, we can therefore be rather confident about the results obtained from the logistic-normal model, and about the presence of a moderate treatment effect on the CGI scores of the schizophrenic patients.

It should be pointed out that our sensitivity analysis is not a robust alternative to the classical generalized linear mixed model, but rather an alternative to the lack of robustness of the generalized linear mixed model. Indeed, given that we consider different choices for the random-effects distribution, we expect the outcome to be optimal when the true random-effects distribution is very similar to one of them. The

idea is then to see how sensitive are our conclusions with respect to the distributional assumptions for the random effects. Similar results obtained under different assumptions will increase our confidence, different ones will increase our caution.

It is becoming clear that there probably will not be a general, easy answer on how to deal with this type of model misspecification. If marginal summaries are adequate for the analysis, then one could focus more on fitting marginal models. For instance, tools such as Generalized Estimating Equations (GEE) are known to be robust against misspecification of the association structure. Further, Heagerty and Kurland (2001) have observed from their simulations that when a marginal regression structure is specified, rather than a conditional mean structure, the corresponding maximum likelihood estimates are much less susceptible to bias resulting from a misspecified random-effects distribution. On the other hand, when the subject-specific effects are of main interest, perhaps in some specific situations, good alternative models can be found by using, for example, random-effects distributions conjugate to the distribution of the outcome, as suggested by Lee and Nelder (1996) and Lee, Nelder and Pawitan (2006). One such special case given by the Poisson-gamma model will be the topic of our next chapter. However, we believe that in general, the proposed diagnostic tools in Chapters 9 and 10, together with the ability to consider several random-effects distributions, would allow for a useful and, arguably, necessary sensitivity analysis.

Chapter 13

Conjugacy

In the previous chapter, we proposed a sensitivity analysis as a plausible alternative to the lack of robustness of the generalized linear mixed model to misspecification of the random-effects distribution. Ideally, one would prefer to find a family of models that exhibits certain degree of robustness to this misspecification, like we observed in the linear mixed model. It is not clear at the moment if such a family exists, but in the present chapter we speculate that the use of a conjugate distribution for the random effects could increase robustness. Actually, in linear mixed models the normal density used for the random effect is the conjugate distribution of the normal likelihood one encounters in that setting. In this chapter, we will study another example where conjugacy and robustness come together: the Poisson-gamma model.

13.1 The Poisson-Gamma Model

As before, let y_{ij} denote the jth measurement for the ith subject, with $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Conditional on a random intercept b_i for subject i, it is assumed that all responses y_{ij} are independent with Poisson density

$$f(y_{ij}|b_i) = \frac{1}{y_{ij}!}\mu_{ij}^{y_{ij}} \exp(-\mu_{ij}), \tag{13.1}$$

where μ_{ij} is modeled as

$$\mu_{ij} = b_i \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}). \tag{13.2}$$

In this expression, $\boldsymbol{x}_{ij}$ denotes a p-dimensional vector of covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of fixed parameters, and the random intercept b_i is assumed to follow a gamma distribution specified by

$$f(b_i|\lambda_b) = \left(\frac{1}{\lambda_b}\right)^{1/\lambda_b} \frac{1}{\Gamma(1/\lambda_b)} b_i^{(1/\lambda_b)-1} \exp(-b_i/\lambda_b). \tag{13.3}$$

Note that μ_{ij} could also be written as $\mu_{ij} = \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + b_i)$. In this case, the random intercept b_i would be assumed to follow a log-gamma distribution.

Given that the gamma distribution is conjugate to the conditional Poisson distribution of the outcomes, we can easily derive a closed-form expression for the marginal model implied by (13.1)–(13.3). Indeed, recall that the marginal distribution of the vector of outcomes $\boldsymbol{y}_i$ for subject i follows from

$$f_i(\boldsymbol{y}_i) = \int f_i(\boldsymbol{y}_i|b_i) f(b_i) db_i. \tag{13.4}$$

In this case,

$$\begin{aligned} f_i(\boldsymbol{y}_i|b_i) &= \prod_{j=1}^{n_i} f(y_{ij}|b_i) \\ &= \left(\prod_{j=1}^{n_i} \frac{1}{y_{ij}!} \mu_{ij}^{y_{ij}}\right) \exp\left(-\sum_{j=1}^{n_i} \mu_{ij}\right) \\ &= \left\{\prod_{j=1}^{n_i} \frac{1}{y_{ij}!} [b_i \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})]^{y_{ij}}\right\} \exp\left[-b_i \sum_{j=1}^{n_i} \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\right]. \end{aligned}$$

Using this expression, it follows that

$$\begin{aligned} f_i(\boldsymbol{y}_i) &= \int \left\{\prod_{j=1}^{n_i} \frac{1}{y_{ij}!} [b_i \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})]^{y_{ij}}\right\} \exp\left[-b_i \sum_{j=1}^{n_i} \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\right] \\ &\qquad \times \left(\frac{1}{\lambda_b}\right)^{1/\lambda_b} \frac{1}{\Gamma(1/\lambda_b)} b_i^{(1/\lambda_b)-1} \exp(-b_i/\lambda_b) db_i \\ &= \left\{\prod_{j=1}^{n_i} \frac{1}{y_{ij}!} [\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})]^{y_{ij}}\right\} \left(\frac{1}{\lambda_b}\right)^{1/\lambda_b} \frac{1}{\Gamma(1/\lambda_b)} \\ &\qquad \times \int b_i^{\sum_j y_{ij} + (1/\lambda_b) - 1} \exp\{-b_i [\sum_j \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}) + (1/\lambda_b)]\} db_i \end{aligned}$$

For simplicity of notation, let $\mu_{i\cdot} = \sum_j \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})$ and $y_{i\cdot} = \sum_j y_{ij}$, and let $\alpha =$

$y_{i.} + (1/\lambda_b)$ and $\gamma = [\mu_{i.} + (1/\lambda_b)]^{-1}$ then

$$\begin{aligned} f_i(\boldsymbol{y}_i) &= \left\{\prod_{j=1}^{n_i} \frac{1}{y_{ij}!}[\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})]^{y_{ij}}\right\}\left(\frac{1}{\lambda_b}\right)^{1/\lambda_b}\frac{1}{\Gamma(1/\lambda_b)} \\ &\quad \times \frac{\Gamma[y_{i.} + (1/\lambda_b)]}{[\mu_{i.} + (1/\lambda_b)]^{y_{i.}+(1/\lambda_b)}}\int \frac{b_i^{\alpha-1}\exp(-b_i/\gamma)}{\gamma^\alpha\Gamma(\alpha)}db_i. \end{aligned}$$

Observe that the integral in the last part of this equation corresponds to a gamma probability density function with parameters α and γ. As a result,

$$f_i(\boldsymbol{y}_i) = \left(\prod_{j=1}^{n_i} \frac{1}{y_{ij}!}[\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})]^{y_{ij}}\right)\left(\frac{1}{\lambda_b}\right)^{1/\lambda_b}\frac{1}{\Gamma(1/\lambda_b)}\frac{\Gamma[y_{i.} + (1/\lambda_b)]}{[\mu_{i+} + (1/\lambda_b)]^{y_{i.}+(1/\lambda_b)}}. \quad (13.5)$$

Finally, since $y! = \Gamma(y+1)$, it follows that the marginal density $f_i(\boldsymbol{y}_i)$, induced by (13.1)-(13.3), can be written as

$$f_i(\boldsymbol{y}_i) = \left\{\prod_{j=1}^{n_i} \frac{1}{\Gamma(y_{ij}+1)}[\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})]^{y_{ij}}\right\}\left(\frac{1}{\lambda_b}\right)^{1/\lambda_b}\frac{1}{\Gamma(1/\lambda_b)}\frac{\Gamma[y_{i.} + (1/\lambda_b)]}{[\mu_{i.} + (1/\lambda_b)]^{y_{i.}+(1/\lambda_b)}}. \quad (13.6)$$

Using this closed-form expression for the marginal likelihood, we can prove the following theorem.

Theorem 13.1 (Consistency Poisson-Gamma Model) *Consider the model given by (13.1)-(13.3). Let h be the true density of b_i, such that $h(b_i) \neq f(b_i, \lambda_b)$ and $E_h(b_i) = 1$. Further, let G be the marginal distribution of $\boldsymbol{Y}$ induced by h. If $\boldsymbol{\beta}_g$ denotes the real values of the parameter $\boldsymbol{\beta}$, and $\widehat{\boldsymbol{\beta}}_n$ the maximum likelihood estimator of $\boldsymbol{\beta}$ associated with model (13.1)–(13.3), then, under Assumptions 5.1–5.3, $\widehat{\boldsymbol{\beta}}_n$ satisfies*

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}_g. \quad (13.7)$$

Proof

First, note that, under the *true* model, $\boldsymbol{y}_i$, conditional on $\boldsymbol{b}_i$, follows a Poisson distri-

bution with mean $\mathrm{E}(\boldsymbol{y}_i|b_i) = b_i \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g)$. Moreover, marginally we have that

$$\begin{aligned}
\mathrm{E}_g(\boldsymbol{y}_i) &= \int \boldsymbol{y}_i g(\boldsymbol{y}_i) d\boldsymbol{y}_i \\
&= \int \boldsymbol{y}_i \left[\int f_i(\boldsymbol{y}_i|b_i) h(b_i) db_i \right] d\boldsymbol{y}_i \\
&= \int \left[\int \boldsymbol{y}_i f_i(\boldsymbol{y}_i|b_i) d\boldsymbol{y}_i \right] h(b_i) db_i \\
&= \int \mathrm{E}(\boldsymbol{y}_i|b_i) h(b_i) db_i \\
&= \int b_i \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g) h(b_i) db_i \\
&= \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g) \int b_i h(b_i) db_i \\
&= \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g) \mathrm{E}_h(b_i) \\
&= \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g).
\end{aligned}$$

The previous expression illustrates that, like in linear mixed models, the parameters in the mean structure of the Poisson-gamma model have both a marginal and a hierarchical interpretation. Now, let $\boldsymbol{\xi} = (\boldsymbol{\beta}, \lambda_b)$ denote the vector of model parameters. From the maximum likelihood theory in misspecified models described in Chapter 5, it is known that the maximum likelihood estimator $\widehat{\boldsymbol{\xi}}_n \xrightarrow{P} \boldsymbol{\xi}^*$, which minimizes the KLIC

$$I(g : f; \boldsymbol{\xi}) = \mathrm{E}_g \left[\log \frac{g(\boldsymbol{y}_i)}{f(\boldsymbol{y}_i, \boldsymbol{\xi})} \right]. \tag{13.8}$$

The Assumptions 5.1–5.3 ensure that the KLIC is well-defined and that its minimum is unique. Observe that Expression (13.8) can also be written as

$$I(g : f; \theta) = \int g(\boldsymbol{y}_i) \log g(\boldsymbol{y}_i) d\boldsymbol{y}_i - \int g(\boldsymbol{y}_i) \log f(\boldsymbol{y}_i, \boldsymbol{\xi}) d\boldsymbol{y}_i, \tag{13.9}$$

and that minimizing I with respect to $\boldsymbol{\xi}$ is equivalent to maximizing

$$I_1 = \int g(\boldsymbol{y}_i) \log f(\boldsymbol{y}_i, \boldsymbol{\xi}) d\boldsymbol{y}_i. \tag{13.10}$$

After substituting the marginal density (13.6) of $\boldsymbol{y}_i$ in this expression, it follows that

$$\begin{aligned}
I_1 &= \int g(\boldsymbol{y}_i)\{\sum_j[(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})y_{ij} - \log\Gamma(y_{ij}+1)] - \frac{1}{\lambda_b}\log\lambda_b - \log\Gamma(1/\lambda_b) \\
&\quad + \log\Gamma[y_{i.} + (1/\lambda_b)] - [y_{i.} + (1/\lambda_b)]\log[\mu_{i.} + (1/\lambda_b)]\}\, d\boldsymbol{y}_i \\
&= \int g(\boldsymbol{y}_i)\sum_j[(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})y_{ij}]d\boldsymbol{y}_i - \int g(\boldsymbol{y}_i)\sum_j[\log\Gamma(y_{ij}+1)]d\boldsymbol{y}_i - \frac{1}{\lambda_b}\log\lambda_b \\
&\quad - \log\Gamma(1/\lambda_b) + \int g(\boldsymbol{y}_i)\log\Gamma[y_{i.} + (1/\lambda_b)]d\boldsymbol{y}_i \\
&\quad - \int g(\boldsymbol{y}_i)[y_{i.} + (1/\lambda_b)]\log[\mu_{i.} + (1/\lambda_b)]d\boldsymbol{y}_i \\
&= \int g(\boldsymbol{y}_i)\sum_j[(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})y_{ij}]d\boldsymbol{y}_i - \int g(\boldsymbol{y}_i)\sum_j[\log\Gamma(y_{ij}+1)]d\boldsymbol{y}_i - \frac{1}{\lambda_b}\log\lambda_b \\
&\quad - \log\Gamma(1/\lambda_b) + \int g(\boldsymbol{y}_i)\log\Gamma[y_{i.} + (1/\lambda_b)]d\boldsymbol{y}_i \\
&\quad - \log[\mu_{i.} + (1/\lambda_b)]\int g(\boldsymbol{y}_i)y_{i.}d\boldsymbol{y}_i - \frac{1}{\lambda_b}\log[\mu_{i.} + (1/\lambda_b)] \\
&= \sum_j(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\mathrm{E}_g(y_{ij}) - \sum_j \mathrm{E}_g[\log\Gamma(y_{ij}+1)] - \frac{1}{\lambda_b}\log\lambda_b - \log\Gamma(1/\lambda_b) \\
&\quad + \mathrm{E}_g\{\log\Gamma[y_{i.} + (1/\lambda_b)]\} - \log[\mu_{i.} + (1/\lambda_b)][\sum_j \mathrm{E}_g(y_{ij}) + (1/\lambda_b)].
\end{aligned}$$

To obtain $\boldsymbol{\xi}^*$, we need to determine the derivatives of I_1 with respect to the parameters of interest, i.e.,

$$\begin{aligned}
\frac{\partial I_1}{\partial\boldsymbol{\beta}} &= \sum_j \mathrm{E}_g(y_{ij})\boldsymbol{x}_{ij} - [\sum_j \mathrm{E}_g(y_{ij}) + (1/\lambda_b)]\frac{\partial}{\partial\boldsymbol{\beta}}\log[\mu_{i.} + (1/\lambda_b)] \\
&= \sum_j \mathrm{E}_g(y_{ij})\boldsymbol{x}_{ij} - \frac{\sum_j \mathrm{E}_g(y_{ij}) + (1/\lambda_b)}{\mu_{i.} + (1/\lambda_b)}\frac{\partial}{\partial\boldsymbol{\beta}}[\mu_{i.} + (1/\lambda_b)] \\
&= \sum_j \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g)\boldsymbol{x}_{ij} - \frac{\sum_j \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g) + (1/\lambda_b)}{\sum_j \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}) + (1/\lambda_b)}\sum_j \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\boldsymbol{x}_{ij}, \quad (13.11)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial I_1}{\partial\lambda_b} &= \frac{1}{\lambda_b^2}\log\lambda_b - \frac{1}{\lambda_b^2} + \frac{1}{\lambda_b^2}\frac{\Gamma'(1/\lambda_b)}{\Gamma(1/\lambda_b)} + \frac{1}{\lambda_b^2}\log[\mu_{i.} + (1/\lambda_b)] + \frac{\sum_j \mathrm{E}_g(y_{ij}) + 1/\lambda_b}{\lambda_b^2[\mu_{i.} + (1/\lambda_b)]} \\
&\quad - \int g(\boldsymbol{y}_i)\frac{\Gamma'[y_{i.} + (1/\lambda_b)]}{\lambda_b^2\Gamma[y_{i.} + (1/\lambda_b)]}d\boldsymbol{y}_i. \quad (13.12)
\end{aligned}$$

Evaluating this expression in $\boldsymbol{\beta} = \boldsymbol{\beta}_g$ leads to $\left.\frac{\partial I_1}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_g} = 0$ and to

$$\begin{aligned}\left.\frac{\partial I_1}{\partial \lambda_b}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_g} &= \frac{1}{\lambda_b^2}\left(\log\lambda_b - 1 + \Psi(1/\lambda_b) - \mathrm{E}_g\{\Psi[y_{i.} + (1/\lambda_b)]\} + \log[\mu_{i.} + (1/\lambda_b)]\right) \\ &\quad + \frac{\sum_j \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g) + (1/\lambda_b)}{\lambda_b^2 \sum_j \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_g) + (1/\lambda_b)} \\ &= \frac{1}{\lambda_b^2}\left(\log\lambda_b + \Psi(1/\lambda_b) - \mathrm{E}_g\{\Psi[y_{i.} + (1/\lambda_b)]\} + \log[\mu_{i.} + (1/\lambda_b)]\right), \end{aligned} \tag{13.13}$$

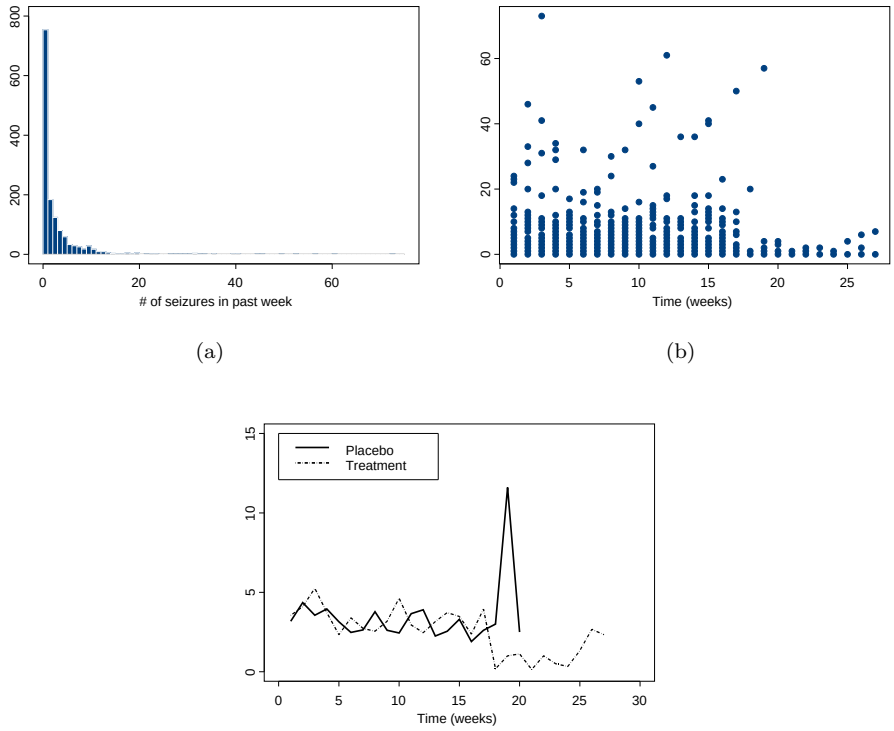
where $\Psi(.) = \frac{\Gamma'(.)}{\Gamma(.)}$ represents the digamma function.

Note that the system of equations given by (13.11) and (13.12) represents a system of $p+1$ equations in $p+1$ variables. For $\boldsymbol{\beta} = \boldsymbol{\beta}_g$ the system reduces to a single equation (13.13), in one single variable λ_b. Let λ_b^* denote the value of λ_b for wich the right hand side of Expression (13.13) becomes zero, then, $(\boldsymbol{\beta}_g, \lambda_b^*)$ is a solution to the system (13.11)-(13.12). Under our set of assumptions, it then follows that $\boldsymbol{\beta}_g = \boldsymbol{\beta}^*$, and therefore the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_n$, based on the Poisson-gamma model given by (13.1)–(13.3) is a consistent estimator for $\boldsymbol{\beta}_g$, even when the distribution of the random intercept is misspecified. □

This theorem suggests that the maximum likelihood estimators of the mean structure parameters in the Poisson-gamma model are consistent, even when the distribution of b_i is misspecified, as far as all the other aspects of the model are correctly specified. However, the same result can not be reproduced for the variance component λ_b. Therefore, caution is still needed in studies where the variance component is of main interest. Following a different approach, a similar result was obtained by Lawless (1987). In the following section, we will illustrate Theorem 13.1 via a practical case study containing repeated counts of epileptic seizures.

13.2 Application: The Epilepsy Data

The data considered in this section are coming from a randomized, double-blind, parallel group, multi-center study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AEDs (Molenberghs and Verbeke, 2005). The first 12-week period served as a stabilization period for the use



(a) (b)

(c)

Figure 13.1: *Epilepsy data. (a) Frequency plot of the number of epileptic seizures, over all visits and both treatment groups. (b) Observed number of epileptic seizures at each visit. (c) Average evolution of the number of epileptic seizures over time.*

of the AED's. After this baseline period, 45 patients were assigned to the placebo group, 44 to the active (new) treatment group. Patients were then measured on a weekly basis. The study was considered double-blind during the first 16 weeks, after which the patients were entered into a long-term open-extension study. Some were followed for up to 27 weeks.

The outcome of interest is given by the number of epileptic seizures experienced during the last week, i.e., since the last time the outcome was measured. Of interest is whether or not the additional new treatment reduces the number of epileptic seizures. Figure 13.1(a) shows a frequency plot of the number of epileptic seizures, over all

Table 13.1: *Epilepsy data. Parameter estimates, standard errors and p-values for the regression coefficients in model (13.14), where the random intercept b_i is assumed to follow a gamma distribution with parameter λ_b, as specified in (13.3).*

Parameter	Estimate (s.e.)	p-value
Fixed effects		
β_0	1.211 (0.061)	< 0.001
β_1	-0.166 (0.075)	0.029
β_2	-0.014 (0.004)	< 0.001
β_3	0.002 (0.006)	0.726
Variance components		
λ_b	1.142 (0.103)	

visits, whereas Figure 13.1(b) displays the observed number of epileptic seizures at each visit. From these graphs it is clear that the distribution of the outcomes is very skewed, with up to 73 seizures in one week time. Further, Figure 13.1(c) shows the average evolution of the number of epileptic seizures over time, by treatment group. Note that the unstable behavior can be explained by the presence of extreme values, but is also the result of the fact that very little observations are available at some of the visits, especially past week 16, i.e., the end of the actual double-blind period.

Let y_{ij} represent the number of epileptic seizures experienced by patient i during week j. Further, let $z_i = 1$ (0) denote the treatment (control) group and let t_j denote the time-point at which y_{ij} was measured, $t_j = 1, 2, \ldots$ until at most 27. We will use the following model to analyze the data

$$\begin{aligned} Y_{ij}|b_i &\sim \text{Poisson}(\mu_{ij}), \\ \log(\mu_{ij}) &= \beta_0 + \beta_1 z_i + \beta_2 t_j + \beta_3 z_i t_j + \log(b_i) \end{aligned} \tag{13.14}$$

where the random intercept b_i is assumed to follow a gamma distribution as specified in (13.3), with parameter λ_b.

Recall from the sensitivity analysis in Chapter 12 that this model can easily be fitted via the SAS procedure NLMIXED, using probability integral transformations (Nelson *et al.*, 2006). Fitting this model required nonadaptive Gaussian quadrature with 61 quadrature points. The results are summarized in Table 13.1. Assuming that (13.14) is correctly specified, then Theorem 13.1 would guarantee the consistency of

our maximum likelihood estimates, and we could be confident about the fixed effects estimates presented in this table, even in the presence of possible random-effects misspecification.

Still, it should be noted that the use of the results presented in Theorem 13.1 is limited. Clearly, in its present format, the theorem is valid only when a random intercept is the only random effect considered. It would be of interest to study whether there exists a "multivariate" extension of the theorem, with a generalization of the univariate gamma distribution, which would allow for more than one random effect. However, different formulations of a multivariate gamma are available in the literature (for an overview of these distributions, we refer to Kotz, Balakrishnan and Johnson, 2000). Therefore, further research should focus on finding that specific generalization, which would allow the results presented in Theorem 13.1 to be extended to Poisson models with more than one random effect.

Additionally, given the current lack of a general model that can guard against the effects of random-effects misspecification, it would also be of interest to find other examples, like the Poisson-gamma model, in which conjugacy between the random-effects distribution and the conditional distribution of the outcome could preserve the consistency of the maximum likelihood estimators under this type of misspecification. In this context, the h-likelihood models proposed by Lee and Nelder (1996) seem to provide a promising approach and deserve further attention.

Chapter 14

Concluding Remarks and Further Research

14.1 Concluding Remarks

Does misspecification of the random-effects distribution affect the maximum likelihood estimators in generalized linear mixed models? Although data analysts often assume that the choice of the random-effects distribution is not crucial for the quality of their inferences, we have shown in this work that this is not a generally valid truth. First, the estimates of the variance components are always subject to considerable bias. Even by increasing both the number of subjects as well as the number of repeated observations per subject, consistency cannot be guaranteed. Clearly this bias can have severe consequences in applications in which the association structure is of main interest. It could also provoke misleading results when we want to predict subject-specific trajectories or use subject-specific parameters for classification purposes.

On the other hand, the linear predictors seem to be less affected. When the variance of the random effects is sufficiently small, the observed bias is generally negligible. However, caution is still necessary when the random effects show a lot of variability, or when complicated covariance structures are used. Note that large random-effects variances are not exceptional in clinical trials, as illustrated by our

case study. Indeed, one could expect little variability in the response, for instance, when a placebo control group is used, whereas a more variable outcome pattern is expected in the treated group. In such a scenario, the linear predictors, including the treatment effect, could be subject to considerable bias under misspecification.

Although this bias in the linear predictors can be considerably reduced in some settings by increasing the number of repeated observations per subject, in some situations this may simply not be possible. For instance, when the response of interest in a clinical trial is invasive for the patient and/or very expensive, it may be unethical and/or costly to subject a patient to many measurement occasions. Given that the heavily biased variance components are the only available tool to study the variability of the true distribution, this can make it difficult to evaluate whether or not problems in the linear predictors will occur.

Finally, a topic which has not received a lot of attention in the literature concerns the power and type I error related with commonly used inferential procedures like the Wald test. We have seen from our simulations that these statistical concepts can be affected in important ways by random-effects misspecification, regardless the variance of the random effects. For instance, the type I error related with the fixed intercept was found to be severely inflated, even for small variance of the random effect. In this case, increasing the number of repeated observations per subject did not help to overcome this issue. Fortunately, not all parameters are affected in the same way. Indeed, we have found that the type I error associated with a test for a covariate's effect will not be asymptotically affected, as far as this variable is not included in the random-effects structure.

Clearly, in the light of these results, detecting those cases in which the random-effects misspecification can have a serious impact on our model inferences is of the utmost importance. Therefore, in this work, we have developed a family of diagnostic tools, along the ideas introduced by White (1982). These include a number of tests based on the eigenvalues of the matrix $B_n(\boldsymbol{\xi}_0)A^{-1}(\boldsymbol{\xi}_0)$, which corresponds to an identity matrix under a correctly specified model. We have also proposed two variants of Whites Information Matrix Test. Simulations with these tools have shown that the Sandwich Estimator Test and the Modified Information Matrix Test, introduced in Chapter 10, showed the best overall performance in detecting misspecification of the random-effects structure.

The availability of such tools obviously raises the question on how to proceed when facing possible misspecification. In this work, we proposed to counter the lack

of robustness of the generalized linear mixed model by incorporating a few non-normal random-effects distributions in a sensitivity analysis framework. We argued that we can feel confident about the maximum likelihood estimates, if these are all similar, irrespective of the distribution used to obtain them. On the other hand, if the results vary considerably, then they are obviously sensitive to the distributional assumptions for the random effects, and caution is still needed.

14.2 Further Research

Although we have tried to provide a complete picture of the impact of random-effects misspecification on the maximum likelihood estimation in generalized linear mixed models, some aspects still deserve some further attention.

For instance, in our case study, the main response obtained using the CGI-scale is by definition an ordinal variable with 7 categories. Clearly, collapsing it into a binary response considerably reduces the amount of available information and can lead to a loss of efficiency. Still, we have focused our analysis and simulations on the logistic-normal model for binary data, as this easy categorization is often of main interest from a clinician's point of view. Given that binary data convey very little information, we can consider this example as a worst-case scenario to study the impact of the misspecification. On the other hand, it would also be of interest to study whether similar results are obtained with ordinal mixed models, which take into account the entire response scale.

Another prominent situation of non-Gaussian outcomes is given by the case of repeated counts. Indeed, the Poisson model has specific features, such as the existence of closed-form solutions for the mean and the variance in the marginal model. Also, the nature of overdispersion is quite different between the binary and Poisson model. Therefore, the Poisson case deserves also its own treatment.

Further, the diagnostic tools we have presented in Chapters 9 and 10 use the marginal distribution of the responses to assess the validity of the model. It is a wide known fact in mixed model literature that different random-effects distributions can generate similar marginal distributions. It is then not clear whether our diagnostic tools would, for instance, detect cases in which the normal random effects assumption is much poorer than another assumption, unless the marginal fits were quite different. Additionally, it is still not very clear how the random-effects misspecification affects the Hessian or the matrix of cross-product derivatives. Therefore, promising research

may focus on diagnostic tools in terms of the conditional distribution of the responses, or at the level of the random effects.

Furthermore, incomplete data is almost an unavoidable problem in longitudinal studies. However, it is also a problem which has received very little attention in the present work. For instance, it would be of interest to study how the performance of the diagnostic tools may be affected by the several missing data mechanisms.

Finally, future research should also be directed towards robust models under random-effects misspecification. One research angle is provided by the multivariate extension of the robust Poisson-gamma model. However, it would also be of interest to study whether conjugacy between the random-effects distribution and the conditional response distribution will lead to robust maximum likelihood estimates in other settings as well. In that sense, the h-likelihood approach, by Lee and Nelder (1996), could to be a promising alternative.

References

Agresti, A. (2002). *Categorical data analysis*, 2nd edition. Hoboken, N.J.: Wiley.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis* **47**, 639–653.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117-128.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M.G., and Vangeneugden, T. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: Canonical correlation approach. *Biometrics* **60**, 845–853.

Alonso, A., Litière, S., and Molenberghs, G. (2007). A family of tests to detect misspecifications in the random effects distribution of generalized linear mixed models. *Manuscript submitted for publication.*

Anderson, T.W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* **34**, 122–148.

Bernardo, J.N. and Smith, A.F.M. (1994). *Bayesian Theory.* Chichester: Wiley.

Böhning D. (2000). *Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping and Others.* Monographs on Statistics and Applied Probability **81**, London: Chapman & Hall/CRC.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints.* New York: Springer-Verlag.

Chen, C.F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society, Series B* **97**, 540–546.

Chen, J., Zhang, D., and Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random-effects distribution. *Biostatistics* **3**, 347–360.

Corcoran, C., Walker, E., Huot, R., Mittal, V., Tessner, K., Kestler, L., and Malaspina, D. (2003). The stress cascade and schizophrenia: etiology and onset. *Schizophrenia Bulletin* **29**, 671–692.

Day, R., Nielsen, J.A., Korten, A., Ernberg, G., *et al.* (1987). Stressful life events preceding the acute onset of schizophrenia: a cross-national study from the World Health Organization. *Culture, Medicine and Psychiatry* **11**, 123-205.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Diggle, P.J., Heagerty, P., Liang, K-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, New York: Oxford University Press.

Fieuws, S., Spiessens, B., and Draney, K. (2004). Mixture models. In *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, P. De Boeck and M. Wilson (eds), 317–340. Statistics for Social Science and Public Policy, New York: Springer-Verlag.

Fraser, D.A.S and McDunnough, P. (1984). Further remarks on the asymptotic normality of likelihood and conditional analysis. *The Canadian Journal of Statistics* **12**, 183–190.

Fu, J.C. and Kass, R.E. (1988). The exponential rate of convergence of posterior distributions. *Annals of the Institute of Statistical Mathematics* **40**, 683–691.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995) *Bayesian Data Analysis*. London: Chapman & Hall.

Ghosh, J.K., Ghosal, S., and Samanta, T. (1994). Stability and convergence of the posterior in non-regular problems. In *Statistical Decision Theory and Related*

Topics V, S.S. Gupta and J.O. Berger (eds), 183–199. New York: Springer-Verlag.

Girschick, M.A. (1939). On the sampling theory of roots of determinantal equations. *Annals of Mathematical Statistics* **10**, 203–224.

Hall, D.B. and Præstgaard, J.T. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika* **88**, 739-751.

Harrison, P.J. and Owen, M.J. (2003). Genes for schizophrenia? Recent findings and their pathophysiological implications. *Lancet* **361**, 417-419.

Heagerty, P.J. and Kurland, B.F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88**, 973–985.

Johnson, R.A., and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis.* Prentice Hall. 4th ed.

Kenward, M.G. and Molenberghs, G. (2007). *Missing Data in Clinical Studies*, Chichester: Wiley.

Kleinman, K., Lazarus, R., and Platt, R. (2004). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology* **159**, 217–224.

Kotz, S., Balakrishnan, N., Johnson, N.L. (2000) *Continuous Multivariate Distributions. Volume I: Models and Applications*, New York: Wiley.

Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73**, 805–811.

Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15**, 209–225.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics* **50**, 325-335.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B* **58**, 619–678.

Lee, Y., Nelder, J.A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood.* Boca Raton: Chapman & Hall/CRC.

Lee, K.J. and Thompson, S.G. (2007). Flexible parametric models for random effects distributions. *Statistics in Medicine* DOI:10.1002/sim.2897

Litière, S., Alonso, A., and Molenberghs, G. (2007a). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics* **63**, 1038–1044.

Litière, S., Alonso, A., and Molenberghs, G. (2007b). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine* **00**, 000-000.

Litière, S., Alonso, A., and Molenberghs, G. (2007c). A sandwich-estimator test for misspecification in mixed-effects models. *Manuscript submitted for publication.*

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* London: Chapman & Hall.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Nelson, K.P., Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J., Parzen, M., and Strawderman, R. (2006). Use of the probability integral transformation to fit nonlinear mixed-effect models with nonnormal random effects. *Journal of Computational and Graphical Statistics* **15**, 39–57.

Neuhaus, J.M., Hauck, W.W., and Kalbfleisch, J.D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755–762.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2007). Shared parameter models under random-effects misspecification. *Biometrika* **00**, 000–000.

Silvapulle, M.J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association* **90**, 342-349.

Sweeting, T.J. and Adekola, A.D. (1987). Asymptotic posterior normality for stochastic processes revisited. *Journal of the Royal Statistical Society, Series B* **49**, 215–222.

Sweeting, T.J. (1992). On asymptotic posterior normality in the multiparameter case. In *Bayesian Statistics 4*, J.M. Bernardo, J.O. DeGroot, D.V. Lindley, and A.F.M. Smith (eds), 755-762. Amsterdam: North-Holland.

Tchetgen, E.J., and Coull, B.A. (2006). A diagnostic test for the mixing distribution in a generalized linear mixed model. *Biometrika* **93**, 1003–1010.

Tempelman, R.J. (1998). Generalized linear mixed models in dairy cattle breeding. *Journal of Dairy Science* **81**, 1428–1444.

Verbeke, G. and Lesaffre, E. (1994). Large sample properties of the maximum likelihood estimators in linear mixed models with misspecified random-effects distributions. Technical Report, Report #1996.1 Biostatistical Centre for Clinical Trials, Catholic University of Leuven, Belgium.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217-221.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis* **53**, 541–556.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59**, 254-262.

Waagepetersen, R. (2006). A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. *Scandinavian Journal of Statistics* **33**, 721–731.

Waternaux, C.M. (1976). Asymptotic distribution of the sample roots for a non-normal population. *Biometrika* **63**, 639–645.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Appendix A

Case Study: the Data

subject	Z	Y0	Y1	Y2	Y4	Y6	Y8	CGI0	CGI1	CGI2	CGI4	CGI6	CGI8
1	0	0	0	0	0	0	.	5	5	5	5	5	.
2	1	0	0	0	1	0	0	4	5	5	2	5	6
3	0	0	0	0	0	.	.	7	6	5	6	.	.
4	1	0	0	0	0	0	.	6	6	5	5	6	.
5	0	0	0	0	.	.	.	6	7	7	.	.	.
6	1	0	0	0	0	.	.	6	6	5	6	.	.
7	0	0	0	0	0	0	0	5	4	5	5	5	5
8	1	1	1	1	.	.	.	3	3	3	.	.	.
9	0	0	0	0	.	.	.	5	4	5	.	.	.
10	0	0	0	0	0	0	0	5	4	4	4	5	5
11	1	0	0	0	0	.	.	6	6	6	7	.	.
12	1	0	0	0	0	.	.	5	5	5	5	.	.
13	0	1	1	1	1	1	1	3	2	2	2	2	2
14	0	0	0	0	0	0	0	4	4	4	4	5	5
15	1	0	1	1	1	1	1	4	3	2	2	3	3
16	0	0	0	0	0	.	.	5	5	4	5	.	.
17	1	1	1	1	1	1	1	3	2	2	2	2	2
18	1	0	0	0	0	0	1	6	4	4	4	4	3
19	0	0	0	1	0	0	1	4	4	3	4	4	3
20	1	0	0	0	0	0	0	4	4	4	4	4	4
21	1	0	0	0	0	.	.	5	4	4	5	.	.

subject	Z	Y0	Y1	Y2	Y4	Y6	Y8	CGI0	CGI1	CGI2	CGI4	CGI6	CGI8
22	0	0	0	.	.	.	.	5	5	.	.	.	.
23	1	0	1	1	1	1	1	4	3	3	3	3	3
24	0	0	0	0	0	0	.	5	5	5	6	6	.
25	1	0	0	0	0	.	.	5	4	6	5	.	.
26	0	0	0	0	0	1	1	4	4	4	4	3	3
27	0	0	0	0	0	.	.	5	4	4	4	.	.
28	0	0	.	0	0	0	0	5	.	4	5	4	4
29	1	0	0	0	0	0	0	5	4	4	4	4	5
30	1	0	0	0	0	1	1	5	5	5	5	3	3
31	0	0	0	0	0	0	0	5	4	4	4	4	5
32	0	0	0	.	.	.	.	6	6	.	.	.	.
33	1	0	0	0	.	.	.	4	4	4	.	.	.
34	0	0	0	0	0	0	0	5	5	5	5	5	5
35	1	0	0	0	0	0	0	6	5	5	5	5	5
36	1	0	0	0	0	0	.	6	6	6	5	6	.
37	0	0	0	0	.	.	.	5	5	5	.	.	.
38	0	0	0	1	0	.	.	4	4	3	4	.	.
39	1	0	0	0	.	.	.	5	5	5	.	.	.
40	1	0	0	1	1	.	.	5	5	3	3	.	.
41	0	0	0	0	0	.	.	5	4	4	4	.	.
42	0	0	0	0	0	.	.	6	6	6	6	.	.
43	1	0	0	0	0	0	0	6	6	6	6	6	6
44	1	0	0	0	0	0	0	6	5	6	5	5	5
45	0	0	0	0	0	0	0	5	5	5	4	4	4
46	0	0	0	.	.	.	.	6	6	.	.	.	.
47	1	0	0	0	.	.	.	4	4	6	.	.	.
48	0	0	0	0	0	.	.	5	6	6	5	.	.
49	1	1	1	1	.	.	.	3	3	3	.	.	.
50	0	0	0	.	.	.	.	5	5	.	.	.	.
51	1	0	0	0	0	1	1	5	4	4	4	3	3
52	1	0	0	0	0	0	0	5	4	4	4	4	4
53	0	0	0	0	.	.	.	4	5	5	.	.	.
54	1	0	1	1	1	1	1	5	3	3	2	2	2
55	1	1	0	1	0	1	1	3	4	3	4	3	3
56	0	0	1	1	1	1	1	7	3	3	3	3	3
57	0	0	0	1	1	1	1	4	4	3	3	2	2

subject	Z	Y0	Y1	Y2	Y4	Y6	Y8	CGI0	CGI1	CGI2	CGI4	CGI6	CGI8
58	0	0	0	0	.	.	.	6	5	6	.	.	.
59	1	0	0	1	1	1	1	5	5	3	3	3	3
60	0	0	0	.	.	.	.	5	5	.	.	.	.
61	1	0	0	0	1	1	0	6	5	4	3	3	4
62	0	0	0	.	.	.	.	5	5	.	.	.	.
63	1	0	0	.	.	.	.	4	4	.	.	.	.
64	0	0	0	0	0	0	0	5	5	5	4	4	5
65	1	0	0	0	0	.	.	4	4	4	4	.	.
66	1	0	1	1	0	.	.	4	3	3	4	.	.
67	1	0	0	0	0	0	0	5	5	5	4	4	4
68	0	0	0	0	0	.	.	6	5	5	5	.	.
69	1	0	0	0	0	0	0	6	4	4	5	4	4
70	0	0	0	0	.	.	.	7	5	7	.	.	.
71	0	0	0	0	0	.	.	5	5	5	5	.	.
72	0	0	0	0	0	0	0	5	5	5	4	4	4
73	1	0	0	0	1	1	1	4	4	4	3	3	3
74	0	0	0	0	0	0	0	4	5	4	4	4	4
75	1	0	1	1	1	1	1	5	3	3	2	2	2
76	0	0	0	0	0	0	0	5	4	4	4	4	4
77	1	0	0	0	0	0	0	5	4	4	4	4	4
78	1	0	0	0	0	.	.	6	4	4	4	.	.
79	0	0	0	0	0	0	0	6	6	6	5	5	4
80	0	0	0	0	.	.	.	6	5	6	.	.	.
81	1	0	0	0	0	1	1	6	5	4	4	3	3
82	0	1	0	0	0	1	1	3	4	4	4	3	2
83	1	0	1	0	1	1	0	5	3	4	3	3	4
84	0	0	0	0	0	1	.	5	5	4	4	3	.
85	1	0	1	1	1	1	1	4	3	3	3	2	2
86	0	0	0	0	0	0	0	5	5	4	4	5	5
87	1	0	0	0	0	1	1	6	5	4	4	3	2
88	0	0	0	0	0	0	0	4	4	4	4	5	5
89	1	0	0	.	.	.	.	4	5	.	.	.	.
90	0	0	0	.	0	.	.	5	5	.	5	.	.
91	1	0	0	0	.	0	0	4	5	5	.	5	4
92	1	0	0	0	0	.	.	5	5	5	5	.	.
93	0	0	0	0	1	1	1	5	5	5	3	3	3

subject	Z	Y0	Y1	Y2	Y4	Y6	Y8	CGI0	CGI1	CGI2	CGI4	CGI6	CGI8
94	0	1	1	1	1	1	1	2	3	3	3	3	2
95	1	0	0	0	0	0	0	5	5	5	5	5	5
96	0	0	0	0	0	.	.	5	5	5	6	.	.
97	0	0	0	1	1	1	1	5	4	3	3	3	3
98	1	0	0	0	0	0	0	6	6	6	6	6	6
99	0	0	0	0	.	.	.	4	4	5	.	.	.
100	1	0	0	0	0	0	0	6	5	5	5	6	6
101	0	0	0	0	0	0	0	4	4	4	4	4	4
102	1	1	1	1	1	1	1	3	3	3	3	3	3
103	0	0	0	0	1	0	0	5	5	4	3	4	4
104	1	0	1	1	1	1	1	4	3	2	2	2	2
105	0	0	0	.	.	.	.	5	5	.	.	.	.
106	1	0	0	.	.	.	.	4	4	.	.	.	.
107	1	0	0	0	1	.	.	5	5	4	3	.	.
108	0	0	0	0	0	.	.	5	5	4	4	.	.
109	1	0	0	0	0	1	.	4	4	4	4	3	.
110	0	0	0	0	0	0	0	4	4	4	4	4	4
111	1	0	0	0	0	0	.	4	4	4	4	4	.
112	0	0	1	0	0	1	1	5	2	4	4	2	2
113	1	0	0	0	1	0	.	6	5	4	2	4	.
114	1	0	0	.	.	.	.	5	5	.	.	.	.
115	0	0	0	0	0	.	.	4	4	4	5	.	.
116	1	0	0	0	0	.	.	6	6	5	5	.	.
117	0	0	0	.	.	.	.	4	4	.	.	.	.
118	0	1	1	1	1	1	1	3	3	2	2	2	2
119	1	1	1	1	1	1	1	3	3	2	2	2	2
120	0	0	0	0	0	0	0	5	5	4	4	4	4
121	1	0	0	0	0	0	0	5	5	5	5	5	5
122	1	0	0	0	0	0	0	5	5	4	4	4	4
123	0	0	0	0	0	0	1	5	5	4	4	4	3
124	0	0	0	0	0	1	1	4	4	4	4	3	3
125	1	0	0	0	0	1	1	5	5	4	4	3	3
126	0	0	0	0	0	.	.	5	5	5	6	.	.
127	1	0	0	0	0	0	0	4	4	4	4	4	4
128	1	0	0	.	.	.	.	5	5	.	.	.	.

Appendix B

Type I Error under Random-Effects Misspecification

The proof of Theorem 8.1 is as follows. For simplicity of the notation we will work out the proof for $\boldsymbol{x}_{Sij}^{M} = \boldsymbol{x}_{ij}^{M}$. The proof for any other subset $\boldsymbol{x}_{Sij}^{M}$ can be obtained in a similar way.

Given $V(\boldsymbol{b}_i) = D$, there always exists a lower triangular matrix U, such that $\boldsymbol{b}_i = U\boldsymbol{a}_i$, where

$$U = \begin{pmatrix} u_{11} & 0 & \cdots & 0 \\ u_{12} & u_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_{1q} & u_{2q} & \cdots & u_{qq} \end{pmatrix},$$

is the Cholesky decomposition of D, i.e.,

$$V(\boldsymbol{b}_i) = D = UU^T,$$

and $E(\boldsymbol{a}_i) = \boldsymbol{0}$ and $V(\boldsymbol{a}_i) = I$. Therefore $E(\boldsymbol{b}_i) = E(U\boldsymbol{a}_i) = \boldsymbol{0}$ and $V(\boldsymbol{b}_i) = V(U\boldsymbol{a}_i) = UV(\boldsymbol{a}_i)U^T = UU^T = D$. This allows us to write (8.1) as $\theta_{ij} = \eta(\beta_0 + \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T U\boldsymbol{a}_i)$. Let us further denote by H and F the true and the assumed distribution of the random effects. According to White (1982), the maximum likeli-

hood estimator of $\boldsymbol{\xi} = (\beta_0, \boldsymbol{\beta}, U)$ converges to the unique value of $\boldsymbol{\xi}$, denoted by $\boldsymbol{\xi}^* = (\beta_0^*, \boldsymbol{\beta}^*, U^*)$, which minimizes the KLIC, i.e. $\boldsymbol{\xi}^*$ minimizes

$$I(H:F,\boldsymbol{\xi}) = E_{\boldsymbol{x}} E_{\boldsymbol{y}|\boldsymbol{x}} \log\left\{\frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})}\right\}, \tag{B.1}$$

where the expectation is taken with respect to the true model. In the previous expression,

$$\begin{aligned} f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z}) &= \int \prod_j \exp[\varphi^{-1}\{y_j\theta_j^0 - \psi(\theta_j^0)\} + c(y_j,\varphi)] dH(\boldsymbol{a}), \\ f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z}) &= \int \prod_j \exp[\varphi^{-1}\{y_j\theta_j - \psi(\theta_j)\} + c(y_j,\varphi)] dF(\boldsymbol{a}), \end{aligned}$$

with

$$\begin{aligned} \theta_j^0 &= \eta(\beta_0^0 + \boldsymbol{x}_j^T\boldsymbol{\beta}^0 + \boldsymbol{z}_j^T U^0 \boldsymbol{a}), \\ \theta_j &= \eta(\beta_0 + \boldsymbol{x}_j^T\boldsymbol{\beta} + \boldsymbol{z}_j^T U \boldsymbol{a}). \end{aligned}$$

For simplicity in the notation, the subject index i has been omitted from the previous equations. To find $\boldsymbol{\xi}^*$ we have to differentiate (B.1) with respect to β_0, $\boldsymbol{\beta}$ and U. Let us start by determining the derivative of the KLIC with respect to β_0. Since $\boldsymbol{x}$ is independent from $\boldsymbol{\xi}$ we have that

$$\begin{aligned} \frac{\partial}{\partial\beta_0} E_{\boldsymbol{x}} E_{\boldsymbol{y}|\boldsymbol{x}} \log\left\{\frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})}\right\} &= E_{\boldsymbol{x}} \frac{\partial}{\partial\beta_0} E_{\boldsymbol{y}|\boldsymbol{x}} \log\left\{\frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})}\right\} \\ &= E_{\boldsymbol{x}} \frac{\partial}{\partial\beta_0} \int \log\left\{\frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})}\right\} f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z}) dy \\ &= E_{\boldsymbol{x}} \frac{\partial}{\partial\beta_0} \int \log\left\{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})\right\} f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z}) dy \\ &\quad - E_{\boldsymbol{x}} \frac{\partial}{\partial\beta_0} \int \log\left\{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})\right\} f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z}) dy \\ &= -E_{\boldsymbol{x}} \int \frac{\partial}{\partial\beta_0} \left[\log\left\{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})\right\} f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})\right] dy \\ &= -E_{\boldsymbol{x}} \int f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z}) \frac{\partial}{\partial\beta_0} \log\left\{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})\right\} dy \\ &= -E_{\boldsymbol{x}} \int \frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})} \frac{\partial}{\partial\beta_0} f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z}) dy \end{aligned}$$

If we let

$$\lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z}) = \frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})},$$

then

$$\frac{\partial}{\partial\beta_0}E_{\boldsymbol{x}}E_{\boldsymbol{y}|\boldsymbol{x}}\log\left\{\frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})}\right\}=-E_{\boldsymbol{x}}\int\lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})\frac{\partial}{\partial\beta_0}f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})dy. \quad \text{(B.2)}$$

Further, if we study $\dfrac{\partial}{\partial\beta_0}f_F$ in more detail, we obtain that

$$\frac{\partial}{\partial\beta_0}f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})=\int\frac{\partial}{\partial\beta_0}\prod_j\exp\left\{\varphi^{-1}[y_j\theta_j-\psi(\theta_j)]+c(y_j,\varphi)\right\}dF(\boldsymbol{a}). \quad \text{(B.3)}$$

The derivative of a product of functions, such as the one given in (B.3), can easily be determined as

$$\begin{aligned}\left(\prod_{i=1}^n f_i\right)' &= f_1'f_2\ldots f_n+f_1f_2'\ldots f_n+\ldots+f_1f_2\ldots f_n'\\ &= \sum_{i=1}^n\left(f_i'\prod_{j\neq i}f_j\right). \end{aligned}\quad \text{(B.4)}$$

Therefore, applied to (B.3), we find that

$$\begin{aligned}\frac{\partial}{\partial\beta_0}f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})=&\int\sum_i\frac{\partial}{\partial\beta_0}[\exp\left\{\varphi^{-1}[y_i\theta_i-\psi(\theta_i)]+c(y_i,\varphi)\right\}]\\ &\times\prod_{j\neq i}\exp\left\{\varphi^{-1}[y_j\theta_j-\psi(\theta_j)]+c(y_j,\varphi)\right\}dF(\boldsymbol{a}). \end{aligned}\quad \text{(B.5)}$$

On the other hand,

$$\begin{aligned}&\frac{\partial}{\partial\beta_0}\left[\exp\left\{\varphi^{-1}[y_i\theta_i-\psi(\theta_i)]+c(y_i,\varphi)\right\}\right]\\ &\quad= \exp\left\{\varphi^{-1}[y_i\theta_i-\psi(\theta_i)]+c(y_i,\varphi)\right\}\frac{\partial}{\partial\beta_0}\left\{\varphi^{-1}[y_i\theta_i-\psi(\theta_i)]+c(y_i,\varphi)\right\}\\ &\quad= \exp\left\{\varphi^{-1}[y_i\theta_i-\psi(\theta_i)]+c(y_i,\varphi)\right\}\left\{\varphi^{-1}\left[y_i\frac{\partial\theta_i}{\partial\beta_0}-\psi'(\theta_i)\frac{\partial\theta_i}{\partial\beta_0}\right]\right\} \end{aligned}\quad \text{(B.6)}$$

Substituting (B.6) in (B.5) we obtain

$$\begin{aligned}\frac{\partial}{\partial\beta_0}f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})=&\int\sum_i\left\{\varphi^{-1}\left[y_i\frac{\partial\theta_i}{\partial\beta_0}-\psi'(\theta_i)\frac{\partial\theta_i}{\partial\beta_0}\right]\right\}\\ &\times\prod_j\exp\left\{\varphi^{-1}[y_j\theta_j-\psi(\theta_j)]+c(y_j,\varphi)\right\}dF(\boldsymbol{a}) \end{aligned}\quad \text{(B.7)}$$

Since

$$\begin{aligned}\frac{\partial\theta_i}{\partial\beta_0} &= \frac{\partial}{\partial\beta_0}\eta(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{z}_i^TU\boldsymbol{a}) \\ &= \eta'(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{z}_i^TU\boldsymbol{a}),\end{aligned} \tag{B.8}$$

it is clear that

$$\begin{aligned}\frac{\partial}{\partial\beta_0}f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z}) = \int\sum_i \eta'(\boldsymbol{\xi})\varphi^{-1}[y_i-\psi'(\theta_i)] \\ \times\prod_j\exp\left\{\varphi^{-1}[y_j\theta_j-\psi(\theta_j)]+c(y_j,\varphi)\right\}dF(\boldsymbol{a}).\end{aligned} \tag{B.9}$$

For simplicity of notation we have referred to $\eta'(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{z}_i^TU\boldsymbol{a})$ as $\eta'(\boldsymbol{\xi})$. Finally, substituting this last expression in (B.2) leads to the derivative of the KLIC (B.1) with respect to β_0, given by

$$\begin{aligned}&\frac{\partial}{\partial\beta_0}E_{\boldsymbol{x}}E_{\boldsymbol{y}|\boldsymbol{x}}\left\{\log\frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0,\boldsymbol{x},\boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{z})}\right\} = -E_{\boldsymbol{x}}\int\lambda(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})\times \\ &\int\sum_i\eta'(\boldsymbol{\xi})\varphi^{-1}[y_i-\psi'(\theta_i)]\prod_j\exp\left\{\varphi^{-1}[y_j\theta_j-\psi(\theta_j)]+c(y_j,\varphi)\right\}dF(\boldsymbol{a})dy.\end{aligned} \tag{B.10}$$

Similar calculations are needed to determine the derivative of the KLIC (B.1) with respect to $\boldsymbol{\beta}$. However, they will require an expression for $\dfrac{\partial\theta_i}{\partial\boldsymbol{\beta}}$. First, let $\boldsymbol{c}$ and $\boldsymbol{x}$ be two $n\times 1$ vectors such that $\boldsymbol{c}^T\boldsymbol{x}$ is a scalar. By definition, the derivative of $\boldsymbol{c}^T\boldsymbol{x}$ by $\boldsymbol{x}$ is given by

$$\frac{\partial\boldsymbol{c}^T\boldsymbol{x}}{\partial\boldsymbol{x}} = \begin{pmatrix}\dfrac{\partial\boldsymbol{c}^T\boldsymbol{x}}{\partial x_1}\\ \dfrac{\partial\boldsymbol{c}^T\boldsymbol{x}}{\partial x_2}\\ \vdots\\ \dfrac{\partial\boldsymbol{c}^T\boldsymbol{x}}{\partial x_n}\end{pmatrix} = \boldsymbol{c}. \tag{B.11}$$

Therefore,

$$\begin{aligned}\frac{\partial\theta_i}{\partial\boldsymbol{\beta}} &= \frac{\partial}{\partial\boldsymbol{\beta}}\eta(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{z}_i^TU\boldsymbol{a}) \\ &= \eta'(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{z}_i^TU\boldsymbol{a})\boldsymbol{x}_i,\end{aligned} \tag{B.12}$$

and the derivative $\frac{\partial}{\partial \boldsymbol{\beta}} f_F$ is given by

$$\frac{\partial}{\partial \boldsymbol{\beta}} f_F(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{x}, \boldsymbol{z}) = \int \sum_i \boldsymbol{x}_i^T \eta'(\boldsymbol{\xi}) \varphi^{-1}[y_i - \psi'(\theta_i)] \times \prod_j \exp\left\{\varphi^{-1}[y_j\theta_j - \psi(\theta_j)] + c(y_j, \varphi)\right\} dF(\boldsymbol{a}). \quad \text{(B.13)}$$

Therefore, the derivative of the KLIC (B.1) with respect to $\boldsymbol{\beta}$ is determined by

$$\frac{\partial}{\partial \boldsymbol{\beta}} E_{\boldsymbol{x}} E_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \log \frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0, \boldsymbol{x}, \boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{x}, \boldsymbol{z})} \right\} = -E_{\boldsymbol{x}} \int \lambda(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) \times$$
$$\int \sum_i \boldsymbol{x}_i \eta'(\boldsymbol{\xi}) \varphi^{-1}[y_i - \psi'(\theta_i)] \prod_j \exp\left\{\varphi^{-1}[y_j\theta_j - \psi(\theta_j)] + c(y_j, \varphi)\right\} dF(\boldsymbol{a}) dy. \quad \text{(B.14)}$$

Finally, determining the derivative of the KLIC (B.1) with respect to U will require

$$\begin{aligned} \frac{\partial \theta_i}{\partial U} &= \frac{\partial}{\partial U} \eta(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T U \boldsymbol{a}) \\ &= \eta'(\boldsymbol{\xi}) \frac{\partial (\boldsymbol{z}_i^T U \boldsymbol{a})}{\partial U}. \end{aligned} \quad \text{(B.15)}$$

Since

$$\begin{aligned} \boldsymbol{z}_i^T U \boldsymbol{a} &= (z_{i1} \cdots z_{iq}) \begin{pmatrix} u_{11} & 0 & \cdots & 0 \\ u_{12} & u_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_{1q} & u_{q2} & \cdots & u_{qq} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{pmatrix} \\ &= (z_{i1} \cdots z_{iq}) \begin{pmatrix} u_{11}a_1 \\ u_{12}a_1 + u_{22}a_2 \\ \cdots \\ u_{1q}a_1 + u_{2q}a_2 + \ldots u_{qq}a_q \end{pmatrix} \\ &= \sum_{k=1}^{q} z_{ik} \cdot \sum_{l \le k} u_{lk} a_l, \end{aligned}$$

this implies that

$$\frac{\partial \boldsymbol{z}_i^T U \boldsymbol{a}}{\partial u_{lk}} = \begin{cases} z_{ik} a_l & \text{if } u_{lk} \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{(B.16)}$$

Furthermore, the derivative of a function f w.r.t. a matrix A is defined as

$$\frac{\partial f}{\partial A} = \begin{pmatrix} \frac{\partial f}{\partial a_{11}} & \cdots & \frac{\partial f}{\partial a_{1q}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial a_{q1}} & \cdots & \frac{\partial f}{\partial a_{qq}} \end{pmatrix}. \tag{B.17}$$

Therefore, the derivative of $\boldsymbol{z}_i^T U \boldsymbol{a}$ with respect to U is given by

$$\frac{\partial \boldsymbol{z}_i^T U \boldsymbol{a}}{\partial U} = \begin{pmatrix} z_{i1}a_1 & 0 & \cdots & 0 \\ z_{i2}a_1 & z_{i2}a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ z_{iq}a_1 & z_{iq}a_2 & \cdots & z_{iq}a_q \end{pmatrix} = Q_i. \tag{B.18}$$

Substituting (B.18) in (B.15) leads to

$$\frac{\partial \theta_i}{\partial U} = \eta'(\boldsymbol{\xi}) Q_i. \tag{B.19}$$

This results in the derivative of the KLIC (B.1) with respect to U being equal to

$$\frac{\partial}{\partial U} E_{\boldsymbol{x}} E_{\boldsymbol{y}|\boldsymbol{x}} \log \left\{ \frac{f_H(\boldsymbol{y}|\boldsymbol{\xi}_0, \boldsymbol{x}, \boldsymbol{z})}{f_F(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{x}, \boldsymbol{z})} \right\} = -E_{\boldsymbol{x}} \int \lambda(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) \cdot$$
$$\int \sum_i Q_i \eta'(\boldsymbol{\xi}) \varphi^{-1} [y_i - \psi'(\theta_i)] \prod_j \exp\{\varphi^{-1}[y_j\theta_j - \psi(\theta_j)] + c(y_j, \varphi)\} dF(\boldsymbol{a}) dy. \tag{B.20}$$

Now $\boldsymbol{\xi}^*$, which minimizes the KLIC (B.1), is the solution to the following system of equations

$$\begin{cases} E_{\boldsymbol{x}} \int \lambda(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) \int \sum_i k_i(\boldsymbol{x}, \boldsymbol{a}) dF(\boldsymbol{a}) dy = 0 \\ E_{\boldsymbol{x}} \int \lambda(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) \int \sum_i \boldsymbol{x}_i k_i(\boldsymbol{x}, \boldsymbol{a}) dF(\boldsymbol{a}) dy = 0 \\ E_{\boldsymbol{x}} \int \lambda(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) \int \sum_i Q_i k_i(\boldsymbol{x}, \boldsymbol{a}) dF(\boldsymbol{a}) dy = 0 \end{cases} \tag{B.21}$$

where $k_i(\boldsymbol{x}, \boldsymbol{a}) = \eta'(\boldsymbol{\xi}) \varphi^{-1} [y_i - \psi'(\theta_i)] \prod_j \exp\{\varphi^{-1}[y_j\theta_j - \psi(\theta_j)] + c(y_j, \varphi)\}$. The rest of the proof is based partly on the assumption that $\boldsymbol{\xi}^*$ is unique (A3 in White, 1982). Therefore, if we assume that the p_0-dimensional vector $\boldsymbol{\beta}^{M*} = 0$ and if we can find a solution for (B.21) under this assumption, then we have found the unique minimum of the KLIC.

Recall that $E(\boldsymbol{x}) = 0$. If $\boldsymbol{\beta}^{M0} = \boldsymbol{\beta}^{M*} = 0$ then it is clear that $\lambda(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z})$, $\eta(\beta_0^* + \boldsymbol{x}_i^T \boldsymbol{\beta}^* + \boldsymbol{z}_i^T U^* \boldsymbol{a})$, $\eta'(\beta_0^* + \boldsymbol{x}_i^T \boldsymbol{\beta}^* + \boldsymbol{z}_i^T U^* \boldsymbol{a})$, $\psi(\theta_i^*)$, $\psi'(\theta_j^*)$ and $h_i(\boldsymbol{x}, \boldsymbol{a})$ are functions

which are independent of $\boldsymbol{x}_i^M$. In this case the second equation in (B.21) can also be written as

$$E_{\boldsymbol{x}}\left[\begin{array}{l} \int \lambda(\boldsymbol{y}|\boldsymbol{x}^R) \int \sum_i \boldsymbol{x}_i^M k_i(\boldsymbol{x}^R, \boldsymbol{a}) dF(\boldsymbol{a}) dy \\ \\ \int \lambda(\boldsymbol{y}|\boldsymbol{x}^R) \int \sum_i \boldsymbol{x}_i^R k_i(\boldsymbol{x}^R, \boldsymbol{a}) dF(\boldsymbol{a}) dy \end{array} = 0\right], \tag{B.22}$$

or equivalently as the two equations

$$E_{\boldsymbol{x}}\left[\int \lambda(\boldsymbol{y}|\boldsymbol{x}^R) \int \sum_i \boldsymbol{x}_i^M k_i(\boldsymbol{x}^R, \boldsymbol{a}) dF(\boldsymbol{a}) dy\right] = 0, \tag{B.23}$$

and

$$E_{\boldsymbol{x}}\left[\int \lambda(\boldsymbol{y}|\boldsymbol{x}^R) \int \sum_i \boldsymbol{x}_i^R k_i(\boldsymbol{x}^R, \boldsymbol{a}) dF(\boldsymbol{a}) dy\right] = 0. \tag{B.24}$$

First note that, if $K(\boldsymbol{y}|\boldsymbol{x}) = \int \sum_i \boldsymbol{x}_i^M k_i(\boldsymbol{x}^R, \boldsymbol{a}) dF(\boldsymbol{a})$, then

$$\begin{aligned} E_{\boldsymbol{x}} \int \lambda(\boldsymbol{y}|\boldsymbol{x}^R) K(\boldsymbol{y}|\boldsymbol{x}) dy &= \int \left[\int \lambda(\boldsymbol{y}|\boldsymbol{x}^R) K(\boldsymbol{y}|\boldsymbol{x}) dy\right] f(\boldsymbol{x}) dx \\ &= \int \int \lambda(\boldsymbol{y}|\boldsymbol{x}^R) K(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}) dx dy \\ &= \int \left[\int \lambda(\boldsymbol{y}|\boldsymbol{x}^R) K(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}) dx\right] dy \\ &= \int E_{\boldsymbol{x}}[\lambda(\boldsymbol{y}|\boldsymbol{x}^R) K(\boldsymbol{y}|\boldsymbol{x})] dy. \end{aligned} \tag{B.25}$$

Second, if $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ are independent and identically distributed according to a function f then, for any function g

$$\begin{aligned} E_{\boldsymbol{x}}[g(\boldsymbol{x})] &= \int \int g(\boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_1, \boldsymbol{x}_2) dx_1 dx_2 \\ &= \int \int g(\boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_1) f(\boldsymbol{x}_2) dx_1 dx_2 \\ &= \int \left\{\int g(\boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_1) dx_1\right\} f(\boldsymbol{x}_2) dx_2 \\ &= E_{\boldsymbol{x}_2}\{E_{\boldsymbol{x}_1}[g(\boldsymbol{x})]\} \end{aligned}$$

Taking into account that $\boldsymbol{x}^M$ and $\boldsymbol{x}^R$ are independent, applied to (B.23) we obtain

that

$$
\begin{aligned}
& E_{\boldsymbol{x}}\left[\int \lambda(\boldsymbol{y}|\boldsymbol{x}^R)\int \sum_i \boldsymbol{x}_i^M k_i(\boldsymbol{x}^R,\boldsymbol{a})dF(\boldsymbol{a})dy\right] \\
&\quad = \int E_{\boldsymbol{x}}\left[\lambda(\boldsymbol{y}|X^R)\int \sum_i \boldsymbol{x}_i^M k_i(\boldsymbol{x}^R,\boldsymbol{a})dF(\boldsymbol{a})\right]dy \\
&\quad = \int E_{\boldsymbol{x}^R}\left\{E_{\boldsymbol{x}^M}\left[\lambda(\boldsymbol{y}|\boldsymbol{x}^R)\int \sum_i \boldsymbol{x}_i^M k_i(\boldsymbol{x}^R,\boldsymbol{a})dF(\boldsymbol{a})\right]\right\}dy \\
&\quad = \int E_{\boldsymbol{x}^R}\left\{\lambda(\boldsymbol{y}|\boldsymbol{x}^R)E_{\boldsymbol{x}^M}\left[\int \sum_i \boldsymbol{x}_i^M k_i(\boldsymbol{x}^R,\boldsymbol{a})dF(\boldsymbol{a})\right]\right\}dy \\
&\quad = \int E_{\boldsymbol{x}^R}\left\{\lambda(\boldsymbol{y}|\boldsymbol{x}^R)\sum_i E_{\boldsymbol{x}^M}(\boldsymbol{x}_i^M)\int k_i(\boldsymbol{x}^R,\boldsymbol{a})dF(\boldsymbol{a})\right\}dy. \qquad \text{(B.26)}
\end{aligned}
$$

Therefore,

$$
\int E_{\boldsymbol{x}^R}\left\{\lambda(\boldsymbol{y}|\boldsymbol{x}^R)\sum_i E_{\boldsymbol{x}^M}(\boldsymbol{x}_i^M)\int k_i(\boldsymbol{x}^R,\boldsymbol{a})dF(\boldsymbol{a})\right\}dy = 0. \qquad \text{(B.27)}
$$

Since $E_{\boldsymbol{x}^M}(\boldsymbol{x}_i^M) = 0$ it follows that the p_0 dimensional left-hand side of (B.27) becomes zero. This means that the second equation in (B.21) leads to $p - p_0$ equations. Together with the first and third expression in (B.21), this lead to a total of $1 + (p - p_0) + \frac{(q)(q+1)}{2}$ equations. The number of parameters also corresponds to $1 + (p - p_0) + \frac{q(q+1)}{2}$. Thus, when $\boldsymbol{\beta}^{M0} = \boldsymbol{\beta}^{M*} = 0$ we can find a solution for the system of $1 + (p - p_0) + \frac{q(q+1)}{2}$ equations (B.21) with $1 + (p - p_0) + \frac{q(q+1)}{2}$ parameters, and this solution is unique. Therefore when $\boldsymbol{\beta}^{M0} = 0$, then also $\boldsymbol{\beta}^{M*} = 0$. And since the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_n^M$, based on a GLMM with a misspecified random-effects distribution, consistently estimates $\boldsymbol{\beta}^{M*} = 0$, this implies that (8.2) holds. □

Appendix C

Implementation of the Diagnostic Tools

In this appendix, we briefly illustrate, using some exemplary SAS code, how the diagnostic tools presented in Chapters 9 and 10 can be calculated for the example of the case study. The first step to obtain the test statistics and corresponding p-values consists in running the statistical model to obtain the maximum likelihood estimates and the Hessian. For, instance, consider model (4.4) for the analysis of the schizophrenia data.

```
proc nlmixed data=new qpoints=50 hess;
  parms beta0=1 beta1=1 beta2=1 s2u1=1;
  eta = beta0 + beta1*Z + beta2*time + u;
  expeta=exp(eta);
  p=expeta/(1+expeta);
  model Y ~ binary(p);
  random u ~ normal(0,s2u1) subject=subject;
  ods output ParameterEstimates=pe Hessian=hess;
run;
```

Next, we run the following macro, which will determine each subjects contribution to $A_n(\boldsymbol{\xi})$ and $B_n(\boldsymbol{\xi})$. Note that the subjects in the data require an identification number, going from 1 to n, the total number of subjects.

```
%macro individual(data=, n=);

data glmmgrad; set _NULL_; run;
data hessind; set _NULL_; run;

%do i = 1 %to &n;
data hulp;
  set &data;
  if subject ^= &i then delete;
run;
proc nlmixed data=hulp qpoints=50 maxiter=0 hess;
  parms / data=pe;
  eta = beta0 + beta1*Z + beta2*time + u;
  expeta=exp(eta);
  p=expeta/(1+expeta);
  model Y ~ binary(p);
  random u ~ normal(0,s2u1) subject=subject;
  ods output ParameterEstimates=pe1 Hessian=hess1;
run;
data glmmgrad; set glmmgrad pe1; run;
data hessind; set hessind hess1; run;
%end;

%mend;

%individual(data=new, n=128);
```

Note that running this macro will lead to a repetition of the following error:

```
ERROR: QUANEW Optimization cannot be completed.
ERROR: QUANEW needs more than 0 iterations or 500 function calls
```

This does not reflect an error in the program, but is a result of the NLMIXED model fitting, with fixed parameters, to each subject in the data set separately.

The files `glmmgrad` and `hessind` contain the first and second order derivatives of the likelihood for each subject in the data set. We can now use IML to calculate the test statistics.

```
proc iml; reset noprint;
n=128;

/* determining A */
use hess; read all into hess;
na = ncol(hess);
np = nrow(hess);
A = -hess[,2:na]; An = A/n; print An;

/* determining B */
use glmmgrad; read all into gradient;
i=1; B = {0};
begin=1;
do while (i <= n);
    end = begin -1 + np;
    first = gradient[begin:end, 2];
    B = B + first*first`;
    begin = end + 1;
    i = i+1;
end;
Bn = B/n; print Bn;

/* the determinant tests */
delta1 = log(det(-Bn*inv(An)));
test_det1 = (n/(2*np))*(delta1**2);
prob_det1 = 1 - PROBCHI(test_det1,1);
print delta1 test_det1 prob_det1;
delta2 = det(Bn)*det(-inv(An));
test_det2 = (n/(2*np))*((delta2-1)**2);
prob_det2 = 1 - PROBCHI(test_det2,1);
print delta2 test_det2 prob_det2;

/* the determinant-trace tests */
delta_dt = trace(Bn)/trace(-An) - det(Bn)/det(-An);
Eb = eigval(-An);
```

```
sigma = sum((Eb/trace(-An) - 1)##2);
test_dettr = (n/2)*(delta_dt**2)/sigma;
prob_dettr = 1-probchi(test_dettr,1);
print delta_dt test_dettr prob_dettr;

/* the SET */

* mean of bi;
meanBi = shape(t(Bn), ncol(Bn)**2,1);

* covariance matrix of bi;
begin=1; varBi={0};
do s=1 to n;
    end = begin -1 + np;
    first = gradient[begin:end, 2];
    Prb = shape(t(first*first‘), ncol(first*first‘)**2, 1);
    varBi = varBi + (Prb-meanBi)*(Prb-meanBi)‘;
    begin = end + 1;
end;
varBi = varBi/(n*(n-1));

* matrix Delta as defined in (10.1);
delta = J(np, np**2, 0);
i = 1;
do while (i <= np);
    j = (i-1)*np + i;
    delta[i, j] = 1;
    i = i+1;
end;

* the test;
Vn = inv(An)*Bn*inv(An);
vd = vecdiag(Vn + inv(An));
Cv = delta * (inv(An)@inv(An))*VarBi*(inv(An)@inv(An)) * t(delta);
```

```
test_SET = t(vd)* inv(Cv) * vd;
prob_SET = 1- PROBCHI(test_SET,np);
print vd test_SET prob_SET;

/* the MIMT */

use hessind; read all into hessind;

meanAi = shape(t(An), ncol(An)**2,1);
begin=1; varAi={0}; covAB={0}; covBA={0};
do s=1 to n;
    end = begin -1 + np;
    first = gradient[begin:end, 2];
    Prb = shape(t(first*first‘), ncol(first*first‘)**2, 1);
    second = -hessind[begin:end, 2:5];
    Pra = shape(t(second), ncol(second)**2, 1);
    varAi = varAi + (Pra-meanAi)*(Pra-meanAi)‘;
    covAB = covAB + (Pra-meanAi)*(Prb-meanBi)‘;
    covBA = covBA + (Prb-meanBi)*(Pra-meanAi)‘;
    begin = end + 1;
end;
varAi = varAi/(n*(n-1));
covAB = covAB/(n*(n-1));
covBA = covBA/(n*(n-1));

Dn = vecdiag(An + Bn);
Cd = delta * (varAi + varBi + covAB + covBA) * t(delta) ;

test_MIMT = t(Dn)*inv(Cd)*Dn;
prob_MIMT = 1- PROBCHI(test_MIMT,np);
print Dn test_MIMT prob_MIMT;
quit; run;
```

Appendix D

The Information Matrix Test for Linear Mixed Models

In linear mixed models, the normal random-effects distribution is conjugate to the normal distribution of the outcome. Consequently, in this setting, the marginal likelihood is available in closed form. Hence, first, second and third order derivatives of the corresponding loglikelihood, necessary to calculate the Information Matrix Test statistic $\Im(N)$ defined in Theorem 5.5, can easily be obtained analytically.

Consider the marginal model, induced by a linear mixed model as described in Section 3.3, such that $\boldsymbol{y}_i \sim \mathrm{N}(X_i\boldsymbol{\beta}, V_i)$, where $V_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$. Let p represent the number of parameters, and q the dimension of D. Further, let $\boldsymbol{\xi}$ represent the $u = p + \frac{q(q+1)}{2} + 1$ dimensional vector containing all parameters of interest: the fixed effects $\boldsymbol{\beta}$, the vector $\boldsymbol{d} = (d_{11}, d_{12}, \ldots, d_{1q}, d_{22}, \ldots, d_{2q}, \ldots, d_{qq})^T$ containing all $q(q+1)/2$ variance components in the upper triangular part of D, and σ^2. The marginal likelihood of this model, on subject level, is then given by

$$f(\boldsymbol{y}_i, \boldsymbol{\xi}) = (2\pi)^{-n_i/2} |V_i|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - X_i\boldsymbol{\beta})^T V_i^{-1} (\boldsymbol{y}_i - X_i\boldsymbol{\beta}) \right\}, \tag{D.1}$$

and the corresponding marginal loglikelihood function follows as

$$\log f(\boldsymbol{y}_i, \boldsymbol{\xi}) = -\frac{n_i}{2} \ln(2\pi) - \frac{1}{2} \ln |V_i| - \frac{1}{2} (\boldsymbol{y}_i - X_i\boldsymbol{\beta})' V_i^{-1} (\boldsymbol{y}_i - X_i\boldsymbol{\beta}). \tag{D.2}$$

Recall that the IMT statistic $\Im(N)$ is based on $d_k(\boldsymbol{y}_i, \boldsymbol{\xi})$ and the covariance matrix

$C_n(\boldsymbol{\xi})$. This first component was defined as

$$d_k(\boldsymbol{y}_i, \boldsymbol{\xi}) = \left\{ \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k} \right\}^2 + \frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{(\partial \xi_k)^2}. \tag{D.3}$$

where $k = 1, \ldots, u$, whereas Expression (5.14) for $C_n(\boldsymbol{\xi})$, requires the calculation of

$$\frac{\partial d_k(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_\ell} = 2 \frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_\ell \partial \xi_k} \frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k} + \frac{\partial^3 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_\ell (\partial \xi_k)^2}. \tag{D.4}$$

Therefore, to determine $\Im(N)$, we need expressions for

- $\dfrac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$
- $\dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^T \otimes \partial \boldsymbol{\xi}}$, and
- the derivative of the diagonal elements of $\dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^T \otimes \partial \boldsymbol{\xi}}$ with respect to ξ_ℓ.

However, before we start to work these out, we first provide in the next section an overview of some mathematical tools which will be used throughout this appendix.

D.1 Some properties of matrix derivatives

Chain rule for matrices
If $z = f_1(\boldsymbol{r})$ and $\boldsymbol{r} = f_2(\boldsymbol{x})$, where z is a scalar, and $\boldsymbol{r}$ and $\boldsymbol{x}$ are vectors, then

$$\frac{\partial z}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{r}^T}{\partial \boldsymbol{x}} \cdot \frac{\partial z}{\partial \boldsymbol{r}}. \tag{D.5}$$

Derivative of a quadratic form
If A is a $n \times n$ matrix and $\boldsymbol{x}$ is a $n \times 1$ vector, then

$$\frac{\partial (\boldsymbol{x}^T A \boldsymbol{x})}{\partial \boldsymbol{x}} = A\boldsymbol{x} + A^T \boldsymbol{x}. \tag{D.6}$$

Derivatives of an inner product

If A is a $n \times n$ matrix and $\boldsymbol{x}$ is a $n \times 1$ vector, then

$$\begin{aligned} \frac{\partial(A\boldsymbol{x})}{\partial \boldsymbol{x}} &= vec(A), && \text{(D.7)} \\ \frac{\partial(A\boldsymbol{x})}{\partial \boldsymbol{x}^T} &= A, && \text{(D.8)} \\ \frac{\partial(\boldsymbol{x}^T A)}{\partial \boldsymbol{x}^T} &= (vec(A^T))^T, && \text{(D.9)} \\ \frac{\partial(\boldsymbol{x}^T A)}{\partial \boldsymbol{x}} &= A. && \text{(D.10)} \\ & && \text{(D.11)} \end{aligned}$$

Derivative of the natural logarithm of a determinant

For any symmetric or non-symmetric matrix A and for a scalar y:

$$\frac{\partial \ln |A|}{\partial y} = \text{tr}\left(A^{-1}\frac{\partial A}{\partial y}\right). \tag{D.12}$$

Derivative of the trace of a matrix

First, recall the following properties of the trace of a matrix. If A, B and C are matrices such that ABC is a square matrix, then

$$\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB) \tag{D.13}$$

Further, if A and B are two $n \times n$ matrices, then

$$\begin{aligned} \text{tr}(A + B) &= \text{tr}(A) + \text{tr}(B), && \text{(D.14)} \\ \text{tr}(A') &= \text{tr}(A) && \text{(D.15)} \end{aligned}$$

Now, let y be a scalar. Then

$$\frac{\partial \text{tr}(A)}{\partial y} = \text{tr}\left(\frac{\partial A}{\partial y}\right) \tag{D.16}$$

Derivative of an inverse matrix

If A is a square, non-singular matrix, and y is a scalar, then

$$\frac{\partial A^{-1}}{\partial y} = -A^{-1}\frac{\partial A}{\partial y}A^{-1}. \tag{D.17}$$

D.2 First-Order Derivatives

First, observe that

$$\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \dfrac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{d}} \\ \dfrac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \sigma^2} \end{pmatrix}.$$

In this section, we will work out the first-order derivative for each of these components separately.

With respect to $\boldsymbol{\beta}$

The first-order derivative of the marginal loglikelihood with respect to $\boldsymbol{\beta}$ is given by

$$\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} = -\frac{1}{2}\frac{\partial \ln |V_i|}{\partial \boldsymbol{\beta}} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})^T V_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta}). \tag{D.18}$$

Since V_i does not depend on $\boldsymbol{\beta}$, the first term in (D.18) disappears. Further, let us denote by $\boldsymbol{r}_i$ the vector $\boldsymbol{y}_i - X_i\boldsymbol{\beta}$ then, applying the chain rule (D.5), we obtain

$$\frac{\partial(\boldsymbol{r}_i^T V_i^{-1}\boldsymbol{r}_i)}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{r}_i^T}{\partial \boldsymbol{\beta}} \cdot \frac{\partial(\boldsymbol{r}_i^T V_i^{-1}\boldsymbol{r}_i)}{\partial \boldsymbol{r}_i}.$$

The first factor in this expression corresponds to

$$\frac{\partial \boldsymbol{r}_i^T}{\partial \boldsymbol{\beta}} = \frac{\partial(\boldsymbol{y}_i - X_i\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}} = -\frac{\partial \boldsymbol{\beta}^T X_i^T}{\partial \boldsymbol{\beta}} \overset{(D.10)}{=} -X_i^T,$$

while the second factor can easily be obtained following the rule for the derivative of a quadratic form (D.6), i.e.,

$$\frac{\partial(\boldsymbol{r}_i^T V_i^{-1}\boldsymbol{r}_i)}{\partial \boldsymbol{r}_i} = V_i^{-1}\boldsymbol{r}_i + (V_i^{-1})^T\boldsymbol{r}_i = 2V_i^{-1}\boldsymbol{r}_i.$$

Substituting these results in (D.18) now leads to the following expression for the first-order derivative of the marginal loglikelihood with respect to $\boldsymbol{\beta}$

$$\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} = X_i^T V_i^{-1}\boldsymbol{r}_i. \tag{D.19}$$

With respect to $\boldsymbol{d}$

Recall that this subset of $\boldsymbol{\xi}$ contains the elements in the upper triangular part of D. To simplify notation, we will work out the first-order derivative of the marginal loglikelihood with respect to $\boldsymbol{d}$, for each element separately. Therefore, for each $1 \leq a \leq b \leq q$, we need to determine

$$\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial d_{ab}} = -\frac{1}{2}\frac{\partial \ln |V_i|}{\partial d_{ab}} - \frac{1}{2}\frac{\partial}{\partial d_{ab}}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})^T V_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta}). \tag{D.20}$$

The first term in (D.20) can be written as

$$\frac{\partial \ln |V_i|}{\partial d_{ab}} \overset{(D.12)}{=} \operatorname{tr}\left(V_i^{-1}\frac{\partial V_i}{\partial d_{ab}}\right). \tag{D.21}$$

Note that

$$\begin{aligned}
V_i &= Z_i D Z_i^T + \sigma^2 I_{n_i} \\
&= \begin{pmatrix} Z_{i11} & \dots & Z_{i1q} \\ \vdots & \ddots & \vdots \\ Z_{in_i1} & \dots & Z_{in_iq} \end{pmatrix} \begin{pmatrix} d_{11} & \dots & d_{1q} \\ \vdots & \ddots & \vdots \\ d_{q1} & \dots & d_{qq} \end{pmatrix} \begin{pmatrix} Z_{i11} & \dots & Z_{in_i1} \\ \vdots & \ddots & \vdots \\ Z_{i1q} & \dots & Z_{in_iq} \end{pmatrix} + \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix} \\
&= \begin{pmatrix} \sum_{j=1}^q Z_{i1j}d_{j1} & \dots & \sum_{j=1}^q Z_{i1j}d_{jq} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^q Z_{in_ij}d_{j1} & \dots & \sum_{j=1}^q Z_{in_ij}d_{jq} \end{pmatrix} \begin{pmatrix} Z_{i11} & \dots & Z_{in_i1} \\ \vdots & \ddots & \vdots \\ Z_{i1q} & \dots & Z_{in_iq} \end{pmatrix} + \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix} \\
&= \begin{pmatrix} \sum_{k=1}^q Z_{i1k}(\sum_{j=1}^q Z_{i1j}d_{jk}) + \sigma^2 & \dots & \sum_{k=1}^q Z_{in_ik}(\sum_{j=1}^q Z_{i1j}d_{jk}) \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^q Z_{i1k}(\sum_{j=1}^q Z_{in_ij}d_{jk}) & \dots & \sum_{k=1}^q Z_{in_ik}(\sum_{j=1}^q Z_{in_ij}d_{jk}) + \sigma^2 \end{pmatrix}
\end{aligned}$$

If we define $\delta_{lm} = 1$ if $l = m$ and $\delta_{lm} = 0$ otherwise, then

$$V_i = \left\{ \sum_{k=1}^q Z_{imk}(\sum_{j=1}^q Z_{ilj}d_{jk}) + \sigma^2\delta_{lm} \right\}_{1 \leq l,m \leq n_i}. \tag{D.22}$$

As a result, when $a = b$, then it follows that

$$\begin{aligned}
\frac{\partial V_i}{\partial d_{aa}} &= \{Z_{ima}Z_{ila}\}_{lm} \\
&= \begin{pmatrix} Z_{i1a}Z_{i1a} & \dots & Z_{i1a}Z_{in_ia} \\ \vdots & \ddots & \vdots \\ Z_{in_ia}Z_{i1a} & \dots & Z_{in_ia}Z_{in_ia} \end{pmatrix} \\
&= \begin{pmatrix} Z_{i1a} \\ \vdots \\ Z_{in_ia} \end{pmatrix} \begin{pmatrix} Z_{i1a} & \dots & Z_{in_ia} \end{pmatrix} \\
&= Z_{i.a}Z_{i.a}^T, \qquad \text{(D.23)}
\end{aligned}$$

where $Z_{i.a}$ represents the ath column of Z_i. On the other hand, when $a < b$, we find that

$$\begin{aligned}
\frac{\partial V_i}{\partial d_{ab}} &= \{Z_{ima}Z_{ilb} + Z_{imb}Z_{ila}\}_{lm} \\
&= Z_{i.b}Z_{i.a}^T + Z_{i.a}Z_{i.b}^T. \qquad \text{(D.24)}
\end{aligned}$$

Therefore, combining (D.23) and (D.24), leads to the following result for $a \leq b$

$$\frac{\partial V_i}{\partial d_{ab}} = Z_{i.b}Z_{i.a}^T + (1 - \delta_{ab})Z_{i.a}Z_{i.b}^T = U_{ab}. \qquad \text{(D.25)}$$

Substituted in (D.21), this leads to

$$\begin{aligned}
\frac{\partial \ln |V_i|}{\partial d_{ab}} &= \operatorname{tr}(V_i^{-1}U_{ab}) \\
&= \operatorname{tr}(V_i^{-1}Z_{i.b}Z_{i.a}^T) + (1 - \delta_{ab})\operatorname{tr}(V_i^{-1}Z_{i.a}Z_{i.b}^T) \\
&\overset{(D.13)}{=} \operatorname{tr}(Z_{i.a}^T V_i^{-1}Z_{i.b}) + (1 - \delta_{ab})\operatorname{tr}(Z_{i.b}^T V_i^{-1}Z_{i.a}) \\
&\overset{(D.15)}{=} Z_{i.a}^T V_i^{-1}Z_{i.b} + (1 - \delta_{ab})Z_{i.a}^T V_i^{-1}Z_{i.b} \\
&= (2 - \delta_{ab})Z_{i.a}^T V_i^{-1}Z_{i.b}. \qquad \text{(D.26)}
\end{aligned}$$

The second term in (D.20) corresponds to

$$\frac{\partial(\boldsymbol{r}_i^T V_i^{-1}\boldsymbol{r}_i)}{\partial d_{ab}} = \boldsymbol{r}_i^T \frac{\partial V_i^{-1}}{\partial d_{ab}}\boldsymbol{r}_i \qquad \text{(D.27)}$$

Given (D.17) for the derivative of an inverse, it follows that

$$\frac{\partial V_i^{-1}}{\partial d_{ab}} = -V_i^{-1}U_{ab}V_i^{-1}, \qquad \text{(D.28)}$$

As a result, the first-order derivative of $\log f(\boldsymbol{y}_i, \boldsymbol{\xi})$ with respect to the elements of $\boldsymbol{d}$, corresponds to $(a \leq b)$

$$\frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial d_{ab}} = -\frac{1}{2}(2 - \delta_{ab}) Z_{i.a}^T V_i^{-1} Z_{i.b} + \frac{1}{2} \boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1} \boldsymbol{r}_i. \tag{D.29}$$

With respect to σ^2

Finally, the first-order derivative of $\log f(\boldsymbol{y}_i, \boldsymbol{\xi})$ with respect to σ^2 is given by

$$\frac{\partial \log f(\boldsymbol{y}, \boldsymbol{\xi})}{\partial \sigma^2} = -\frac{1}{2} \frac{\partial \ln |V_i|}{\partial \sigma^2} - \frac{1}{2} \frac{\partial}{\partial \sigma^2} (\boldsymbol{y}_i - X_i \boldsymbol{\beta})^T V_i^{-1} (\boldsymbol{y}_i - X_i \boldsymbol{\beta}). \tag{D.30}$$

The first term in this expression can be worked out as

$$\frac{\partial \ln |V_i|}{\partial \sigma^2} \overset{(D.12)}{=} \operatorname{tr}\left(V_i^{-1} \frac{\partial V_i}{\partial \sigma^2} \right) = \operatorname{tr}(V_i^{-1} I_{n_i}) = \operatorname{tr}(V_i^{-1}), \tag{D.31}$$

whereas the second term in (D.30) follows from

$$\begin{aligned} \frac{\partial (\boldsymbol{r}_i^T V_i^{-1} \boldsymbol{r}_i)}{\partial \sigma^2} &= \boldsymbol{r}_i^T \frac{\partial V_i^{-1}}{\partial \sigma^2} \boldsymbol{r}_i \\ &\overset{(D.17)}{=} -\boldsymbol{r}_i^T V_i^{-1} \frac{\partial V_i}{\partial \sigma^2} V_i^{-1} \boldsymbol{r}_i \\ &= -\boldsymbol{r}_i^T V_i^{-2} \boldsymbol{r}_i. \end{aligned} \tag{D.32}$$

Substituting (D.31) and (D.32) in (D.30), leads to

$$\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \sigma^2} = -\frac{1}{2} \operatorname{tr}(V_i^{-1}) + \frac{1}{2} \boldsymbol{r}_i^T V_i^{-2} \boldsymbol{r}_i. \tag{D.33}$$

D.3 Second-Order Derivatives

In this section, we will focus on the second-order derivative of the marginal loglikelihood $\log f(\boldsymbol{y}_i, \boldsymbol{\xi})$. Given that

$$\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^T \otimes \partial \boldsymbol{\xi}} = \begin{pmatrix} \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}^T \otimes \partial \boldsymbol{\beta}} & \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{d}^T \otimes \partial \boldsymbol{\beta}} & \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \sigma^2 \partial \boldsymbol{\beta}} \\ \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}^T \otimes \partial \boldsymbol{d}} & \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{d}^T \otimes \partial \boldsymbol{d}} & \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \sigma^2 \partial \boldsymbol{d}} \\ \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}^T \partial \sigma^2} & \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{d}^T \partial \sigma^2} & \dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \sigma^2 \partial \sigma^2} \end{pmatrix}, \tag{D.34}$$

it can be easily seen that the calculation of $d_k(\boldsymbol{y}_i, \boldsymbol{\xi})$ in (D.3) is based on the diagonal elements of this matrix. Further, the elements in the upper triangular part of (D.34) are required to determine the first term of (D.4).

With respect to $\boldsymbol{\beta}$

The second-order derivative of the marginal likelihood with respect to $\boldsymbol{\beta}^T$ and $\boldsymbol{\beta}$ is obtained by

$$\begin{aligned}
\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}^T \otimes \partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}^T}\left(\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}}\right) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}^T}\left[X_i^T V_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})\right] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}^T}\left(-X_i^T V_i^{-1} X_i \boldsymbol{\beta}\right) \\
&\overset{(D.8)}{=} -X_i^T V_i^{-1} X_i. \qquad \text{(D.35)}
\end{aligned}$$

With respect to $\boldsymbol{d}$ and $\boldsymbol{\beta}$

First note that

$$\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{d}^T \otimes \partial \boldsymbol{\beta}} = \left[\begin{array}{ccc} \frac{\partial}{\partial d_{11}}\left(\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}}\right) & \cdots & \frac{\partial}{\partial d_{qq}}\left(\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}}\right) \end{array}\right], \quad \text{(D.36)}$$

so that we can work out this derivative for each element d_{ab} separately. Therefore, for $1 \leq a \leq b \leq q$:

$$\begin{aligned}
\frac{\partial}{\partial d_{ab}}\left(\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}}\right) &= \frac{\partial}{\partial d_{ab}}\left(X_i^T V_i^{-1} \boldsymbol{r}_i\right) \\
&= X_i^T \frac{\partial V_i^{-1}}{\partial d_{ab}} \boldsymbol{r}_i \\
&\overset{(D.17)}{=} -X_i^T V_i^{-1} \frac{\partial V_i}{\partial d_{ab}} V_i^{-1} \boldsymbol{r}_i \\
&= -X_i^T V_i^{-1} U_{ab} V_i^{-1} \boldsymbol{r}_i \qquad \text{(D.37)}
\end{aligned}$$

With respect to σ^2 and $\boldsymbol{\beta}$

The second-order derivative of the marginal likelihood with respect to σ^2 and $\boldsymbol{\beta}$ is

given by

$$\begin{aligned}\frac{\partial^2 \log f(\boldsymbol{y},\boldsymbol{\xi})}{\partial\sigma^2\partial\boldsymbol{\beta}} &= \frac{\partial}{\partial\sigma^2}\left(X_i^T V_i^{-1}\boldsymbol{r}_i\right)\\ &= X_i^T\frac{\partial V_i^{-1}}{\partial\sigma^2}\boldsymbol{r}_i\\ &\overset{(D.17)}{=} -X_i^T V_i^{-2}\boldsymbol{r}_i \end{aligned} \tag{D.38}$$

With respect to $\boldsymbol{d}$

Recall that the matrix containing the second-order derivatives of the marginal likelihood with respect to the elements of $\boldsymbol{d}$ can be written as

$$\frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial\boldsymbol{d}^T\otimes\partial\boldsymbol{d}} = \begin{pmatrix} \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11}\partial d_{11}} & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{12}\partial d_{11}} & \cdots & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq}\partial d_{11}} \\ \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11}\partial d_{12}} & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{12}\partial d_{12}} & \cdots & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq}\partial d_{12}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11}\partial d_{qq}} & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{12}\partial d_{qq}} & \cdots & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq}\partial d_{qq}} \end{pmatrix}. \tag{D.39}$$

The elements of this matrix are obtained, for each combination of $1 \le c \le f \le q$ and $1 \le a \le b \le q$, by determining

$$\frac{\partial}{\partial d_{cf}}\left(\frac{\partial \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{ab}}\right) = \frac{\partial}{\partial d_{cf}}\left[-\frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-1} Z_{i.b} + \frac{1}{2}\boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1}\boldsymbol{r}_i\right]. \tag{D.40}$$

The first term in (D.40) can be written as

$$\begin{aligned}\frac{\partial}{\partial d_{cf}}\left[-\frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-1} Z_{i.b}\right] &= -\frac{1}{2}(2-\delta_{ab})Z_{i.a}^T\frac{\partial V_i^{-1}}{\partial d_{cf}}Z_{i.b}\\ &\overset{(D.17)}{=} \frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-1}\frac{\partial V_i}{\partial d_{cf}}V_i^{-1}Z_{i.b}\\ &= \frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-1} U_{cf} V_i^{-1} Z_{i.b}, \end{aligned} \tag{D.41}$$

while the second term in (D.40) follows as

$$\begin{aligned}\frac{\partial}{\partial d_{cf}}\left[\frac{1}{2}\boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1}\boldsymbol{r}_i\right] &= \frac{1}{2}\boldsymbol{r}_i^T\frac{\partial V_i^{-1}}{\partial d_{cf}}U_{ab}V_i^{-1}\boldsymbol{r}_i + \frac{1}{2}\boldsymbol{r}_i^T V_i^{-1}U_{ab}\frac{\partial V_i^{-1}}{\partial d_{cf}}\boldsymbol{r}_i\\ &\overset{(D.17)}{=} -\frac{1}{2}\boldsymbol{r}_i^T V_i^{-1}(U_{cf}V_i^{-1}U_{ab} + U_{ab}V_i^{-1}U_{cf})V_i^{-1}\boldsymbol{r}_i. \end{aligned} \tag{D.42}$$

Substituted in (D.40) this leads to

$$\begin{aligned}\frac{\partial}{\partial d_{cf}}\left(\frac{\partial \log f(\boldsymbol{y},\boldsymbol{\xi})}{\partial d_{ab}}\right) &= \frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-1}U_{cf}V_i^{-1}Z_{i.b} \\ &\quad -\frac{1}{2}\boldsymbol{r}_i^T V_i^{-1}(U_{cf}V_i^{-1}U_{ab}+U_{ab}V_i^{-1}U_{cf})V_i^{-1}\boldsymbol{r}_i. \quad \text{(D.43)}\end{aligned}$$

With respect to σ^2 and $\boldsymbol{d}$

Observe that

$$\frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial\sigma^2 \otimes \partial \boldsymbol{d}} = \left[\begin{array}{ccc} \frac{\partial}{\partial\sigma^2}\left(\frac{\partial \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11}}\right) & \cdots & \frac{\partial}{\partial\sigma^2}\left(\frac{\partial \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq}}\right)\end{array}\right]^T, \quad \text{(D.44)}$$

Therefore, the second-order derivative of the marginal loglikelihood with respect to σ^2 and the elements d_{ab} of $\boldsymbol{d}$ $(1 \le a \le b \le q)$ is given by

$$\frac{\partial}{\partial\sigma^2}\left(\frac{\partial \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{ab}}\right) = \frac{\partial}{\partial\sigma^2}\left[-\frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-1}Z_{i.b} + \frac{1}{2}\boldsymbol{r}_i^T V_i^{-1}U_{ab}V_i^{-1}\boldsymbol{r}_i\right]. \quad \text{(D.45)}$$

This expression can be worked out as a combination of

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\left[-\frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-1}Z_{i.b}\right] &= -\frac{1}{2}(2-\delta_{ab})Z_{i.a}^T\frac{\partial V_i^{-1}}{\partial\sigma^2}Z_{i.b} \\ &= \frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-2}Z_{i.b}, \quad \text{(D.46)}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\left[\frac{1}{2}\boldsymbol{r}_i^T V_i^{-1}U_{ab}V_i^{-1}\boldsymbol{r}_i\right] &= \frac{1}{2}\boldsymbol{r}_i^T\frac{\partial V_i^{-1}}{\partial\sigma^2}U_{ab}V_i^{-1}\boldsymbol{r}_i + \frac{1}{2}\boldsymbol{r}_i^T V_i^{-1}U_{ab}\frac{\partial V_i^{-1}}{\partial\sigma^2}\boldsymbol{r}_i \\ &= -\frac{1}{2}\boldsymbol{r}_i^T V_i^{-1}(V_i^{-1}U_{ab}+U_{ab}V_i^{-1})V_i^{-1}\boldsymbol{r}_i. \quad \text{(D.47)}\end{aligned}$$

Therefore, (D.45) can be written as

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\left(\frac{\partial \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{ab}}\right) &= \frac{1}{2}(2-\delta_{ab})Z_{i.a}^T V_i^{-2}Z_{i.b} \\ &\quad -\frac{1}{2}\boldsymbol{r}_i^T V_i^{-1}(V_i^{-1}U_{ab}+U_{ab}V_i^{-1})V_i^{-1}\boldsymbol{r}_i. \quad \text{(D.48)}\end{aligned}$$

With respect to σ^2

Finally, the second-order derivative of the marginal likelihood with respect to σ^2 is given by

$$\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \sigma^2 \partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2}\mathrm{tr}(V_i^{-1}) + \frac{1}{2}\boldsymbol{r}_i^T V_i^{-2}\boldsymbol{r}_i\right] \tag{D.49}$$

The first term in (D.49) corresponds to

$$\frac{\partial}{\partial \sigma^2}\left[-\frac{1}{2}\mathrm{tr}(V_i^{-1})\right] \overset{(D.16)}{=} -\frac{1}{2}\mathrm{tr}\left(\frac{\partial V_i^{-1}}{\partial \sigma^2}\right) = \frac{1}{2}\mathrm{tr}(V_i^{-2}). \tag{D.50}$$

Next, it can be easily seen that the second term in (D.49) can be written as

$$\frac{\partial}{\partial \sigma^2}\left(\frac{1}{2}\boldsymbol{r}_i^T V_i^{-2}\boldsymbol{r}_i\right) = -\boldsymbol{r}_i^T V_i^{-3}\boldsymbol{r}_i. \tag{D.51}$$

Therefore,

$$\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \sigma^2 \partial \sigma^2} = \frac{1}{2}\mathrm{tr}(V_i^{-2}) - \boldsymbol{r}_i^T V_i^{-3}\boldsymbol{r}_i. \tag{D.52}$$

D.4 Third-Order Derivatives

The matrix of third-order derivatives $\dfrac{\partial}{\partial \boldsymbol{\xi}^T}\left[\dfrac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \xi_k^2}\right]_k$ required in the second term in (D.4) can be written as

$$\begin{pmatrix}
\frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_1 \partial \beta_1^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_1 \partial \beta_p^2} & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_1 \partial d_{11}^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_1 \partial d_{qq}^2} & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_1 (\partial \sigma^2)^2} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_p \partial \beta_1^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_p \partial \beta_p^2} & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_p \partial d_{11}^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_p \partial d_{qq}^2} & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \beta_p (\partial \sigma^2)^2} \\
\frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11} \partial \beta_1^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11} \partial \beta_p^2} & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11} \partial d_{11}^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11} \partial d_{qq}^2} & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{11} (\partial \sigma^2)^2} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq} \partial \beta_1^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq} \partial \beta_p^2} & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq} \partial d_{11}^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq} \partial d_{qq}^2} & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{qq} (\partial \sigma^2)^2} \\
\frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \sigma^2 \partial \beta_1^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \sigma^2 \partial \beta_p^2} & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \sigma^2 \partial d_{11}^2} & \cdots & \frac{\partial^3 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \sigma^2 \partial d_{qq}^2} & \frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial \sigma^2 (\partial \sigma^2)^2}
\end{pmatrix}.$$

In what follows, we will work out the expressions for the derivatives in this matrix.

With respect to $\boldsymbol{\beta}$

First observe that, for $t = 1, \ldots, p$,

$$\frac{\partial}{\partial \beta_t}\left(\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}^T \otimes \partial \boldsymbol{\beta}}\right) = \frac{\partial}{\partial \beta_t}(-X_i^T V_i^{-1} X_i) = 0.$$

Therefore, it can be easily seen that, for $k = 1, \ldots, p$,

$$\frac{\partial^3 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \beta_t \partial \beta_k^2} = 0.$$

Further, for $a \leq b$,

$$\frac{\partial^3 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial \beta_t \partial d_{ab}^2} = \frac{\partial}{\partial \beta_t}\left[\frac{1}{2}(2 - \delta_{ab}) Z_{i.a}^T V_i^{-1} U_{ab} V_i^{-1} Z_{i.b} - \boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1} U_{ab} V_i^{-1} \boldsymbol{r}_i\right]. \tag{D.53}$$

It can be easily seen that the derivative of the first term in (D.53) is equal to zero. The second term can be written as

$$\begin{aligned} &\frac{\partial}{\partial \beta_t}\left[-\boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1} U_{ab} V_i^{-1} \boldsymbol{r}_i\right] \\ &\quad = -\frac{\partial \boldsymbol{r}_i^T}{\partial \beta_t} V_i^{-1} U_{ab} V_i^{-1} U_{ab} V_i^{-1} \boldsymbol{r}_i - \boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1} U_{ab} V_i^{-1} \frac{\partial \boldsymbol{r}_i}{\partial \beta_t}. \end{aligned} \tag{D.54}$$

Note that

$$\begin{aligned} \frac{\partial \boldsymbol{r}_i^T}{\partial \beta_t} &= \frac{\partial (\boldsymbol{y}_i - X_i \boldsymbol{\beta})^T}{\partial \beta_t} \\ &= -\frac{\partial (\boldsymbol{\beta}^T X_i^T)}{\partial \beta_t} \\ &= -(\ X_{i1t} \quad \cdots \quad X_{in_i t}\ ) = -X_{i.t}^T, \end{aligned}$$

and similarly

$$\frac{\partial (\boldsymbol{y}_i - X_i \boldsymbol{\beta})}{\partial \beta_t} = -X_{i.t}.$$

Therefore,

$$\frac{\partial}{\partial \beta_t}\left(\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{\partial d_{ab}^2}\right) = X_{i.t}^T V_i^{-1} U_{ab} V_i^{-1} U_{ab} V_i^{-1} \boldsymbol{r}_i + \boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1} U_{ab} V_i^{-1} X_{i.t}. \tag{D.55}$$

Finally,

$$\frac{\partial}{\partial \beta_t}\left(\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{(\partial \sigma^2)^2}\right) = \frac{\partial}{\partial \beta_t}\left[\frac{1}{2}\mathrm{tr}(V_i^{-2}) - \boldsymbol{r}_i^T V_i^{-3} \boldsymbol{r}_i\right]. \tag{D.56}$$

The first term in this expression does not depend on β_t and therefore vanishes. Therefore, (D.56) can be written as

$$\begin{aligned}\frac{\partial}{\partial\beta_t}\left(\frac{\partial^2\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{(\partial\sigma^2)^2}\right) &= -\frac{\partial\boldsymbol{r}_i^T}{\partial\beta_t}V_i^{-3}\boldsymbol{r}_i-\boldsymbol{r}_i^TV_i^{-3}\frac{\partial\boldsymbol{r}_i}{\partial\beta_t}\\ &= X_{i.t}^TV_i^{-3}\boldsymbol{r}_i+\boldsymbol{r}_i^TV_i^{-3}X_{i.t}\end{aligned}\tag{D.57}$$

With respect to $\boldsymbol{d}$

Note that, for $1\le c\le f\le q$,

$$\frac{\partial}{\partial d_{cf}}\left[\frac{\partial^2\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial\beta_k^2}\right]=\operatorname{diag}\left\{\frac{\partial}{\partial d_{cf}}\left[\frac{\partial^2\log f(\boldsymbol{y},\boldsymbol{\xi})}{\partial\boldsymbol{\beta}^T\otimes\partial\boldsymbol{\beta}}\right]\right\}.\tag{D.58}$$

Since

$$\begin{aligned}\frac{\partial}{\partial d_{cf}}\left[\frac{\partial^2\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial\boldsymbol{\beta}^T\otimes\partial\boldsymbol{\beta}}\right] &= \frac{\partial}{\partial d_{cf}}(-X_i^TV_i^{-1}X_i)\\ &= -X_i^T\frac{\partial V_i^{-1}}{\partial d_{cf}}X_i\\ &= X_i^TV_i^{-1}U_{cf}V_i^{-1}X_i,\end{aligned}\tag{D.59}$$

if follows that

$$\frac{\partial}{\partial d_{cf}}\left[\frac{\partial^2\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial\beta_k^2}\right]=\operatorname{diag}\left(X_i^TV_i^{-1}U_{cf}V_i^{-1}X_i\right).\tag{D.60}$$

Next,

$$\frac{\partial^3\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial d_{cf}\partial d_{ab}^2}=\frac{\partial}{\partial d_{cf}}\left[\frac{1}{2}(2-\delta_{ab})Z_{i.a}^TV_i^{-1}U_{ab}V_i^{-1}Z_{i.b}-\boldsymbol{r}_i^TV_i^{-1}U_{ab}V_i^{-1}U_{ab}V_i^{-1}\boldsymbol{r}_i\right].\tag{D.61}$$

The first term in (D.61) can be worked out as

$$\begin{aligned}&\frac{\partial}{\partial d_{cf}}\left[\frac{1}{2}(2-\delta_{ab})Z_{i.a}^TV_i^{-1}U_{ab}V_i^{-1}Z_{i.b}\right]\\ &\quad=\frac{1}{2}(2-\delta_{ab})Z_{i.a}^T\frac{\partial V_i^{-1}}{\partial d_{cf}}U_{ab}V_i^{-1}Z_{i.b}+\frac{1}{2}(2-\delta_{ab})Z_{i.a}^TV_i^{-1}U_{ab}\frac{\partial V_i^{-1}}{\partial d_{cf}}Z_{i.b}\\ &\quad=-\frac{1}{2}(2-\delta_{ab})Z_{i.a}^TV_i^{-1}(U_{cf}V^{-1}U_{ab}+U_{ab}V^{-1}U_{cf})V_i^{-1}Z_{i.b},\end{aligned}\tag{D.62}$$

while the second term in (D.61) follows from

$$\begin{aligned}
&\frac{\partial}{\partial d_{cf}}\left[-\boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1} U_{ab} V_i^{-1} \boldsymbol{r}_i\right] \\
&\quad = -\boldsymbol{r}_i^T \frac{\partial V_i^{-1}}{\partial d_{cf}} U_{ab} V_i^{-1} U_{ab} V_i^{-1} \boldsymbol{r}_i - \boldsymbol{r}_i^T V_i^{-1} U_{ab} \frac{\partial V_i^{-1}}{\partial d_{cf}} U_{ab} V_i^{-1} \boldsymbol{r}_i \\
&\qquad -\boldsymbol{r}_i^T V_i^{-1} U_{ab} V_i^{-1} U_{ab} \frac{\partial V_i^{-1}}{\partial d_{cf}} \boldsymbol{r}_i \\
&\quad = \boldsymbol{r}_i^T V_i^{-1} (U_{cf} V^{-1} U_{ab} V^{-1} U_{ab} + U_{ab} V^{-1} U_{cf} V^{-1} U_{ab} \\
&\qquad + U_{ab} V^{-1} U_{ab} V^{-1} U_{cf}) V_i^{-1} \boldsymbol{r}_i. \qquad \text{(D.63)}
\end{aligned}$$

Finally,

$$\frac{\partial}{\partial d_{cf}}\left[\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{\xi})}{(\partial \sigma^2)^2}\right] = \frac{\partial}{\partial d_{cf}}\left[\frac{1}{2}\mathrm{tr}(V_i^{-2}) - \boldsymbol{r}_i^T V_i^{-3} \boldsymbol{r}_i\right]. \qquad \text{(D.64)}$$

First,

$$\begin{aligned}
\frac{\partial}{\partial d_{cf}}\left[\frac{1}{2}\mathrm{tr}(V_i^{-2})\right] &\overset{(D.16)}{=} \frac{1}{2}\mathrm{tr}\left[\frac{\partial V_i^{-2}}{\partial d_{cf}}\right] \\
&= \frac{1}{2}\mathrm{tr}\left(V_i^{-1}\frac{\partial V_i^{-1}}{\partial d_{cf}}\right) + \frac{1}{2}\mathrm{tr}\left(\frac{\partial V_i^{-1}}{\partial d_{cf}} V_i^{-1}\right) \\
&= -\frac{1}{2}\mathrm{tr}\left(V_i^{-1} V_i^{-1} U_{cf} V_i^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(V_i^{-1} U_{cf} V_i^{-1} V_i^{-1}\right) \\
&\overset{(D.13)}{=} -\mathrm{tr}\left(V_i^{-1} U_{cf} V_i^{-1} V_i^{-1}\right) \\
&= -\mathrm{tr}\left\{V_i^{-1}[Z_{i.c} Z_{i.f}^T + (1-\delta_{cf}) Z_{i.f} Z_{i.c}^T] V_i^{-1} V_i^{-1}\right\} \\
&\overset{(D.13,D.15)}{=} -\mathrm{tr}(Z_{i.f}^T V_i^{-3} Z_{i.c}) - (1-\delta_{cf})\mathrm{tr}(Z_{i.f}^T V_i^{-3} Z_{i.c}) \\
&= -(2-\delta_{cf}) Z_{i.f}^T V_i^{-3} Z_{i.c}. \qquad \text{(D.65)}
\end{aligned}$$

Second,

$$\begin{aligned}
\frac{\partial}{\partial d_{cf}}\left[-\boldsymbol{r}_i^T V_i^{-3} \boldsymbol{r}_i\right] &= -\boldsymbol{r}_i^T \frac{\partial V_i^{-1}}{\partial d_{cf}} V_i^{-2} \boldsymbol{r}_i - \boldsymbol{r}_i^T V_i^{-1} \frac{\partial V_i^{-1}}{\partial d_{cf}} V_i^{-1} \boldsymbol{r}_i - \boldsymbol{r}_i^T V_i^{-2} \frac{\partial V_i^{-1}}{\partial d_{cf}} \boldsymbol{r}_i \\
&= \boldsymbol{r}_i^T V_i^{-1} (U_{cf} V_i^{-2} + V_i^{-1} U_{cf} V_i^{-1} + V_i^{-2} U_{cf}) V_i^{-1} \boldsymbol{r}_i. \qquad \text{(D.66)}
\end{aligned}$$

With respect to σ^2

As before we can work out the derivative of the marginal loglikelihood with respect

to σ^2 and β_k as

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\left[\frac{\partial^2\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial\beta_k^2}\right] &= \operatorname{diag}\left\{\frac{\partial}{\partial\sigma^2}\left[\frac{\partial^2\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial\boldsymbol{\beta}^T\otimes\partial\boldsymbol{\beta}}\right]\right\}\\ &= \operatorname{diag}\left[\frac{\partial(-X_i^TV_i^{-1}X_i)}{\partial\sigma^2}\right]\\ &= \operatorname{diag}\left[-X_i^T\frac{\partial V_i^{-1}}{\partial\sigma^2}X_i\right]\\ &= \operatorname{diag}\left[X_i^TV_i^{-2}X_i\right]. \end{aligned} \tag{D.67}$$

Next, it can be easily seen that

$$\frac{\partial^3\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{\partial\sigma^2\partial\beta_k^2} = \frac{\partial}{\partial\sigma^2}\left[\frac{1}{2}(2-\delta_{ab})Z_{i.a}^TV_i^{-1}U_{ab}V_i^{-1}Z_{i.b} - \boldsymbol{r}_i^TV_i^{-1}U_{ab}V_i^{-1}U_{ab}V_i^{-1}\boldsymbol{r}_i\right] \tag{D.68}$$

The first part of this expression corresponds to

$$\begin{aligned}&\frac{\partial}{\partial\sigma^2}\left[\frac{1}{2}(2-\delta_{ab})Z_{i.a}^TV_i^{-1}U_{ab}V_i^{-1}Z_{i.b}\right]\\ &\quad= \frac{1}{2}(2-\delta_{ab})Z_{i.a}^T\frac{\partial V_i^{-1}}{\partial\sigma^2}U_{ab}V_i^{-1}Z_{i.b} + \frac{1}{2}(2-\delta_{ab})Z_{i.a}^TV_i^{-1}U_{ab}\frac{\partial V_i^{-1}}{\partial\sigma^2}Z_{i.b}\\ &\quad= -\frac{1}{2}(2-\delta_{ab})Z_{i.a}^TV_i^{-1}(V_i^{-1}U_{ab}+U_{ab}V_i^{-1}),\end{aligned} \tag{D.69}$$

whereas the second part can be written as

$$\begin{aligned}&\frac{\partial}{\partial\sigma^2}\left[-\boldsymbol{r}_i^TV_i^{-1}U_{ab}V_i^{-1}U_{ab}V_i^{-1}\boldsymbol{r}_i\right]\\ &\quad= -\boldsymbol{r}_i^T\frac{\partial V_i^{-1}}{\partial\sigma^2}U_{ab}V_i^{-1}U_{ab}V_i^{-1}\boldsymbol{r}_i - \boldsymbol{r}_i^TV_i^{-1}U_{ab}\frac{\partial V_i^{-1}}{\partial\sigma^2}U_{ab}V_i^{-1}\boldsymbol{r}_i\\ &\qquad -\boldsymbol{r}_i^TV_i^{-1}U_{ab}V_i^{-1}U_{ab}\frac{\partial V_i^{-1}}{\partial\sigma^2}\boldsymbol{r}_i\\ &\quad= \boldsymbol{r}_i^TV_i^{-1}(V^{-1}U_{ab}V^{-1}U_{ab}+U_{ab}V^{-2}U_{ab}+U_{ab}V^{-1}U_{ab}V^{-1})V_i^{-1}\boldsymbol{r}_i.\end{aligned} \tag{D.70}$$

Finally, the last derivative to be considered here corresponds to

$$\frac{\partial}{\partial\sigma^2}\left[\frac{\partial^2\log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{(\partial\sigma^2)^2}\right] = \frac{\partial}{\partial\sigma^2}\left[\frac{1}{2}\operatorname{tr}(V_i^{-2}) - \boldsymbol{r}_i^TV_i^{-3}\boldsymbol{r}_i\right]. \tag{D.71}$$

The first term in this expression follows from

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\left[\frac{1}{2}\operatorname{tr}(V_i^{-2})\right] &= \frac{1}{2}\operatorname{tr}\left(V_i^{-1}\frac{\partial V_i^{-1}}{\partial\sigma^2} + \frac{\partial V_i^{-1}}{\partial\sigma^2}V_i^{-1}\right)\\ &= -\operatorname{tr}(V_i^{-3})\end{aligned} \tag{D.72}$$

and the second term from

$$\frac{\partial}{\partial\sigma^2}\left[-\boldsymbol{r}_i^T V_i^{-3}\boldsymbol{r}_i\right] \quad = \quad 3\boldsymbol{r}_i^T V_i^{-4}\boldsymbol{r}_i. \tag{D.73}$$

Therefore,

$$\frac{\partial}{\partial\sigma^2}\left[\frac{\partial^2 \log f(\boldsymbol{y}_i,\boldsymbol{\xi})}{(\partial\sigma^2)^2}\right] = -\mathrm{tr}(V_i^{-3}) + 3\boldsymbol{r}_i^T V_i^{-4}\boldsymbol{r}_i. \tag{D.74}$$

Appendix E

Alternative Information Matrix Tests

This appendix can be considered as a supplement to Chapter 10, containing the power of the diagnostic tools to detect misspecification of the random-effects structure, for samples of $n = 100$ and 350 subjects.

Table E.1: *Power of the determinant tests $\delta_{d1}(n)$ and $\delta_{d2}(n)$, the determinant-trace test $\delta_{dt}(n)$, the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ to detect model misspecification, when a logistic-normal model is assumed, but the variance of the random intercept depends on a binary cluster-level covariate, $[b_{i0}|z_i = 0] \sim N(0, \sigma_0^2)$ and $[b_{i0}|z_i = 1] \sim N(0, \sigma_1^2)$. (sample size $n = 100$ and 350).*

σ_1	σ_0	n	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$	$\delta_s(n)$	$\Im_m(n)$
0.5	0.5	100	0.140	0.037	0.066	0.156	0.298
		350	0.080	0.064	0.094	0.042	0.076
	1.0	100	0.111	0.026	0.046	0.143	0.256
		350	0.076	0.066	0.096	0.080	0.402
	2.0	100	0.144	0.038	0.048	0.420	0.842
		350	0.150	0.070	0.046	0.874	1.000
	3.0	100	0.314	0.076	0.038	0.770	0.988
		350	0.500	0.334	0.186	1.000	1.000
1.0	0.5	100	0.138	0.028	0.088	0.134	0.380
		350	0.146	0.076	0.182	0.068	0.518
	1.0	100	0.106	0.042	0.068	0.156	0.190
		350	0.062	0.034	0.062	0.030	0.044
	2.0	100	0.114	0.026	0.028	0.206	0.382
		350	0.088	0.034	0.038	0.424	0.912
	3.0	100	0.166	0.048	0.034	0.512	0.896
		350	0.246	0.126	0.066	0.980	1.000
2.0	0.5	100	0.182	0.026	0.168	0.166	0.872
		350	0.286	0.176	0.486	0.554	1.000
	1.0	100	0.136	0.024	0.076	0.142	0.612
		350	0.132	0.058	0.182	0.286	0.926
	2.0	100	0.078	0.026	0.026	0.092	0.086
		350	0.070	0.044	0.050	0.020	0.026
	3.0	100	0.090	0.038	0.032	0.124	0.186
		350	0.088	0.048	0.032	0.150	0.412
3.0	0.5	100	0.352	0.068	0.314	0.422	0.998
		350	0.650	0.496	0.844	0.970	1.000
	1.0	100	0.220	0.052	0.134	0.302	0.924
		350	0.322	0.204	0.458	0.852	1.000
	2.0	100	0.148	0.042	0.042	0.126	0.236
		350	0.070	0.032	0.044	0.122	0.500
	3.0	100	0.106	0.044	0.032	0.090	0.082
		350	0.056	0.046	0.040	0.030	0.034

Table E.2: *Power of the determinant tests $\delta_{d1}(n)$ and $\delta_{d2}(n)$, the determinant-trace test $\delta_{dt}(n)$, the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ to detect model misspecification, when a logistic-normal model is assumed, but the data are generated using both a random intercept and slope ($b_{ij} = b_{i0} + b_{i1}t_j$), with variance σ_0^2 and σ_1^2, respectively. (sample size $n = 100$ and 350).*

σ_1	σ_0	n	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$	$\delta_s(n)$	$\Im_m(n)$
0.2.0	0.5	100	0.101	0.029	0.059	0.134	0.238
	0.5	350	0.098	0.066	0.114	0.040	0.086
	1.0	100	0.084	0.028	0.056	0.132	0.178
	1.0	350	0.072	0.050	0.078	0.026	0.038
	2.0	100	0.094	0.050	0.052	0.084	0.096
	2.0	350	0.076	0.056	0.062	0.040	0.048
	3.0	100	0.116	0.058	0.034	0.082	0.092
	3.0	350	0.068	0.058	0.038	0.022	0.026
0.5	0.5	100	0.096	0.034	0.068	0.130	0.256
	0.5	350	0.066	0.054	0.096	0.050	0.114
	1.0	100	0.078	0.050	0.068	0.104	0.132
	1.0	350	0.038	0.042	0.062	0.016	0.044
	2.0	100	0.060	0.054	0.056	0.080	0.100
	2.0	350	0.046	0.062	0.050	0.014	0.020
	3.0	100	0.078	0.058	0.052	0.050	0.076
	3.0	350	0.068	0.084	0.070	0.010	0.028
0.8	0.5	100	0.059	0.039	0.061	0.137	0.214
	0.5	350	0.068	0.092	0.112	0.136	0.190
	1.0	100	0.078	0.050	0.080	0.118	0.146
	1.0	350	0.058	0.114	0.098	0.054	0.128
	2.0	100	0.068	0.092	0.094	0.046	0.056
	2.0	350	0.090	0.134	0.134	0.022	0.046
	3.0	100	0.070	0.096	0.076	0.046	0.078
	3.0	350	0.124	0.184	0.144	0.020	0.036
1.0	0.5	100	0.050	0.048	0.070	0.194	0.226
	0.5	350	0.1.0	0.152	0.146	0.412	0.510
	1.0	100	0.046	0.074	0.092	0.114	0.150
	1.0	350	0.144	0.234	0.218	0.130	0.238
	2.0	100	0.074	0.108	0.104	0.050	0.088
	2.0	350	0.194	0.254	0.228	0.018	0.086
	3.0	100	0.084	0.138	0.124	0.038	0.040
	3.0	350	0.242	0.348	0.278	0.024	0.064

Table E.3: *Power of the determinant tests $\delta_{d1}(n)$ and $\delta_{d2}(n)$, the determinant-trace test $\delta_{dt}(n)$, the SET $\delta_s(n)$ and the MIMT $\Im_m(n)$ to detect model misspecification, when a logistic-normal model is assumed, but the data are generated using autocorrelated random effects b_{ij} such that $cov(b_{ij}, b_{ik}) = \sigma^2\rho^{|t_{ij}-t_{ik}|}$. (sample size $n = 100$ and 350).*

ρ	σ	n	$\delta_{d1}(n)$	$\delta_{d2}(n)$	$\delta_{dt}(n)$	$\delta_s(n)$	$\Im_m(n)$
0.5	0.5	100	0.114	0.051	0.085	0.147	0.217
	0.5	350	0.077	0.063	0.080	0.029	0.049
	1.0	100	0.064	0.064	0.092	0.073	0.156
	1.0	350	0.105	0.158	0.187	0.049	0.105
	2.0	100	0.129	0.237	0.275	0.096	0.163
	2.0	350	0.658	0.758	0.794	0.446	0.182
	3.0	100	0.292	0.452	0.482	0.142	0.148
	3.0	350	0.946	0.970	0.984	0.884	0.410
0.7	0.5	100	0.122	0.042	0.071	0.134	0.211
	0.5	350	0.093	0.095	0.130	0.023	0.077
	1.0	100	0.051	0.076	0.090	0.100	0.182
	1.0	350	0.158	0.230	0.272	0.052	0.080
	2.0	100	0.222	0.367	0.397	0.096	0.132
	2.0	350	0.868	0.916	0.922	0.674	0.382
	3.0	100	0.555	0.727	0.741	0.236	0.210
	3.0	350	0.998	0.998	0.998	0.968	0.840
0.9	0.5	100	0.122	0.047	0.087	0.139	0.172
	0.5	350	0.080	0.072	0.094	0.033	0.074
	1.0	100	0.056	0.056	0.078	0.086	0.134
	1.0	350	0.070	0.094	0.126	0.018	0.050
	2.0	100	0.154	0.290	0.292	0.052	0.108
	2.0	350	0.602	0.688	0.696	0.170	0.196
	3.0	100	0.420	0.592	0.564	0.082	0.152
	3.0	350	0.982	0.992	0.986	0.714	0.780

Appendix F

The Bayesian Central Limit Theorem

In this appendix we provide a heuristic overview of the large sample approximation of the posterior random-effects distribution. The following proceedings are based on the asymptotic inference results discussed in Bernardo and Smith (1994, Section 5.3).

In what follows, we will leave out the subject index i to simplify the notation. First, observe that the posterior distribution of the random effects can be written as

$$\begin{aligned} f(\boldsymbol{b}|\boldsymbol{y}) &\propto f(\boldsymbol{b}) \prod_{j=1}^{m} f(y_j|\boldsymbol{b}) \\ &\propto \exp[\log f(\boldsymbol{b}) + \log f(\boldsymbol{y}|\boldsymbol{b})], \end{aligned} \tag{F.1}$$

where m refers to the number of repeated observations. The logarithmic terms in (F.1) can be expanded around their maxima $\boldsymbol{b}_0$ and $\widehat{\boldsymbol{b}}_m$, determined by setting $\nabla \log f(\boldsymbol{b}) = \boldsymbol{0}$ and $\nabla \log f(\boldsymbol{y}|\boldsymbol{b}) = \boldsymbol{0}$, respectively. This leads to

$$\begin{aligned} \log f(\boldsymbol{b}) &= \log f(\boldsymbol{b}_0) - \frac{1}{2}(\boldsymbol{b}-\boldsymbol{b}_0)^T H_0 (\boldsymbol{b}-\boldsymbol{b}_0) + R_0 \\ \log f(\boldsymbol{y}|\boldsymbol{b}) &= \log f(\boldsymbol{y}|\widehat{\boldsymbol{b}}_m) - \frac{1}{2}(\boldsymbol{b}-\widehat{\boldsymbol{b}}_m)^T H(\widehat{\boldsymbol{b}}_m)(\boldsymbol{b}-\widehat{\boldsymbol{b}}_m) + R_m. \end{aligned}$$

In these expressions, R_0 and R_m denote remainder terms, and the Hessian matrices

H_0 and $H(\widehat{\boldsymbol{b}}_m)$ are defined as

$$H_0 = -\frac{\partial^2 \log f(\boldsymbol{b})}{\partial b_k \partial b_\ell}\bigg|_{\boldsymbol{b}=\boldsymbol{b}_0} \qquad H(\widehat{\boldsymbol{b}}_m) = -\frac{\partial^2 \log f(\boldsymbol{y}|\boldsymbol{b})}{\partial b_k \partial b_\ell}\bigg|_{\boldsymbol{b}=\widehat{\boldsymbol{b}}_m}.$$

Assuming that R_0 and R_m are sufficiently small for large m, and ignoring constants of proportionality, the posterior random-effects distribution can therefore be approximated by

$$\begin{aligned} f(\boldsymbol{b}|\boldsymbol{y}) &\propto \exp[-\frac{1}{2}(\boldsymbol{b}-\boldsymbol{b}_0)^T H_0(\boldsymbol{b}-\boldsymbol{b}_0) - \frac{1}{2}(\boldsymbol{b}-\widehat{\boldsymbol{b}}_m)^T H(\widehat{\boldsymbol{b}}_m)(\boldsymbol{b}-\widehat{\boldsymbol{b}}_m)] \\ &\propto \exp[-\frac{1}{2}(\boldsymbol{b}-\boldsymbol{b}_m)^T H_m(\boldsymbol{b}-\boldsymbol{b}_m)], \end{aligned}$$

where

$$\begin{aligned} H_m &= H_0 + H(\widehat{\boldsymbol{b}}_m) \\ \boldsymbol{b}_m &= H_m^{-1}[H_0\boldsymbol{b}_0 + H(\widehat{\boldsymbol{b}}_m)\widehat{\boldsymbol{b}}_m]. \end{aligned}$$

This result suggests that for large m, $f(\boldsymbol{b}|\boldsymbol{y})$ will resemble a multivariate normal distribution with mean $\boldsymbol{b}_m$ and covariance matrix H_m^{-1}. Further, observe that for large m, $H(\widehat{\boldsymbol{b}}_m)$ will become the dominant term in H_m. Consequently, for m sufficiently large, $f(\boldsymbol{b}|\boldsymbol{y})$ will also be well approximated by $\mathrm{N}_q(\boldsymbol{b}|\widehat{\boldsymbol{b}}_m, H^{-1}(\widehat{\boldsymbol{b}}_m))$, where q refers to the dimension of $\boldsymbol{b}$.

There is a large literature available on the regularity conditions required to justify mathematically this important result. In what follows, we will base our account on Bernardo and Smith (1994) and Chen (1985). Let us assume that $\boldsymbol{b} \in \Theta \subseteq \Re^q$ and that $\{f_m(\boldsymbol{b}), m = 1, 2, \ldots\}$ contains a sequence of posterior densities for $\boldsymbol{b}$, typically of the form $f_m(\boldsymbol{b}) = f(\boldsymbol{b}|y_1, \ldots, y_m)$, based on the parametric model $f(\boldsymbol{y}|\boldsymbol{b})$ and the prior $f(\boldsymbol{b})$.

Further, define $L_m(\boldsymbol{b}) = \log f_m(\boldsymbol{b})$, and assume that, for every m, there is a strict local maximum $\boldsymbol{b}_m$ of $f_m(\boldsymbol{b})$, satisfying

$$L'_m(\boldsymbol{b}_m) = \nabla L_m(\boldsymbol{b})|_{\boldsymbol{b}=\boldsymbol{b}_m} = 0,$$

and implying a positive-definite

$$\Sigma_m = (-L''_m(\boldsymbol{b}_m))^{-1},$$

where

$$\{L''_m(\boldsymbol{b}_m)\}_{k\ell} = \left[\frac{\partial^2 L_m(\boldsymbol{b})}{\partial b_k \partial b_\ell}\right]\bigg|_{\boldsymbol{b}=\boldsymbol{b}_m}.$$

Finally, let $|\boldsymbol{b}| = (\boldsymbol{b}^T\boldsymbol{b})^{1/2}$ and $B_\varepsilon(\boldsymbol{b}^*) = \{\boldsymbol{b} \in \Theta; |\boldsymbol{b} - \boldsymbol{b}^*| < \varepsilon\}$, then the following three basic conditions are sufficient to ensure a valid normal approximation for $f_m(\boldsymbol{b})$ in a small neighbourhood of $\boldsymbol{b}_m$ as m becomes large.

1. *Steepness*: Let $\bar{\sigma}_m^2$ be the largest eigenvalue of Σ_m. Then $\bar{\sigma}_m^2 \to 0$ as $m \to \infty$.

2. *Smoothness*: For any $\delta > 0$, there exists M and $\varepsilon > 0$ such that, for any $m > M$ and $\boldsymbol{b} \in B_\varepsilon(\boldsymbol{b}_m)$, $L_m''(\boldsymbol{b})$ exists and satisfies

$$I - A(\delta) \leq L_m''(\boldsymbol{b})[L_m''(\boldsymbol{b}_m)]^{-1} \leq I + A(\delta),$$

   where I is the $q \times q$ identity matrix, and $A(\delta)$ is a $q \times q$ symmetric positive-semi-definite matrix whose largest eigenvalue tends to zero as $\delta \to 0$.

3. *Concentration*: For any $\varepsilon > 0$, $\int_{B_\varepsilon(\boldsymbol{b}_m)} f_m(\boldsymbol{b})d\boldsymbol{b} \to 1$ as $m \to \infty$.

Essentially, the *steepness* and *smoothness* condition ensure that, for large m, inside a small neighbourhood of $\boldsymbol{b}_m$ the function $f_m(\boldsymbol{b})$ becomes highly peaked and behaves like the multivariate normal kernel $\exp[-\frac{1}{2}(\boldsymbol{b}-\boldsymbol{b}_m)^T\Sigma_m^{-1}(\boldsymbol{b}-\boldsymbol{b}_m)]$. The *concentration* condition ensures that the probability outside any neighbourhood of $\boldsymbol{b}_m$ becomes negligible.

Samenvatting

Longitudinale gegevens kunnen omschreven worden als bepaalde kenmerken van een individu of een groep, die herhaaldelijk gemeten worden over tijd. Aangezien metingen van eenzelfde individu gewoonlijk sterker verwant zijn dan metingen van verschillende individuen, moet een geldige analyse rekening houden met dit aspect, de associatie dus. Verschillende modelfamilies kunnen hiervoor aangesproken worden. Wanneer men bijvoorbeeld slechts geïnteresseerd is aan het gemiddelde gedrag van de populatie, kan men gebruik maken van de zogeheten *marginale* modellen. Als men daarentegen de associatie tussen de herhaalde metingen wenst te bestuderen, of men is geïnteresseerd aan effecten specifiek voor elk individu, worden *subject-specifieke* termen aan het model toegevoegd. Dergelijke parameters, die trouwens niet geobserveerd worden, blijven dan constant voor een gegeven individu, maar verschillen van individu tot individu. Dergelijke modellen worden gevat onder de noemer *random-effecten* of *individu-specifieke* modellen.

Naargelang het soort informatie bestaan er verschillende random-effecten modellen. Het *linear mixed model* (LMM; Verbeke and Molenberghs, 2000) wordt algemeen aanvaard als dé basis voor de analyse van normaal verdeelde responsen. Voor niet-normaal verdeelde gegevens wordt vaak gebruik gemaakt van het *generalized linear mixed model* (GLMM; Molenberghs and Verbeke, 2005). Het fitten van dit type modellen is gebaseerd op de klassieke *meest aannemelijke schattingstechnieken* (of *maximum likelihood estimation*), en vereist het maximaliseren van de *marginale aannemelijkheidfunctie* of *marginale likelihood*. Deze wordt verkregen door de random effecten over hun veronderstelde distributie uit de conditionele likelihood te integreren.

Het correct schatten van de parameters, en het bijhorende testen van hypothesen, hangen uiteraard af van de veronderstelling dat het model, en bijgevolg ook de verdeling van de random effecten, correct gespecificeerd is. Aangezien random effecten

niet geobserveerd worden, is het moeilijk om deze veronderstelling te valideren. In het geval van normaal verdeelde uitkomsten werd reeds aangetoond dat, in het LMM, de meest aannemelijke schatters (of *maximum likelihood estimators*) van de parameters in het model asymptotisch onvertekend zijn, zelfs wanneer de verdeling van de random effecten verkeerdelijk gespecificeerd is (Verbeke and Lesaffre, 1997). Of dit ook het geval is voor de GLMM, is de kernvraag die we met dit werk proberen te beantwoorden.

Aan de hand van simulaties met een logistiek regressiemodel bestuderen we de gevolgen van het verkeerd specificeren van de verdeling van de random effecten. Deze simulaties geven aan dat de meest aannemelijke schatters in GLMM niet langer asymptotisch onvertekend zijn. De schatters van de variantie componenten lijken altijd beïnvloed te worden door zo'n verkeerde onderstelling. Dit kan aanzienlijke gevolgen hebben in studies waar het correct schatten van de associatiestructuur van belang is. Verder kan dit resulteren in misleidende conclusies, wanneer men individu-specifieke profielen probeert te voorspellen. Anderzijds blijken de lineaire predictoren minder beïnvloed te worden. Wanneer de variabiliteit van de random effecten klein is, is ook de resulterende vertekening gering. Niettemin moet men voorzichtig zijn bij het interpreteren van de resultaten wanneer de random effecten veel variabiliteit vertonen of wanneer ingewikkelde covariantiestructuren gebruikt worden. Dergelijke situaties zijn niet uitzonderlijk in klinische studies waar men weinig veranderlijke responsprofielen kan observeren in de placebo groep, terwijl meer variabele profielen verwacht worden in de behandelingsgroep. In dit geval kunnen de lineaire predictoren dan ook onderhevig zijn aan een aanzienlijke vertekening. Tenslotte stellen we vast dat de type I fout en de kracht van vaak gebruikte testen, zoals de Wald test, beïnvloed kunnen worden door de misspecificatie.

Deze resultaten doen natuurlijk vragen rijzen over hoe men zich het beste kan beschermen tegen de gevolgen van het verkeerd onderstellen van de verdeling van de random effecten. Op de eerste plaats tonen we aan dat de type I fout, geassocieerd met een test voor de aanwezigheid van een effect, asymptotisch niet beïnvloed zal worden, zolang dit effect geen deel uitmaakt van de random structuur. Wanneer, in dit geval, wordt vastgesteld dat een effect significant is, kunnen we dan ook vrij zeker zijn van de aanwezigheid van dit effect. Verder stellen we een familie van diagnostische testen voor, gebaseerd op de ideeën van White (1982). Uit simulaties blijkt dat vooral de *Sandwich Estimator Test* (SET) en de *Modified Information Matrix Test* (MIMT) het beste in staat zijn om de verkeerd gespecificeerde random effecten verdeling vast

te stellen. Daarnaast illustreren we hoe de voorgestelde diagnostische testen ook gebruikt kunnen worden om meer algemene vormen van model misspecificatie vast te stellen.

Als men, na het toepassen van de testen, geconfronteerd wordt met een significant resultaat is het tevens belangrijk te weten hoe hiermee om te gaan. In de statistiek wordt dikwijls vastgesteld dat schatters en de conclusies, gebaseerd op deze schatters, robuust zijn tegen afwijkingen van de onderstellingen waarop ze rusten, althans wanneer de steekproef voldoende groot is. In het geval van GLMM blijkt een grote steekproef alleen niet voldoende te zijn. In dit werk tonen we aan dat, als zowel het aantal individuen, als het aantal metingen per individu voldoende talrijk zijn, dan de schatters van de lineaire predictoren robuust zullen zijn tegen het verkeerd specificeren van de verdeling van de random effecten. Dit resultaat kan echter niet uitgebreid worden naar de parameters in de random structuur. Desondanks is het niet altijd mogelijk om voldoende herhaalde metingen te verzamelen. Als, bijvoorbeeld, het opmeten van een kenmerk in een klinische studie heel oncomfortabel en/of heel duur is, kan het onethisch en/of kostelijk zijn om een deelnemer aan meerdere metingen te onderwerpen.

Als anderzijds de steekproef onvoldoende groot is om op asymptotische argumenten te vertrouwen, is het belangrijk te beschikken over alternatieve methodologie. In dit werk stellen we voor om het gebrek aan robuustheid van de schatters in GLMM op te vangen door een aantal niet-normale verdelingen voor de random effecten te integreren in een sensitiviteitsanalyse. Als de schatters gelijkaardig zijn, onafhankelijk van de gekozen random-effecten verdeling, dan kunnen we vrij zeker zijn van de verkregen resultaten. Als de resultaten echter aanzienlijk verschillen, dan zijn de schatters duidelijk gevoelig voor de keuze van de verdeling van de random effecten. Voorzichtigheid is dan geboden.