ABSTRACT

Title of dissertation: GOODNESS OF FIT TESTS FOR GENERALIZED LINEAR MIXED MODELS

Min Tang, Doctor of Philosophy, 2010

Dissertation directed by: Professor Eric V. Slud
Department of Mathematics
Dr. Ruth M. Pfeiffer
National Cancer Institute

Generalized Linear mixed models (GLMMs) are widely used for regression analysis of data, continuous or discrete, that are assumed to be clustered or correlated. Assessing model fit is important for valid inference. We therefore propose a class of chi-squared goodness-of-fit tests for GLMMs. Our test statistic is a quadratic form in the differences between observed values and the values expected under the estimated model in cells defined by a partition of the covariate space. We show that this test statistic has an asymptotic chi-squared distribution. We study the power of the test through simulations for two special cases of GLMMs, linear mixed models (LMMs) and logistic mixed models. For LMMs, we further derive the analytical power of the test under contiguous local alternatives and compare it with simulated empirical power. Three examples are used to illustrate the proposed test.

GOODNESS OF FIT TESTS FOR GENERALIZED LINEAR MIXED MODELS

by

Min Tang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:
Dr. Eric V. Slud, Co-Advisor
Dr. Ruth M. Pfeiffer, Co-Advisor
Dr. Paul J. Smith
Dr. Abram M. Kagan
Dr. Wolfgang Losert

Dedication

To my parents and Jun.

Acknowledgments

It is a great honor for me to thank all the people who made this thesis possible.

First and foremost, I owe my deepest gratitude to my two coadvisors, Dr. Eric V. Slud and Dr. Ruth M. Pfeiffer.

I am grateful to Dr. Eric V. Slud for his expert guidance and continuous support during my research and study at University of Maryland, College Park. Dr. Slud has been a significant presence in my life. His perpetual energy and his rigor and enthusiasm in research have been motivating me through my whole doctoral study. His insights have strengthened this study significantly. I will always be thankful for his wisdom, knowledge, deep concern and constant encouragement.

I would like to show my greatest gratitude and sincere thanks to Dr. Ruth M. Pfeiffer who offered me the predoctoral training at National Cancer Institute (NCI) which led me to the world of Biostatistics and cancer research. Her rigor and passion on research influenced me a lot. I am deeply indebted to her for her being always ready to help and her precious time on guiding me through my research. I could not have finished this thesis within four years without her invaluable help. She has also made available her support in a number of other ways, such as revising my CV, giving suggestions to my job talk rehearsal. She is the best mentor I have ever met. It has been an honor for me to work with her.

I am grateful to all my committee members Dr. Paul J. Smith, Dr. Abram Kagan and Dr. Wolfgang Losert for their valuable comments and suggestions to this thesis. I would like to thank Dr. Paul J. Smith for all his generous help and

guidance on taking courses in my entire graduate period. I also want to thank Dr. Abram M. Kagan for his interest and discussion with me on the MLE consistency proof in this thesis.

My special thanks go to Dr. Partha Lahiri for initially recommending me to the Biostatistics Branch in NCI.

Many thanks to all my colleagues and collaborators involved in the projects I have been working on at NCI. I would also show my sincere thanks to all my friends at the Mathematics Department, especially Ziliang Li for all the discussions on my thesis, Tinghui Yu for his generous help on my statistical questions, Ning Jiang for his encouragement during my qualify exam stage, Huilin Li for her help and sharing information during my stay at NCI and Lingyan Cao, Yong Zheng, Guanhua Lu, Denise, Shihua Wen, Shu Zhang, Wei Guo, Neung Ha and Rongrong Wang, for their sincere help and all the good time with them during my graduate study.

I am grateful to my cousins Lifang Liu and Juan Tang who provided tremendous help when I was applying for graduate school, who took care of my parents in China when I was pursuing my Ph.D. in the US, who have been standing by my side at all stages of my life. I also want to thank my good friend Caixia Kou and all my other friends in China for their constant supports.

My deepest gratitude goes to my family for their unflagging love and support throughout my life.

Finally, I am forever indebted to Jun. His encouragement and companionship have turned my journey through graduate school into a pleasure. This dissertation is simply impossible without them.

Table of Contents

List of Tables

List of Figures

List of Notation and Abbreviations

Ω	a space of outcomes of an experiment
P	a probability measure on Ω
$\mathcal{R}$	the real line, $(-\infty, \infty)$
$\mathcal{R}^+$	the non-negative real line, $[0, \infty)$
$\mathcal{Z}^+$	the set of positive integer numbers
$I_{\{A\}}$	indicator function, where A is either a set or an event defined by a property enclosed in brackets, defined in terms of a random variable
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$E(X)$	mean of the random variable X
$Var(X)$	variance of the random variable X
$\sigma\{\cdot\}$	the σ-algebra generated by random variables in $\{\cdot\}$
$E(Y\|X)$	the conditional expectation of Y w.r.t. a σ-algebra generated by X
$Var(Y\|X)$	the conditional variance of Y given X
i.i.d.	independent and identically distributed
$A \approx B$	$A - B \xrightarrow{P} 0$
$d(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$	Euclidean distance $\|\theta - \theta_0\|_2$ between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$
$\mathbf{a}^T$	transpose of the vector or matrix $\mathbf{a}$
$\mathbf{I}_N$	$N \times N$ identity matrix
$\mathbf{1}_n$	a $n \times 1$ vector of all 1s
$\mathbf{a}^{\otimes 2}$	Kronecker product of vector $\mathbf{a}$, $\mathbf{a}\mathbf{a}^T$
$\mathbf{1}_n^{\otimes 2}$	the $n \times n$ matrix of all 1s
A^{-1}	inverse of matrix A
$tr(A)$	trace of matrix A, defined to be the sum of the diagonal elements of matrix A
$\|A\|$	Euclidean norm of matrix A
$\xrightarrow{P}$	convergence in probability
$\xrightarrow{D}$	convergence in distribution
$o_p(1)$	a random variable sequence $X_n = o_p(1)$ means $X_n \xrightarrow{P} 0$

Chapter 1

Introduction

1.1 Background

The linear mixed model (LMM) (McCulloch and Searle, 2001) extends the standard linear regression model by including random effects in addition to the usual fixed effects in the linear predictors. LMMs can be expressed as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is a vector of observations, $\mathbf{X}$ is a matrix of known covariates, $\boldsymbol{\beta}$ is a vector of unknown fixed regression coefficients which are called fixed effects, $\mathbf{Z}$ is a known matrix, $\boldsymbol{\alpha}$ is a vector of unknown random effects and $\boldsymbol{\epsilon}$ is a vector of unobservable random errors. By incorporating random effects, LMMs can accommodate clustered or correlated or longitudinal data. For example, in medical studies, various measurements are often collected from the same individual over time. It is then reasonable to assume that the observations for the same individual are correlated. Examples of applying LMM to longitudinal data can be found for example in Laird and Ware (1982), Weiss (2005, Chapter 9) and Lee et al. (2006). Generalized linear mixed models (GLMMs) further extend the LMM family to discrete or categorical exponential family data. Examples include logistic mixed models for binomial data and Poisson mixed models for count data. Data examples of GLMMs are given in McCullagh and Nelder (1989, Section 14.5) and GLMMs have numerous applications in medical research and the survey area. Jiang and Lahiri (2006) gave a

very good general literature review of prediction based on GLMMs, with particular application to small-area estimation.

There are various methods of estimating GLMM parameters. The method of maximum likelihood is widely used. A full maximum likelihood analysis requires numerical integration techniques to calculate the log-likelihood function and thus the distribution of the random effects needs to be known. Jiang (1998b) proposed estimating equations that apply to GLMMs not necessarily having a block-diagonal covariance matrix structure. Jiang (1999) proposed a method of inference which in many ways resembles the method of least squares in linear models and relies on weak distributional assumptions about random effects. In this thesis, we focus on the maximum likelihood method for parameter estimation in GLMMs.

Developments in model fitting algorithms and their implementations in statistical packages have greatly facilitated the applications of LMMs and GLMMs. The commonly used functions for mixed modeling in the statistical software package SAS, version 9.2, are PROC MIXED, PROC NLMIXED and PROC GLIMMIX. The commonly used functions for mixed modeling in R, version 2.11.1, are

- linear mixed models: aov(), lme() in library(nlme), lmer() in library(lme4);
- generalized linear mixed models: glmmPQL() in library(MASS), glmer() in library(lme4), MCMCglmm() in library(MCMCglmm).

Two important steps in modeling are selecting a model and checking its fit. Frequently, model selection is done by comparing nested models, via likelihood ratio, wald or score tests, as part of model building and there are approaches for com-

paring non-nested models (Cox, 1961; Godfrey, 1988). AIC (Akaike's information criterion), BIC (Bayesian information criterion) and other model selection principles (Rao and Wu, 1989; Shao, 1997) focus on selection of covariates. Rao, Wu et al. (2001) gave a concise review on the subject of the statistical model selection.

These methods select the best statistical model from a set of potential models chosen by the researcher, given the observed data. Even though the finally selected model may be the best in the class of potential models, it might still not provide a good fit to the data. Thus once a model is selected, its fit should be assessed. This is often done by checking residuals and formal goodness of fit tests. There are various diagnostics and graphical techniques in assessing goodness of fit of models. Lin, Wei and Ying (2002) developed objective and informative model-checking techniques for a variety of statistical models and data structures, including generalized linear models with independent or dependent observations, by taking the cumulative sums of residuals over certain coordinates. Lange and Ryan (1989) described a graphical method for checking distributional assumptions about the random effects in random effects models. Park and Lee (2004) proposed residual plots to investigate the goodness of fit for repeated measure data, where they mainly focus on the mean model diagnostics. Jacqmin-Gadda et al. (2007) discussed the fit of a linear mixed model through Cholesky residuals and conditional residuals. Pan and Lin (2005) developed graphical and numerical methods for checking the adequacy of generalized linear mixed models, by comparing the cumulative sums of residuals with certain Gaussian processes. Regarding formal tests for the model adequacy, goodness of fit tests for generalized linear models for fixed effects can be found in Chapter 4

in Agresti (2002). However, the literature for formally assessing the overall fit for GLMMs is limited. Some procedures to assess model misspecification have been proposed. Testing for the presence of random effects in LMMs has been discussed by Self and Liang (1987) and by Crainiceanu and Ruppert (2004). Jiang (2001) and Ritz (2004) assessed the distributional assumptions for the random effects in LMMs. Claeskens and Hart (2009) proposed formal tests for testing the normality of random effects and/or error terms in LMMs. Khuri, Mathew and Sinha (1998) presented derivations of both exact and optimal tests regarding variance component models, as well as guidance on using such tests for hypothesis testing for the fixed effect part. These are separate tools for checking fixed-effect specification or for separately checking the residuals or forms of the random effect specification.

Testing the overall adequacy of a proposed model has been discussed in the literature for several types of models with fixed effects. Tsiatis (1980) proposed a goodness-of-fit test to test the overall fit of a logistic regression model. The test is originally established based on the efficient scores test and after simplification, it is reduced to a quadratic form of observed counts minus the expected counts in regions of the covariate space. However, for logistic mixed models, with the presence of random effects, this efficient scores test can not be simplified as a quadratic form of observed counts minus the expected counts because of the integrals involved in the likelihood function. For survival data, Schoenfeld (1980) presented a class of omnibus chi-squared goodness of fit tests for the proportional hazards regression model. Slud and Kedem (1994) adapted the idea of Schoenfeld to generalized linear time series models and discussed fixed effects binary-response models with time-

dependent covariates. Kedem and Fokianos (2002) extended this approach to various other goodness of fit tests based on categorical time series residuals. In this thesis, we adopt the idea of Schoenfeld (1980) and develop a class of goodness of fit tests for GLMMs by comparing the observed and expected values computed from the model within cells of a partition of the covariate space. This class of goodness of fit tests primarily assesses the adequacy of fit of the fixed effects part in the presence of random effects.

1.2 Overview of thesis

In Chapter 2, we present the linear mixed models (LMMs). We adopt the idea of Schoenfeld (1980) and propose a class of goodness of fit tests for testing the statistical adequacy of the selected LMM. We study two classes of LMMs, the general LMM with additive random effects $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{r=1}^{R} \mathbf{Z}_r\boldsymbol{\alpha}_r + \boldsymbol{\varepsilon}$, where the random effects $\boldsymbol{\alpha}_r, r = 1, \ldots, R$ are normally distributed and in a moderate to large sample setting, the two-level LMM $y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i + \epsilon_{ij}$, that is, the LMM with one random intercept, where no distributional assumption is made on the random effect α_i or the error term ϵ_{ij}. For this two-level LMM, $i = 1, \ldots, m$, m denotes the number of clusters, and $j = 1, \ldots, n_i$, n_i denotes the size of cluster i. To deal with technical issues, the covariate matrix $\mathbf{X}$ is is assumed in different settings to be a matrix either of fixed constants or of random variables. We propose a test statistic based on differences between observations and their expected values computed under the model aggregated over cells of a partition of the covariate space.

We first discuss assumptions needed, and then derive the asymptotic properties of this test statistic as the total number of observations N tends to infinity under the null hypothesis and under local alternatives. For the two-level LMM, $N = \sum_{i=1}^{m} n_i = (\sum_{i=1}^{m} n_i/m)m$. Under the assumption of the existence of $\sum_{i=1}^{m} n_i/m$, N tending to infinity is equivalent to m tending to infinity. For the general LMM with additive normal random effects, we estimate parameters using maximum likelihood estimators (MLEs) and assume that the covariate matrix $\mathbf{X}$ is fixed and nonrandom. For the two-level LMM with no distributional assumptions made on the random effect or the error term, we assume that $(\mathbf{x}_i, n_i)$, where i is the cluster index, are i.i.d random vectors and estimate the parameters using least squares and method of moments. In Chapter 2, we also check the theoretical power in simulations, and study the impact of choice of cell partitions on the test as well as the robustness of the test with respect to the error distribution. We illustrate this test in three real datasets.

In Chapter 3, we extend the test to random intercept generalized linear mixed models (GLMMs) and derive its theoretical power. To do that, we also prove the MLE consistency of GLMMs under certain assumptions. The covariate matrix $\mathbf{X}$ in this Chapter is considered to be random variables. Again, we conduct simulations to assess factors with an impact on the power of the test in GLMMs.

In Chapter 4, we discuss the applicability of the results and point to future research directions.

There is a separate technical appendix at the end of each chapter. Appendix A contains three lemmas cited in this thesis.

Chapter 2

Goodness of fit tests for linear mixed models

2.1 Linear mixed models (LMMs)

A linear mixed model (LMM) has the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{Y}_{N\times 1}$ is the vector of observations; $\mathbf{X}_{N\times p}$ is the design matrix for the fixed effects part of the model; $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown fixed effects parameters; $\mathbf{u}$ is a vector of random effects and $\boldsymbol{\varepsilon}$ is a vector of errors. Typically $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are assumed to be independent of each other and each independently normally distributed with mean 0 and unknown variances.

In this thesis, we only consider the LMM (2.1) with $\mathbf{Z}\boldsymbol{\alpha} = \sum_{r=1}^{R} \mathbf{Z}_r\boldsymbol{\alpha}_r$, i.e.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{r=1}^{R} \mathbf{Z}_r\boldsymbol{\alpha}_r + \boldsymbol{\varepsilon}. \tag{2.2}$$

Here $\mathbf{Z}_r$, an $N \times m_r$ matrix of constants, is the design matrix for the random effect $\boldsymbol{\alpha}_r$, $r = 1, \ldots, R$. The quantity $\boldsymbol{\alpha}_r$ is an $m_r \times 1$ random vector, $r = 1, \ldots, R$. Also, components of $\boldsymbol{\alpha}_r$ are i.i.d. within the vector, $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_R$ are independent and are also independent of $\boldsymbol{\varepsilon}$. In this thesis, $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_R$ are always assumed normal

except for the 2-level LMM case discussed in Section 2.2.2 where no distributional assumptions are made on either the random effect or the error term. Let $\boldsymbol{\psi} = (\sigma_\epsilon^2, \sigma_1^2, \ldots, \sigma_R^2)$, the parameter vector of all variance components and let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi})$. Let $\mathbf{G}_r = \mathbf{Z}_r\mathbf{Z}_r^T$, $r = 1, \ldots, R$ and $\mathbf{G}_0 = \mathbf{I}_N$. The $\mathbf{X}$ matrix can be fixed or random. To deal with technical issues, $\mathbf{X}$ is considered to be fixed in Section 2.2.1 and to be random in Section 2.2.2.

As an important case of model (2.2), we now introduce the linear mixed model with a single random effect:

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i + \epsilon_{ij}, \;\; i = 1, \ldots, m, \;\; j = 1, \ldots, n_i, \tag{2.3}$$

where the $1 \times p$ vector $\mathbf{x}_{ij}^T = (1, x_{ij1}, \ldots, x_{ij(p-1)})$ denotes covariates for fixed effects for the jth observation within the ith cluster. The cluster specific random effects $\alpha_i \sim N(0, \sigma_a^2)$ are assumed to be independent of the error terms ϵ_{ij}, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. To accommodate an intercept term in the model, the first entry in $\mathbf{x}_{ij}$ is 1. We write $N = \sum_{i=1}^m n_i$ and $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$ denotes the vector of observations for the ith cluster.

Under these assumptions, $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, with a block diagonal covariance matrix $\mathbf{V}$, where each of the m $n_i \times n_i$ blocks has the structure

$$\mathbf{V}_i = \begin{pmatrix} \sigma_a^2 + \sigma_\epsilon^2 & \cdots & \sigma_a^2 \\ \vdots & \ddots & \vdots \\ \sigma_a^2 & \cdots & \sigma_a^2 + \sigma_\epsilon^2 \end{pmatrix}_{n_i \times n_i}. \tag{2.4}$$

For asymptotic analysis of the LMM model (2.3), we always assume that m goes to infinity, thus N also goes to infinity.

The following assumptions are made on model (2.2) throughout Chapter 2.

Assumption 2.1 *The true parameter point* $\boldsymbol{\theta_0} = (\boldsymbol{\beta}_0, \boldsymbol{\psi}_0)$ *is an interior point of* $\boldsymbol{\Theta} = (\mathcal{R}^p, (\mathcal{R}^+)^{R+1})$. *For the 2-level LMM (2.3),* $R = 1$.

Assumption 2.2 *The covariate matrix* $\boldsymbol{X}$ *can be either fixed or random. If* $\boldsymbol{X}$ *is assumed to be fixed, then it is assumed to have full rank* p. *If* $\boldsymbol{X}$ *is assumed to be a matrix of random variables which is the 2-level LMM (2.3) discussed in Section 2.2.2, then* $(\boldsymbol{x}_i, n_i)$ *are assumed i.i.d. with* $\|E(\boldsymbol{x}_i^{\otimes 2})\| < \infty$ *and* $E(n_i^2) < \infty$.

Assumption 2.3 *When the covariate matrix* $\boldsymbol{X}$ *is considered to be fixed, we always assume that, with* $E_1, \ldots, E_l$ *being a cell partition of the covariate space, for* $l = 1, \ldots, L$, $N^{-1} \sum_{k=1}^{N} I_{\{\boldsymbol{x}_k \in E_l\}} \mathbf{x}_k^T$ *exists.*

Remark 2.1 *For the 2-level LMM with fixed covariate matrix* $\boldsymbol{X}$, *Assumption 2.3 is equivalent to the existence of* $N^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} \mathbf{x}_{ij}^T$. *When the covariate matrix* $\boldsymbol{X}$ *is considered to be random, the existence of* $N^{-1} \sum_{k=1}^{N} I_{\{\boldsymbol{x}_k \in E_l\}} \mathbf{x}_k^T$ *is ensured by Assumption 2.2.* □

2.2 Goodness of fit test statistic

2.2.1 Test statistic and its asymptotic properties for LMMs when parameters are estimated by maximum likelihood

2.2.1.1 LMM with a single random effect

In this Section, we discuss the 2-level LMM (2.3) with normality assumptions on both the random effect and the error term. The covariate matrix $\mathbf{X}$ is considered to be fixed. We derive our test statistic for the setting where the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}) = (\boldsymbol{\beta}, \sigma_a^2, \sigma_\epsilon^2)$ are estimated by the maximum likelihood. Since we assume normality both for the random intercept term α_i and for the error term ϵ_{ij}, we can use the result of Miller (1977) to show the consistency and asymptotic normality of the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$. The following assumptions are made for the two-level LMM.

Assumption 2.4 $\boldsymbol{J}_{\beta\beta} = \lim_{N\to\infty} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}/N$ *exists and is positive definite;*

Here the positive definiteness assumption for $\mathbf{J}_{\beta\beta}$ is equivalent to the assumption that $\mathbf{X}$ has full rank.

Assumption 2.5 *The* 2×2 *matrix* $\boldsymbol{M}$ *with elements defined below exists and is positive definite;*

$$[\boldsymbol{M}]_{st} = \frac{1}{2} \lim_{N\to\infty} \left(tr\, \boldsymbol{V}^{-1} \boldsymbol{G}_s \boldsymbol{V}^{-1} \boldsymbol{G}_t \right) /N, \quad s,t = 0,1,$$

where $\boldsymbol{G}_0 = \boldsymbol{I}$ *is the* $N \times N$ *identity matrix and* $\boldsymbol{G}_1 = \boldsymbol{1}^{\otimes 2}$ *is the* $N \times N$ *matrix of*

all 1*s*.

Under Assumptions 2.1 – 2.5, based on Miller (1977), the maximum likelihood estimators (MLEs) for model (2.3) exist and are consistent.

To test the goodness of fit of the LMM (2.3), we first divide the covariate space into L disjoint regions $E_1, \ldots, E_L$. We compute the observed and expected sums in each region E_l as

$$f_l = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} y_{ij}, \tag{2.5}$$

$$e_l(\boldsymbol{\beta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} E(y_{ij}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} \mathbf{x}_{ij}^T \boldsymbol{\beta}, \tag{2.6}$$

where I denotes the indicator function.

Remark 2.2 *The cell partition can also be based on covariates not included in model (2.3). In this case, if we let* $\boldsymbol{x}_{ij}$ *denote the vector of all available covariates and* $\boldsymbol{x}_{ij}^*$ *only includes covariates used in the regression, then we would define* $e_l(\boldsymbol{\beta}^*) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} (\boldsymbol{x}_{ij}^*)^T \boldsymbol{\beta}^*$*, where* $\boldsymbol{\beta}^*$ *corresponds to the coefficients of* $\boldsymbol{x}_{ij}^*$*. But no matter which kind of cell partition we choose, we employ the expressions in (2.5) and (2.6) in the whole thesis for notational simplicity.* □

With the notation $\mathbf{f} = (f_1, \ldots, f_L)$ and $\mathbf{e}(\boldsymbol{\beta}) = (e_1, \ldots, e_L)$, the observed minus the expected vector is

$$\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0) = \begin{pmatrix} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_1\}} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}_0) \\ \vdots \\ \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_L\}} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}_0) \end{pmatrix}. \tag{2.7}$$

Since the true parameter vector $\boldsymbol{\beta}_0$ is unknown, we replace $\boldsymbol{\beta}_0$ in (2.7) by its MLE $\hat{\boldsymbol{\beta}}$ for Theorem 2.3. The proof of this Theorem is given in Section 2.6.1. The following assumption is made to ensure the existence of components of the variance covariance matrix for the test statistic.

Assumption 2.6 $\lim_{K\to\infty} \limsup_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} I_{\{n_i^2 \geq K\}}\, n_i^2 \to 0.$

Theorem 2.3 *For model (2.3), let $E_1, \ldots, E_L$ constitute a disjoint partition of the covariate space generated by $\mathbf{X}$. Under Assumptions 2.1 – 2.6, as $N \to \infty$,*

$$\sqrt{N} \begin{pmatrix} (\mathbf{f} - \boldsymbol{e}(\boldsymbol{\beta}_0))/N \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^T),$$

where with the $(L+p) \times N$ matrix $\boldsymbol{D}$ given in equation (2.29),

$$\boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^T = \begin{pmatrix} \boldsymbol{H} & \boldsymbol{\Lambda}\boldsymbol{J}_{\beta\beta}^{-1} \\ \boldsymbol{J}_{\beta\beta}^{-1}\boldsymbol{\Lambda}^T & \boldsymbol{J}_{\beta\beta}^{-1} \end{pmatrix}_{(L+p)\times(L+p)}. \tag{2.8}$$

The entries of $\boldsymbol{H}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{J}_{\beta\beta}$ are given below

$$\boldsymbol{H}_{lk} = \sigma_a^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_k\}} \right) \left(\sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} \right) \right], \tag{2.9}$$

$$\boldsymbol{H}_{ll} = \sigma_\epsilon^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} + \sigma_a^2 \frac{1}{N} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} \right)^2, \tag{2.10}$$

$$\boldsymbol{\Lambda}_l = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} \mathbf{x}_{ij}^T, \tag{2.11}$$

$$\boldsymbol{J}_{\beta\beta} = \lim_{N\to\infty} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}/N. \tag{2.12}$$

Consistent estimators for these quantities are given in Corollary 2.5 below.

Remark 2.4 *Assumption 2.3 ensures the existence of* $\boldsymbol{\Lambda}$ *in (2.8). Assumptions 2.6 and 2.4 ensure the existence of* $\boldsymbol{H}$*, the limiting variance covariance matrix for* $(\boldsymbol{f} - \boldsymbol{e}(\boldsymbol{\beta}_0))/N$*, and* $\boldsymbol{J}_{\beta\beta}$ *in (2.8).* □

Corollary 2.5 *Consistent estimators for elements in the block matrix* $\boldsymbol{DVD}^T$ *in Theorem 2.3 are:*

$$\hat{\boldsymbol{H}}_{lk} = \hat{\sigma}_a^2 \frac{1}{N} \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_k\}} \right) \left(\sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} \right) \right],$$

$$\hat{\boldsymbol{H}}_{ll} = \hat{\sigma}_\epsilon^2 \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} + \hat{\sigma}_a^2 \frac{1}{N} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} \right)^2,$$

$$\hat{\boldsymbol{\Lambda}}_l = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}} \mathbf{x}_{ij}^T, \quad \hat{\boldsymbol{I}}_{\beta\beta} = \boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}/N,$$

where $\hat{\boldsymbol{H}}_{lk}, \hat{\boldsymbol{H}}_{ll}$ *are estimators for off-diagonal and diagonal elements of* $\boldsymbol{H}$*, and* $\hat{\boldsymbol{\Lambda}}_l$ *estimates the* l*-th row of* $\boldsymbol{\Lambda}$*.*

Remark 2.6 *If* $\boldsymbol{X}$ *is random, under the more restrictive assumption that* $\boldsymbol{x}_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i$ *are i.i.d. and are independent of* n_i*, the diagonal and off-diagonal elements of* $\boldsymbol{H}$ *given in (2.9) and (2.10) are*

$$\boldsymbol{H}_{lk} = \sigma_a^2 \cdot \frac{E(n_1^2 - n_1)}{E(n_1)} P\left(\boldsymbol{x}_{11} \in E_l\right) P\left(\boldsymbol{x}_{11} \in E_k\right), \quad \forall\ l \neq k$$

and

$$\boldsymbol{H}_{ll} = (\sigma_\epsilon^2 + \sigma_a^2) P(\boldsymbol{x}_{11} \in E_l) + \sigma_a^2 \frac{E(n_1^2 - n_1)}{E(n_1)} \left[P(\boldsymbol{x}_{11} \in E_l)\right]^2.$$

In this case, another way of estimating the diagonal and off-diagonal elements of $\boldsymbol{H}$ *is to use*

$$\hat{\boldsymbol{H}}_{lk} = \hat{\sigma}_a^2 \frac{\hat{E}(n_1^2) - \hat{E}(n_1)}{\hat{E}(n_1)} \hat{P}(\boldsymbol{x}_{11} \in E_l)\hat{P}(\boldsymbol{x}_{11} \in E_k)$$

and

$$\hat{\boldsymbol{H}}_{ll} = (\hat{\sigma}_\epsilon^2 + \hat{\sigma}_a^2)\hat{P}(\boldsymbol{x}_{11} \in E_l) + \hat{\sigma}_a^2 \frac{\hat{E}(n_1^2 - n_1)}{\hat{E}(n_1)} \left[\hat{P}(\boldsymbol{x}_{11} \in E_l)\right]^2,$$

where $\hat{E}(n_1^2) = \sum_{i=1}^m n_i^2/m$, $\hat{E}(n_1) = \sum_{i=1}^m n_i/m$ *and* $\hat{P}(\boldsymbol{x}_{11} \in E_l) = \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}}/N$. *In the R code for simulations and data analysis below, we use the estimators in Corollary 2.5 to estimate* $\boldsymbol{H}$ *and in calculating the theoretical power in the analytical power study Section, the estimators in this Remark are applied.* □

Corollary 2.7 *Under Assumptions 2.1 – 2.6,* $(\boldsymbol{f} - \boldsymbol{e}(\hat{\boldsymbol{\beta}}))/\sqrt{N} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma})$, *where* $\boldsymbol{\Sigma} = \boldsymbol{H} - \boldsymbol{\Lambda}\boldsymbol{J}_{\beta\beta}^{-1}\boldsymbol{\Lambda}^{\mathbf{T}}$ *is an* $L \times L$ *matrix and can be replaced by its consistent estimator* $\hat{\boldsymbol{\Sigma}}^0 = \hat{\boldsymbol{H}} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{J}}_{\beta\beta}^{-1}\hat{\boldsymbol{\Lambda}}^T$ *based on Corollay 2.5.*

We compute Singular Value Decomposition for $\hat{\boldsymbol{\Sigma}}^0$. For each eigenvalue of $\hat{\boldsymbol{\Sigma}}^0$, we compare it with a small preset threshold, such as 10^-4 ζ. For any eigenvalue less than ζ, we instead set this eigenvalue to be 0 and reconstruct the $\hat{\boldsymbol{\Sigma}}^0$ matrix using the non-zero eigenvalues and their corresponding eigenvectors. We denote this reconstructed matrix as $\hat{\boldsymbol{\Sigma}}$. Based on Corollary 5.3 given in the Appendix, $P(\text{rank }(\hat{\boldsymbol{\Sigma}}) = \text{rank }(\boldsymbol{\Sigma})) \to 1$.

Our goodness of fit test statistic is then given by the following quadratic form

$$T = \frac{1}{N}(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\beta}}))^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\beta}})), \tag{2.13}$$

where $\hat{\boldsymbol{\Sigma}}^{-1}$ denotes the Moore-Penrose pseudoinverse of $\hat{\boldsymbol{\Sigma}}$. Under the null hypothesis, T has an asymptotic central χ^2_k distribution, where $k = \text{rank}(\hat{\boldsymbol{\Sigma}}) = \text{rank}(\boldsymbol{\Sigma})$ for large N, based on Corollary 5.3.

2.2.1.2 LMM with additive random effects

We consider next the LMM with additive random effects (2.2), that is,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{r=1}^{R} \mathbf{Z}_r \boldsymbol{\alpha}_r + \boldsymbol{\varepsilon}. \tag{2.14}$$

The covariate matrix $\mathbf{X}$ is considered to be fixed numbers in this Section. This is model (1) in Miller (1977). We first state and comment on the assumptions Miller (1977) made to ensure consistency and asymptotic normality of the MLE for $\boldsymbol{\theta}$ in (2.2).

Assumption 2.7 ***A.1*** *The partitioned matrix* $[\boldsymbol{X} : \boldsymbol{Z}_r]$ *has rank greater than* p, $r = 1, \ldots, R$.

A.2 *The matrices* $\boldsymbol{G}_0, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_R$ *are linearly independent; that is,* $\sum_{r=0}^{R} \tau_r \boldsymbol{G}_r = 0$ *implies* $\tau_r = 0$, $r = 0, 1, \ldots, R$.

A.3 N *and each* m_r, $r = 1, \ldots, R$, *tend to infinity.*

A.4 *Let* $m_0 = N$. *Then for each* $s, t = 0, 1, \ldots, R$, *either* $\lim_{N\to\infty} m_s/m_t = \rho_{st}$ *or*

$\lim_{N\to\infty} m_t/m_s = \rho_{ts}$ *exists. If* $\rho_{st} = 0$*, then let* $\rho_{ts} = \infty$ *for notational convenience.*

Without loss of generality, let $\boldsymbol{Z}_r$ *be labeled so that for* $s < t$*,* $\rho_{st} > 0$*; i.e., the* m_r *are in decreasing order of magnitude. Generate a partition of the integers* $0, 1, \ldots, R$*,* $\boldsymbol{S}_0, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_c$*, so that for indices* r *in the same set* $\boldsymbol{S}_s$*, the associated* m_r*'s have the same order of magnitude. Such a partition is generated as follows:*

i) $r_0 = 0$*;* $\boldsymbol{S}_0 = \{0\}$*;* $r_1 = 1$*.*

ii) For $s = 1, 2, \ldots$*, it is true that* $r_s \in \boldsymbol{S}_s$*. Then for* $r = r_s+1, r_s+2, \ldots$*, include* r *in* $\boldsymbol{S}_s$ *until* $\rho_{r_s,r} = \infty$*; call the first value of* r *where this occurs* r_{s+1}*; then* $r_{s+1} \in \boldsymbol{S}_{s+1}$*.*

iii) Continue as in step ii until R *has been placed in a set. Call this set* $\boldsymbol{S}_c$*.*

There are then $c + 1$ *sets in partitions,* $\boldsymbol{S}_0, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_c$*, and* $\boldsymbol{S}_s = \{r_s, \ldots, r_{s+1} - 1\}$*.*

For each $r = 1, 2, \ldots, R$*,* $r \in \boldsymbol{S}_s$ *for some* $s = 1, 2, \ldots, c$*. Define sequences* K_r *(depending on* N*) as follows:*

$K_r = rank[\boldsymbol{Z}_{r_s} : \boldsymbol{Z}_{r_s+1} : \cdots : \boldsymbol{Z}_R] - rank[\boldsymbol{Z}_{r_s} : \cdots : \boldsymbol{Z}_{r-1} : \boldsymbol{Z}_{r+1} : \cdots : \boldsymbol{Z}_R]$,

$r = 1, 2, \ldots, R$,

$K_0 = N - rank[\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_R]$.

(The K_r *so defined are closely related to the degrees of freedom of sums of squares in the analysis of variance.)*

A.5 *Each of the* $\lim_{N\to} K_r/m_r$*,* $r = 1, \ldots, R$ *exists and is positive.*

Let $\boldsymbol{V}_0 = \sum_{r=1}^{R} \sigma_r^2 \boldsymbol{Z}_r$ *be the true covariance matrix.*

A.6 *There exists a sequence* K_{R+1} *(depending on* N*) increasing to infinity such that the* $p \times p$ *matrix* $\boldsymbol{C}_0$ *defined by* $\boldsymbol{C}_0 = \lim_{N\to\infty}[\boldsymbol{X}' \boldsymbol{V}_0^{-1} \boldsymbol{X}]/K_{R+1}$ *exists and is positive definite.*

Define the $(R+1)\times(R+1)$ *matrix* $\boldsymbol{C}_1$ *by*

$$[\boldsymbol{C}_1]_{st} = \frac{1}{2}\lim_{N\to\infty}[tr\,\boldsymbol{V}_0^{-1}\boldsymbol{G}_s\,\boldsymbol{V}_0^{-1}\boldsymbol{G}_t]/K_s^{\frac{1}{2}}K_t^{\frac{1}{2}}, \quad s,t=0,1,\ldots,R.$$

A.7 *Each of the limits used in defining* $[\boldsymbol{C}_1]_{st}$ *exists,* $s,t=0,1,\ldots,R$. *The matrix* $\boldsymbol{C}_1$ *is positive definite.*

Remark 2.8 *Assumption* $A.1$ *requires that the fixed effects not be confounded with any of the random effects.* $A.2$ *requires that the random effects not be confounded with each other. For the LMM (2.3) or hierarchical linear mixed models with nested blocks such as (2.18), the matrices* $\boldsymbol{Z}_r, r=1,\ldots,R$, *consist only of* 0*'s or* 1*'s and satisfy* $A.1-A.2$. *Assumptions* $A.1-A.2$ *are sufficient to guarantee identifiability of the MLE* $\hat{\boldsymbol{\theta}}$. *Assumptions* $A.3-A.7$, *which correspond to Assumptions* $3.1-3.5$ *in Miller (1977), are used to ensure the consistency of the MLE. Assumption* $A.3$ *is natural and necessary for the consistency property of MLE estimators of both* $\boldsymbol{\beta}$ *and the variance components* σ_ϵ^2 *and* $\sigma_r^2, r=1,\ldots,R$, *because the sample size used to estimate* $\boldsymbol{\beta}$ *and* σ_ϵ^2 *is* N *and the sample size used to estimate* σ_r^2 *is* m_r. *Assumptions* $A.6-A.7$ *are used to establish the existence and positive definiteness of the limiting variance-covariance matrix of the MLE* $\hat{\boldsymbol{\theta}}$. □

In addition to Assumption 2.7, taken from Miller (1977), we also require the following Assumption to ensure the existence of components in the variance covariance matrix for the test statistic.

Assumption 2.8 $\boldsymbol{H} = \lim_{N\to\infty} \boldsymbol{F}\boldsymbol{V}\boldsymbol{F}^{'}$ *exists and is positive definite, with* $\boldsymbol{F}$ *given in (2.16).*

Remark 2.9 *Assumption 2.3 ensures the existence of* $\boldsymbol{\Lambda}$ *in (2.17). Assumption 2.8 ensures the existence and positive definiteness of* $\boldsymbol{H}$ *in* $\boldsymbol{\Sigma}$. $\boldsymbol{H}$ *is the limiting variance covariance matrix for* $(\boldsymbol{f} - \boldsymbol{e}(\boldsymbol{\beta}_0))/N$, *which involves empirical moments of the covariates and empirical moments of cluster sizes at different levels. At the end of this Section, we give specific forms of* $\boldsymbol{H}$ *and ways of estimating* $\boldsymbol{H}$ *for the special case of 3-level LMM (2.18).* □

We next state our goodness of fit test for model (2.2) in Theorem 2.10 below. The details of the proof are given in Section 2.6.3. The covariate space is divided into L disjoint regions $E_1, \ldots, E_L$. This partition can also be based on covariates not included in model (2.2). As discussed in Remark 2.2 in Section 2.2.1.1, for notational simplicity, the following notation applies whether or not the cell partition is based only on covariates in the model.

For $l = 1, 2, \ldots, L$, define

$$f_l = \sum_{k=1}^{N} I_{\{\mathbf{x}_k \in E_l\}} y_k,$$

$$e_l(\boldsymbol{\beta}) = \sum_{k=1}^{N} I_{\{\mathbf{x}_k \in E_l\}} E(y_k) = \sum_{k=1}^{N} I_{\{\mathbf{x}_k \in E_l\}} \mathbf{x}_k^T \boldsymbol{\beta}.$$

Let $\mathbf{f} = (f_1, \ldots, f_L)$, $\mathbf{e}(\boldsymbol{\beta}) = (e_1(\boldsymbol{\beta}), \ldots, e_L(\boldsymbol{\beta}))$.

Theorem 2.10 *For model (2.2), let* $E_1, \ldots, E_L$ *constitute a disjoint partition of the*

covariate space generated by $\mathbf{X}$. *Under Assumptions 2.1 , 2.2 and 2.7 – 2.8,*

$$(\boldsymbol{f} - \boldsymbol{e}(\hat{\boldsymbol{\beta}}))^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{f} - \boldsymbol{e}(\hat{\boldsymbol{\beta}}))/\sqrt{N} \xrightarrow{\mathcal{D}} \chi^2_k, \tag{2.15}$$

where $\hat{\boldsymbol{\beta}}$ *is the MLE,* $\hat{\boldsymbol{\Sigma}}$ *is the reconstructed matrix by applying Singular Value Decomposition on a consistent estimator of* $\boldsymbol{\Sigma} = \boldsymbol{H} - \boldsymbol{\Lambda} \boldsymbol{J}_{\beta\beta}^{-1} \boldsymbol{\Lambda}^T$, $\hat{\boldsymbol{\Sigma}}^{-1}$ *denotes the Moore-Penrose pseudoinverse of* $\hat{\boldsymbol{\Sigma}}$, $k = rank(\hat{\boldsymbol{\Sigma}}) = rank(\boldsymbol{\Sigma})$. *Here* $\boldsymbol{J}_{\beta\beta} = \lim_{N\to\infty} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}/N$, $\boldsymbol{H} = \lim_{N\to\infty} \boldsymbol{F}\boldsymbol{V}\boldsymbol{F}^T$, *with*

$$\boldsymbol{F} = \frac{1}{\sqrt{N}} \begin{pmatrix} I_{\{\boldsymbol{x}_1 \in E_1\}} \cdots I_{\{\boldsymbol{x}_N \in E_1\}} \\ \vdots \\ I_{\{\boldsymbol{x}_1 \in E_L\}} \cdots I_{\{\boldsymbol{x}_N \in E_L\}} \end{pmatrix} \tag{2.16}$$

and

$$\boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_1^T \\ \vdots \\ \Lambda_L^T \end{pmatrix}_{L\times p} = \lim_{N\to\infty} \begin{pmatrix} \frac{1}{N} \sum_{k=1}^N I_{\{\boldsymbol{x}_k \in E_1\}} \mathbf{x}_k^T \\ \vdots \\ \frac{1}{N} \sum_{k=1}^N I_{\{\boldsymbol{x}_k \in E_L\}} \mathbf{x}_k^T \end{pmatrix}. \tag{2.17}$$

Remark 2.11 *The detailed steps used in deriving the matrix* $\hat{\boldsymbol{\Sigma}}$ *with small-eigenvalue eigenspaces project to* 0 *are exactly analogous to those described after Corollary 2.7 in Section 2.2.1.1. For the special case of a 2-level LMM (2.3), an explicit form of* $\boldsymbol{H}$, *which also follows from (2.16), and two consistent estimators were provided in Section 2.2.1.1 under the more restrictive assumption that* $\boldsymbol{x}_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i$ *are i.i.d. random variables and are independent of* n_i. *The explicit form*

of $\boldsymbol{H}$ *is also given for the 3-level LMM (2.18) in this Remark. Even if the alternate forms of the estimators for* $\boldsymbol{H}$ *and* $\boldsymbol{\Lambda}$ *are used, we still need to do the small-eigenvalue thresholding in defining* $\hat{\boldsymbol{\Sigma}}$.

For the 3-level hierarchical block nested LMM

$$y_{ijt} = \boldsymbol{x}_{ijt}^T\boldsymbol{\beta} + u_i + v_{ij} + \epsilon_{ijt},\ i = 1, \ldots, m;\ j = 1, \ldots, n_i;\ t = 1, \ldots, n_{ij}, \qquad (2.18)$$

the matrix $\boldsymbol{V}$ *has the structure*

$$\boldsymbol{V} = \begin{pmatrix} \begin{pmatrix} a & \cdots & b \\ \vdots & \ddots & \vdots \\ b & \cdots & a \end{pmatrix}_{n_{i1}\times n_{i1}} & \cdots & \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_1^2 \\ \vdots & \ddots & \vdots \\ \sigma_1^2 & \cdots & \sigma_1^2 \end{pmatrix}_{n_{i1}\times n_{i,n_{i1}}} \\ & \ddots & \\ \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_1^2 \\ \vdots & \ddots & \vdots \\ \sigma_1^2 & \cdots & \sigma_1^2 \end{pmatrix}_{n_{i1}\times n_{i,n_{i1}}} & \cdots & \begin{pmatrix} a & \cdots & b \\ \vdots & \ddots & \vdots \\ b & \cdots & a \end{pmatrix}_{n_{i,n_{i1}}\times n_{i,n_{i1}}} \end{pmatrix},$$

where $\sigma_1^2 = Var(u_i), \sigma_2^2 = Var(v_{ij}), \sigma_0^2 = Var(\epsilon_{ijt}), a = \sigma_0^2 + \sigma_1^2 + \sigma_2^2, b = \sigma_1^2 + \sigma_2^2$. *In this case,* $\forall\ l \neq k$, *the off-diagonal elements of* $\boldsymbol{H}$ *are*

$$\begin{aligned} \boldsymbol{H}_{lk} &= \sigma_1^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n_i} \sum_{h=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijh} \in E_l\}} \right) \left(\sum_{j=1}^{n_i} \sum_{h=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijh} \in E_k\}} \right) \right] \\ &+ \sigma_2^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\sum_{h=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijh} \in E_l\}} \right) \left(\sum_{h=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijh} \in E_k\}} \right). \end{aligned}$$

For $l = 1, \cdots, L$, the diagonal elements of $\boldsymbol{H}$ are

$$\begin{aligned}\boldsymbol{H}_{ll} &= \sigma_0^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijk} \in E_l\}} + \sigma_1^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijk} \in E_l\}} \right)^2 \\ &+ \sigma_2^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\sum_{k=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijk} \in E_l\}} \right)^2 .\end{aligned}$$

Similar to the 2-level LMM, under the assumption that the covariate vectors are i.i.d. random variables and the cluster sizes are independent of the covariate vectors, the $\boldsymbol{H}_{lk}$ and $\boldsymbol{H}_{ll}$ here can be expressed as functions of moments of n_1, n_{11} and $\boldsymbol{x}_{111}$. This can be done by applying Law of Large Numbers theory and by taking iterated conditional expectations first conditioning on $\{n_1, n_{11}\}$, which is similar to what was done in (2.9), (2.10). Because of the complexity of these functions in the 3-level LMM model (2.18), we recommend directly estimating $\boldsymbol{H}_{lk}$ as

$$\begin{aligned}\hat{\boldsymbol{H}}_{lk} &= \hat{\sigma}_1^2 \frac{1}{N} \sum_{i=1}^{m} \left[\left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijk} \in E_l\}} \right) \left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijk} \in E_k\}} \right) \right] \\ &+ \hat{\sigma}_2^2 \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\sum_{k=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijk} \in E_l\}} \right) \left(\sum_{k=1}^{n_{ij}} I_{\{\boldsymbol{x}_{ijk} \in E_k\}} \right) .\end{aligned}$$

This also applies to estimating $\boldsymbol{H}_{ll}$.

□

2.2.2 Test statistic and its asymptotic properties for two-level LMM with parameters estimated by least squares and method of moments

We consider the LMM (2.3), but now only require that $E(\alpha_i) = E(\epsilon_{ij}) = 0, Var(\alpha_i) = \sigma_a^2, Var(\epsilon_{ij}) = \sigma_\epsilon^2$, instead of assuming normality of α_i and ϵ_{ij}. The covariate vectors $\mathbf{x}_{ij}$ are considered to be random variables in this Section and $(\mathbf{x}_i, n_i)$ are assumed to be i.i.d. This model, also called Nested Error Regression Model, is widely used and studied in small area estimation (Prasad and Rao, 1990).

We estimate $\boldsymbol{\beta}$ by the generalized least squares estimator

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (\mathbf{X}^T\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\tilde{\mathbf{V}}^{-1}\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}) + o_p(1), \quad \text{as } N \to \infty, \end{aligned}$$

where $\mathbf{V}$ is a function of the variance components $\boldsymbol{\psi} = (\sigma_a^2, \sigma_\epsilon^2)$, which are estimated by the method of moments by equating the right-hand sides of

$$E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} - \bar{\mathbf{x}}_{i.}^T\boldsymbol{\beta}\right)^2 + (N-m)\sigma_\epsilon^2 \tag{2.19}$$

$$E\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2\right] = \sum_{i=1}^{m} n_i\left(\bar{\mathbf{x}}_{i.}^T\boldsymbol{\beta} - \bar{\mathbf{x}}_{..}^T\boldsymbol{\beta}\right)^2 + \left(N - \frac{1}{N}\sum_{i=1}^{m} n_i^2\right)\sigma_a^2 + (m-1)\sigma_\epsilon^2 \tag{2.20}$$

respectively to their estimates

$$\text{SSW} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$$

and

$$\text{SSB} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{m} n_i\bar{y}_{i\cdot}^2 - N\bar{y}_{\cdot\cdot}^2$$

respectively, where SSW is the Sum of Squares Within groups and SSB is the Sum of Squares Between groups in the analysis of variance. Because different clusters (over index i) are independent and the three quantities $\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$, $n_i\bar{y}_{i\cdot}^2$ and $\sum_{j=1}^{n_i} y_{ij}$ have finite second moments, SSW/m and SSB/m satisfy Laws of Large Numbers. Equations (2.19), (2.19) and (2.20) are estimating equations for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_a^2, \sigma_\epsilon^2)$, which can be solved iteratively. The solutions of equations (2.19), (2.19) and (2.20) $\tilde{\boldsymbol{\theta}}$, is consistent.

We first divide the covariate space into L disjoint regions $E_1, \ldots, E_L$ and compute the observed and expected values in each cell E_l as given in (2.5) and (2.6). We then state our goodness of fit test in Theorem 2.12 below with details of the proof in Section 2.6.4.

Theorem 2.12 *For the LMM (2.3) with finite second moments for both α_i and ϵ_{ij}, under Assumptions 2.1 and 2.2, $(\boldsymbol{f}-\boldsymbol{e}(\tilde{\boldsymbol{\beta}}))/\sqrt{N} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{H}-\boldsymbol{\Lambda}\boldsymbol{J}_{\beta\beta}^{-1}\boldsymbol{\Lambda}'$. Thus $(\boldsymbol{f} - \boldsymbol{e}(\tilde{\boldsymbol{\beta}}))'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{f} - \boldsymbol{e}(\tilde{\boldsymbol{\beta}}))/N \xrightarrow{\mathcal{D}} \chi_k^2$, where $\hat{\boldsymbol{\Sigma}}$ is the reconstructed matrix by applying Singular Value Decomposition on a consistent estimator of $\boldsymbol{\Sigma}$ and $k = rank(\hat{\boldsymbol{\Sigma}})$.*

Based on Corollary 5.3, $k = \text{rank}(\hat{\boldsymbol{\Sigma}}) = \text{rank}(\boldsymbol{\Sigma})$ for large N. Detailed steps used in defining a reconstructed matrix $\hat{\boldsymbol{\Sigma}}$ with all eigenvalues lower-bounded away from 0 are exactly analogous to those following Corollary 2.7 in Section 2.2.1.1. The matrices $\mathbf{J}_{\beta\beta}$, $\boldsymbol{\Lambda}$ and $\mathbf{H}$ are the same as for the two-level LMM (2.3) where normality was assumed for both α_i and ϵ_{ij} and MLEs are used, with formulas given in (2.9), (2.10), (2.11) and (2.12).

2.2.3 Power of the test

For the multi-level LMM (2.2), we derive the theoretical power under local, and more specifically under contiguous alternatives for the test in (2.13) in the situation where some covariates that influence the outcome y have been omitted from model (2.2). Let $\mathbf{X}$ be the true $N \times p$ covariate matrix and $\mathbf{X}^*$ be a submatrix of $\mathbf{X}$ of dimension $N \times p^*$ used in model (2.2), with $p^* < p$. Let the null hypothesis be $H_0 : \boldsymbol{\theta}_N = \boldsymbol{\theta}_0$. We assess the power of T under the alternative

$$H_1 : \boldsymbol{\theta}_N = \boldsymbol{\theta}_0 + \frac{\mathbf{a}}{\sqrt{N}}, \tag{2.21}$$

with $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\psi}_0)$, where several components of $\boldsymbol{\beta}_0$ are 0. We use the vector $\boldsymbol{\beta}_0^*$ to denote the non-zero components of $\boldsymbol{\beta}_0$. The indexing of $\boldsymbol{\beta}_0^*$ as a sub-vector of $\boldsymbol{\beta}_0$ corresponds to the same index subset as the columns of $\mathbf{X}^*$ within $\mathbf{X}$.

Based on the derivation for Theorem 2.10, we have that under H_0,

$$\frac{1}{\sqrt{N}}(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\beta}}^*)) \overset{H_0}{\to} N(\mathbf{0}, \boldsymbol{\Sigma}^*),$$

where $\mathbf{\Sigma}^*$ is the limiting variance covariance matrix of $(\mathbf{f}-\mathbf{e}(\hat{\boldsymbol{\beta}}^*))/\sqrt{N}$. By checking the condition (2.14) in Le Cam's third lemma in Section 2.6.5, we find that under the alternative hypothesis H_1 in (2.21),

$$\frac{1}{\sqrt{N}}(\mathbf{f}-\mathbf{e}(\hat{\boldsymbol{\beta}}^*)) \stackrel{H_1}{\to} N(\boldsymbol{\tau}, \mathbf{\Sigma}^*),$$

where

$$\boldsymbol{\tau} = \lim_{N\to\infty}\{\mathbf{\Lambda} - \mathbf{\Lambda}^*[(\mathbf{X}^*)^T V^{-1}\mathbf{X}^*]^{-1}[(\mathbf{X}^*)^T V^{-1}\mathbf{X}]\}\mathbf{a}, \tag{2.22}$$

with $\mathbf{\Lambda}$ given by expression (2.17) and $\mathbf{\Lambda}^*$ corresponds to the same expression, but computed using $\mathbf{X}^*$ and $\boldsymbol{\beta}^*$.

Thus under H_1, T^* has a limiting noncentral χ^2 distribution

$$T^* = \frac{1}{N}[\mathbf{f}-\mathbf{e}(\hat{\boldsymbol{\beta}}^*)]^T(\mathbf{\Sigma}^*)^{-1}[\mathbf{f}-\mathbf{e}(\hat{\boldsymbol{\beta}}^*)] \stackrel{H_1}{\to} \chi^2_k(\delta), \tag{2.23}$$

where $k = \text{rank}(\mathbf{\Sigma}^*)$ and the non centrality parameter is $\delta = \boldsymbol{\tau}^T(\mathbf{\Sigma}^*)^{-1}\boldsymbol{\tau}$. For a given type I error level α, the power is thus $P(T^* > \chi^2_{k,\alpha})$, where $\chi^2_{k,\alpha}$ is the $1-\alpha$ quantile of the central χ^2_k distribution and P denotes the non central $\chi^2_k(\delta)$ distribution. We then substitute all parameter values in $\mathbf{\Sigma}^*$ with their MLEs and reconstruct this consistent estimator of $\mathbf{\Sigma}$ by applying Singular Value Decomposition as explained in the context after Corollary 2.7 in Section 2.2.1.1. With $\hat{\mathbf{\Sigma}}^*$ denoting the reconstructed matrix,

$$\hat{T^*} = \frac{1}{N}[\mathbf{f}-\mathbf{e}(\hat{\boldsymbol{\beta}}^*)]^T(\hat{\mathbf{\Sigma}}^*)^{-1}[\mathbf{f}-\mathbf{e}(\hat{\boldsymbol{\beta}}^*)] \stackrel{H_1}{\to} \chi^2_k(\hat{\delta}), \tag{2.24}$$

where $k = \text{rank}(\hat{\boldsymbol{\Sigma}}) = \text{rank}(\boldsymbol{\Sigma})$ for large N, based on Corollary 5.3.

We now compute $\boldsymbol{\tau}$ and $\boldsymbol{\Sigma}^*$ explicitly for two-level LMMs for the setting of three covariates $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})$ that are from a multivariate normal distribution (2.25) (as studied in the simulation Section 2.3.1), where $\mathbf{x}_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i$ are i.i.d., and $\mathbf{x}_{ij}$ and n_i are independent. We assume $\mathbf{Y} \sim N(\mathbf{X}^T\boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ and $\mathbf{V}$ is given in (2.4), but then omit $\mathbf{x}_3$ in fitting the model, leading to $\mathbf{X}^* = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2)$. Here $\mathbf{a}$ in $\boldsymbol{\tau}$ is equal to $(0, 0, 0, \beta_3)$. With derivation details given in Sections 2.6.5.1 and 2.6.5.2,

$$
\begin{aligned}
\boldsymbol{\Sigma}^* &= \lim_{N\to\infty} \frac{1}{N}(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}^* \\
&= \frac{1}{\sigma_\epsilon^2}\begin{bmatrix} 1 & Ex_1 & Ex_2 \\ Ex_1 & Ex_1^2 & E(x_1x_2) \\ Ex_2 & E(x_1x_2) & Ex_2^2 \end{bmatrix} - \frac{1}{E(n)}\frac{\sigma_a^2}{\sigma_\epsilon^2}\begin{bmatrix} c_1 & c_1Ex_1 & c_1Ex_2 \\ c_1Ex_1 & h_1 & h_2 \\ c_1Ex_2 & h_2 & h_3 \end{bmatrix},
\end{aligned}
$$

where $c_1 = E_n\big[n^2/(\sigma_\epsilon^2 + n\sigma_a^2)\big]$, $c_2 = E_n\big[n/(\sigma_\epsilon^2 + n\sigma_a^2)\big]$, $c_3 = c_1 - c_2$, $h_1 = c_2Ex_1^2 + c_3(Ex_1)^2$, $h_2 = c_2E(x_1x_2) + c_3Ex_1Ex_2$ and $h_3 = c_2Ex_2^2 + c_3(Ex_2)^2$.

And the component $(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}$ in $\boldsymbol{\tau}$ satisfies

$$\begin{aligned}
&\frac{1}{N}(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X} \\
\xrightarrow{P} \quad &\frac{1}{\sigma_\epsilon^2}\begin{bmatrix} 1 & Ex_1 & Ex_2 & Ex_3 \\ Ex_1 & Ex_1^2 & E(x_1x_2) & E(x_1x_3) \\ Ex_2 & E(x_1x_2) & Ex_2^2 & E(x_2x_3) \end{bmatrix} \\
- \quad &\frac{1}{E(n)}\frac{\sigma_a^2}{\sigma_\epsilon^2}\begin{bmatrix} c_1 & c_1Ex_1 & c_1Ex_2 & c_1Ex_3 \\ c_1Ex_1 & h_1 & h_2 & h_4 \\ c_1Ex_2 & h_2 & h_3 & h_5 \end{bmatrix},
\end{aligned}$$

where $h_4 = c_2E(x_1x_3) + c_3Ex_1Ex_3$ and $h_5 = c_2E(x_2x_3) + c_3Ex_2Ex_3$.

When the cell partition $E_1, \ldots, E_l$ is based on the omitted covariate $\mathbf{x}_3$, the elements of $\boldsymbol{\Lambda}$ in $\boldsymbol{\tau}$ are

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{x_{ij,3}\in E_l\}} &\xrightarrow{P} \int_{E_l} f_3(x_3)dx_3 \\
\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{x_{ij,3}\in E_l\}}x_{ij,1} &\xrightarrow{P} \int_{x_1}\int_{E_l} x_1 f_{(x_1,x_3)}(x_1,x_3)dx_3dx_1 \\
\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{x_{ij,3}\in E_l\}}x_{ij,3} &\xrightarrow{P} \int_{E_l} x_3 f_3(x_3)dx_3.
\end{aligned}$$

With $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ jointly normal, which is the Scenario I in Section 2.3.1, $F_3(x_3)$ and $f_{(x_1,x_3)}(x_1, x_3)$ are the corresponding normal and bivariate normal densities.

Based on the above quantities, we next study the impact of the magnitude of the variance components σ_a^2 and σ_ϵ^2 and the correlations ρ_{13} and ρ_{23} in (2.25) between the omitted covariate x_3 and the covariates in the model (x_1 and x_2) on

the theoretical power when the cell partition is based on quantiles of the omitted covariate x_3 with $L = 8$ cells. For $\rho_{13} = .5$ and $\rho_{23} = .6$, Figure 2.1 plots the theoretical power against $\beta_3/(\sigma_a^2 + \sigma_\epsilon^2)$ for three choices of $(\sigma_a^2, \sigma_\epsilon^2)$ and varying β_3 on the x-axis. For any fixed pair of $(\sigma_a^2, \sigma_\epsilon^2)$, the power of the test not surprisingly increases as a function of β_3, the coefficient of the omitted covariate x_3. This observation can be made by a Taylor Expansion to the theoretical power formula. Let $G(x, \delta) = P_{\chi^2_{k,\ \delta}}([\chi^2_{k,\ \alpha}, \infty))$ be the theoretical power, then by a Taylor Expansion to $G(x, \delta)$ around $\delta = 0$, we get

$$G(x, \delta) \approx G(x, \delta) + \frac{\partial}{\partial \delta} G(x, \delta)\ \delta.$$

Thus the theoretical power $G(x, \delta)$ for δ close to zero is approximately a linear function of $\delta = \boldsymbol{\tau}^T (\boldsymbol{\Sigma}^*)^{-1} \boldsymbol{\tau}$, which is a function of β_3^2.

For any fixed β_3, the power increases when the random effect σ_a^2 decreases compared to the error term σ_ϵ^2. Figure 2.2 plots the power for fixed $\sigma_a^2 = 1$, $\sigma_\epsilon^2 = .25$ and different choices of (ρ_{13}, ρ_{23}). It shows that the power of test increases as $\rho_{13}^2 + \rho_{23}^2$ decreases. When we set $\rho_{13} = 0$ and $\rho_{23} = 0$, that is, when x_3 is correlated with neither x_1 nor x_2, we can see from Figure 2.3 that the power is not affected by the individual values of σ_a^2 and σ_ϵ^2, as long as $\sigma_a^2 + \sigma_\epsilon^2$ is fixed.

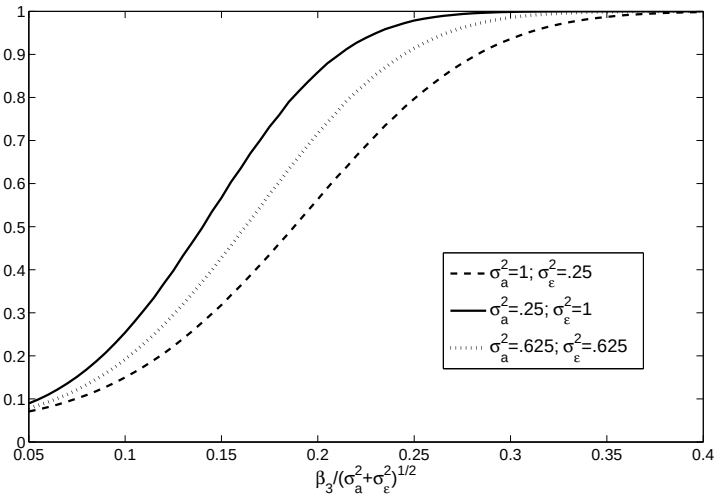

Figure 2.1: The impact of $(\sigma_a^2, \sigma_\epsilon^2)$ on analytical power (Scenario I, LMM)

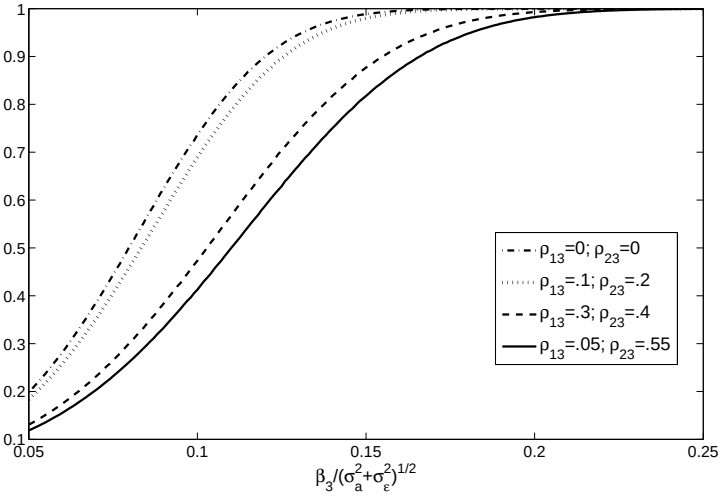

Figure 2.2: The impact of (ρ_{13}, ρ_{23}) on analytical power (Scenario I, LMM)

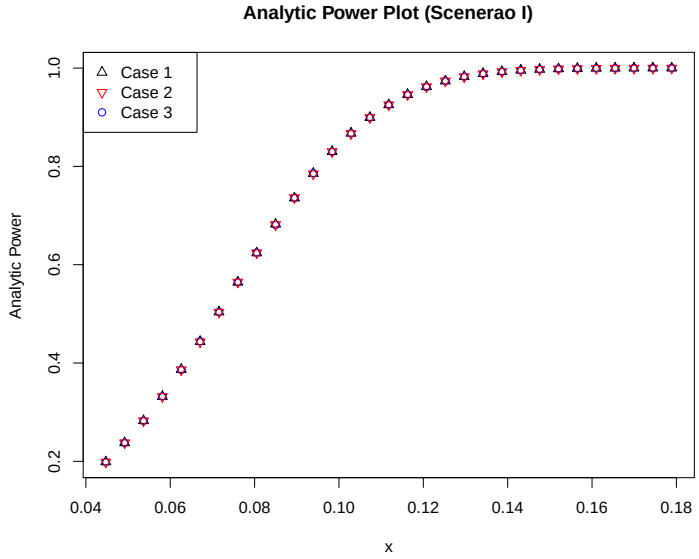

Figure 2.3: Analytical power plot for $\rho_{31} = \rho_{32} = 0$, x-axis is $\beta_3/(\sigma_a^2 + \sigma_\epsilon^2)$. Case 1: $\sigma_a^2 = .25, \sigma_\epsilon^2 = 1$; Case 2: $\sigma_a^2 = .625, \sigma_\epsilon^2 = .625$; Case 3: $\sigma_a^2 = 1, \sigma_\epsilon^2 = .25$.

2.3 Simulations

2.3.1 Normally distributed covariates (Scenario I)

We simulate data from the following setting. For $m = 500$, we first generate cluster sizes $n_i \sim$ uniform on $\{2, 3, 4, 5\}$ and compute $N = \sum_{i=1}^{m} n_i$. Thus the total number of observations in each repetition is always around $N = 1750$. We then draw N covariates $\mathbf{x}_{ij} = (x_{1ij}, x_{2ij}, x_{3ij})$, fixed in the sense that they are simulated only once for the entire simulation of $K = 1000$ repetitions, independently from a multivariate normal distribution,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim \mathbf{N} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \rho_{13} \\ 0 & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} \right). \tag{2.25}$$

Given $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ and parameters $\boldsymbol{\beta}$, σ_a and σ_ϵ, we generate $\mathbf{Y}$ from a multivariate normal distribution $\mathbf{Y} \sim N(\mathbf{X}'\boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{V}$ is given in (2.4).

We first do simulations to show that our goodness of fit test statistic (2.13) indeed has asymptotic χ^2 distribution. We choose $\rho_{13} = \rho_{23} = 0$ and set true parameter values $\sigma_a = 1$, $\sigma_\epsilon = .5$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, 1, .25)$. We then fit model (2.3) with all covariates that influence the response y. The number of cells L in the computation of T is twelve based on empirical quantiles of x_1 and x_2. With the number of iterations being 1000, we then have 1000 test statistic values $N^{-1}(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\beta}}))'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\beta}}))$. Figure 2.4 gives the histogram of these 1000 independently calculated test statistics which turns out to be close to χ^2_{12}, with p value from the Kolmogorov-Smirnov test being around .5 and p value from Pearson's chi-square goodness of fit test being around .27. The simulation result coincides with the theory we claim for (2.13).

Let α be the level of significance. We show in Table 2.1 the empirical size of the test, which is the proportion of iterations that have p values less than or equal to α. For example, the first row of Table 2.1 says that 4.7% of the 1000 simulations have p-values less than or equal to .05. The third column is the standard deviation of the corresponding empirical size (ES), which is calculated as $\sqrt{ES(1-ES)/1000}$.

We also check the size of the test under various choices of cell partition based on $\mathbf{X}$. We choose $\rho_{13} = \rho_{23} = 0$ in (2.25) and let $\sigma_a = 1$, $\sigma_\epsilon = .5$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, 1, 1)$, and fit model (2.3) with all covariates $\mathbf{X}$ in the model. Cell partitions in the computation of T are always based on empirical quantiles of each generated covariate matrix in each repetition. For example, the first row in Table 2.2 means

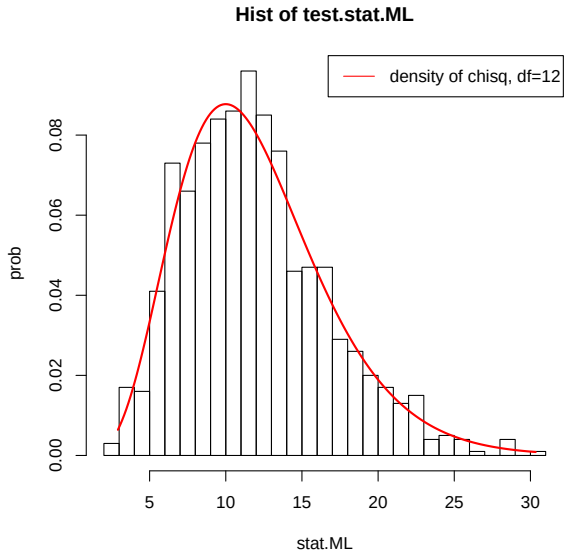

Figure 2.4: Empirical and Asymptotic distribution of the test statistic 2.13 in Scenario I with correlations $\rho_{13} = \rho_{23} = 0$.

Table 2.1: Empirical size of the test under different α levels (LMM) $m = 500, E(N) = 1750, \beta_3 = .25, \rho_{13} = \rho_{23} = 0, K = 1000$.

significance level α	Empirical Size(ES)	Standard Deviation of ES
0.05	0.047	0.0067
0.1	0.097	0.0094
0.2	0.189	0.0124
0.3	0.279	0.0142
0.4	0.385	0.0154
0.5	0.505	0.0158
0.6	0.604	0.0155
0.7	0.703	0.0144
0.8	0.797	0.0127

that in each of $K = 1000$ repetitions, the cell partition is based on the each generated x_1 with number of cells being 8. The second row in Table 2.2 means that in each of $K = 1000$ repetitions, the cell partition is based on both the generated x_1 and x_2 using cross tabulation. Table 2.2 shows empirical sizes (Emp. Size) were all close to the nominal α levels of 0.05 and 0.1 for all choices of cell partitions.

Table 2.2: Empirical size of the test under different cell partitions (LMM). $m = 500, E(N) = 1750, \beta_3 = 1, \rho_{13} = \rho_{23} = 0, K = 1000$.

L	α	Emp. Size	α	Emp. Size
8 (x_1)	0.05	0.056	0.1	0.097
3×4 (x_1, x_2)	0.05	0.058	0.1	0.109
5×4 (x_1, x_3)	0.05	0.048	0.1	0.088
6×7 (x_2, x_3)	0.05	0.050	0.1	0.097

To assess the power of the test, we fit model (2.3) to the data without including x_3 among the covariates. We then use a cell partition based on the empirical quantiles of the omitted x_3 with $L = 8$ cells when the number of clusters $m = 500$ or 50. For $m = 20$, we use fixed cell boundaries, based on theoretical quantiles of the distribution of x_3, to divide x_3 into $L = 8$ cells. We set $(\rho_{13}, \rho_{23}) = (.5, .6)$, $\sigma_a = 1$, $\sigma_\epsilon = .5$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, 1, .25)$. For a given design matrix $\mathbf{X}$, we compute the theoretical power (Theo.Pow.), the mean estimated theoretical power (Theo.Pow.hat), and the mean empirical power (Empi.Pow.n) for $K = 1000$ iterations. We compute the theoretical power of T^* in (2.23) based on the asymptotic χ^2 distribution with the true values of the variance components and empirical moments for $\mathbf{X}$ used in the calculation of the non-centrality parameter (2.22). We compare these values to the estimated theoretical power, that is computed based on the asymptotic χ^2 distribution with estimated variance components and empirical moments for $\mathbf{X}$ in (2.22). We repeat the power computations for $D = 500$ randomly generated matrices $\mathbf{X}$. Table 2.3 shows means and variances of the 500 distinct power estimates (each based on K=1000 iterations) varying over the design matrices for $m = 500, 50$ and $m = 20$ clusters. The theoretical power, the empirical power and the estimated theoretical power agree very well, even when m is small.

Table 2.3: Power and robustness study for Scenario I (LMM).
$L = 8, K = 1000, D = 500, (\rho_{13}, \rho_{23}) = (.5, .6), \sigma_a = 1, \sigma_\epsilon = .5$.

Power	$m = 500, EN = 1750$		$m = 50, EN = 175$		$m = 20, EN = 70$	
$\beta_3 = .25$	mean	stan.dev.	mean	stan.dev.	mean	stan.dev.
Theo.Pow.	.798	.0392	.122	.0236	.084	.0175
Theo.Pow.hat	.796	.0383	.127	.0241	.089	.0184
Empi.Pow.n	.796	.0388	.113	.0227	.062	.0181
Empi.Pow.t_3	.797	.0383	.113	.0234	.061	.0184
Empi.Pow.t_5	.797	.0375	.113	.0237	.061	.0180
Power	$m = 500, EN = 1750$		$m = 50, EN = 175$		$m = 20, EN = 70$	
$\beta_3 = .8$	mean	stan.dev.	mean	stan.dev.	mean	stan.dev.
Theo.Pow.	1	0	.853	.0967	.533	.1842
Theo.Pow.hat	1	0	.827	.0895	.506	.1443
Empi.Pow.n	1	0	.827	.0954	.439	.1586
Empi.Pow.t_3	1	0	.828	.0942	.443	.1629
Empi.Pow.t_5	1	.0001	.827	.0941	.440	.1591

However, only for $m = 500$ was there adequate power to detect lack of fit when $\beta_3 = 0.25$, which was substantially smaller than the coefficients $\beta_1 = \beta_2 = 1$ of x_1, and x_2, the covariates included in the model. When the effect of the omitted covariate was larger, $\beta_3 = 0.8$, the test statistic had approximately 80% power even for $m = 50$ clusters.

2.3.1.1 Impact of choice of the cell partition on power

As is true for Pearson's chi-square test, the choice of cell partition plays an important role for our goodness-of-fit test. We now illustrate the impact of the cell partition on the power of our test.

We let $\rho_{13} = .5$ and $\rho_{23} = 0$ and set $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, 1, .15)$, $\sigma_a = 1$ and $\sigma_\epsilon = .5$, for different choices of cell partitions. Again, we generate y from a model whose proper specification includes x_1, x_2 and x_3 but then analyze the

Table 2.4: Impact of cell partition on empirical power for Scenario I (LMM). $m = 500, \beta_3 = .15, \sigma_a = 1, \sigma_\epsilon = .5, K = 1000$.

Parti	$\rho_{13} = 0, \rho_{23} = 0$		$\rho_{13} = 0.2, \rho_{23} = 0.3$		$\rho_{13} = 0.4, \rho_{23} = 0.5$	
	$L = 12$	$L = 42$	$L = 12$	$L = 42$	$L = 12$	$L = 42$
x_1	0.059	0.044	0.055	0.055	0.058	0.051
x_2	0.052	0.044	0.044	0.046	0.041	0.053
x_3	0.991	0.871	0.907	0.708	0.539	0.292
x_1, x_2	0.041	0.044	0.053	0.047	0.060	0.053
x_1, x_3	0.962	0.860	0.922	0.737	0.566	0.360
x_2, x_3	0.961	0.871	0.889	0.723	0.569	0.368

results of omitting x_3 in the subsequent model fitting. We choose six different cell partitions: partitions based on only x_1, only x_2, only x_3, both x_1 and x_2, both x_1 and x_3, or both x_2 and x_3. For all cell partitions, we use empirical quartiles based on data to divide the covariates. The number of replications in our simulation study is $K = 1000$. The results in Table 2.4 show that a lack of fit is detecTable by our test statistic only when the cell partition involves the omitted covariate, x_3. Similar results are observed when x_3 is independent of x_1 and x_2.

2.3.1.2 Robustness of T with respect to error distribution

In Table 2.3 we also assessed the impact of misspecification of the error distribution on the power of the test statistic. Using the same setting as in the power calculations given above, we generated ϵ from a t distribution with $k = 3$ or 5 degree of freedom instead of from a $N(0, \sigma_\epsilon^2)$. We rescaled the variance of ϵ so that $Var(\sigma_\epsilon t_k \sqrt{k-2}/\sqrt{k}) = \sigma_\epsilon^2$ to ensure that the noise had the same variance as in the normal case. The power of the test under a t-distribution is virtually the same as the power of the test with a normal error distribution. For example, for $m = 50$

the power was 0.83 for the normal error distribution and for t-distributions with 3 and 5 degrees of freedom (Table 2.3). Thus, by comparing the last three rows of Table 2.3, we can see that our test is very robust with respect to symmetric error distributions.

2.3.1.3 A summary parameter related to the power

The power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is not true. The rejection ratio in Table 1, which is defined as total number of iterations with p value $< .05$ over the total number of simulation iterations, can be considered as an estimator of the power for our goodness-of-fit test. We performed many simulations to show that the rejection ratio is closely related to a summary parameter Δ which is defined as the ratio of two variances. Suppose the true model is $y_{ij} = h(x_{1ij}, x_{2ij}, x_{3ij}) + \alpha_i + \epsilon_{ij}$, where the true covariates that impact y are x_1, x_2 and x_3. But we fit the data only using covariates x_1 and x_2. Then the summary parameter

$$\Delta = \frac{Var[h(x_1, x_2, x_3) - \prod_{x_1,x_2} h(x_1, x_2, x_3)]}{Var(\alpha_i + \epsilon_{ij})}, \tag{2.26}$$

where $\prod_{x_1,x_2} h(x_1, x_2, x_3) = \hat{E}(h(X_1, X_2, X_3)|X_1, X_2)$ is the linear projection.

One simulation scenario we considered is similar to the one we already explained in Section 2.3.1, where both the random effect α_i and the error term ϵ_{ij} are generated from the normal distribution. The true model is defined to have covariate

vector generated from the multivariate normal distribution:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 12 \\ 2 \\ 24 \end{bmatrix}, \begin{bmatrix} 1 & 0 & .5 \\ 0 & .5 & .8 \\ .5 & .8 & 2 \end{bmatrix} \right).$$

Under this joint distribution, one verifies that $x_3|x_1, x_2 \sim \mathcal{N}(.5x_1+1.6x_2+14.8, 0.47)$. But we fit the data only using x_1 and x_2, omitting x_3. Then the summary parameter (2.26) is simplified to be $\beta_3^2 Var(x_3|x_1, x_2)/Var(\epsilon) = .47\beta_3^2/(\sigma_a^2 + \sigma_\epsilon^2)$. With the covariate set $\mathbf{x}$ fixed for each iteration, we change the magnitude of the coefficient vector $\boldsymbol{\beta}$ and the two variance components σ_a^2 and σ_ϵ^2. For each set of parameters $\{\boldsymbol{\beta}, \sigma_a^2, \sigma_\epsilon^2\}$, there is a corresponding summary parameter Δ, which we plotted on the x-axis. We then ran simulations with 1000 iterations to get the corresponding rejection ratio, which we plotted on the y-axis. The simulation results in Figure 2.5 indicate that the empirical power of the test increases as the summary parameter Δ increases and this is true for three different cell partitions. As we can see from the graph, there is a nearly linear relationship between the estimated power and the defined summary parameter Δ and we can get the slope from the Taylor expansion of the theoretical power around 0.

We then compared what we found here with what we saw by using the analytic power formula for the special case when (x_1, x_2, x_3) comes from a multivariate normal (2.25), the simulation scenario we carefully studied in Section 2.3.1. In this case, the numerator of Δ in (2.26) becomes $Var(x_3|x_1, x_2) = 1 - (\rho_{13}^2 + \rho_{23}^2)$. Thus,

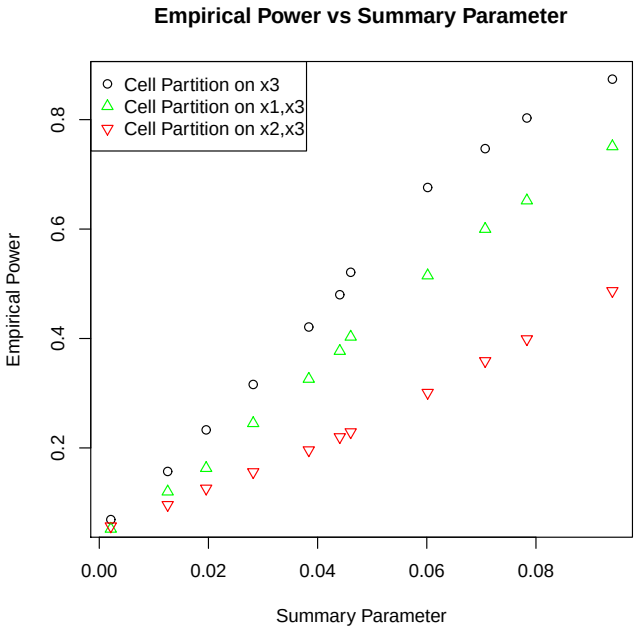

Figure 2.5: Empirical power vs summary parameter Δ

$\Delta = (1 - (\rho_{13}^2 + \rho_{23}^2))/(\sigma_a^2 + \sigma_\epsilon^2)$. This agrees with what we found from Figure 2.2, i.e. the power of test increases as $\rho_{13}^2 + \rho_{23}^2$ decreases. However, this Δ doesn't include other information we got in Section 2.3.1, such as the way that the power increases as β_3 increase and the relationship of the power to the relative magnitude of σ_a^2 and σ_ϵ^2, given fixed $\sigma_a^2 + \sigma_\epsilon^2$.

2.3.2 Normally distributed interacting covariates (Scenario II)

We again generate y from a linear model that depends on three covariates x_1, x_2 and x_3. However, we now let x_1 and x_2 arise from a bivariate normal distribution with mean $(1, 0.5)$, with variances $\sigma_1^2 = \sigma_2^2 = 1$ and covariance $\rho_{12} = 0$, and define x_3 as their product, i.e. $x_3 = x_1 x_2$. Again, we choose $m = 500$ and generate the cluster size n_i from a uniform distribution on $\{2, 3, 4, 5\}$. We let $\sigma_a = 1$, $\sigma_\epsilon = .5$,

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, 1, .2)$ and use empirical quantiles of the covariates to define the cell partition. First, we fit model (2.3) with all covariates $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ that influence the response y and check the size of the test. Again, the empirical sizes were all close to the nominal α level under different choices of cell partitions, as shown in Table 2.5. We then fit model (2.3) without x_3 and study the power of

Table 2.5: Empirical size of the test for Scenario II under different cell partitions with standard deviations in brackets (LMM).
$m = 500, E(N) = 1750, \beta_3 = .2, \sigma_a = 1, \sigma_\epsilon = .5, \rho_{12} = 0, K = 1000.$

L	α	Emp. Size	α	Emp. Size
$8(x_1)$	0.05	0.048 (0.0068)	0.1	0.102 (0.0096)
$3\times4(x_1, x_2)$	0.05	0.045 (0.0066)	0.1	0.085 (0.0088)
$5\times4(x_1, x_3)$	0.05	0.060 (0.0075)	0.1	0.098 (0.0094)
$6\times7(x_2, x_3)$	0.05	0.051 (0.0070)	0.1	0.100 (0.0095)

the test using different cell partitions. Results in Table 2.6 indicate that the test has adequate power only when the cell partition is based on x_1 and x_2, or on the omitted interaction term $x_3 = x_1x_2$, but not if the cell partition is based on either x_1 or x_2 alone, for $\rho_{12} = 0$ and $\rho_{12} = .3$.

We next study the impact of the magnitude of the variance components σ_a^2 and σ_ϵ^2 on the theoretical power when the cell partition is based on the omitted covariate

Table 2.6: Impact of cell partition on empirical power for Scenario II (LMM).
$m = 500, E(N) = 1750, \beta_3 = .2, \sigma_a = 1, \sigma_\epsilon = .5, K = 1000.$

Partition	$\rho_{12} = 0$		$\rho_{12} = 0.3$	
	$L = 12$	$L = 42$	$L = 12$	$L = 42$
x_1	0.049	0.049	0.256	0.182
x_2	0.038	0.041	0.273	0.173
x_3	0.893	0.771	0.859	0.749
x_1, x_2	0.991	0.966	0.989	0.975
x_1, x_3	0.843	0.938	0.885	0.939
x_2, x_3	0.936	0.912	0.956	0.928

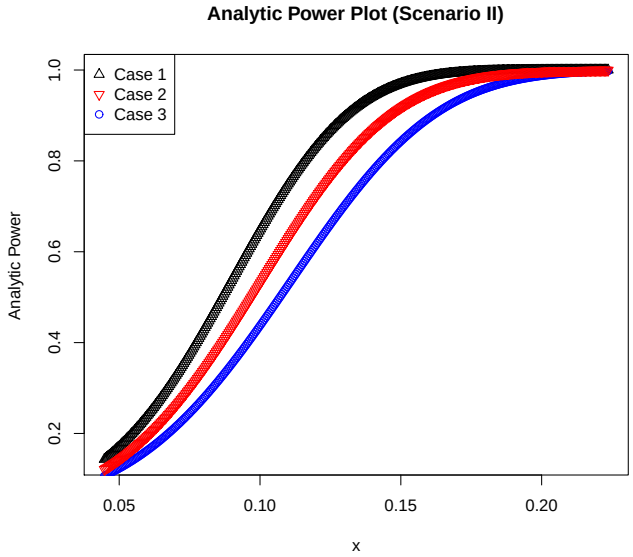

Figure 2.6: The impact of $(\sigma_a^2, \sigma_\epsilon^2)$ on analytical power, $\rho_{12} = 0$, x-axis is $\beta_3/(\sigma_a^2 + \sigma_\epsilon^2)$. Case 1: $\sigma_a^2 = .25, \sigma_\epsilon^2 = 1$; Case 2: $\sigma_a^2 = .625, \sigma_\epsilon^2 = .625$; Case 3: $\sigma_a^2 = 1, \sigma_\epsilon^2 = .25$.

x_3 with $L = 8$ cells using fixed cell boundaries. For $\rho_{12} = 0$, Figure 2.6 plots the theoretical power against $\beta_3/(\sigma_a^2 + \sigma_\epsilon^2)$ for three choice of $(\sigma_a^2, \sigma_\epsilon^2)$ and varying β_3 on the x-axis. Our conclusions are consistent with what we saw in Section 2.3.1. For any fixed pair of $(\sigma_a^2, \sigma_\epsilon^2)$, the power of the test increases as a function of β_3, the coefficient of the omitted covariate x_3. For any fixed β_3, the power increases when the random effect σ_a^2 decreases compared to the error term σ_ϵ^2.

2.4 Data examples

2.4.1 Birth weight data

These data were obtained from the book [19] and consist of fetal birth weights of 432 boys from 108 families of size 4, and BMI of the mother, age of the mother,

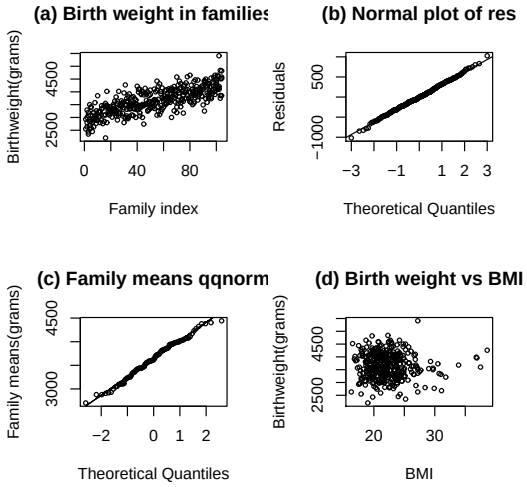

Figure 2.7: Normality assumption checking for birth weight data.

birth order, gestation time and a family indicator. Lee et al. [19] analyzed this data set using a linear mixed model with only an intercept term. To use all covariates, we excluded individuals who had missing covariates and modified possible recording errors on the covariate order. The complete data set consists of 104 families with 370 individuals. Family sizes differed for different families, but all family sizes were four or less. We repeated the graphical analysis in Chapter 5 of [19] on the modified data and got similar results. There is a strong familial effect of birth weight and the within-family variation is normally distributed. We conclude that the weight data satisfy assumptions for (2.3). We tried various covariate combinations for the fixed part in (2.3), including interactions. Based on the significance of the covariates and BIC (or AIC) criteria, the best model is the LMM with 3 covariates: intercept, mom.age and gestation. Residual plots further confirm the normality assumptions in our LMM. We then apply the goodness of fit test to our model. We use the cell partition based on the covariate order, which is not in our final model, and obtain

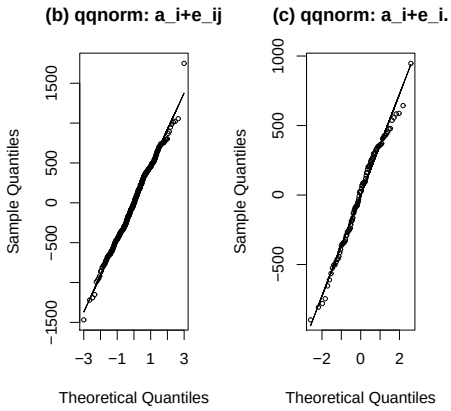

Figure 2.8: Normality checking for residuals of final model for birth weight data

the following number of observations in the cells: 61, 74, 79, 80, 31, 45. The value of the test statistic is $T = 2.44$, which corresponds to a p value of .87 for a chi-square distribution with 6 degrees of freedom. However, when we tested the fit of the intercept only model in [19] using the same cell partition, we also found that it had adequate fit to the data with $T = 4.97$. Thus both of these two models are judged adequate by our test.

2.4.2 Alcohol data

These data come from a Women's Alcohol Study [8], where 53 healthy, non-smoking postmenopausal women completed a random-order, three-period (8-week treatment for each period) study with a crossover design in which each woman received 0, 15 or 30 g of alcohol per day. Participants were not told the amount of alcohol they were consuming and each controlled feeding period was preceded by a two to five week washout period during which time the participant consumed no alcohol. During the controlled feeding period, all food and beverages for the

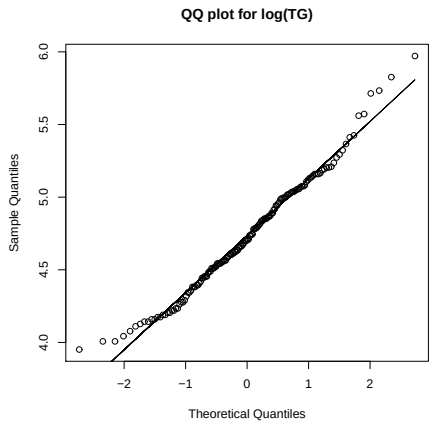

Figure 2.9: Normality checking for log(TG) for Alcohol data.

participants were supplied by the study investigators.

For each woman in the data set, three blood measurements for the three periods are recorded, corresponding to the three randomized alcohol intake periods. We assessed the association of Plasma Triglycerides level (TG) with alcohol intake. Other relevant covariates were race, age, height, weight and BMI (Body Mass Index). A log transformation of TG met the normality assumption by a formal goodness of fit check on the response variable. We then tried all possible covariate combinations for the fixed-effect terms of (2.3), including interactions. Assessed by the significance of the covariates and BIC (or AIC) criteria, our final model included an intercept, age, BMI and alcohol. The residual plots further confirmed the normality assumptions in (2.3). To apply our goodness of fit test on this model, we use the cell partition based on the covariates height and weight, which are not in our final model, and obtain the following number of observations in the cells: $19, 19, 13, 17, 15, 11, 13, 15, 14, 22$. The value of the test statistic is $T = 8.98$, which corresponds to a p values of .53 for a chi-square distribution with 10 degrees of

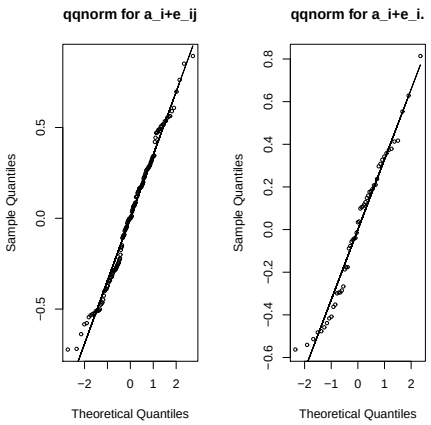

Figure 2.10: Normality checking for residuals of the final model for Alcohol data

freedom. Thus our final model fitted well and we concluded, not surprisingly, that alcohol intake affects Plasma Triglycerides levels.

2.4.3 Factors impacting thyroglobulin levels in an iodine deficient population

On April 26, 1986, an accident at the Chernobyl power plant located in northwestern Ukraine, close to the border with Belarus, released large amounts of radioactive materials including iodine-131 (I-131) into the atmosphere from the destroyed reactor. Deposition of these radioactive materials resulted in serious contamination of the territory and exposed its residents. Because the thyroid gland concentrates iodine, the doses to the thyroid due to consumption of I-131 contaminated milk were much greater than those to any other organs. The National Cancer Institute, NIH is involved in a cohort study in Belarusian individuals exposed to the accident [34]. While the main objective of the study is to evaluate the relationship between I-131

doses and risk of thyroid cancer, investigators were also interested in describing the levels of iodine in this population, as historical data suggest that the study area could be mildly iodine deficient and iodine deficiency impacts absorption of I-131.

We therefore evaluated the relationship between levels of serum thyroglobulin (TG), a sensitive marker of iodine deficiency, and patients' characteristics including age, sex, thyroid volume and other demographic and clinical variables that might reflect or influence the intake of dietary iodine and were identified in a initial screen of variables.

We restrict our example to men from four of the five study regions, who had complete information on the covariates. After excluding observations with $TG > 80$, $\log(TG)$ was normally distributed. The final dataset was comprised of $m = 933$ individuals, of whom 404 had a single TG measurement, 484 had two, 42 three and 3 four TG measurements during follow-up, resulting in a total of $N = 1510$ observations.

An initial screening of the variables, one at a time by simple linear regression, indicated that age at time of exam, age at time of the accident, rural or urban residence, smoking status, urinary iodine levels, serum thyroid-stimulating hormone (TSH) levels, serum anti-thyroglobulin antibody (ATG) levels, thyroid volume, presence of thyroid nodules, presence of goiter and presence of any thyroid abnormality may impact TG levels.

We fit various of the LMM (2.3) model, including combinations of those covariates and their interaction terms using PROC GLIMMIX (SAS 9.2). A model (Model 1) that included all the variables mentioned above, with the exception of

presence of nodules, and in addition included an interaction term of ATG levels with presence of any thyroid abnormality that was marginally significant (Wald p-value $p = 0.054$), had a log-likelihood value of 1625.2. The random effect variance estimate for Model 1 was $\sigma_a^2 = 0.29$ and the error variance estimate was $\sigma_\epsilon^2 = 0.25$. A second model (Model 2) that had no interaction term, but included presence of nodules, resulted in a log-likelihood of 1621.3. The variance component estimates were similar to model 1, $\sigma_a^2 = 0.29$ and $\sigma_\epsilon^2 = 0.27$. However, as models 1 and 2 are not nested, we could not compare them using a likelihood ratio test.

To assess the fit of Models 1 and 2, we formed the cell partition for the test based on $L = 8$ cells defined by quantiles of ATG and presence of any thyroid abnormality. Based on a chi-squared test statistic with eight degrees of freedom, there was no indication of lack of fit for either model, with corresponding p-values $p = 0.32$ and $p = 0.40$ for Models 1 and 2 respectively. We repeated the calculation of the test statistic for a second cell partition based on presence of nodules and presence of goiter, resulting in $L = 4$ cells, with corresponding p-values $p = 0.19$ and $p = 0.70$ for Models 1 and 2 respectively. These results suggested that both models provided an adequate fit to the data.

When we checked a third model, that included neither presence of nodules not the interaction term (log-likelihood 1621.5), there was also no evidence of lack of fit based on eight ($p = 0.39$) or 4 degrees of freedom chi square tests ($p = .21$). However, with a model that included all the covariates, the interaction term of ATG levels with presence of any thyroid abnormality but excluded presence of goiter, we did detect a lack of fit in the cell partition defined by presence of nodules and goiter,

with $p = 0.006$, while there was no evidence of lack of fit based on the $L = 8$ cell partition ($p = 0.28$). This highlights the importance of presence of any thyroid abnormality in the model. However, omitting the interaction term of this variable with ATG levels does not affect fit to the data.

2.5 Discussion

Schoenfeld (1980) presented a class of chi-squared goodness of fit tests for the proportional hazards regression model. We adapted this idea and proposed a class of goodness of fit tests for testing the statistical adequacy of a linear mixed model. We described the asymptotic properties of the test when parameters were estimated and developed its theoretical power under the local, or contiguous, alternative. We assessed factors that impact the power, the impact of choice of cell partitions on the test as well as the robustness of the test with respect to symmetric error distribution in simulations. We found that when a specific covariate that is associated with outcome is omitted, especially interaction terms or a covariate correlated with covariates already in the model, cell partitions based on the omitted covariate result in adequate power of the test. However, if the cell partition is based only on covariates already in the model, this test has no power to detect any lack of model fit. We also found that the estimated theoretical power calculated using Le Cam's third lemma was reliable at least when the number of clusters m is above 50. However, when m is very small, it may be advisable to rely on the empirical power computed through simulations. Our test was also very robust to violations of the normality

assumption of the error distribution.

This goodness of fit test can be used to test the statistical adequacy of the finally selected LMM in real application. The test statistic is very easy to implement, as all that is needed in order to apply the test are the final model parameter estimates and their variance covariance matrix, which are standard outputs from any statistical software. As a note of caution, in applying the test one needs to check the rank of the estimated variance covariance matrix $\hat{\boldsymbol{\Sigma}}$ in (2.13) to ensure the correct degrees of freedom for the test statistic.

In Chapter 3, we extend this test statistic to assess the fit of generalized linear mixed models.

2.6 Technical details for Chapter 2

2.6.1 Proof of Theorem 2.3

Proof: Let $\mathbf{J}$ be the limit of the sample information matrix per observation,

$$\mathbf{J} = \lim_{N\to\infty} \frac{1}{N} \begin{pmatrix} -\frac{\partial^2 l}{\partial\beta_i\partial\beta_j} & -\frac{\partial^2 l}{\partial\beta_i\partial\psi_j} \\ -\frac{\partial^2 l}{\partial\beta_i\partial\psi_j} & -\frac{\partial^2 l}{\partial\psi_i\partial\psi_j} \end{pmatrix} = \begin{bmatrix} \mathbf{J}_{\beta\beta} & \mathbf{J}_{\beta\psi} \\ \mathbf{J}_{\beta\psi}^T & \mathbf{J}_{\psi\psi} \end{bmatrix}. \tag{2.27}$$

The MLE $\hat{\boldsymbol{\theta}}$ in model (2.3) is consistent as follows from results by Miller (1977). By Taylor series expansion of the score function $\mathbf{S}(\hat{\boldsymbol{\theta}})$, we obtain

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx \left(-\frac{1}{N}\frac{\partial S(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)^{-1} \frac{1}{\sqrt{N}} S(\boldsymbol{\theta}_0) \approx \mathbf{J}^{-1}\frac{1}{\sqrt{N}} S(\boldsymbol{\theta}_0). \tag{2.28}$$

Under model (2.3), $\mathbf{J}_{\beta\psi} = 0$ in (2.27) (Wand 2007). The Fisher information is therefore a block diagonal matrix and $\mathbf{J}^{-1} = \begin{bmatrix} \mathbf{J}_{\beta\beta}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{\psi\psi}^{-1} \end{bmatrix}$. As $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{V})$, the score function for $\boldsymbol{\beta}$, which is the first p components of $S(\boldsymbol{\theta})$, is

$$S_\beta(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\beta})l(\boldsymbol{\theta}) = \mathbf{X}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

By extracting the first p components of (2.28), we have

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approx \mathbf{J}_{\beta\beta}^{-1}\frac{1}{\sqrt{N}}S_\beta(\boldsymbol{\theta}_0) = \mathbf{J}_{\beta\beta}^{-1}\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)/\sqrt{N}.$$

Thus,

$$\begin{aligned} \sqrt{N}\begin{pmatrix} (\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0))/N \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} &\approx \begin{pmatrix} N^{-1/2} \left[I_{\{\mathbf{x}_{11}\in E_1\}} \cdots I_{\{\mathbf{x}_{mnm}\in E_1\}}\right] \\ \vdots \\ N^{-1/2} \left[I_{\{\mathbf{x}_{11}\in E_L\}} \cdots I_{\{\mathbf{x}_{mnm}\in E_L\}}\right] \\ N^{-1/2}\ \mathbf{J}_{\beta\beta}^{-1}\mathbf{X}^T\mathbf{V}^{-1} \end{pmatrix} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0) \\ &= \mathbf{D}_{(L+p)\times N}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0), \end{aligned} \tag{2.29}$$

which is a linear combination of Gaussian random variables.

Therefore, as $N \to \infty$,

$$\sqrt{N}\begin{pmatrix} (\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0))/N \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \xrightarrow{D} N(\mathbf{0}, \mathbf{D}\mathbf{V}\mathbf{D}^T),$$

where

$$\mathbf{DVD}^T = \begin{pmatrix} \mathbf{H} & \mathbf{\Lambda}\mathbf{J}_{\beta\beta}^{-1} \\ \mathbf{J}_{\beta\beta}^{-1}\mathbf{\Lambda}^T & \mathbf{J}_{\beta\beta}^{-1} \end{pmatrix},$$

with

$$\mathbf{\Lambda} = \begin{pmatrix} \Lambda_1^T \\ \vdots \\ \Lambda_L^T \end{pmatrix}_{L\times p} = \lim_{N\to\infty} \begin{pmatrix} N^{-1}\sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_1\}}\mathbf{x}_{ij}^T \\ \vdots \\ N^{-1}\sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_L\}}\mathbf{x}_{ij}^T \end{pmatrix}, \tag{2.30}$$

and $\mathbf{H}$ is a symmetric matrix of dimension $L \times L$. For $l = 1, \cdots, L-1$; $k = l+1, \cdots, L$, the off-diagonal elements of $\mathbf{H}$ are

$$\mathbf{H}_{lk} = \sigma_a^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^m \left[\left(\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}} \right) \left(\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_k\}} \right) \right] \tag{2.31}$$

Similarly, for $l = 1, \cdots, L$, the diagonal elements of $\mathbf{H}$ are

$$\mathbf{H}_{ll} = \sigma_\epsilon^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}} + \sigma_a^2 \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^m \left(\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}} \right)^2. \tag{2.32}$$

Remark 2.13 *If $\boldsymbol{x}_{ij}$ are random variables, under Assumption 2.2,*

$\Lambda_l = E_{\{\boldsymbol{x}_1, n_1\}}(\sum_{j=1}^{n_1} I_{\{\boldsymbol{x}_{1j} \in E_l\}} \boldsymbol{x}_{1j})/E(n_1)$, *and*

$$\begin{aligned}
\boldsymbol{H}_{lk} &= \sigma_a^2 \cdot \lim_{m\to\infty} \frac{1}{\left(\sum_{i=1}^{m} n_i\right)/m} \cdot \frac{1}{m} \sum_{i=1}^{m} \left[\left(\sum_{s=1}^{n_i} I_{\{\boldsymbol{x}_{is} \in E_l\}} \right) \left(\sum_{t=1}^{n_i} I_{\{\boldsymbol{x}_{it} \in E_k\}} \right) \right] \\
&= \frac{\sigma_a^2}{E(n_1)} \cdot E_{(\boldsymbol{x}_1, n_1)} \left[\left(\sum_{s=1}^{n_1} I_{\{\boldsymbol{x}_{1s} \in E_l\}} \right) \left(\sum_{t=1}^{n_1} I_{\{\boldsymbol{x}_{1t} \in E_k\}} \right) \right] \\
&= \frac{\sigma_a^2}{E(n_1)} \cdot E_{n_1} \left\{ E_{\boldsymbol{x}_1} \left[\left(\sum_{s=1}^{n_1} I_{\{\boldsymbol{x}_{1s} \in E_l\}} \right) \left(\sum_{t=1}^{n_1} I_{\{\boldsymbol{x}_{1t} \in E_k\}} \right) \Big| n_1 \right] \right\}.
\end{aligned}$$

The expressions of $\boldsymbol{H}_{lk}$ *and* $\boldsymbol{H}_{ll}$ *depend on the joint distribution of* $\boldsymbol{x}_i$ *and* n_i*. Under the more restrictive assumption that* $\boldsymbol{x}_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i$ *are i.i.d. and are independent of* n_i*, then* $\boldsymbol{H}_{lk}$ *and* $\boldsymbol{H}_{ll}$ *can be further simplified as*

$$\begin{aligned}
\boldsymbol{H}_{lk} &= \frac{\sigma_a^2}{E(n_1)} \cdot E_{n_1} \left\{ n_1 E_{\boldsymbol{x}_{11}} \left(I_{\{\boldsymbol{x}_{11} \in E_l\}} I_{\{\boldsymbol{x}_{11} \in E_k\}} \right) + (n_1^2 - n_1) E \left(I_{\{\boldsymbol{x}_{11} \in E_l\}} \right) E \left(I_{\{\boldsymbol{x}_{11} \in E_k\}} \right) \right\} \\
&= \sigma_a^2 \cdot \frac{E(n_1^2 - n_1)}{E(n_1)} P \left(I_{\{\boldsymbol{x}_{11} \in E_l\}} \right) P \left(I_{\{\boldsymbol{x}_{11} \in E_k\}} \right)
\end{aligned}$$

and

$$\boldsymbol{H}_{ll} = (\sigma_\epsilon^2 + \sigma_a^2) E\big(I_{\{\boldsymbol{x}_{11} \in E_l\}}\big) + \sigma_a^2 \frac{E(n_1^2 - n_1)}{E(n_1)} \Big[P(I_{\{\boldsymbol{x}_{11} \in E_l\}}) \Big]^2.$$

□

2.6.2 Proof of Corollary 2.7

Proof: Under the asymptotic normality of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, with $A \approx B$ denoting $A - B \xrightarrow{P} 0$,

$$\begin{aligned}
\frac{1}{\sqrt{N}}\left(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\beta}})\right) &= \frac{1}{\sqrt{N}}\left(\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0)\right) + \frac{1}{\sqrt{N}}\left(\mathbf{e}(\boldsymbol{\beta}_0) - \mathbf{e}(\hat{\boldsymbol{\beta}})\right) \\
&\approx \frac{1}{\sqrt{N}}\left(\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0)\right) - \frac{1}{\sqrt{N}}\nabla\mathbf{e}(\boldsymbol{\beta}_0)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \\
&\approx \frac{1}{\sqrt{N}}\left(\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0)\right) - \boldsymbol{\Lambda}\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \\
&= (\mathbf{I} \mid -\boldsymbol{\Lambda})\sqrt{N}\begin{pmatrix} (\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0))/N \\ \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix}.
\end{aligned}$$

Since $N^{-1/2}\left(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\beta}})\right)$ is a linear combination of components in $\sqrt{N}\begin{pmatrix} (\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0))/N \\ \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix}$, thus

$$\frac{1}{\sqrt{N}}\left(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\beta}})\right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2.33}$$

with $\boldsymbol{\Sigma} = \mathbf{H} - \boldsymbol{\Lambda}\mathbf{J}_{\beta\beta}^{-1}\boldsymbol{\Lambda}^T$.

2.6.3 Proof of Theorem 2.10

Since the 2-level LMM model (2.3) is a special case of the LMM model (2.2) under normality assumptions for both random effects and the error term, the proof of Theorem 2.10 for model (2.2) is quite similar to the proof of Theorem 2.3 for model (2.3) as stated in Section 2.6.1. Two results that we need to use for the proof

of Theorem 2.10 are, again, the MLE consistency (Miller, 1977) and the fact that the off-diagonal matrix $\mathbf{J}_{\beta\psi}$ of $\mathbf{J}$, the limit of the sample information matrix per observation, is 0 (Wand 2007, equation (3)). Key steps of the proof are first to do a Taylor expansion to the MLE $\hat{\boldsymbol{\theta}}$, and then to use the fact that the response vector $\mathbf{Y}$ is normally distributed to show the asymptotic normality of the observed minus estimated expected vector. In this case, $\mathbf{H} = \lim_{N\to\infty} \mathbf{FVF}^T$ is a symmetric $L \times L$ matrix, with

$$\mathbf{F} = \frac{1}{\sqrt{N}} \begin{pmatrix} I_{\{\mathbf{x}_1 \in E_1\}} \cdots I_{\{\mathbf{x}_N \in E_1\}} \\ \vdots \\ I_{\{\mathbf{x}_1 \in E_L\}} \cdots I_{\{\mathbf{x}_N \in E_L\}} \end{pmatrix}. \tag{2.34}$$

and

$$\boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_1^T \\ \vdots \\ \Lambda_L^T \end{pmatrix}_{L\times p} = \lim_{N\to\infty} \begin{pmatrix} \frac{1}{N}\sum_{k=1}^N I_{\{\mathbf{x}_k \in E_1\}} \mathbf{x}_k^T \\ \vdots \\ \frac{1}{N}\sum_{k=1}^N I_{\{\mathbf{x}_k \in E_L\}} \mathbf{x}_k^T \end{pmatrix}. \tag{2.35}$$

2.6.4 Proof of Theorem 2.12

We use the multivariate Central Limit Theorem to prove Theorem 2.12.

Proof: Let the $n_i \times p$ covariate matrix for the i-th cluster be

$$\mathbf{x}_i = \begin{pmatrix} 1 & x_{i1,1} & \cdots & x_{i1,p-1} \\ \vdots & \vdots & \vdots & \\ 1 & x_{in_i,1} & \cdots & x_{in_i,p-1} \end{pmatrix},$$

then with $\mathbf{V}_i$ given in (2.4),

$$\begin{aligned}\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \sqrt{N}(\mathbf{X}^T\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\tilde{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0) \\ &\approx \left(\frac{\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}}{N}\right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{m} \mathbf{x}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_0).\end{aligned}$$

Let $z_{il} = \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}(y_{ij} - E(y_{ij})) = \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}_0)$,

$i = 1, \ldots, m;\ l = 1, \ldots, L$. Then

$$\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0) = \begin{pmatrix} \sum_{i=1}^m z_{i1} \\ \vdots \\ \sum_{i=1}^m z_{iL} \end{pmatrix},$$

and our test statistic is based on a quadratic form in

$$\begin{aligned} & (\mathbf{f} - \mathbf{e}(\tilde{\boldsymbol{\beta}}))/\sqrt{N} \\ = \ & (\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0))/\sqrt{N} + \left(\mathbf{e}(\boldsymbol{\beta}_0) - \mathbf{e}(\tilde{\boldsymbol{\beta}})\right)/\sqrt{N} \\ \approx \ & (\mathbf{f} - \mathbf{e}(\boldsymbol{\beta}_0))/\sqrt{N} - \nabla\mathbf{e}(\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)/\sqrt{N} \\ \approx \ & \frac{1}{\sqrt{N}} \begin{pmatrix} \sum_{i=1}^m z_{i1} \\ \vdots \\ \sum_{i=1}^m z_{iL} \end{pmatrix} - \frac{1}{\sqrt{N}} \frac{\bigtriangledown\mathbf{e}(\boldsymbol{\beta}_0)}{N} \left(\frac{\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}}{N}\right)^{-1} \sum_{i=1}^{m} \mathbf{x}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_0) \\ = \ & \sum_{i=1}^{m} \frac{1}{\sqrt{N}} \left[\begin{pmatrix} z_{i1} \\ \vdots \\ z_{iL} \end{pmatrix} - \frac{\bigtriangledown\mathbf{e}(\boldsymbol{\beta}_0)}{N} \left(\frac{\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}}{N}\right)^{-1} \mathbf{x}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_0) \right]. \end{aligned}$$

Let $\tilde{\mathbf{\Lambda}} = N^{-1}\nabla\mathbf{e}(\boldsymbol{\beta}_0)$, $\tilde{\mathbf{J}}_{\beta\beta} = N^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$. Then under Assumption 2.4, $\tilde{\mathbf{J}}_{\beta\beta} \xrightarrow{P} \mathbf{J}_{\beta\beta}$, and under Assumption 2.2, $\tilde{\mathbf{\Lambda}} \xrightarrow{P} \mathbf{\Lambda}$ (Remark 2.1), with $\mathbf{\Lambda}$ given in (2.30).

We next show that $(\mathbf{f}-\mathbf{e}(\tilde{\boldsymbol{\beta}}))/\sqrt{N}$ has a limiting Gaussian distribution by using the multivariate Central Limit Theorem. For any constant vector $\mathbf{C} = (C_1, \ldots, C_L)$, since the inverse of $\mathbf{V}_i$ given in equation (2.4) is

$$\mathbf{V}_i = \mathbf{I}_{n_i}/\sigma_\epsilon^2 - \sigma_a^2/(\sigma_\epsilon^2(\sigma_\epsilon^2 + n_i\sigma_a^2))\mathbf{1}^{\otimes 2},$$

we have

$$\begin{aligned}
& \mathbf{C}^T N^{-1/2}\left(\mathbf{f} - \mathbf{e}(\tilde{\boldsymbol{\beta}})\right) \\
\approx\ & \sum_{i=1}^{m} \frac{1}{\sqrt{N}}\left[\sum_{l=1}^{L} C_l z_{il} - \mathbf{C}^T\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}_{\beta\beta}^{-1}\mathbf{x}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_0)\right] \\
=\ & \sum_{i=1}^{m} \frac{1}{\sqrt{N}}\Big[\sum_{l=1}^{L} C_l \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}_0) - \\
& \mathbf{C}^T\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}_{\beta\beta}^{-1}\mathbf{x}_i^T\left(\frac{1}{\sigma_\epsilon^2}\mathbf{I}_{n_i} - \frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2 + n_i\sigma_a^2)}\mathbf{1}^{\otimes 2}\right)(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_0)\Big] \\
=\ & \sum_{i=1}^{m} \frac{1}{\sqrt{N}}\Big\{\sum_{j=1}^{n_i}\left[\sum_{l=1}^{L} C_l I_{\{\mathbf{x}_{ij}\in E_l\}}\right](\alpha_i + \epsilon_{ij}) - \\
& \mathbf{C}^T\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}_{\beta\beta}^{-1}\sum_{j=1}^{n_i}\left[\frac{1}{\sigma_\epsilon^2}\mathbf{x}_{ij} - \frac{n_i\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2 + n_i\sigma_a^2)}\bar{\mathbf{x}}_{i\cdot}\right](\alpha_i + \epsilon_{ij})\Big\} \\
=\ & \sum_{i=1}^{m}\left\{\frac{1}{\sqrt{N}}\sum_{j=1}^{n_i}\left[\sum_{l=1}^{L} C_l I_{\{\mathbf{x}_{ij}\in E_l\}} - \mathbf{C}^T\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}_{\beta\beta}^{-1}\left(\frac{1}{\sigma_\epsilon^2}\mathbf{x}_{ij} - \frac{n_i\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2 + n_i\sigma_a^2)}\bar{\mathbf{x}}_{i\cdot}\right)\right]\right\}\alpha_i \\
+\ & \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{\sqrt{N}}\left[\sum_{l=1}^{L} C_l I_{\{\mathbf{x}_{ij}\in E_l\}} - \mathbf{C}^T\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}_{\beta\beta}^{-1}\left(\frac{1}{\sigma_\epsilon^2}\mathbf{x}_{ij} - \frac{n_i\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2 + n_i\sigma_a^2)}\bar{\mathbf{x}}_{i\cdot}\right)\right]\epsilon_{ij} \\
=\ & \sum_{i=1}^{m} c_{i,n_i}\alpha_i + \sum_{s=1}^{N} w_s\epsilon_s,
\end{aligned}$$

where the double index (i, j) is placed in one-to-one correspondence with the single index s. Because $\{\alpha_i\}_{i=1}^m$ are i.i.d and $\{\epsilon_s\}_{s=1}^N$ are i.i.d, we can show both of the above sums have limiting normal distributions as $m \to \infty$, by checking the conditions (a) and (b) of Lemma 5.1 in the Appendix. First we bound

$$
\begin{aligned}
& |c_{i,n_i}| \\
= \; & \frac{1}{\sqrt{N}} \left| \sum_{j=1}^{n_i} \left\{ \sum_{l=1}^{L} C_l I_{[\mathbf{x}_{ij} \in E_l]} - \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{I}}_{\beta\beta}^{-1} \left(\frac{1}{\sigma_\epsilon^2} \mathbf{x}_{ij} - \frac{n_i \sigma_a^2}{\sigma_\epsilon^2 (\sigma_\epsilon^2 + n_i \sigma_a^2)} \bar{\mathbf{x}}_{i.} \right) \right\} \right| \\
= \; & \frac{1}{\sqrt{N}} \left| \sum_{j=1}^{n_i} \sum_{l=1}^{L} C_l I_{[\mathbf{x}_{ij} \in E_l]} - \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{I}}_{\beta\beta}^{-1} \left(\frac{1}{\sigma_\epsilon^2} n_i \bar{\mathbf{x}}_{i.} - \frac{n_i \sigma_a^2}{\sigma_\epsilon^2 (\sigma_\epsilon^2 + n_i \sigma_a^2)} n_i \bar{\mathbf{x}}_{i.} \right) \right| \\
= \; & \frac{1}{\sqrt{\sum_{i=1}^m n_i/m}} \frac{1}{\sqrt{m}} \left| \sum_{l=1}^{L} C_l \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} - \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{I}}_{\beta\beta}^{-1} \frac{n_i \sigma_a^2}{\sigma_\epsilon^2 + n_i \sigma_a^2} \frac{1}{\sigma_a^2} \bar{\mathbf{x}}_{i.} \right| \\
\leq \; & \frac{1}{\sqrt{\sum_{i=1}^m n_i/m}} \frac{1}{\sqrt{m}} \left(n_i \left| \sum_{l=1}^{L} C_l \right| + \left| \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{I}}_{\beta\beta}^{-1} \frac{1}{\sigma_a^2} \bar{\mathbf{x}}_{i.} \right| \right) \\
\to \; & \frac{1}{\sqrt{E(n)}} \cdot 0 \quad \forall i = 1, \ldots, m
\end{aligned}
$$

Next,

$$
\begin{aligned}
& \sum_{i=1}^{m} c_{i,n_i}^2 \\
= \; & \frac{1}{\sum_{i=1}^m n_i/m} \frac{1}{m} \sum_{i=1}^{m} \left\{ \sum_{l=1}^{L} C_l \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} - \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{I}}_{\beta\beta}^{-1} \frac{n_i \sigma_a^2}{\sigma_\epsilon^2 + n_i \sigma_a^2} \frac{1}{\sigma_a^2} \bar{\mathbf{x}}_{i.} \right\}^2 \\
\leq \; & \frac{1}{\sum_{i=1}^m n_i/m} \frac{1}{m} \sum_{i=1}^{m} 2 \left\{ \left(\sum_{l=1}^{L} C_l \right)^2 n_i^2 + \left(\frac{1}{\sigma_a^2} \right)^2 \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{I}}_{\beta\beta}^{-1} \bar{\mathbf{x}}_{i.} \bar{\mathbf{x}}_{i.}^T \tilde{\mathbf{I}}_{\beta\beta}^{-1} \tilde{\mathbf{\Lambda}}^T \mathbf{C} \right\} \\
\xrightarrow{\mathcal{P}} \; & \frac{2}{E(n)} \left\{ \left(\sum_{l=1}^{L} C_l \right)^2 E n^2 + \frac{1}{\sigma_a^4} \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{I}}_{\beta\beta}^{-1} \left[\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \bar{\mathbf{x}}_{i.} \bar{\mathbf{x}}_{i.}^T \right] \tilde{\mathbf{I}}_{\beta\beta}^{-1} \tilde{\mathbf{\Lambda}}^T \mathbf{C} \right\} \\
= \; & \text{a finite constant.} \qquad (2.36)
\end{aligned}
$$

Thus condition (a) in Lemma 5.1 holds, i.e.

$$\max_{1\le i\le m} |c_{i,n_i}| \xrightarrow{\mathcal{P}} 0, \quad \text{as} \quad m \to \infty.$$

Since $(\mathbf{x}_i, n_i)$ are assumed i.i.d., and based on (2.36), the c^2_{i,n_i} terms can be written as $g(\mathbf{x}_i, n_i)/m$ after removing the factor $m/(\sum_{i=1}^m n_i)$. This function g is the same across i, with $E(g(\mathbf{x}_i, n_i)) < \infty$. Thus the Law of Large Numbers Theorem holds for $(\mathbf{x}_i, n_i)$, condition (b) in Lemma 5.1 holds, i.e.

$$\sum_{i=1}^{m} c^2_{i,n_i} \xrightarrow{\mathcal{P}} \text{a finite constant.}$$

Then $\sum_{i=1}^m c_{i,n_i}\alpha_i$ is asymptotically normally distributed since both conditions (a) and (b) in Lemma 5.1 are satisfied.

We do the same checking for $\sum_{s=1}^N w_s\epsilon_s$.

$$\begin{aligned}
|w_s| &= \frac{1}{\sqrt{N}} \left| \sum_{l=1}^{L} C_l I_{\{\mathbf{x}_{ij}\in E_l\}} - \mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}^{-1}_{\beta\beta} \left(\frac{1}{\sigma^2_\epsilon}\mathbf{x}_{ij} - \frac{n_i\sigma^2_a}{\sigma^2_\epsilon(\sigma^2_\epsilon + n_i\sigma^2_a)}\bar{\mathbf{x}}_{i.} \right) \right| \\
&\le \frac{1}{\sqrt{N}} \left[\left| \sum_{l=1}^{L} C_l I_{\{\mathbf{x}_{ij}\in E_l\}} \right| + \left| \mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}^{-1}_{\beta\beta}\frac{1}{\sigma^2_\epsilon}\mathbf{x}_{ij} \right| + \left| \mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}^{-1}_{\beta\beta}\frac{n_i\sigma^2_a}{\sigma^2_\epsilon(\sigma^2_\epsilon + n_i\sigma^2_a)}\bar{\mathbf{x}}_{i.} \right| \right] \\
&\le \frac{1}{\sqrt{N}} \left[\sum_{l=1}^{L} |C_l| + \frac{1}{\sigma^2_\epsilon} \left| \mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}^{-1}_{\beta\beta}\mathbf{x}_{ij} \right| + \left| \mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{I}}^{-1}_{\beta\beta}\bar{\mathbf{x}}_{i.} \right| \right] \\
&\xrightarrow{\mathcal{P}} 0 \quad \forall s = 1, \ldots, N. \qquad (2.37)
\end{aligned}$$

Based on (2.37) and the fact that $(x+y+z)^2 \leq 3(x^2+y^2+z^2)$, we have

$$\begin{aligned}
\sum_{s=1}^{N} w_s^2 &= \frac{1}{N}\sum_{s=1}^{N}\left[\sum_{l=1}^{L} C_l I_{\{\mathbf{x}_{ij}\in E_l\}} - \mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{J}}_{\beta\beta}^{-1}\left(\frac{1}{\sigma_\epsilon^2}\mathbf{x}_{ij} - \frac{n_i\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}\bar{\mathbf{x}}_{i.}\right)\right]^2 \\
&\leq \frac{1}{N}\sum_{s=1}^{N}\left[\sum_{l=1}^{L}|C_l| + \frac{1}{\sigma_\epsilon^2}\left|\mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{J}}_{\beta\beta}^{-1}\mathbf{x}_{ij}\right| + \frac{1}{\sigma_\epsilon^2}\left|\mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{J}}_{\beta\beta}^{-1}\bar{\mathbf{x}}_{i.}\right|\right]^2 \\
&\leq \frac{1}{N}\sum_{s=1}^{N} 3\left[\left(\sum_{l=1}^{L}|C_l|\right)^2 + \left(\frac{1}{\sigma_\epsilon^2}\right)^2\left|\mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{J}}_{\beta\beta}^{-1}\mathbf{x}_{ij}\right|^2 + \left(\frac{1}{\sigma_\epsilon^2}\right)^2\left|\mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{J}}_{\beta\beta}^{-1}\bar{\mathbf{x}}_{i.}\right|^2\right] \\
&\xrightarrow{\mathcal{P}} 3\left(\sum_{l=1}^{L}|C_l|\right)^2 + 3\left(\frac{1}{\sigma_a^2}\right)^2\left|\mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{J}}_{\beta\beta}^{-1}\left(\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{x}_{ij}^{'}\right)\tilde{\mathbf{J}}_{\beta\beta}^{-1}\tilde{\mathbf{\Lambda}}'\mathbf{C}\right| \\
&+ 3\left(\frac{1}{\sigma_a^2}\right)^2\left|\mathbf{C}'\tilde{\mathbf{\Lambda}}\tilde{\mathbf{J}}_{\beta\beta}^{-1}\left(\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{m} n_i\bar{\mathbf{x}}_{i.}\bar{\mathbf{x}}_{i.}^{'}\right)\tilde{\mathbf{J}}_{\beta\beta}^{-1}\tilde{\mathbf{\Lambda}}'\mathbf{C}\right| \\
&= \text{a finite constant.}
\end{aligned}$$

Therefore, we have

$$\max_{1\leq s\leq N}|w_s| \xrightarrow{\mathcal{P}} 0, \quad \text{as} \quad N\to\infty.$$

With the same argument that we did on $\sum_{i=1}^{m} c_{i,n_i}\alpha_i$, we get the asymptotic normality of $\sum_{s=1}^{N} w_s\epsilon_s$ by checking conditions (a) and (b) of Lemma 5.1.

We have shown above that both $\sum_{i=1}^{m} c_{i,n_i}\alpha_i$ and $\sum_{s=1}^{N} w_s\epsilon_s$ are asymptotically normal. Because α_i and ϵ_{ij} are independent, the sums $\sum_{i=1}^{m} c_{i,n_i}\alpha_i$ and $\sum_{s=1}^{N} w_s\epsilon_s$ are conditionally independent given $(\mathbf{x}_i, n_i)$. These two sums are jointly normal and asymptotically uncorrelated. Therefore they are asymptotically independent. Thus the limiting distribution of $\mathbf{C}'N^{-1/2}(\mathbf{f}-\mathbf{e}(\tilde{\boldsymbol{\beta}}))$, which is asymptotically the sum of these two quantities, is normal. Moreover, for any constant vector $\mathbf{C}$, its limiting

variance is of the form $\mathbf{C}^T\mathbf{\Sigma}\mathbf{C}$ with the same fixed $\mathbf{\Sigma}$,

$$\begin{aligned}\mathbf{\Sigma} &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{m} Var\begin{pmatrix}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_1\}}(y_{ij}-\mathbf{x}_{ij}\boldsymbol{\beta}_0)\\ \vdots \\ \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_L\}}(y_{ij}-\mathbf{x}_{ij}\boldsymbol{\beta}_0)\end{pmatrix} - \\ &\quad \lim_{N\to\infty}\left[\frac{\bigtriangledown\mathbf{e}(\boldsymbol{\beta}_0)}{N}\right]\left[\frac{\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}}{N}\right]^{-1}\left[\frac{\bigtriangledown\mathbf{e}(\boldsymbol{\beta}_0)}{N}\right]^T \\ &= \mathbf{H}-\mathbf{\Lambda}\mathbf{J}_{\beta\beta}^{-1}\mathbf{\Lambda}^T.\end{aligned}$$

Therefore, $N^{-1/2}(\mathbf{f}-\mathbf{e}(\tilde{\boldsymbol{\beta}})) \xrightarrow{\mathcal{D}} N(\mathbf{0},\mathbf{\Sigma})$, and $(\mathbf{f}-\mathbf{e}(\tilde{\boldsymbol{\beta}}))^T\mathbf{\Sigma}^{-1}(\mathbf{f}-\mathbf{e}(\tilde{\boldsymbol{\beta}}))/N \xrightarrow{\mathcal{D}} \chi^2_k$, where $k=\text{rank}(\mathbf{\Sigma})$. We replace $\mathbf{\Sigma}$ with $\hat{\mathbf{\Sigma}}$, the reconstructed matrix by applying Singular Value Decomposition on a consistent estimator of $\mathbf{\Sigma}$. One such consistent estimator of $\mathbf{\Sigma}$ is to replace all parameters in $\mathbf{\Sigma}$ with their MLEs. Based on Lemma 5.3, rank $(\hat{\mathbf{\Sigma}})$ = rank $(\mathbf{\Sigma})$ for large N. Thus

$$(\mathbf{f}-\mathbf{e}(\tilde{\boldsymbol{\beta}}))^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{f}-\mathbf{e}(\tilde{\boldsymbol{\beta}}))/N \xrightarrow{\mathcal{D}} \chi^2_k.$$

2.6.5 Derivation of the power of T

We derive the power of the test for multi level LMM (2.1) under contiguous alternatives, based on Le Cam's third lemma (Van der Vaart, 2000), as stated below.

Lemma 2.14 ***(Le Cam's third lemma)*** *Let P_m and Q_m be two measures on a measurable space, corresponding to a null distribution under investigation, and an*

alternative hypothesis respectively. If

$$(W_m, \log \frac{dQ_m}{dP_m}) \overset{\mathcal{P}_m}{\to} N_{L+1} \left(\begin{bmatrix} \mu \\ -\sigma_\epsilon^2/2 \end{bmatrix}, \begin{bmatrix} \Sigma & \tau \\ \tau^T & \sigma_\epsilon^2 \end{bmatrix} \right), \tag{2.38}$$

then $W_m \overset{\mathcal{Q}_m}{\to} N_L(\mu + \tau, \Sigma)$. □

Let

$$H_0 : \boldsymbol{\theta}_N = \boldsymbol{\theta}_0,$$

$$H_1 : \boldsymbol{\theta}_N = \boldsymbol{\theta}_0 + \frac{\mathbf{a}}{\sqrt{N}},$$

where $\mathbf{a}$ is a constant vector, $\mathbf{a}/\sqrt{N} \to \mathbf{0}$, as $N \to \infty$. Thus $\boldsymbol{\theta}_N \to \boldsymbol{\theta}_0$, as $N \to \infty$. By Taylor expansion, under Theorem 5.21 in Van der Vaart (2000),

$$\begin{aligned} \log \frac{dQ_N}{dP_N} &= \log \frac{\text{Likelihood}(\boldsymbol{\theta}_N; \mathbf{Y}, \mathbf{X})}{\text{Likelihood}(\boldsymbol{\theta}_0; \mathbf{Y}, \mathbf{X})} \\ &\triangleq \log \frac{L(\boldsymbol{\theta}_N)}{L(\boldsymbol{\theta}_0)} \\ &\approx (\nabla \log(L(\boldsymbol{\theta}_0)))^T \frac{\mathbf{a}}{\sqrt{N}} + \frac{1}{2} \frac{\mathbf{a}^T}{\sqrt{N}} \left(\nabla^{\otimes 2} \log(L(\boldsymbol{\theta}_0)) \right) \frac{\mathbf{a}}{\sqrt{N}} \\ &\approx (\mathbf{S}_N(\boldsymbol{\theta}_0))^T \frac{\mathbf{a}}{\sqrt{N}} - \frac{1}{2} \mathbf{a}^T \mathbf{J}(\boldsymbol{\theta}_0) \mathbf{a}, \end{aligned}$$

where

$$\begin{aligned}\mathbf{S}_N(\boldsymbol{\theta}_0) &= \nabla \log(L(\boldsymbol{\theta}_0)) \\ &= \begin{bmatrix} \mathbf{X}^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0) \\ -\frac{1}{2}tr(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_a^2}) + \frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0)^T\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_a^2}\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0) \\ -\frac{1}{2}tr(\mathbf{V}^{-1}) + \frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0)^T\mathbf{V}^{-1}\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0) \end{bmatrix}\end{aligned} \tag{2.39}$$

is the score function for all observations (Chapter 6, McCulloch and Searle, 2001). Under Assumptions 2.4 and 2.5, we get the existence of

$$\mathbf{J}(\boldsymbol{\theta}_0) = \lim_{N\to\infty} -\nabla^{\otimes 2}\log(L(\boldsymbol{\theta}_0))/N,$$

the limit of the sample Fisher information per observation and

$$\lim_{N\to\infty} Var(\mathbf{S}_N(\boldsymbol{\theta}_0)/\sqrt{N}) = \mathbf{J}(\boldsymbol{\theta}_0).$$

Thus

$$\log\frac{dQ_N}{dP_N} \overset{\mathcal{P}_N}{\to} N(-\frac{1}{2}\mathbf{a}^T\mathbf{J}(\boldsymbol{\theta}_0)\mathbf{a}, \mathbf{a}^T\mathbf{J}(\boldsymbol{\theta}_0)\mathbf{a}).$$

For the special case when we fit a reduced model to the data, using $\mathbf{X}^*_{N\times p^*}$ instead of $\mathbf{X}_{N\times p}$ with $p^* < p$, we only estimate the coefficient $\boldsymbol{\beta}^*$ corresponding to $\mathbf{X}^*$. The $\mathbf{e}(\cdot)$ function in (2.7)

$$e_l(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}E(y_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}\mathbf{x}_{ij}^T\boldsymbol{\beta}$$

has $\mathbf{R}^p$ as its domain. Let function $\mathbf{e}^*(\cdot)$ have the same meaning of function $\mathbf{e}(\cdot)$, but with $\mathbf{R}^{p^*}$ as its domain.

$$e_l^*(\boldsymbol{\beta}^*) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}} E^*(y_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}} (\mathbf{x}_{ij}^*)^T \boldsymbol{\beta}^*.$$

Let $W_N = (\mathbf{f} - \mathbf{e}^*(\hat{\boldsymbol{\beta}}^*))/\sqrt{N}$ be the first vector component of (2.38). Under the null hypothesis P_N, W_N is asymptotically normal, $W_N \rightarrow N(\mathbf{0}, \boldsymbol{\Sigma}^*)$, based on Corollary 2.7.

Next, we compute the variance-covariance matrix $\boldsymbol{\Sigma}$ in (2.38), which is equivalent to the variance-covariance matrix of $\mathbf{a}^T\mathbf{S}_N(\theta_0)/\sqrt{N}$ and $(\mathbf{f} - \mathbf{e}^*(\hat{\boldsymbol{\beta}}^*))/\sqrt{N}$.

$$\begin{aligned}
\frac{1}{\sqrt{N}}(\mathbf{f} - \mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)) &\approx \frac{1}{\sqrt{N}}(\mathbf{f} - \mathbf{e}^*(\boldsymbol{\beta}_0^*)) - \frac{1}{\sqrt{N}}\nabla\mathbf{e}^*(\boldsymbol{\beta}_0^*)(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0^*) \\
&\approx \frac{1}{\sqrt{N}}(\mathbf{f} - \mathbf{e}^*(\boldsymbol{\beta}_0^*)) - \boldsymbol{\Lambda}^*\sqrt{N}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0^*) \\
&\approx \frac{1}{\sqrt{N}}(\mathbf{f} - \mathbf{e}^*(\boldsymbol{\beta}_0^*)) - \boldsymbol{\Lambda}^*(\mathbf{J}_{\beta\beta}^*)^{-1}(\mathbf{X}^*)^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta}_0^*)/\sqrt{N} \\
&= \frac{1}{\sqrt{N}}(\mathbf{A} - \mathbf{B}) \cdot (\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta}_0^*),
\end{aligned}$$

where $\mathbf{J}_{\beta\beta}^*$ denotes the information matrix when the second derivative for the log-likelihood function is taken with respect to $\boldsymbol{\beta}^*$, and

$$\mathbf{A} = \begin{bmatrix} I_{\{\mathbf{x}_{11}\in E_1\}} \cdots I_{\{\mathbf{x}_{mn_m}\in E_1\}} \\ \vdots \\ I_{\{\mathbf{x}_{11}\in E_L\}} \cdots I_{\{\mathbf{x}_{mn_m}\in E_L\}} \end{bmatrix}, \quad \mathbf{B} = \boldsymbol{\Lambda}^*(\mathbf{J}_{\beta\beta}^*)^{-1}(\mathbf{X}^*)^T\mathbf{V}^{-1}.$$

Thus,

$$\begin{aligned} & Cov(\frac{\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)}{\sqrt{N}}, \log\frac{dQ_n}{dP_n}) \\ = \; & Cov(\frac{\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)}{\sqrt{N}}, \frac{\mathbf{a}^T\mathbf{S}_N(\boldsymbol{\theta}_0)}{\sqrt{N}}) \\ = \; & \frac{1}{N}Cov(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*), \mathbf{a}_1^T\mathbf{S}_\beta + a_2\mathbf{S}_{\sigma_a^2} + a_3\mathbf{S}_{\sigma_\epsilon^2}). \end{aligned} \tag{2.40}$$

Under equation (2.39), since both $tr(\mathbf{V}^{-1}(\partial\mathbf{V}/\partial\sigma_a^2))$ and $tr(\mathbf{V}^{-1})$ are constants, we have

$$\begin{aligned} & Cov\left(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*), \mathbf{S}_{\sigma_a^2}\right) \\ = \; & Cov\left((\mathbf{A}-\mathbf{B})\times(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*), \frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0)^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma_a^2}\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0)\right) \\ = \; & (\mathbf{A}-\mathbf{B})\; Cov\left(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*, \frac{1}{2}(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*)^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma_a^2}\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*)\right) \\ = \; & 0, \end{aligned}$$

because $Cov(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*, \frac{1}{2}(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*)^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma_a^2}\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*)) = 0$.

Similarly, we get

$$Cov(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*), \mathbf{S}_{\sigma_\epsilon^2}) = 0.$$

Therefore (2.40) becomes

$$
\begin{aligned}
& Cov(\frac{\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)}{\sqrt{N}}, \log\frac{dQ_n}{dP_n}) \\
= & \frac{1}{N}\, Cov\left(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*), \mathbf{a}_1^T\mathbf{S}_\beta\right) \\
= & \frac{1}{N}\, Cov\left((\mathbf{A}-\mathbf{B})\times(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*), \mathbf{a}_1^T\mathbf{X}^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*)\right) \\
= & \frac{1}{N}\, (\mathbf{A}-\mathbf{B})\, Var(\mathbf{Y})\, \mathbf{V}^{-1}\mathbf{X}\mathbf{a}_1 \\
= & \frac{1}{N}\, (\mathbf{A}-\mathbf{B})\mathbf{X}\mathbf{a}_1 \\
= & \left\{\boldsymbol{\Lambda} - \frac{1}{N}\boldsymbol{\Lambda}^*(\mathbf{J}^*_{\beta\beta})^{-1}\left[(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}\right]\right\}\mathbf{a}_1 \\
= & \left\{\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^*\left[(\mathbf{X}^*)^T\mathbf{V}^{-1}(\mathbf{X}^*)\right]^{-1}\left[(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}\right]\right\}\mathbf{a}_1. \qquad (2.41)
\end{aligned}
$$

Since both $\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)$ and $\mathbf{a}_1^T\mathbf{S}_\beta$ can be written as a matrix multiply by the same normal vector $\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0 = \mathbf{Y}-\mathbf{X}^*\boldsymbol{\beta}_0^*$, we easily get the asymptotic joint normality of $\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)$ and $\mathbf{a}_1^T\mathbf{S}_\beta$. Because $\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)$ is asymptotically uncorrelated with both $\mathbf{S}_{\sigma_a^2}$ and $\mathbf{S}_{\sigma_\epsilon^2}$ as shown in the above context, we also get the asymptotic jointly normality of $\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)$ and $\mathbf{a}^T\mathbf{S}_N(\boldsymbol{\theta}_0)$. □

We next calculate the limits of the terms in
$Cov\left(N^{-1/2}(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\beta}}^*)), \log(dQ_n/dP_n)\right)$ in Section 2.6.5.1 and 2.6.5.2.

2.6.5.1 Limit of $(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}$ and $(\mathbf{X}^*)^T\mathbf{V}^{-1}(\mathbf{X}^*)$ in (2.22)

The analytical expressions in (2.41) calculated in this subsection as well as in Subsection 2.6.5.2 are used to get the theoretical powers in the settings discussed in the simulation Section 2.3, where $\mathbf{x}_{ij}, i = 1,\ldots,m; j = 1,\ldots,n_i$ are i.i.d., and $\mathbf{x}_{ij}$

and n_i are independently generated.

Note that $(\mathbf{X}^*)^T\mathbf{V}^{-1}(\mathbf{X}^*)$ is a submatrix of $(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}$. To get the limit of $(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}/N$ for the setting in the simulation Section, we assume $\mathbf{X} = (1, x_1, x_2, x_3)$ and $\mathbf{X}^* = (1, x_1, x_2)$. The ith block matrix (2.4) can then be written as $\sigma_\epsilon^2\mathbf{I}_{n_i}+\sigma_a^2\mathbf{1}_{n_i}^{\otimes 2}$, with its inverse matrix being $\mathbf{I}_{n_i}/\sigma_\epsilon^2 - \sigma_a^2/(\sigma_\epsilon^2(\sigma_\epsilon^2 + n_i\sigma_a^2))\mathbf{1}_{n_i}^{\otimes 2}$. Thus

$$
\begin{aligned}
&\frac{1}{N}(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X} \\
= \;& \frac{1}{N}\begin{bmatrix} & \vdots & \\ 1 & x_{i1,1} & x_{i1,2} \\ & \vdots & \\ 1 & x_{in_i,1} & x_{in_i,2} \\ & \vdots & \end{bmatrix}^T \left[\mathrm{diag}(\sigma_\epsilon^2\mathbf{I}_{n_i} + \sigma_a^2\mathbf{1}_{n_i}^{\otimes 2})\right]^{-1} \begin{bmatrix} & \vdots & & \\ 1 & x_{i1,1} & x_{i1,2} & x_{i1,3} \\ & \vdots & & \\ 1 & x_{in_i,1} & x_{in_i,2} & x_{in_i,3} \\ & \vdots & & \end{bmatrix} \\
= \;& \frac{1}{N}\sum_{i=1}^{m} \begin{bmatrix} 1 & \cdots & 1 \\ x_{i1,1} & \cdots & x_{in_i,1} \\ x_{i1,2} & \cdots & x_{in_i,2} \end{bmatrix} \left[\frac{1}{\sigma_\epsilon^2}\mathbf{I}_{n_i} - \frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2 + n_i\sigma_a^2)}\mathbf{1}_{n_i}^{\otimes 2}\right] \\
& \times \begin{bmatrix} 1 & x_{i1,1} & x_{i1,2} & x_{i1,3} \\ & \vdots & & \\ 1 & x_{in_i,1} & x_{in_i,2} & x_{in_i,3} \end{bmatrix} \\
= \;& \frac{1}{\sigma_\epsilon^2}\frac{1}{N}\sum_{i=1}^{m} \begin{bmatrix} 1 & \cdots & 1 \\ x_{i1,1} & \cdots & x_{in_i,1} \\ x_{i1,2} & \cdots & x_{in_i,2} \end{bmatrix} \begin{bmatrix} 1 & x_{i1,1} & x_{i1,2} & x_{i1,3} \\ & \vdots & & \\ 1 & x_{in_i,1} & x_{in_i,2} & x_{in_i,3} \end{bmatrix} - \\
& \frac{1}{N}\sum_{i=1}^{m} \frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2 + n_i\sigma_a^2)} \begin{bmatrix} 1 & \cdots & 1 \\ x_{i1,1} & \cdots & x_{in_i,1} \\ x_{i1,2} & \cdots & x_{in_i,2} \end{bmatrix} \mathbf{1}_{n_i}^{\otimes 2} \begin{bmatrix} 1 & x_{i1,1} & x_{i1,2} & x_{i1,3} \\ & \vdots & & \\ 1 & x_{in_i,1} & x_{in_i,2} & x_{in_i,3} \end{bmatrix}.
\end{aligned}
$$

Here the first sum is

$$\frac{1}{\sigma_\epsilon^2}\frac{1}{N}\sum_{i=1}^{m}\begin{bmatrix} 1 & \cdots & 1 \\ x_{i1,1} & \cdots & x_{in_i,1} \\ x_{i1,2} & \cdots & x_{in_i,2} \end{bmatrix}\begin{bmatrix} 1 & x_{i1,1} & x_{i1,2} & x_{i1,3} \\ & \vdots & & \\ 1 & x_{in_i,1} & x_{in_i,2} & x_{in_i,3} \end{bmatrix}$$

$$= \frac{1}{\sigma_\epsilon^2}\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\begin{bmatrix} 1 & x_{ij,1} & x_{ij,2} & x_{ij,3} \\ x_{ij,1} & x_{ij,1}^2 & x_{ij,1}x_{ij,2} & x_{ij,1}x_{ij,3} \\ x_{ij,2} & x_{ij,1}x_{ij,2} & x_{ij,2}^2 & x_{ij,2}x_{ij,3} \end{bmatrix}.$$

As $N \to \infty$, its limit is

$$\frac{1}{\sigma_\epsilon^2}\begin{bmatrix} 1 & Ex_1 & Ex_2 & Ex_3 \\ Ex_1 & Ex_1^2 & E(x_1x_2) & E(x_1x_3) \\ Ex_2 & E(x_1x_2) & Ex_2^2 & E(x_2x_3) \end{bmatrix}.$$

The second sum is

$$\frac{1}{N}\sum_{i=1}^{m}\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}\begin{bmatrix} 1 & \cdots & 1 \\ x_{i1,1} & \cdots & x_{in_i,1} \\ x_{i1,2} & \cdots & x_{in_i,2} \end{bmatrix}\mathbf{1}_{n_i}^{\otimes 2}\begin{bmatrix} 1 & x_{i1,1} & x_{i1,2} & x_{i1,3} \\ & \vdots & & \\ 1 & x_{in_i,1} & x_{in_i,2} & x_{in_i,3} \end{bmatrix}$$

$$= \frac{1}{N}\sum_{i=1}^{m}\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}\times$$

$$\begin{bmatrix} n_i^2 & n_i\sum_{j=1}^{n_i}x_{ij,1} & n_i\sum_{j=1}^{n_i}x_{ij,2} & n_i\sum_{j=1}^{n_i}x_{ij,3} \\ n_i\sum_{j=1}^{n_i}x_{ij,1} & (\sum_{j=1}^{n_i}x_{ij,1})^2 & d_1 & d_2 \\ n_i\sum_{j=1}^{n_i}x_{ij,2} & d_1 & (\sum_{j=1}^{n_i}x_{ij,2})^2 & d_3 \end{bmatrix}$$

where $d_1 = (\sum_{j=1}^{n_i} x_{ij,1})(\sum_{j=1}^{n_i} x_{ij,2})$, $d_2 = (\sum_{j=1}^{n_i} x_{ij,1})(\sum_{j=1}^{n_i} x_{ij,3})$,

$d_3 = (\sum_{j=1}^{n_i} x_{ij,2})(\sum_{j=1}^{n_i} x_{ij,3})$ and

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{m}\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}n_i^2 &= \frac{1}{(\sum_{i=1}^{m} n_i)/m}\cdot\frac{1}{m}\sum_{i=1}^{m}\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}n_i^2 \\
&\xrightarrow{P} \frac{1}{E(n)}\cdot E\big[\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n\sigma_a^2)}\cdot n^2\big] \\
&= \frac{1}{E(n)}\cdot\frac{\sigma_a^2}{\sigma_\epsilon^2}\cdot c_1,
\end{aligned}$$

where $c_1 = E(n^2/(\sigma_\epsilon^2+n\sigma_a^2))$.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{m}\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}n_i\sum_{j=1}^{n_i}x_{ij,1} &= \frac{1}{(\sum_{i=1}^{m} n_i/m)}\cdot\frac{1}{m}\sum_{i=1}^{m}\frac{n_i\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}\left(\sum_{j=1}^{n_i}x_{ij,1}\right) \\
&\xrightarrow{P} \frac{1}{E(n)}\cdot E_{(\mathbf{x},n)}\left[\frac{n\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n\sigma_a^2)}\left(\sum_{j=1}^{n}x_{j,1}\right)\right] \\
&= \frac{1}{E(n)}\cdot E_n\left\{E_{\mathbf{x}}\left[\frac{n\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n\sigma_a^2)}\left(\sum_{j=1}^{n}x_{j,1}\right)\Big|n\right]\right\} \\
&= \frac{1}{E(n)}\cdot E_n\left\{\frac{n^2\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n\sigma_a^2)}\cdot EX_1\right\} \\
&= \frac{EX_1}{E(n)}\cdot\frac{\sigma_a^2}{\sigma_\epsilon^2}\cdot c_1.
\end{aligned}$$

$$
\begin{aligned}
& \frac{1}{N}\sum_{i=1}^{m}\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}\left(\sum_{j=1}^{n_i}x_{ij,1}\right)\left(\sum_{j=1}^{n_i}x_{ij,2}\right) \\
= & \frac{1}{(\sum_{i=1}^{m}n_i)/m}\cdot\frac{1}{m}\sum_{i=1}^{m}\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}\left(\sum_{j=1}^{n_i}x_{ij,1}\right)\left(\sum_{j=1}^{n_i}x_{ij,2}\right) \\
\xrightarrow{P} & \frac{1}{E(n)}\cdot E_{(\mathbf{x},n)}\left\{\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n\sigma_a^2)}\left(\sum_{j=1}^{n}x_{j,1}\right)\left(\sum_{j=1}^{n}x_{j,2}\right)\right\} \\
= & \frac{1}{E(n)}\cdot E_n\left\{E_{\mathbf{x}}\left[\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n\sigma_a^2)}\left(\sum_{j=1}^{n}x_{j,1}\right)\left(\sum_{j=1}^{n}x_{j,2}\right)\Big|n\right]\right\} \\
= & \frac{1}{E(n)}\cdot E_n\left\{\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n\sigma_a^2)}\left[nE(x_1x_2)+(n^2-n)\cdot Ex_1\cdot Ex_2\right]\right\} \\
= & \frac{1}{E(n)}\cdot\frac{\sigma_a^2}{\sigma_\epsilon^2}\left\{E_n\left(\frac{n}{\sigma_\epsilon^2+n\sigma_a^2}\right)\cdot E\left(x_1x_2\right)+E_n\left(\frac{n^2-n}{\sigma_\epsilon^2+n\sigma_a^2}\right)\cdot Ex_1Ex_2\right\}.
\end{aligned}
$$

Let $c_2 = E_n\left[n/(\sigma_\epsilon^2+n\sigma_a^2)\right]$, then the above limit is

$$
\frac{1}{E(n)}\cdot\frac{\sigma_a^2}{\sigma_\epsilon^2}\left\{c_2\cdot E(x_1x_2)+c_3\cdot Ex_1Ex_2\right\},
$$

where $c_3 = c_1 - c_2$.

Combining the above expressions, we have the limit for the second component of $\frac{1}{N}(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}$:

$$
\frac{1}{E(n)}\cdot\frac{\sigma_a^2}{\sigma_\epsilon^2}\begin{bmatrix} c_1 & c_1Ex_1 & c_1Ex_2 & c_1Ex_3 \\ c_1Ex_1 & h_1 & h_2 & h_4 \\ c_1Ex_2 & h_2 & h_3 & h_5 \end{bmatrix},
$$

with $h_1 = c_2Ex_1^2 + c_3(Ex_1)^2$, $h_2 = c_2E(x_1x_2) + c_3Ex_1Ex_2$, $h_3 = c_2Ex_2^2 + c_3(Ex_2)^2$, $h_4 = c_2E(x_1x_3) + c_3Ex_1Ex_3$ and $h_5 = c_2E(x_2x_3) + c_3Ex_2Ex_3$.

Finally,

$$\begin{aligned}
\frac{1}{N}(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X} &= \frac{1}{\sigma_\epsilon^2}\frac{1}{N}\sum_{i=1}^{m}(\mathbf{X}_i^*)^T\mathbf{X}_i - \frac{1}{N}\sum_{i=1}^{m}\frac{\sigma_a^2}{\sigma_\epsilon^2(\sigma_\epsilon^2+n_i\sigma_a^2)}(\mathbf{X}_i^*)^T\mathbf{1}_{n_i}^{\otimes 2}\mathbf{X}_i \\
&\xrightarrow{P} \frac{1}{\sigma_\epsilon^2}\begin{bmatrix} 1 & Ex_1 & Ex_2 & Ex_3 \\ Ex_1 & Ex_1^2 & E(x_1x_2) & E(x_1x_3) \\ Ex_2 & E(x_1x_2) & Ex_2^2 & E(x_2x_3) \end{bmatrix} \\
&\quad - \frac{1}{E(n)}\frac{\sigma_a^2}{\sigma_\epsilon^2}\begin{bmatrix} c_1 & c_1Ex_1 & c_1Ex_2 & c_1Ex_3 \\ c_1Ex_1 & h_1 & h_2 & h_4 \\ c_1Ex_2 & h_2 & h_3 & h_5 \end{bmatrix},
\end{aligned}$$

where $c_1 = E_n\big[n^2/(\sigma_\epsilon^2+n\sigma_a^2)\big]$, $c_2 = E_n\big[n/(\sigma_\epsilon^2+n\sigma_a^2)\big]$ and $c_3 = c_1 - c_2$, $h_4 = c_2E(x_1x_3) + c_3Ex_1Ex_3$ and $h_5 = c_2E(x_2x_3) + c_3Ex_2Ex_3$.

Since $(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}^*$ is a submatrix of $(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}$,

$$\begin{aligned}
\mathbf{\Sigma}^* &= \lim_{N\to\infty}\frac{1}{N}(\mathbf{X}^*)^T\mathbf{V}^{-1}\mathbf{X}^* \\
&= \frac{1}{\sigma_\epsilon^2}\begin{bmatrix} 1 & Ex_1 & Ex_2 \\ Ex_1 & Ex_1^2 & E(x_1x_2) \\ Ex_2 & E(x_1x_2) & Ex_2^2 \end{bmatrix} - \frac{1}{E(n)}\frac{\sigma_a^2}{\sigma_\epsilon^2}\begin{bmatrix} c_1 & c_1Ex_1 & c_1Ex_2 \\ c_1Ex_1 & h_1 & h_2 \\ c_1Ex_2 & h_2 & h_3 \end{bmatrix}.
\end{aligned}$$

2.6.5.2 Limit of Λ in (2.22)

We discuss the case when the cell partition is based on $\mathbf{x}_3$. Again, the results got from this Section is used to calculate the theoretical power in the scenarios of Section 2.3.

Let E_l be the lth cell of the cell partition, $l = 1, \ldots, L$, then the elements of Λ are:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{x_{ij,3}\in E_l\}} &\xrightarrow{P} \int_{E_l} f_3(x_3)dx_3 \\
\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{x_{ij,3}\in E_l\}}x_{ij,1} &\xrightarrow{P} \int_{x_1}\int_{E_l} x_1 f_{(x_1,x_3)}(x_1,x_3)dx_3dx_1 \\
\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{x_{ij,3}\in E_l\}}x_{ij,3} &\xrightarrow{P} \int_{E_l} x_3 f_3(x_3)dx_3.
\end{aligned}$$

When $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ are jointly normal, which is the Scenario I in Section 2.3.1, $F_3(x_3)$ and $f_{(x_1,x_3)}(x_1, x_3)$ are the corresponding normal and bivariate normal densities.

When $x_3 = x_1x_2$ and x_1 and x_2 independent, i.e. $\rho_{12} = 0$, which is the Scenario II in Section 2.3.2, we can calculate $F_3(x_3)$ and $f_{(x_1,x_3)}(x_1, x_3)$ as follows:

$$\begin{aligned}
F_3(z) &= P(x_1x_2 \le z) \\
&= P(x_1x_2 \le z, x_1 > 0) + P(x_1x_2 \le z, x_1 < 0) \\
&= \iint\limits_{\{x_2<z/x_1, x_1>0\}} f_{(x_1x_2)}(x_1,x_2)dx_1dx_2 + \iint\limits_{\{x_2>z/x_1, x_1<0\}} f_{(x_1x_2)}(x_1,x_2)dx_1dx_2 \\
&= \int_{x_1=0}^{\infty}\int_{x_2=-\infty}^{\frac{z}{x_1}} f_{x_1}(x_1)f_{x_2}(x_2)dx_1dx_2 + \int_{x_1=-\infty}^{0}\int_{x_2=\frac{z}{x_1}}^{\infty} f_{x_1}(x_1)f_{x_2}(x_2)dx_1dx_2 \\
&= \int_0^{\infty} F_2\left(\frac{z}{x_1}\right) f_1(x_1)dx_1 + \int_{-\infty}^{0}\left[1 - F_2\left(\frac{z}{x_1}\right)\right] f_1(x_1)dx_1
\end{aligned}$$

$$f_3(z) = \frac{dF_3(z)}{dz} = \int_0^{\infty} f_2\left(\frac{z}{x_1}\right)\frac{1}{x_1}f_1(x_1)dx_1 - \int_{-\infty}^{0} f_2\left(\frac{z}{x_1}\right)\frac{1}{x_1}f_1(x_1)dx_1.$$

To get the joint distribution of $(x_1, x_3) = (x_1, x_1x_2)$,

let $\begin{cases} y_1 = x_1 \\ y_2 = x_1 x_2 \end{cases}$, then equivalently, $\begin{cases} x_1 = y_1 \\ x_2 = \frac{y_2}{y_1} \end{cases}$.

So the Jacobian matrix $\mathcal{J} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -\frac{y_2}{y_1^2} & \frac{1}{y_1} \end{vmatrix} = \frac{1}{|y_1|}$

Thus

$$\begin{aligned} f_{(y_1,y_2)}(y_1, y_2) &= f_{(x_1,x_2)}\left(y_1, \frac{y_2}{y_1}\right) \cdot \mathcal{J} \\ &= f_{x_1}(y_1) \cdot f_{x_2}\left(\frac{y_2}{y_1}\right) \cdot \frac{1}{|y_1|} \end{aligned}$$

Similarly, the joint distribution of (x_2, x_3) is

$$\begin{aligned} f_{(x_2,x_3)}(x_2, x_3) &= f_{(x_1,x_2)}\left(\frac{x_3}{x_2}, x_2\right) \cdot \frac{1}{|x_2|} \\ &= f_{x_1}\left(\frac{x_3}{x_2}\right) f_{x_2}(x_2) \cdot \frac{1}{|x_2|} \end{aligned}$$

Remark 2.15 *The limiting terms calculated in Section 2.6.5.1 and 2.6.5.2 are used to calculate the theoretical power in Section 2.2.3. Figure 2.1, 2.2 and 2.3 are plotted by making use of these limit expressions.* □

Chapter 3

Goodness of fit tests for generalized linear mixed models

3.1 Generalized linear mixed models (GLMMs)

For $i = 1, \ldots, m$, and $j = 1, \ldots, n_i$, let $\mathbf{x}_{ij}$ (with the first component being 1) and y_{ij} be the covariate and outcome value for the jth subject in cluster i respectively, and let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$. The GLMM has the following form: $E(y_{ij}|\mathbf{u}_i) = g(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{w}_{ij}^T\mathbf{u}_i)$, where $g(.)$ is a known strictly monotonic and differentiable function, the *i.i.d.* cluster random effects $\mathbf{u}_i$ are assumed to be from a known probability distribution, $f_{\mathbf{u}}(\mathbf{u}_i)$, with unknown parameter vector $\boldsymbol{\nu} = (\sigma_1^2, \ldots, \sigma_s^2)$, and $\mathbf{w}_{ij}$ are covariates for the random effects. We assume that the conditional distribution of y_{ij} given $\lambda_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{w}_{ij}^T\mathbf{u}_i$ and ϕ, follows a distribution from the exponential family with density function $f_{y|\lambda}(y|\lambda, \phi) = \exp\{[yQ(\lambda) - b(\lambda)]a(\phi) + c(y, \phi)\}$, where ϕ is generally an unknown parameter related to $Var(y_{ij})$. In this Chapter, we restrict the exponential family to the canonical form, that is $Q(\lambda) = \lambda$. One can always define a transformed parameter to convert an exponential family to canonical form. The response variables y_{ij} are conditionally independent given the random effect $\mathbf{u}_i$. The covariates $\mathbf{x}_{ij}$ are usually treated as fixed in practice. However, to deal with technical issues arising when we prove the consistency of the maximum likelihood estimators and the asymptotic properties of the test statistic, we assume throughout this chapter that $(\mathbf{x}_i, n_i)$ are i.i.d.

The likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\nu}, \phi)$ is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_{y|\lambda}(y_{ij}|\lambda_{ij}) dF(\lambda_{ij}).$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is the solution to $S(\boldsymbol{\theta}) = 0$, where the S denotes the score function

$$S(\boldsymbol{\theta}) = \frac{\partial log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{m} S_i(\boldsymbol{\theta}).$$

In what follows, we will focus on GLMMs with a single random intercept, given by

$$E(y_{ij}|\alpha_i) = g(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \alpha_i), \quad \alpha_i \sim N(0, \sigma^2). \tag{3.1}$$

In this case, the parameter vector $\boldsymbol{\nu}$ for variance components reduces to a single parameter σ^2.

The linear mixed model, which is an important case of GLMMs, has been carefully studied in Chapter 2. We now describe another two special cases of the GLMMs with random intercept.

3.1.1 Mixed-effects logistic models

For mixed-effects logistic models with a cluster specific random intercept, y_{ij} given p_{ij}, where $p_{ij} = P(y_{ij} = 1)$, comes from a binomial distribution: $y_{ij} \sim$

Binom$(1, p_{ij})$ and $\text{logit}(p_{ij}) = \log(p_{ij}/(1-p_{ij})) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i$, where the i.i.d. cluster random effects α_i are assumed to be $N(0, \sigma^2)$. The response variables y_{ij} are conditionally independent given α_i. The unknown parameter vector is now $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

The marginal probability of the response for the ith cluster, conditionally given $(\mathbf{x}_{ij}, n_i)$, under the random intercept logistic mixed model

$$y_{ij} \sim \text{Binom}(1, p_{ij}), \quad \text{logit}(p_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i, \quad \alpha_i \sim N(0, \sigma^2) \tag{3.2}$$

is

$$P(Y_{ij} = y_{ij}) = \int \prod_{j=1}^{n_i} p_{ij}^{y_{ij}}\, q_{ij}^{1-y_{ij}}\, dF(\alpha_i),$$

where $q_{ij} = 1 - p_{ij}$.

The means and variance-covariances for $y_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, n_i,$, conditionally given $(\mathbf{x}_{ij}, n_i)$, are

$$E(y_{ij}) = E(E(y_{ij}|\alpha_i)) = E(p_{ij}) = \int p_{ij}\, dF(\alpha_i),$$

$$\begin{aligned} Var(y_{ij}) &= E[Var(y_{ij}|\alpha_i)] + Var[E(y_{ij}|\alpha_i)] \\ &= E[p_{ij}(1-p_{ij})] + Var(p_{ij}) = E(p_{ij}) - [E(p_{ij})]^2, \end{aligned}$$

$$Cov(y_{is}, y_{it}) = \int p_{is}p_{it}\, dF(\alpha_i) - E(p_{is})E(p_{it}),\ \ \forall s \neq t.$$

3.1.2 Mixed-effects Poisson models

For the mixed-effects Poisson regression model with a cluster specific random effect, responses y_{ij} given μ_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, follow Poisson distributions with means μ_{ij} and $g^{-1}(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i$, where the i.i.d. cluster random effects α_i are assumed $N(0, \sigma^2)$ and g is a known monotonic and differentiable link function. When g^{-1} is the log function, this link function is canonical, transforming the mean μ_{ij} to the natural exponential parameter. The response variables y_{ij} are conditionally independent given the random effects α_i. The unknown parameter vector in this case is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

The marginal probability of the response for the ith cluster, conditionally given $(\mathbf{x}_{ij}, n_i)$, under the mixed-effect Poisson regression model is

$$P(Y_{ij} = y_{ij}) = \int \prod_{j=1}^{n_i} \frac{\mu_{ij}^{y_{ij}} \exp(-\mu_{ij})}{y_{ij}!} dF(\alpha_i).$$

When the link function is canonical, as we assume from now on, the means and variance-covariances for the response variables y_{ij}, $i = 1, \ldots, m$; $j = 1, \ldots, n_i$, conditionally given $(\mathbf{x}_{ij}, n_i)$, are

$$E(y_{ij}) = E[E(y_{ij}|\mu_{ij})] = E(\mu_{ij}) = \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \sigma^2/2),$$

$$\begin{aligned} Var(y_{ij}) &= E[Var(y_{ij}|\mu_{ij})] + Var[E(y_{ij}|\mu_{ij})] \\ &= \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \sigma^2/2) + \exp(2\mathbf{x}_{ij}^T\boldsymbol{\beta} + 2\sigma^2) - \exp(2\mathbf{x}_{ij}^T\boldsymbol{\beta} + \sigma^2), \end{aligned}$$

$$Cov(y_{is}, y_{it}) = [\exp(2\sigma^2) - \exp(\sigma^2)] \exp(\mathbf{x}_{is}^T\boldsymbol{\beta} + \mathbf{x}_{it}^T\boldsymbol{\beta}), \qquad \forall \quad s \neq t.$$

3.2 Proof of the consistency of MLE for GLMMs

Let $f(\mathbf{y}_i; \boldsymbol{\theta})$ be the likelihood function for the ith cluster, $i = 1, \ldots, m$. The log likelihood function for the whole set of observations is $\sum_{i=1}^{m} \log f(\mathbf{y}_i; \boldsymbol{\theta})$. The normalized score function is

$$\mathbf{S}_m(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \nabla \log f(\mathbf{y}_i; \boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} (\partial/\partial\boldsymbol{\theta}) \log f(\mathbf{y}_i; \boldsymbol{\theta}).$$

The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ solves $\mathbf{S}_m(\boldsymbol{\theta}) = 0$. We always assume natural and canonical parameterization. With probability 1 as m gets large, the MLE exists and is unique (Bickel and Doksum 2006).

Let $B_\epsilon(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq \epsilon\}$ be the ϵ-neighborhood of the true parameter vector $\boldsymbol{\theta}_0$, which is an open convex Borel set in $\mathcal{R}^p$. Let

$$\tilde{M}_m(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \log f(\mathbf{y}_i; \boldsymbol{\theta}), \quad \tilde{M}(\boldsymbol{\theta}) = E[\log f(\mathbf{y}_i; \boldsymbol{\theta})]$$

and

$$\mathbf{J}(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = -E_{\boldsymbol{\theta}_0}[\nabla^{\otimes 2} \log f(\mathbf{y}_1, \mathbf{x}_1, n_1; \boldsymbol{\theta})].$$

The following set of assumptions (Assumption 3.1) are used for the consistency proof of the MLE $\hat{\boldsymbol{\theta}}$ as well as for the asymptotic properties of the test statistic that will be discussed in the Section following. We comment on this set of Assumptions in Remark 3.1 below.

Assumption 3.1 ***B.0*** $(\boldsymbol{x}_i, n_i)$ *are i.i.d.*

B.1 $\tilde{M}(\boldsymbol{\theta})$ *and* $\boldsymbol{J}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$ *exist for all* $\boldsymbol{\theta}$*;*

B.2 $\boldsymbol{J}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)$ *is positive definite;*

B.3 $\boldsymbol{J}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$ *is continuous at* $\boldsymbol{\theta}_0$ *as a function of* $\boldsymbol{\theta}$*;*

B.4 $(\partial/\partial\boldsymbol{\theta})\nabla^{\otimes 2} \log f(\boldsymbol{y}_1, \boldsymbol{x}_1, n_1; \boldsymbol{\theta})$ *exists and is integrable .*

B.5 *Derivatives and expectations are interchangeable for* $\log f(\boldsymbol{y}_1, \boldsymbol{x}_1, n_1; \boldsymbol{\theta})$ *up to the third derivative;*

B.6 *The true parameter point* $\boldsymbol{\theta_0} = (\boldsymbol{\beta}_0, \sigma_0^2)$ *is an interior point of* $\boldsymbol{\Theta} = (\mathcal{R}^p, \mathcal{R}^+)$*;*

Remark 3.1 ***B.1*** *in Assumption 3.1 ensures that we can apply the Law of Large Numbers (LLN) to* $(m^{-1}) \sum_{i=1}^{m} \log f(\boldsymbol{y}_i, \boldsymbol{x}_i, n_i; \boldsymbol{\theta})$ *and* $-(m^{-1}) \sum_{i=1}^{m} \nabla^{\otimes 2} \log f(\boldsymbol{y}_i, \boldsymbol{x}_i, n_i; \boldsymbol{\theta})$*. In order for* ***B.1*** *–* ***B.5*** *to hold, assumptions are needed on* $(\boldsymbol{x}_i, n_i)$*. We provide sufficient assumptions on* $(\boldsymbol{x}_i, n_i)$ *and check* ***B.2*** *for the random intercept logistic mixed model stated as Lemma 3.8 in Section 3.6.1. To check* ***B.3*** *and* ***B.5*** *in Assumption 3.1 , based on the Dominated Convergence Theorem, it suffices to show that* $\forall\ \boldsymbol{\theta} \in \boldsymbol{\Theta}$*, there exists* $B_\epsilon(\boldsymbol{\theta})$*, such that*

$$-E_{\boldsymbol{\theta}_0}[\sup_{\boldsymbol{\theta}' \in B_\epsilon(\boldsymbol{\theta})} \nabla^{\otimes 2} \log f(\boldsymbol{y}_1, \boldsymbol{x}_1, n_1; \boldsymbol{\theta}')] < \infty. \tag{3.3}$$

We refer to condition (3.3) as the dominatedness condition. We provide sufficient assumptions on $(\boldsymbol{x}_i, n_i)$ *and check (3.3) for both the random intercept logistic mixed model (stated as Lemma 3.9 in Section 3.6.2) and the random intercept Poisson mixed model (stated as Lemma 3.10 in Section 3.6.3).* □

Lemma 3.2 *Under Assumption 3.1 within model (3.1), there exists an open neighborhood* $\mathcal{U}$ *of* $\boldsymbol{\theta}_0$*, such that*

$$P(\tilde{M}_m(\boldsymbol{\theta}) \text{ is a concave function of } \boldsymbol{\theta} \text{ on } \mathcal{U}) \to 1, \quad \text{as } m \to +\infty. \tag{3.4}$$

Proof: Based on the existence of $\tilde{M}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ as stated in **B.1** of Assumption 3.1, by the Law of Large Numbers on $\log f(\mathbf{y}_i, \mathbf{x}_i, n_i; \boldsymbol{\theta})$ and $\nabla^{\otimes 2} \log f(\mathbf{y}_i, \mathbf{x}_i, n_i; \boldsymbol{\theta})$, pointwise for each $\boldsymbol{\theta} \in \Theta$,

$$\tilde{M}_m(\boldsymbol{\theta}) \xrightarrow{P} \tilde{M}(\boldsymbol{\theta}), \tag{3.5}$$

and

$$-\frac{1}{m}\sum_{i=1}^{m} \nabla^{\otimes 2} \log f(\mathbf{y}_i, \mathbf{x}_i, n_i; \boldsymbol{\theta}) \to \mathbf{J}(\boldsymbol{\theta};\boldsymbol{\theta}_0), \quad \text{as } m \to +\infty. \tag{3.6}$$

By **B.2** − **B.3** of Assumption 3.1, there exists $\mathcal{U} = B_\epsilon(\boldsymbol{\theta}_0)$ such that $\forall\ \boldsymbol{\theta} \in \mathcal{U}$, $\mathbf{J}(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ is positive definite. By **B.1** together with **B.4** of Assumption 3.1, the uniform LLN holds for $(m^{-1})\sum_{i=1}^{m} \nabla^{\otimes 2} \log f(\mathbf{y}_i, \mathbf{x}_i, n_i; \boldsymbol{\theta})$ (van der Vaart, 2000, page 271, Example 19.7), i.e.

$$\sup_{\boldsymbol{\theta}\in\mathcal{U}} \| -\frac{1}{m}\sum_{i=1}^{m} \nabla^{\otimes 2} \log f(\mathbf{y}_i, \mathbf{x}_i, n_i; \boldsymbol{\theta}) - \mathbf{J}(\boldsymbol{\theta};\boldsymbol{\theta}_0)\| \to 0, \quad \text{as } m \to +\infty.$$

Then it follows that $\tilde{M}(\boldsymbol{\theta}) = (m^{-1})\sum_{i=1}^{m} \log f(\mathbf{y}_i, \mathbf{x}_i, n_i; \boldsymbol{\theta})$ is a concave function of $\boldsymbol{\theta} \in \mathcal{U}$ with probability approaching 1. □

To show consistency of the MLE, we use two theorems from Andersen and Gill (1982) and van der Vaart (2000) which are stated here before we summarize our result as a theorem.

Theorem 3.3 *(Theorem II.1 in Appendix II of Andersen and Gill (1982))*

Let E be an open convex subset of $\mathcal{R}^p$ and let $M_1, M_2, \ldots,$ be a sequence of random concave functions on E such that $\forall\ \theta \in E$, $M_m(\theta) \xrightarrow{P} M(\theta)$, as $m \to \infty$, where M is some real function on E. Then M is also concave and for all compact $A \subset E$,

$$\sup_{\theta \in A} |M_m(\theta) - M(\theta)| \xrightarrow{P} 0, as\ \ m \to \infty. \tag{3.7}$$

Theorem 3.4 *(van der Vaart 2000, p.45) Let M_m be random functions and let M be a fixed function of θ such that for every $\epsilon > 0$,*

$$\sup_{\theta \in \Theta} |M_m(\theta) - M(\theta)| \xrightarrow{P} 0,$$

$$\sup_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta) < M(\theta_0).$$

Then any sequence of estimators $\hat{\theta}_m$ with $M_m(\hat{\theta}_m) \geq M_m(\theta_0) - o_p(1)$ converges in probability to θ_0.

Theorem 3.5 *Under Assumption 3.1, the MLE $\hat{\boldsymbol{\theta}}$ of the GLMM (3.1) is consistent.*

Proof: Let $M_m(\boldsymbol{\theta}) = \tilde{M}_m(\boldsymbol{\theta}) I_{\{\tilde{M}_m \text{ concave on } \mathcal{U}\}}$. By Lemma 3.2, the MLE $\hat{\boldsymbol{\theta}}$ is also the solution to $\nabla M_m(\boldsymbol{\theta}) = \mathbf{0}$, with probability approaching 1. Together with (3.5), based on Theorem 3.3, we get (3.7), which is the uniform convergence of $M_m(\boldsymbol{\theta})$. Based on Theorem 3.4, the MLE $\hat{\boldsymbol{\theta}}$ is consistent. □

Remark 3.6 *Whether or not the natural parameterization is used, with probability*

approaching 1 for large m, the MLE $\hat{\boldsymbol{\theta}}$ is unique in $B_\epsilon(\boldsymbol{\theta}_0)$ and is the unique solution of the likelihood score equation in that neighborhood. □

3.3 Goodness of fit test for GLMMs

To test the goodness of fit for a proposed GLMM with random intercept, first we partition the covariate space into non-overlapping cells $E_1, \ldots, E_L$. For $l = 1, \ldots, L$, define

$$f_l = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} y_{ij}, \quad e_l(\boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} E(y_{ij}). \tag{3.8}$$

Let $\mathbf{f} = (f_1, \ldots, f_L)$, $\mathbf{e}(\boldsymbol{\theta}) = (e_1(\boldsymbol{\theta}), \ldots, e_L(\boldsymbol{\theta}))$. For simplicity, we denote $E_{\boldsymbol{\theta}}(y_{ij})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \hat{E}(y_{ij})$. The test statistic is based on the observed minus the expected counts,

$$\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_1\}} (y_{ij} - \hat{E}(y_{ij})) \\ \vdots \\ \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_L\}} (y_{ij} - \hat{E}(y_{ij})) \end{pmatrix},$$

a vector of length L.

Theorem 3.7 *For GLMM with random intercept (3.1), let $E_1, \ldots, E_L$ constitute a partition of the covariate space generated by $\mathbf{X}$ into disjoint sets such that $E(\sum_{j=1}^{n_i} I_{\{\boldsymbol{x}_{ij} \in E_l\}}) > 0$ for all $l = 1, \ldots, L$. Under Assumption 3.1 in Section 3.2, as $m \to \infty$,*

$$(\boldsymbol{f} - \boldsymbol{e}(\hat{\boldsymbol{\theta}}))' \hat{\boldsymbol{\Sigma}}_{svd}^{-1} (\boldsymbol{f} - \boldsymbol{e}(\hat{\boldsymbol{\theta}}))/m \xrightarrow{D} \chi_k^2,$$

where $\hat{\boldsymbol{\Sigma}}_{svd}$ *is the reconstructed* $L \times L$ *square matrix by applying Singular Value Decomposition on a consistent estimator* $\hat{\boldsymbol{\Sigma}}$*, given in (3.11), of* $\boldsymbol{\Sigma} = Var(\boldsymbol{\xi}_i)$ *in (3.10).* $k = rank\ (\boldsymbol{\Sigma}) = rank\ (\hat{\boldsymbol{\Sigma}}_{svd})$ *for large* m*.* $\hat{\boldsymbol{\Sigma}}_{svd}^{-1}$ *is the Moore – Penrose pseudoinverse.*

Proof: The notation $A \approx B$ is used to indicate $A - B \xrightarrow{P} 0$. Since the MLE $\hat{\boldsymbol{\theta}}$ is consistent (Section 3.2), by a first order Taylor series expansion of the score function $\mathbf{S}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$, we obtain the approximation up to terms asymptotically negligible in probability,

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx (-\frac{1}{m}\frac{\partial \mathbf{S}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}})^{-1}\frac{1}{m}\mathbf{S}(\boldsymbol{\theta}_0)) = \tilde{\mathbf{J}}_0^{-1}\frac{1}{m}\sum_{i=1}^{m}\mathbf{S}_i(\boldsymbol{\theta}_0),$$

where $\mathbf{S}_i(\boldsymbol{\theta}_0)$ is the score function for the ith cluster and the conditional sample fisher information $\tilde{\mathbf{J}}_0 = -m^{-1}\partial/\partial\boldsymbol{\theta}_0\mathbf{S}(\boldsymbol{\theta}_0)|(\mathbf{x}_i, n_i) \xrightarrow{P} \mathbf{J}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)$. The existence of $\mathbf{J}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)$ is ensured by **B.1** in Assumption 3.1, and its invertibility by **B.2**. For the rest of this proof, we use $\mathbf{J}_0$ to denote $\mathbf{J}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)$.

With $E(y_{ij}) = E_{\alpha_i}[g(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i)]$, differentiation under the integral sign by virtue of **B.5** implies that

$$\frac{\partial E(y_{ij})}{\partial \boldsymbol{\beta}} = E_{\alpha_i}[g^{'}(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i)] \cdot \mathbf{x}_{ij} = \int_{\alpha_i} g^{'}(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i) f_{\alpha_i}\, d\alpha_i \cdot \mathbf{x}_{ij},$$

and

$$\frac{\partial E(y_{ij})}{\partial \sigma^2} = \int_{\alpha_i} g(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i)\frac{\partial f_{\alpha_i}}{\partial \sigma^2}\, d\alpha_i.$$

By applying the LLN to the summands over i defining $\mathbf{e}_l(\boldsymbol{\theta}_0)$, as $N \to \infty$,

$$\tilde{\boldsymbol{\Lambda}} = \frac{1}{m} \bigtriangledown \mathbf{e}(\boldsymbol{\theta}_0) = \begin{pmatrix} m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_1\}} \frac{\partial}{\partial \boldsymbol{\theta}_0} E(y_{ij}) \\ \vdots \\ m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_L\}} \frac{\partial}{\partial \boldsymbol{\theta}_0} E(y_{ij}) \end{pmatrix} \xrightarrow{\mathcal{P}} \begin{pmatrix} \Lambda_1^T \\ \vdots \\ \Lambda_L^T \end{pmatrix} = \boldsymbol{\Lambda}.$$

For $l = 1, \ldots, L$,

$$\Lambda_l^T = E_{(\mathbf{x}_i, n_i)} \left(\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} \frac{\partial E_{\alpha_i}(g({\mathbf{x}_{ij}}^T \boldsymbol{\beta} + \alpha_i))}{\partial \boldsymbol{\theta}_0} \right).$$

Let $z_{il} = \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij} \in E_l\}} (y_{ij} - E(y_{ij}))$, $i = 1, \ldots, m;\ l = 1, \ldots, L$. Then

$$\mathbf{f} - \mathbf{e}(\boldsymbol{\theta}) = \left(\sum_{i=1}^{m} z_{i1}, \ldots, \sum_{i=1}^{m} z_{iL} \right)^T$$

and our test statistic is based on a quadratic form in the vector

$$\begin{aligned} (\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}}))/\sqrt{m} &= (\mathbf{f} - \mathbf{e}(\boldsymbol{\theta}_0))/\sqrt{m} + (\mathbf{e}(\boldsymbol{\theta}_0) - \mathbf{e}(\hat{\boldsymbol{\theta}}))/\sqrt{m} \\ &\approx (\mathbf{f} - \mathbf{e}(\boldsymbol{\theta}_0))/\sqrt{m} - \nabla \mathbf{e}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/\sqrt{m} \\ &\approx \frac{1}{\sqrt{m}} \left(\sum_{i=1}^{m} z_{i1}, \ldots, \sum_{i=1}^{m} z_{iL} \right)^T - \frac{1}{\sqrt{m}} \nabla \mathbf{e}(\boldsymbol{\theta}_0) \tilde{\mathbf{J}}_0^{-1} \frac{1}{m} \sum_{i=1}^{m} S_i(\boldsymbol{\theta}_0) \\ &= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left[(z_{i1}, \ldots, z_{iL})^T - \tilde{\boldsymbol{\Lambda}} \tilde{\mathbf{J}}_0^{-1} S_i(\boldsymbol{\theta}_0) \right] \\ &= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \boldsymbol{\xi_i}. \end{aligned} \tag{3.9}$$

Under the assumption that $(\mathbf{x}_i, n_i)$ are i.i.d. for $i = 1, \ldots, m$, also the variables $\boldsymbol{\xi}_i = \boldsymbol{\xi}_i(\mathbf{y}_i, \mathbf{x}_i, n_i; \boldsymbol{\theta})$ are i.i.d. with mean $\mathbf{0}$. Under **B.1** – **B.3** in Assumption 3.1,

$Var(\boldsymbol{\xi}_i) = \boldsymbol{\Sigma}$ exists. By multivariate Central Limit Theorem,

$$(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}}))/\sqrt{m} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3.10}$$

Therefore

$$\frac{1}{m}(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}}))' \boldsymbol{\Sigma}^{-1} (\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}})) \xrightarrow{D} \chi^2_k,$$

where $k = \text{rank}(\boldsymbol{\Sigma})$. We estimate $\boldsymbol{\Sigma}$ using

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m} \sum_{i=1}^{m} \hat{Var}(\boldsymbol{\xi}_i | \mathbf{x}_i, n_i), \tag{3.11}$$

where $\hat{Var}(\boldsymbol{\xi}_i | \mathbf{x}_i, n_i)$ means that the MLE $\hat{\boldsymbol{\theta}}$ is substituted for $\boldsymbol{\theta}$ in $Var(\boldsymbol{\xi}_i | \mathbf{x}_i, n_i)$.

$\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$ and its simplified form is given in Section 3.6.4. Then by Slutsky's theorem, the test statistic

$$T = \frac{1}{m}(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}}))' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}})) \xrightarrow{D} \chi^2_k, \tag{3.12}$$

We then compute Singular Value Decomposition for $\hat{\boldsymbol{\Sigma}}$. For each eigenvalue of $\hat{\boldsymbol{\Sigma}}$, we compare it with a preset upper bound ζ. For any eigenvalue less than ζ, we instead set this eigenvalue to be 0 and reconstruct the $\hat{\boldsymbol{\Sigma}}$ matrix using the non-zero eigenvalues and their corresponding eigenvectors. We denote this reconstructed matrix as $\hat{\boldsymbol{\Sigma}}_{svd}$. Based on Corollary 5.3 given in the Appendix, $P(\text{rank } (\hat{\boldsymbol{\Sigma}}_{svd}) =$

rank $(\boldsymbol{\Sigma})) \to 1$. The test statistic to be used is

$$\frac{1}{m}(\mathbf{f}-\mathbf{e}(\hat{\boldsymbol{\theta}}))'\hat{\boldsymbol{\Sigma}}_{svd}^{-1}(\mathbf{f}-\mathbf{e}(\hat{\boldsymbol{\theta}})) \xrightarrow{D} \chi_k^2. \tag{3.13}$$

3.3.1 Derivation of the power of T

We derive the power of the test for the 2-level GLMM under contiguous alternatives, based on Le Cam's third lemma (Lemma 2.14).

For a fixed constant vector $\mathbf{a}$, let

$$H_0 : \boldsymbol{\theta}_m = \boldsymbol{\theta}_0,$$

$$H_1 : \boldsymbol{\theta}_m = \boldsymbol{\theta}_0 + \frac{\mathbf{a}}{\sqrt{m}},$$

where one or more components of $\boldsymbol{\beta}_0$ in $\boldsymbol{\theta}_0$ are 0s. Denote the non-zero components of $\boldsymbol{\theta}_0$ as $\boldsymbol{\theta}_0^*$, which is a sub-vector of $\boldsymbol{\theta}_0$. Note that $\boldsymbol{\theta}_m \to \boldsymbol{\theta}_0$, as $m \to \infty$. By Taylor expansion, using Theorem 5.21 in Van der Vaart (2000),

$$\begin{aligned}
\log \frac{dQ_m}{dP_m} &= \log \frac{\text{Likelihood}(\boldsymbol{\theta}_m; \mathbf{Y}, \mathbf{X})}{\text{Likelihood}(\boldsymbol{\theta}_0; \mathbf{Y}, \mathbf{X})} \\
&\triangleq \log \frac{L(\boldsymbol{\theta}_m)}{L(\boldsymbol{\theta}_0)} \\
&\approx (\nabla \log(L(\boldsymbol{\theta}_0)))^T \frac{\mathbf{a}}{\sqrt{m}} + \frac{1}{2}\frac{\mathbf{a}^T}{\sqrt{m}}(\nabla^{\otimes 2} \log(L(\boldsymbol{\theta}_0)))\frac{\mathbf{a}}{\sqrt{m}} \\
&\approx (\mathbf{S}_m(\boldsymbol{\theta}_0))^T \frac{\mathbf{a}}{\sqrt{m}} - \frac{1}{2}\mathbf{a}^T \mathbf{J}_0 \mathbf{a},
\end{aligned}$$

Based on the LLN on $-m^{-1} \sum_{i=1}^m \nabla^{\otimes 2} \log f(\mathbf{y}_i, \mathbf{x}_i, n_i; \boldsymbol{\theta})$ (3.6), with $\mathbf{J}_0 =$

$\mathbf{J}(\boldsymbol{\theta}_0;\boldsymbol{\theta}_0)$,

$$\lim_{m\to\infty} Var(\mathbf{S}_m(\boldsymbol{\theta}_0)/\sqrt{m}) = \mathbf{J}_0.$$

Thus

$$\log \frac{dQ_m}{dP_m} \xrightarrow{\mathcal{P}_m} N(-\frac{1}{2}\mathbf{a}^T\mathbf{J}_0\mathbf{a}, \mathbf{a}^T\mathbf{J}_0\mathbf{a}).$$

Since several components of $\boldsymbol{\theta}_0$ are 0s, under H_0, we fit a reduced model to the data, using $\mathbf{X}^*_{N\times p^*}$ instead of $\mathbf{X}_{N\times p}$ with $p^* < p$. We only estimate the coefficient $\boldsymbol{\beta}^*$ corresponding to $\mathbf{X}^*$. The $\mathbf{e}(\cdot)$ function in (3.8)

$$e_l(\boldsymbol{\theta}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}E(y_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}\int g(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\alpha_i)\phi(\alpha_i)d\alpha_i$$

has $\mathcal{R}^p \times \mathcal{R}^+$ as its domain, where p is the dimension of $\boldsymbol{\beta}$ and $\mathcal{R}^+$ is the domain for σ^2, the variance component for the random effect α_i. Let function $\mathbf{e}^*(\cdot)$ be the same meaning of function $\mathbf{e}(\cdot)$, but has $\mathcal{R}^{p^*} \times \mathcal{R}^+$ as its domain

$$e_l^*(\boldsymbol{\theta}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}E^*(y_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}}\int g\left((\mathbf{x}_{ij}^*)^T\boldsymbol{\beta}^*+\alpha_i\right)\phi(\alpha_i)d\alpha_i.$$

Let $W_m = (\mathbf{f} - \mathbf{e}^*(\hat{\boldsymbol{\theta}}^*))/\sqrt{m}$. Under the null hypothesis H_0, W_m is asymptotically normal, $W_m \to N(\mathbf{0}, \boldsymbol{\Sigma}^*)$, by (3.10). The joint distribution of $\log(dQ_m/dP_m)$ and $(\mathbf{f} - \mathbf{e}^*(\hat{\boldsymbol{\theta}}^*))/\sqrt{m}$ is asymptotically equal in probability to the joint distribution of $\mathbf{a}^T\mathbf{S}_m(\theta_0)/\sqrt{m}$ and $(\mathbf{f} - \mathbf{e}^*(\hat{\boldsymbol{\theta}}^*))/\sqrt{m}$. We next show the jointly normality of these two quantities.

By the same sequence of steps as in (3.9),

$$\frac{1}{\sqrt{m}}(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*)) \approx \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[(z_{i1}^*,\ldots,z_{iL}^*)^T - \tilde{\boldsymbol{\Lambda}}^*(\tilde{\mathbf{J}}_0^*)^{-1}\mathbf{S}_i^*(\boldsymbol{\theta}_0^*)\right],$$

where the $*$ in z_{il}^* and $\tilde{\boldsymbol{\Lambda}}^*$ means that the $\mathbf{X}$ and $\boldsymbol{\theta}$ in z_{il} and $\tilde{\boldsymbol{\Lambda}}$ are replaced with $\mathbf{X}^*$ and $\boldsymbol{\theta}^*$. $\tilde{\mathbf{J}}_0^*$ and $\mathbf{S}_i^*(\boldsymbol{\theta}_0^*)$ denote the information matrix and the score when first and second derivatives for the log-likelihood function are taken with respect to $\boldsymbol{\theta}^*$.

Thus for all constant $(L+1)$-vectors $\mathbf{C}$, with $\mathbf{C}_{\{1:L\}}$ denoting the first L components of $\mathbf{C}$,

$$\begin{aligned}
&\mathbf{C}^T\begin{pmatrix}\left(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*)\right)/\sqrt{m}\\ \mathbf{a}^T\mathbf{S}_m(\theta_0)/\sqrt{m}\end{pmatrix}\\
=\;&\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\sum_{l=1}^{L}\mathbf{C}_l z_{il}^* - \mathbf{C}_{\{1:L\}}^T\tilde{\boldsymbol{\Lambda}}^*(\tilde{\mathbf{J}}_0^*)^{-1}\mathbf{S}_i^*(\boldsymbol{\theta}_0^*) + \mathbf{C}_{L+1}\mathbf{a}^T\mathbf{S}_i(\boldsymbol{\theta}_0)\right].
\end{aligned}$$

This quantity has a limiting Gaussian distribution under Assumption 3.1. Let $\boldsymbol{\tau}$ in (2.38) be the variance-covariance matrix between $\mathbf{a}^T\mathbf{S}_m(\theta_0)/\sqrt{m}$ and $(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*))/\sqrt{m}$. Then by Le Cam's third lemma (Lemma 2.14), under the contiguous alternative H_1,

$$\left(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*)\right)/\sqrt{m} \to N(\boldsymbol{\tau},\boldsymbol{\Sigma}^*).$$

Thus under H_1, with $\delta = \boldsymbol{\tau}^T(\boldsymbol{\Sigma}^*)^{-1}\boldsymbol{\tau}$,

$$\frac{1}{m}\left(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*)\right)^T(\boldsymbol{\Sigma}^*)^{-1}\left(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*)\right) \overset{H_1}{\to} \chi_k^2(\delta),$$

and

$$\frac{1}{m}\left(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*)\right)^T\left(\hat{\boldsymbol{\Sigma}}^*_{svd}\right)^{-1}\left(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*)\right)\overset{H_1}{\to}\chi^2_k(\hat{\delta}),$$

where $\hat{\boldsymbol{\Sigma}}^*_{svd}$ is the reconstructed matrix by applying Singular Value Decomposition on a consistent estimators of $\boldsymbol{\Sigma}^*$, $\hat{\boldsymbol{\tau}}$ is a consistent estimators of $\boldsymbol{\tau}$, $k=\text{rank}(\hat{\boldsymbol{\Sigma}}^*_{svd})=\text{rank}(\boldsymbol{\Sigma}^*)$ (Corollary 2.7) and the non-centrality parameter is $\hat{\delta}=\hat{\boldsymbol{\tau}}^T(\hat{\boldsymbol{\Sigma}}^*_{svd})^{-1}\hat{\boldsymbol{\tau}}$. For GLMMs, it is difficult to get the explicit form of $\boldsymbol{\tau}$, thus $\hat{\boldsymbol{\tau}}$, a consistent estimator of $\boldsymbol{\tau}$, can be taken as the empirical variance -covariance matrix between $\mathbf{a}^T\mathbf{S}_m(\theta_0)/\sqrt{m}$ and $(\mathbf{f}-\mathbf{e}^*(\hat{\boldsymbol{\theta}}^*))/\sqrt{m}$, which is

$$\frac{1}{m}\sum_{i=1}^{m}(z^*_{i1},\ldots,z^*_{iL})^T\left(\mathbf{S}^*_i(\boldsymbol{\theta}^*_0)\right)^T\mathbf{a}-\tilde{\boldsymbol{\Lambda}}^*(\tilde{\mathbf{J}}^*_0)^{-1}\frac{1}{m}\sum_{i=1}^{m}\mathbf{S}_i(\boldsymbol{\theta}_0)\left(\mathbf{S}^*_i(\boldsymbol{\theta}^*_0)\right)^T\mathbf{a},$$

with MLE $\hat{\boldsymbol{\theta}}$ substituted for parameters in the above expression.

For a given type I error level α, the approximate limiting power is thus $P(T^*>\chi^2_{k,\alpha})$, where $\chi^2_{k,\alpha}$ is the $1-\alpha$ quantile of the central χ^2_k distribution and P denotes the non central $\chi^2_k(\hat{\delta})$ distribution.

3.4 Simulations for logistic mixed models

In this Section, we implement the goodness of fit test for the 2-level logistic mixed models in R and study its performance in simulation. The two stages of the model are:

$$y_{ij}|p_{ij}\sim\text{Binom}(1,p_{ij}),\quad i=1,\ldots,m,\quad j=1,\ldots,n_i \tag{3.14}$$

$$\text{logit}(p_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i, \quad \alpha_i \sim N(0, \sigma^2).$$

Data are simulated from the following setting. We choose $m = 500$ clusters and let $n_i = 5$ for all $i = 1, \ldots, m$. Thus the total number of observation $N = \sum_{i=1}^{m} n_i = 2500$. Then N fixed covariates $\mathbf{x}_{ij} = (x_{1ij}, x_{2ij}, x_{3ij})$, were independently drew from a multivariate normal distribution,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \rho_{13} \\ 0 & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} \right). \tag{3.15}$$

Given the variance component parameter σ^2, for $m = 500$, we generate $\alpha_i, i = 1, \ldots, m$ independently from $N(0, \sigma^2)$. Given $\boldsymbol{\beta}$ and the generated $\mathbf{x}_{ij}$, $p_{ij}, i = 1, \ldots, m, \quad j = 1, \ldots, n_i$ are calculated from the equation $\text{logit}(p_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \alpha_i$. The last step of generating data is to independently generate y_{ij} from (3.14).

3.4.1 Computational issues and code checking

Implementing the goodness of fit test for logistic mixed models is more complicated than for linear mixed models because each $E(y_{ij})$ and also the variance and covariance terms involve numerical integration. We used the Gaussian quadrature numerical integration method with 60 quadrature points to calculate those integrals, which are used to calculate the $\hat{\boldsymbol{\Sigma}}$ in our test statistic. The numerical errors introduced by approximating so many integrals through the Gaussian quadrature rule may result in instability when we invert the $\hat{\boldsymbol{\Sigma}}$ matrix. It may even destroy the

invertibility of the matrix, which is used to calculate the test statistics. Thus we computed a Singular Value Decomposition for each estimated $\hat{\mathbf{\Sigma}}$. For each eigenvalue of $\hat{\mathbf{\Sigma}}$, we compare it with a preset upper bound ι (e.g. $\iota = .0005$). For any eigenvalue less than ι, we instead set this eigenvalue to be 0 and reconstruct the $\hat{\mathbf{\Sigma}}$ matrix using the non-zero eigenvalues and their corresponding eigenvectors.

We then did the following checks to make sure that our R code works.

- We compared the empirical variance covariance matrix of $(\mathbf{f} - \hat{\mathbf{e}}(\hat{\boldsymbol{\theta}}))/\sqrt{m}$ with $\hat{\mathbf{\Sigma}}$ in equation (3.11);

- Letting $L(\boldsymbol{\theta})$ denote the likelihood function, and $\mathbf{f}$ and $\mathbf{e}$ be defined in equation (3.8), we compared the empirical variance covariance matrix of $\mathbf{f} - \mathbf{e}(\boldsymbol{\theta}_0)$ and $\nabla_{\boldsymbol{\theta}} log L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, computed based on $30,000$ simulated data sets, with its estimated analytical variance, which involves terms in $\hat{\mathbf{\Sigma}}$ in equation (3.11);

- We checked in simulations that the goodness of fit test statistic (3.12) indeed has an asymptotic χ^2 distribution. We choose $\rho_{13} = \rho_{23} = 0$ and set the true parameter values $\sigma^2 = .5$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (.1, .5, -.5, .5)$. We then fit the logistic mixed model with all covariates that influence the response y. We select $L = 12$ cells in the computation of T in (3.12) based on x_1 and x_2 by using fixed cell boundaries. With the number of iterations K being 5000, we then have 5000 test statistics $(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}}))' \hat{\mathbf{\Sigma}}^{-1} (\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}}))/m$. Figure 3.1 gives the histogram of these 5000 independently calculated test statistics, which is close to χ^2_{11}, with p value from the Kolmogorov-Smirnov test being .9294 and

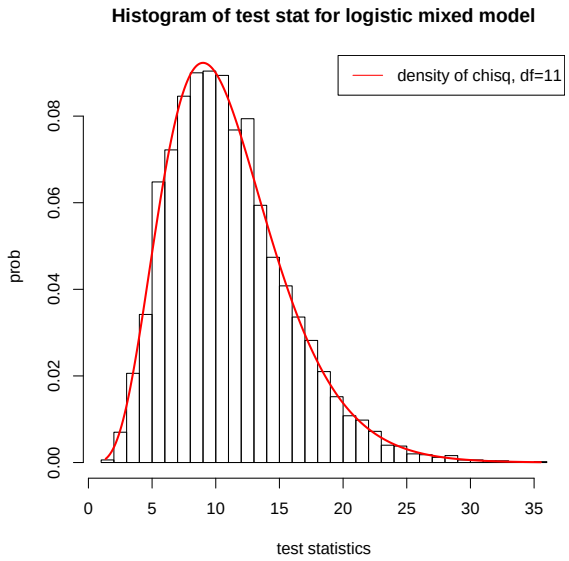

Figure 3.1: Histogram of 1000 test statistics for logistic mixed model

p value from Pearson's chi-square goodness of fit test .40 when the number of cells used is 20. The simulation result agrees with the theory.

3.4.2 Checking the size of the test in simulations

We checked the size of the test under various choices of cell partitions based on $\mathbf{X}$. We chose $\rho_{13} = \rho_{23} = 0$ in (3.15) and let $\sigma^2 = .5$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (.1, .5, -.5, .5)$, and fit the logistic mixed model with all covariates $\mathbf{X}$ in the model. All cell partitions in the computation of T in (3.12) were based on fixed cut offs. Table 3.2 shows that the empirical size estimates (Emp. Size) were close to the nominal α levels of 0.05 and 0.1 for all choices of cell partitions. This closeness was observed with greater consistency when $K = 5000$. Table 3.1 gives the empirical sizes of the test under different nominal α levels when the cell partition is based on

x_1 and the number of cells is $L = 8$. They were all very close. The third column is the standard deviation of the corresponding empirical size (ES), which is calculated as $\sqrt{ES(1-ES)/K}$.

Table 3.1: Empirical size of the test under different α levels (logistic mixed model). $m = 500, n_i = 5, \boldsymbol{\beta} = (.1, .5, -.5, .5), \sigma^2 = .5, \rho_{13} = \rho_{23} = 0, L = 12, K = 5000$.

significance level α	Empirical Size(ES)	Standard Deviation of ES
0.05	0.049	0.0031
0.1	0.099	0.0042
0.2	0.199	0.0056
0.3	0.297	0.0065
0.4	0.406	0.0069
0.5	0.505	0.0071
0.6	0.605	0.0069
0.7	0.702	0.0065
0.8	0.801	0.0056

Table 3.2: Empirical size of the test under different cell partitions (logistic mixed model). $m = 500, n_i = 5, \boldsymbol{\beta} = (.1, .5, -.5, .5), \sigma^2 = .5, \rho_{13} = \rho_{23} = 0$.

L	α	Empirical Size		α	Empirical Size	
		K=1000	K=5000		K=1000	K=5000
8 (x_1)	0.05	0.054	0.0508	0.1	0.1080	0.1034
3×4 (x_1, x_2)	0.05	0.048	0.0492	0.1	0.0950	0.0994
5×4 (x_1, x_3)	0.05	0.044	0.0494	0.1	0.0940	0.0986
6×7 (x_2, x_3)	0.05	0.056	0.0490	0.1	0.1090	0.1012

3.4.3 Simulations to assess empirical power of the test

To assess the power of the test, we fit the logistic mixed model to the data without including x_3 among the covariates. We then tried six different cell partitions based on subsets of the design matrix $\mathbf{X}$ with $L = 12$ and 42 cells. We used fixed cutoffs to do the cell partitions. We set $\sigma^2 = .5$ and $(\beta_0, \beta_1, \beta_2) = (0, .8, -.8)$ for all

Table 3.3: Impact of cell partition on empirical power I (logistic mixed model). $m = 500, \rho_{13} = \rho_{23} = 0, (\beta_0, \beta_1, \beta_2) = (0, .8, -.8), K = 1000.$

Parti	$\beta_3 = .2$		$\beta_3 = .3$		$\beta_3 = .4$	
	$L = 12$	$L = 42$	$L = 12$	$L = 42$	$L = 12$	$L = 42$
x_1	0.048	0.050	0.047	0.049	0.046	.056
x_2	0.055	0.061	0.051	0.067	0.052	.071
x_3	0.787	0.497	0.997	0.939	1	.999
x_1, x_2	0.053	0.054	0.048	0.047	0.047	.051
x_1, x_3	0.724	0.478	0.989	0.918	1	1
x_2, x_3	0.729	0.468	0.992	0.917	1	.999

simulations in this power study Section.

We set $(\rho_{13}, \rho_{23}) = (0, 0)$ and study the impact of the magnitude of β_3 on power. For a given design matrix $\mathbf{X}$, we simulated $K = 1000$ sets of $\mathbf{Y}$ and computed the empirical power of the test over $K = 1000$ iterations for $\beta_3 = .2$, .3 or .4 separately. Table 3.3 shows that with all other settings being the same, the power of the test increases as the magnitude of β_3 increases except with those choices of partition (the first second and fourth) where power is effectively constant at 0.05. To study the impact of (ρ_{13}, ρ_{23}) on power, we then fix $\beta_3 = .3$ and choose three different pairs of (ρ_{13}, ρ_{23}). For each pair of (ρ_{13}, ρ_{23}), we simulated a set of design matrix $\mathbf{X}$. We then simulated $K = 5000$ sets of $\mathbf{Y}$ based on this $\mathbf{X}$ and computed the empirical power of the test over these K iterations. Table 3.4 shows that with all other settings being the same, the power of the test decreases as the correlation between the x_3 and x_1, x_2 increases, where x_3 is the omitted covariate.

From Tables 3.3 and 3.4 we can also see that the choice of cell partition strongly affects the power. When a covariate that should be in the model is omitted, cell partitions based on that omitted covariate result in adequate power of the test.

Table 3.4: Impact of cell partition on empirical power II (logistic mixed model). $m = 500, \boldsymbol{\beta}_3 = (0, .8, -.8, .3), \sigma^2 = .5, K = 5000$.

Parti	$\rho_{13} = 0, \rho_{23} = 0$		$\rho_{13} = 0.2, \rho_{23} = 0.3$		$\rho_{13} = 0.4, \rho_{23} = 0.5$	
	$L = 12$	$L = 42$	$L = 12$	$L = 42$	$L = 12$	$L = 42$
x_1	0.047	0.049	0.046	0.054	0.054	0.058
x_2	0.051	0.067	0.05	0.053	0.047	0.047
x_3	0.997	0.939	0.988	0.883	0.902	0.657
x_1, x_2	0.048	0.047	0.047	0.046	0.046	0.061
x_1, x_3	0.989	0.918	0.974	0.862	0.831	0.630
x_2, x_3	0.992	0.917	0.975	0.850	0.818	0.557

However, if the cell partition is based only on covariates already in the model, this test has low power to detect any lack of model fit. All the above findings and conclusions in the logistic mixed model agree with what we saw earlier in the linear mixed models.

3.5 Discussion

In this Chapter, we extended the goodness of fit tests developed for Linear Mixed Models (LMMs) in Chapter 2 to Generalized Linear Random Intercept Models (GLMMs). We described the asymptotic properties of the tests when parameters were estimated through maximum likelihood. We assessed factors that impact the power and the impact of choice of cell partitions on the test in simulations for logistic mixed models. We obtained conclusions consistent with those found for the LMMs in Chapter 2, that when a specific covariate that is associated with outcome is omitted, cell partitions based on the omitted covariate result in adequate power of the test. However, if the cell partition is based only on covariates already in the model, this test has low power to detect any lack of model fit.

This type of goodness of fit test can be used to test the statistical adequacy of the finally selected GLMM in real applications. All that is needed in order to implement the test are the final model parameter estimates and their variance covariance matrix as well as the estimated means for the outcome y under the model. Implementing logistic mixed models is more complicated that implementing LMMs, as all the means and variances of the outcome y involve integrals. The covariance matrix between the estimated fixed effect parameters and the estimated variance components is not available in R, but is available in PROC nlmixed in SAS. As a note of caution, in applying the test in GLMMs, one needs to check the eigenvalues of the estimated variance covariance matrix $\hat{\mathbf{\Sigma}}$ in (3.11) and eliminate the tiny eigenvalues to make sure its inverse does not blow up. This also helps to get the correct rank of the estimated variance covariance matrix $\hat{\mathbf{\Sigma}}$ in (3.11) to ensure the correct degrees of freedom for the test statistic.

3.6 Technical details for Chapter 3

3.6.1 Checking B.2 in Assumption 3.1 for logistic mixed model

Lemma 3.8 *Let E be the event that $\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}^{\otimes 2}$ has full rank, then the conditional distribution of $(\boldsymbol{b}', c)\nabla_{\beta,\sigma^2} log f(\boldsymbol{y}_i|\{\boldsymbol{x}_{ij}\}_j, n_i)$ given $\{(\boldsymbol{x}_{ij}, n_i)\}$ on the event E is non-degenerate at $\boldsymbol{\theta}_0$.*

Proof: In order to show that $\mathbf{J}(\boldsymbol{\theta}_0;\boldsymbol{\theta}_0)$ is positive definite for the random intercept logistic mixed model, we just need to show that $\forall$ $(\mathbf{b}', c)$ non-trival, where $\mathbf{b}$ is any $p \times 1$ constant vector and c is any constant, $(\mathbf{b}', c)\nabla_{\boldsymbol{\beta},\sigma^2} \log f_i(\boldsymbol{\beta}_0, \sigma_0^2)$ is non-

degenerate. We show this in the random intercept logistic model (3.2), where $p_{ij,0} = 1/(1+e^{-(\mathbf{x}'_{ij}\boldsymbol{\beta}_0+\sigma_0 a_i)})$ and $\{\mathbf{x}_{ij}\}_j = \{\mathbf{x}_{ij}, j = 1,\ldots,n_i\}$. Then with $a_i \sim N(0,1)$

$$
\begin{aligned}
& (\mathbf{b}', c)\nabla_{\beta,\sigma^2} \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i) \\
= \;& \mathbf{b}'\frac{\partial}{\partial\beta}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i) + c\,\frac{\partial}{\partial\sigma^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i) \\
= \;& \mathbf{b}'\sum_{j=1}^{n_i}(y_{ij}-p_{ij,0})\mathbf{x}_{ij} + c\,\frac{1}{2\sigma_0}\sum_{j=1}^{n_i}(y_{ij}-p_{ij,0})a_i \\
= \;& \sum_{j=1}^{n_i}(y_{ij}-p_{ij,0})(\mathbf{b}'\mathbf{x}_{ij} + c\frac{a_i}{2\sigma_0}).
\end{aligned}
$$

Let $g_1(\mathbf{x}_{ij}^T\boldsymbol{\beta}_0, \sigma_0) = \int_{a_i} p_{ij,0}\ \phi(a_i)da_i$, and $g_2(\mathbf{x}_{ij}^T\boldsymbol{\beta}_0, \sigma_0) = \int_{a_i} p_{ij,0}\alpha_i\phi(a_i)da_i$, integrating out a_i then gives

$$
\begin{aligned}
& (\mathbf{b}^T, c)\nabla_{\beta,\sigma^2} \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i) \\
= \;& \int_{a_i}\sum_{j=1}^{n_i}(y_{ij}-p_{ij,0})(\mathbf{b}^T\mathbf{x}_{ij} + c\frac{a_i}{2\sigma_0})\phi(a_i)da_i \\
= \;& \sum_{j=1}^{n_i}\left[y_{ij}\mathbf{b}^T\mathbf{x}_{ij} + y_{ij}\ \frac{c}{2\sigma_0}E\alpha_i - \int_{a_i} p_{ij,0}\ \phi(a_i)da_i\ \mathbf{b}^T\mathbf{x}_{ij} - \frac{c}{2\sigma_0}\ \int_{a_i} p_{ij,0}\alpha_i\phi(a_i)da_i\right] \\
= \;& \sum_{j=1}^{n_i}\left(y_{ij} - g_1(\mathbf{x}_{ij}^T\boldsymbol{\beta}_0, \sigma_0)\right)\mathbf{x}_{ij}^T\mathbf{b} - \frac{c}{2\sigma_0}\sum_{j=1}^{n_i} g_2(\mathbf{x}_{ij}^T\boldsymbol{\beta}_0, \sigma_0), \qquad (3.16)
\end{aligned}
$$

where $\sum_{j=1}^{n_i} g_2(\mathbf{x}_{ij}^T\boldsymbol{\beta}_0, \sigma_0)$ is free of y_{ij}. With the fact that conditional on

$$\{(\mathbf{x}_{ij}, n_i), s.t.\ P(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{x}_{ij}^{\otimes 2} \text{ has full rank}) > 0\},$$

every $\mathbf{y}_i \in \{0,1\}^{n_i}$ has positive probability, (3.16) equals 0 only when both $\mathbf{b} = \mathbf{0}$ and $c = 0$.

3.6.2 Checking B.3 in Assumption 3.1 for Logistic mixed model

Lemma 3.9 *For the random intercept logistic mixed model (3.2), the dominatedness condition (3.3) holds under the assumption that* $\sum_{j=1}^{n_i} \|\boldsymbol{x}_{ij}^{\otimes 2}\|$ *is bounded.*

Proof: With $a_i \sim N(0,1)$, the conditional likelihood for the ith cluster, conditioning on $(\{\mathbf{x}_{ij}\}_j, n_i, a_i)$, is

$$\log f\left(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i\right) = \sum_j \left(\mathbf{x}_{ij}^T\beta + \sigma a_i\right) y_{ij} - \sum_j \log\left(1 + e^{\mathbf{x}_{ij}^T\beta + \sigma a_i}\right).$$

The first derivatives of the conditional likelihood are

$$\frac{\partial \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial \beta} = \sum_j \left(y_{ij} - \frac{1}{1 + e^{-(\mathbf{x}_{ij}^T\beta + \sigma a_i)}}\right)\mathbf{x}_{ij};$$

$$\begin{aligned}\frac{\partial \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial \sigma^2} &= \frac{1}{2\sigma}\frac{\partial \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial \sigma} \\ &= \frac{1}{2\sigma}\sum_j \left(y_{ij} - \frac{1}{1 + e^{-(\mathbf{x}_{ij}^T\beta + \sigma a_i)}}\right) a_i.\end{aligned}$$

The second derivatives of the conditional likelihood are

$$-\frac{\partial^2 \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial(\beta)^2} = \sum_j \frac{e^{\mathbf{x}_{ij}^T\beta + \sigma a_i}}{(1 + e^{\mathbf{x}_{ij}^T\beta + \sigma a_i})^2}\mathbf{x}_{ij}\mathbf{x}_{ij}^T;$$

$$-\frac{\partial^2 \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial\beta\partial\sigma^2} = \frac{1}{2\sigma}\sum_j \frac{e^{\mathbf{x}_{ij}^T\beta + \sigma a_i}}{(1 + e^{\mathbf{x}_{ij}^T\beta + \sigma a_i})^2} a_i\mathbf{x}_{ij};$$

$$
\begin{aligned}
& -\frac{\partial^2 \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial(\sigma^2)^2} \\
= & -\frac{1}{2\sigma}\frac{\partial}{\partial\sigma}\left[\frac{1}{2\sigma}\sum_j\left(y_{ij}-\frac{1}{1+e^{-(\mathbf{x}_{ij}^T\beta+\sigma a_i)}}\right)a_i\right] \\
= & \frac{1}{4\sigma^3}\sum_j\left[a_i\left(y_{ij}-\frac{1}{1+e^{-(\mathbf{x}_{ij}^T\beta+\sigma a_i)}}+\sigma a_i\frac{e^{\mathbf{x}'_{ij}\beta+\sigma a_i}}{(1+e^{\mathbf{x}_{ij}^T\beta+\sigma a_i})^2}\right)\right].
\end{aligned}
$$

We first take expectations with respect to y_{ij}, conditionally given all other, that is, $\{\mathbf{x}_{ij}\}_j, n_i, a_i$, to integrate out y_{ij}.

$$
E_{\theta_0}\left[-\frac{\partial^2}{\partial\beta^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)\right] = \sum_j \frac{e^{\mathbf{x}_{ij}^T\beta+\sigma a_i}}{(1+e^{\mathbf{x}_{ij}^T\beta+\sigma a_i})^2}\mathbf{x}_{ij}\mathbf{x}_{ij}^T, \tag{3.17}
$$

$$
E_{\theta_0}\left[-\frac{\partial^2}{\partial\beta\partial\sigma^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)\right] = \frac{1}{2\sigma}\sum_j \frac{e^{\mathbf{x}_{ij}^T\beta+\sigma a_i}}{(1+e^{\mathbf{x}_{ij}^T\beta+\sigma a_i})^2}a_i\mathbf{x}_{ij}. \tag{3.18}
$$

$$
\begin{aligned}
& E_{\theta_0}\left[-\frac{\partial^2}{\partial(\sigma^2)^2}\log f\left(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i\right)\right] = \\
& \frac{1}{4\sigma^3}\sum_j\left[a_i\left(\frac{1}{1+e^{-(\mathbf{x}_{ij}^T\beta_0+\sigma_0 a_i)}}-\frac{1}{1+e^{-(\mathbf{x}_{ij}^T\beta+\sigma a_i)}}+\sigma a_i\frac{e^{\mathbf{x}_{ij}^T\beta+\sigma a_i}}{(1+e^{\mathbf{x}'_{ij}\beta+\sigma a_i})^2}\right)\right].
\end{aligned}
$$

The two expectations in (3.17) and (3.18) have exactly the same forms as without taking the expectations since there is no term involving y_{ij}. We next integrate out a_i and with $p_{ij} = 1/(1+e^{-(\mathbf{x}_{ij}^T\beta+\sigma a_i)})$, get

$$
E_{\theta_0}\left[-\frac{\partial^2}{\partial\beta^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i)\right] = \sum_j \int_{a_i} p_{ij}(1-p_{ij})dF(a_i)\mathbf{x}_{ij}\mathbf{x}'_{ij},
$$

$$E_{\theta_0}\left[-\frac{\partial^2}{\partial\beta\partial\sigma^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i)\right] = \frac{1}{2\sigma}\sum_j\left(\int_{a_i} p_{ij}(1-p_{ij})a_i dF(a_i)\right)\mathbf{x}_{ij}.$$

$$E_{\theta_0}\left[-\frac{\partial^2}{\partial(\sigma^2)^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i)\right] =$$
$$\frac{1}{4\sigma^3}\sum_j\left(\int_{a_i}\frac{1}{1+e^{-(\mathbf{x}_{ij}^T\beta_0+\sigma_0 a_i)}}a_i dF(a_i) - \int_{a_i} p_{ij}a_i dF(a_i) + \sigma\int_{a_i} p_{ij}(1-p_{ij})a_i^2 dF(a_i)\right).$$

Since p_{ij} is bounded, the dominatedness condition (3.3) will hold by inspection if $\sum_{j=1}^{n_i}\|\mathbf{x}_{ij}^{\otimes 2}\|$ is bounded.

3.6.3 Checking B.3 in Assumption 3.1 for Poisson mixed model

Lemma 3.10 *For the random intercept Poisson mixed model, the dominatedness condition (3.3) holds under the assumption that* $E(\sum_j e^{c\sum_k|x_{ijk}|}) < \infty$, $\forall\ c \leq \max_k|\beta_k| + \epsilon$.

Proof: With $a_i \sim N(0,1)$, the conditional likelihood for the ith cluster, conditioned on $(\{\mathbf{x}_{ij}\}_j, n_i, a_i)$, is

$$\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i) = \sum_j\left[\left(\mathbf{x}_{ij}^T\beta + \sigma a_i\right)y_{ij} - e^{\mathbf{x}_{ij}'\beta+\sigma a_i} - \log(y_{ij}!)\right].$$

The first derivatives of the conditional likelihood are

$$\frac{\partial\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial\beta} = \sum_j\left(y_{ij} - e^{\mathbf{x}_{ij}'\beta+\sigma a_i}\right)\mathbf{x}_{ij};$$

$$\frac{\partial \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial \sigma^2} = \frac{1}{2\sigma}\frac{\partial \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial \sigma}$$
$$= \frac{1}{2\sigma}\sum_j \left(y_{ij} - e^{\mathbf{x}'_{ij}\beta+\sigma a_i}\right) a_i.$$

The second derivatives of the conditional likelihood are

$$-\frac{\partial^2 \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial(\beta)^2} = \sum_j e^{\mathbf{x}'_{ij}\beta+\sigma a_i}\mathbf{x}_{ij}\mathbf{x}'_{ij};$$

$$-\frac{\partial^2 \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial\beta\partial\sigma^2} = \frac{1}{2\sigma}\sum_j e^{\mathbf{x}'_{ij}\beta+\sigma a_i}a_i\mathbf{x}_{ij};$$

$$-\frac{\partial^2 \log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)}{\partial(\sigma^2)^2} = -\frac{1}{2\sigma}\frac{\partial}{\partial\sigma}\left[\frac{1}{2\sigma}\sum_j \left(y_{ij} - e^{\mathbf{x}'_{ij}\beta+\sigma a_i}\right) a_i\right]$$
$$= \frac{1}{4\sigma^3}\sum_j \left[a_i\left(y_{ij} - e^{\mathbf{x}'_{ij}\beta+\sigma a_i}\right) + \sigma a_i^2 e^{\mathbf{x}'_{ij}\beta+\sigma a_i}\right].$$

We first take expectations with respect to y_{ij}, conditionally given $\{\mathbf{x}_{ij}\}_j, n_i, a_i$, to integrate out y_{ij}.

$$E_{\theta_0}\left[-\frac{\partial^2}{\partial\beta^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)\right] = \sum_j e^{\mathbf{x}'_{ij}\beta+\sigma a_i}\mathbf{x}_{ij}\mathbf{x}'_{ij}, \tag{3.19}$$

$$E_{\theta_0}\left[-\frac{\partial^2}{\partial\beta\partial\sigma^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)\right] = \frac{1}{2\sigma}\sum_j e^{\mathbf{x}'_{ij}\beta+\sigma a_i}a_i\mathbf{x}_{ij}. \tag{3.20}$$

$$E_{\theta_0}\left[-\frac{\partial^2}{\partial(\sigma^2)^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i, a_i)\right] = \frac{1}{4\sigma^2}\sum_j a_i^2 e^{\mathbf{x}'_{ij}\beta+\sigma a_i} + \frac{1}{4\sigma^3}\sum_j\left[a_i\left(e^{\mathbf{x}'_{ij}\beta_0+\sigma_0 a_i} - e^{\mathbf{x}'_{ij}\beta+\sigma a_i}\right)\right].$$

The two expectations in (3.19) and (3.20) have exactly the same forms as without taking the expectations since there is no term involving y_{ij}. We next integrate out a_i. By making use of the three equations $E(e^{\sigma a_i}) = e^{\sigma^2/2}$, $E_{a_i}(a_i^2 e^{\sigma a_i}) = (1+\sigma^2)\, e^{\sigma^2/2}$ and $E_{a_i}(a_i e^{\sigma a_i}) = \sigma e^{\sigma^2/2}$, we get

$$E_{\theta_0}\left[-\frac{\partial^2}{\partial\beta^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i)\right] = \left(\sum_j e^{\mathbf{x}'_{ij}\beta}\mathbf{x}_{ij}\mathbf{x}'_{ij}\right) e^{\sigma^2/2}.$$

$$E_{\theta_0}\left[-\frac{\partial^2}{\partial\beta\partial\sigma^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i)\right] = \frac{1}{2}\, e^{\sigma^2/2}\left(\sum_j e^{\mathbf{x}'_{ij}\beta}\mathbf{x}_{ij}\right).$$

$$\begin{aligned} & E_{\theta_0}\left[-\frac{\partial^2}{\partial(\sigma^2)^2}\log f(\mathbf{y}_i|\{\mathbf{x}_{ij}\}_j, n_i)\right] = \frac{1}{4\sigma^2}\left(\sum_j e^{\mathbf{x}'_{ij}\beta}\right) E\left(a_i^2 e^{\sigma a_i}\right) + \\ & \frac{1}{4\sigma^3}\left[\left(\sum_j e^{\mathbf{x}'_{ij}\beta_0}\right) E\left(a_i e^{\sigma_0 a_i}\right) - \left(\sum_j e^{\mathbf{x}'_{ij}\beta}\right) E\left(a_i e^{\sigma a_i}\right)\right] \\ = & \frac{1}{4\sigma^2}\left(\sum_j e^{\mathbf{x}'_{ij}\beta}\right)\, e^{\sigma^2/2}\,(1+\sigma^2) + \\ & \frac{1}{4\sigma^3}\left[\left(\sum_j e^{\mathbf{x}'_{ij}\beta_0}\right)\sigma_0 e^{\sigma_0^2/2} - \left(\sum_j e^{\mathbf{x}'_{ij}\beta}\right)\, \sigma e^{\sigma^2/2}\right] \\ = & \frac{1}{4}e^{\sigma^2/2}\left(\sum_j e^{\mathbf{x}'_{ij}\beta}\right) + \frac{\sigma_0}{4\sigma^3}e^{\sigma_0^2/2}\left(\sum_j e^{\mathbf{x}'_{ij}\beta_0}\right). \end{aligned}$$

(3.3) will hold by inspection if assumptions are imposed ensuring that $\sup_{\beta\in B_\epsilon}(\sum_j e^{\mathbf{x}'_{ij}\beta}(1+\mathbf{x}'_{ij}\mathbf{x}_{ij}))$ is integrable. The following condition,

$E(\sum_j e^{c\sum_k |x_{ijk}|}) < \infty$, $\forall\ c \leq \max_k |\beta_k| + \epsilon$, is sufficient to ensure that.

3.6.4 Simplification of $\hat{\Sigma}$ in equation (3.11)

In this section, we manipulate the expression for the estimated asymptotic variance $\hat{\boldsymbol{\Sigma}}$ in equation (3.11). The notations Var^C, Cov^C, E^C mean the corresponding quantities involve integrals over $\mathbf{y}_i$ alone, conditionally given $\{\mathbf{x}_i, n_i\}_i = \{(\mathbf{x}_i, n_i),\ i = 1, \ldots, m\}$. Because $\tilde{\mathbf{J}}_0$ and $\tilde{\boldsymbol{\Lambda}}$ are functions of only $\{\mathbf{x}_i, n_i\}_i$, conditionally given $\{\mathbf{x}_i, n_i\}_i$,

$$
\begin{aligned}
& Var^C(\boldsymbol{\xi}_i) = Var(\boldsymbol{\xi}_i | \{\mathbf{x}_i, n_i\}_i) \\
= \ & Var^C\left(z_{i1}, \ldots, z_{iL}\right)^T + Var^C(\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\mathbf{S}_i(\boldsymbol{\theta}_0)) - \\
& 2Cov^C\left[\left(z_{i1}, \ldots, z_{iL}\right)^T, \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\mathbf{S}_i(\boldsymbol{\theta}_0)\right] \\
= \ & Var^C\left(z_{i1}, \ldots, z_{iL}\right)^T + \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1} Var^C(\mathbf{S}_i(\boldsymbol{\theta}_0))\tilde{\mathbf{J}}_0^{-1}\tilde{\boldsymbol{\Lambda}}^T - \\
& 2\left(Cov^C\left(z_{i1}, \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\mathbf{S}_i(\boldsymbol{\theta}_0)\right), \ldots, Cov^C\left(z_{iL}, \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\mathbf{S}_i(\boldsymbol{\theta}_0)\right)\right)^T,
\end{aligned}
$$

where $Cov^C\left(z_{il}, \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\mathbf{S}_i(\boldsymbol{\theta}_0)\right) = E^C_{\mathbf{y}_i}\left(z_{il}\mathbf{S}_i^T(\boldsymbol{\theta}_0)\right)\tilde{\mathbf{J}}_0^{-1}\tilde{\boldsymbol{\Lambda}}^T$. We next simplify $E^C_{\mathbf{y}_i}\left(z_{il}\mathbf{S}_i^T(\boldsymbol{\theta}_0)\right)$, with notation $f^C_{\mathbf{y}_i}$ meaning the conditional density of $f_{\mathbf{y}_i}$ given $\{\mathbf{x}_i, n_i\}_i$,

$$
\begin{aligned}
E^C_{\mathbf{y}_i}(z_{il}\mathbf{S}_i^T(\boldsymbol{\theta}_0)) &= \int_{\mathbf{y}_i}\left(z_{il}\ \frac{\nabla f^C_{\mathbf{y}_i}}{f^C_{\mathbf{y}_i}}\right) f^C_{\mathbf{y}_i}\ d\mathbf{y}_i \ = \ \int_{\mathbf{y}_i} z_{il}\ \left(\nabla f^C_{\mathbf{y}_i}\right)\ d\mathbf{y}_i \\
&= \nabla \int_{\mathbf{y}_i} z_{il}\ f^C_{\mathbf{y}_i}\ dy_i \ - \ \int_{\mathbf{y}_i}(\nabla z_{il})\ f^C_{\mathbf{y}_i}\ d\mathbf{y}_i,
\end{aligned}
$$

and the first of these last two integrals is 0 because of the identity:

$$\int_{\mathbf{y}_i} z_{il}\ f^C_{\mathbf{y}_i} d\mathbf{y}_i = E^C_{\mathbf{y}_i}(z_{il}) = 0.$$

Thus

$$\begin{aligned} E^C_{\mathbf{y}_i}\left(z_{il}\mathbf{S}^T_i(\boldsymbol{\theta}_0)\right) &= -\int_{y_i} (\nabla z_{il}) f^C_{y_i} dy_i \\ = \int_{y_i} \left(\sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}} \nabla E^C y_{ij}\right) f^C_{y_i}\ dy_i &= \sum_{j=1}^{n_i} I_{\{\mathbf{x}_{ij}\in E_l\}} \nabla E^C y_{ij}. \end{aligned}$$

The last equality holds because $\nabla E^C y_{ij}$ is not a function of $\mathbf{y}_i$ and $\int_{y_i} f^C_{y_i}\ dy_i = 1$.

By the preceding calculations, an by definition of $\mathbf{e}(\boldsymbol{\theta}_0)$,

$$\begin{pmatrix} \sum_{i=1}^m E^C\left[z_{i1}\mathbf{S}^T_i(\boldsymbol{\theta}_0)\right] \\ \vdots \\ \sum_{i=1}^m E^C\left[z_{iL}\mathbf{S}^T_i(\boldsymbol{\theta}_0)\right] \end{pmatrix} = \nabla\ \mathbf{e}(\boldsymbol{\theta}_0). \tag{3.21}$$

Let $\tilde{\Sigma} = \sum_{i=1}^m Var\left(\boldsymbol{\xi}_i \mid \{\mathbf{x}_i, n_i\}_i\right)/m$ and $\Sigma = \lim_{m\to\infty} \tilde{\Sigma}$. Then

$$
\begin{aligned}
\tilde{\Sigma} &= \frac{1}{m}\sum_{i=1}^{m} Var(\boldsymbol{\xi}_i \mid \{\mathbf{x}_i, n_i\}_i) \\
&= \frac{1}{m}\sum_{i=1}^{m} Var^C\left(z_{i1}, \ldots, z_{iL}\right)^T + \frac{1}{m}\sum_{i=1}^{m} \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1} Var^C(\mathbf{S}_i(\boldsymbol{\theta}_0))\tilde{\mathbf{J}}_0^{-1}\tilde{\boldsymbol{\Lambda}}^T - \\
&\quad \frac{1}{m}\sum_{i=1}^{m} 2\left(Cov^C\left(z_{i1}, \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\mathbf{S}_i(\boldsymbol{\theta}_0)\right), \ldots, Cov^C\left(z_{iL}, \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\mathbf{S}_i(\boldsymbol{\theta}_0)\right)\right)^T \\
&= \frac{1}{m}\sum_{i=1}^{m} Var^C\left(z_{i1}, \ldots, z_{iL}\right)^T + \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\tilde{\boldsymbol{\Lambda}}^T - 2\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\tilde{\boldsymbol{\Lambda}}^T \\
&= \frac{1}{m}\sum_{i=1}^{m} Var^C\left(z_{i1}, \ldots, z_{iL}\right)^T - \tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{J}}_0^{-1}\tilde{\boldsymbol{\Lambda}}^T,
\end{aligned}
$$

The existence of $m^{-1}\ \sum_{i=1}^{m} Var\left(z_{i1}, \ldots, z_{iL}\right)^T$ is ensured by Assumption 3.1.

Under the assumption that $(\mathbf{x}_i, n_i)$ are i.i.d, $\tilde{\boldsymbol{\Sigma}}$ converge in probability to $\boldsymbol{\Sigma}$, the limiting variance covariance matrix of $(\mathbf{f} - \mathbf{e}(\hat{\boldsymbol{\theta}}))/\sqrt{m}$. With $\boldsymbol{\theta}$ in $\tilde{\boldsymbol{\Sigma}}$ replaced by its MLE $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\Sigma}} = \frac{1}{m}\sum_{i=1}^{m} \hat{Var}(\boldsymbol{\xi}_i|\mathbf{x}_i, n_i)$ is a consistent estimator of $\boldsymbol{\Sigma}$.

Chapter 4

Discussion and further research

4.1 Discussion

Schoenfeld (1980) presented a class of omnibus chi-squared goodness of fit tests for the proportional hazards regression model. We adapted this idea and proposed a class of goodness of fit tests for testing the statistical adequacy of a 2-level generalized linear mixed model (GLMM). We described the asymptotic properties of the test when parameters were estimated through maximum likelihood. For a special case of linear mixed models (LMMs), we extended this test to 2-level LMMs with no distributional assumptions for either the random effect α_i or the error term ϵ_{ij}. We also extended the test to multi-level LMMs.

We assessed factors that impact the power and the impact of choice of cell partitions on the test in simulations for both linear mixed models (LMMs) and logistic mixed models. We found that when a specific covariate that is associated with outcome is omitted, cell partitions based on the omitted covariate result in adequate power of the test. However, if the cell partition is based only on covariates already in the model, this test has low power to detect any lack of model fit. For LMMs, we also conducted simulations to show that our test was very robust to violations of the normality assumption of the error distribution when we use symmetric distributions.

For LMMs, we developed the theoretical power of the test under the local

alternative. We then checked this theoretical power, obtained from Le Cam's third lemma, with simulations. We found that the estimated theoretical power calculated using Le Cam's third lemma was reliable at least when the number of clusters m is above 50. However, when m is very small, it may be advisable to rely on the empirical power computed through simulations. For LMMs, we also proposed a criteria parameter Δ that is closely related to the power.

This goodness of fit test can be used to test the statistical adequacy of the finally selected GLMM in real applications. All that is needed in order to implement the test are the final model parameter estimates and their variance covariance matrix as well as the estimated means for the outcome y under the model. It is easy to implement the test in LMMs as these are standard outputs from any statistical software. Implementing logistic mixed models is relatively more complicated, as all the means and variances of the outcome y involve integrals. Also, the covariance matrix between the estimated fixed effect parameters and the estimated variance components is not available in R, but is available in PROC nlmixed in SAS. As a note of caution, in applying the test in GLMMs, one needs to check the eigenvalues of the estimated variance covariance matrix $\hat{\mathbf{\Sigma}}$ in (3.12) and eliminate the tiny eigenvalues to make sure its inverse won't blow up. This also helps to get the correct rank of the estimated variance covariance matrix $\hat{\mathbf{\Sigma}}$ in (2.13) to ensure the correct degrees of freedom for the test statistic.

4.2 Further research

In this thesis, I applied the Schoenfeld's residual approach [30] to test the goodness of fit of classical generalized linear mixed models, via quadratic goodness of fit statistics. I then applied this test to three biomedical data, testing the goodness of fit of two-level linear mixed models. One of the further research will be applying this test to multilevel and longitudinal data once we have these kinds of data. Another aspect of further research will be comparing our proposed goodness of fit test with the existing bootstrap and Bayesian approaches.

Chapter 5

Appendix: Three general lemmas

The following three Lemmas are used in several places in the text.

Lemma 5.1 *Suppose that* $\{u_{in} : \quad n \geq 1, \quad 1 \leq i \leq n\}$ *is a triangular array of independent identically distributed random variables within each row (i.e., across* i*) with mean* 0 *and finite variance* σ_u^2*, and that these variables are independent of the random array* $\{c_{in} : \quad n \geq 1, \quad 1 \leq i \leq n\}$ *which satisfies the additional properties that as* $n \to \infty$

$$\text{(a)} \quad \max_{1 \leq i \leq n} |c_{in}| \to 0 \qquad \text{and} \qquad \text{(b)} \quad \sum_{i=1}^{n} c_{in}^2 \to \kappa$$

in probability, where $\kappa \in (0, \infty)$*. Then* $\sum_{i=1}^{n} c_{in}\, u_{in} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \kappa)$ *as* $n \to \infty$*.*

Proof. $\{\sum_{i=1}^{k} c_{in}\, u_{in}\}_{k=1}^{n}$ is a martingale with respect to the filtration $\mathcal{F}_{kn} = \sigma(\{c_{in}, .u_{in} : \quad 1 \leq i \leq k\})$. The Lemma is an immediate consequence of the Martingale Central Limit Theorem (D. McLeish 1974; P. Hall and C. Heyde 1981). A direct proof from the standard Lindeberg Central Limit Theorem (applied conditionally given $\{c_{in}\}_{i=1}^{n}$) is also not difficult. □

Lemma 5.2 *Let* $\mathbf{\Sigma}$ *be a* $q \times q$ *covariance matrix of rank* k *with smallest non-zero*

eigenvalue λ*, and let* $\zeta \in (0, \lambda)$ *be arbitrary. Let* $\hat{\mathbf{\Sigma}}_N$ *be a sequence of random* $q \times q$ *covariance matrices such that with probability approaching* 1 *as* $N \to \infty$*, the smallest positive eigenvalue of* $\hat{\Sigma}_N$ *is greater than* ζ*. If* $\hat{\mathbf{\Sigma}}_N \xrightarrow{P} \mathbf{\Sigma}$*, then* $P\left(\text{rank } (\hat{\mathbf{\Sigma}}_N) = k\right) \to 1$*, as* $N \to \infty$*.*

Proof. Let $\mathcal{V}$ and $\mathcal{V}^\perp$ be the range and null space of $\mathbf{\Sigma}$ respectively, and $\dim(\mathcal{V}) = \text{rank}(\mathbf{\Sigma}) = k$. Because $\hat{\mathbf{\Sigma}}_N \xrightarrow{P} \mathbf{\Sigma}$, we have that $\forall\ \mathbf{w} \in \mathcal{V}^\perp$ with $\mathbf{w} \neq \mathbf{0}$, $\|\hat{\mathbf{\Sigma}}_N \mathbf{w}\| \xrightarrow{P} 0$, thus

$$P(\|\hat{\mathbf{\Sigma}}_N \mathbf{w}\| \leq \zeta \|\mathbf{w}\|) \to 1.$$

By the hypothesis on $\hat{\mathbf{\Sigma}}_N$ stated in the Lemma,

$$P\left(\left[\ \|\hat{\mathbf{\Sigma}}_N \mathbf{w}\| \leq \zeta \|\mathbf{w}\|\ \right] \bigcap \left[\ \mathbf{w} \neq \mathbf{0}\ \right]\right) \to 0.$$

Thus

$$P(\mathbf{w} \in \{\text{null space of } \hat{\mathbf{\Sigma}}_N\}) \to 1, \quad \forall\ \mathbf{w} \in \mathcal{V}^\perp,$$

where $\mathcal{V}^\perp$ is the null space of $\mathbf{\Sigma}$. Similarly, $\forall\ \mathbf{v} \in \mathcal{V}$ with $\mathbf{v} \neq \mathbf{0}$, $\hat{\mathbf{\Sigma}}_N \mathbf{v} \xrightarrow{P} \mathbf{\Sigma}\mathbf{v}$. Thus

$$\|\hat{\mathbf{\Sigma}}_N \mathbf{v}\| \xrightarrow{P} \|\mathbf{\Sigma}\mathbf{v}\| \geq \lambda \|\mathbf{v}\| > \zeta \|\mathbf{v}\|,$$

which leads to $P(\hat{\mathbf{\Sigma}}_N \mathbf{v} \neq \mathbf{0}) \to 1$. Overall, $\forall\ \mathbf{w} \in \mathcal{V}^\perp$ with $\mathbf{w} \neq \mathbf{0}$ and $\forall\ \mathbf{v} \in \mathcal{V}$ with $\mathbf{v} \neq \mathbf{0}$,

$$P(\mathbf{w} \in \text{null space of } \hat{\mathbf{\Sigma}}_N,\ \mathbf{v} \in \text{range of } \hat{\mathbf{\Sigma}}_N) \to 1.$$

Therefore, $P\left(\text{rank } (\hat{\mathbf{\Sigma}}_N) = k\right) \to 1$, as $N \to \infty$. □

Corollary 5.3 *Let the symmetric matrix* $\hat{\mathbf{\Sigma}}_N^0$ *be a consistent estimator sequence for a* $q \times q$ *covariance matrix* $\mathbf{\Sigma}$ *whose smallest non-zero eigenvalue is* λ. *Let* ζ *be an arbitrary number, where* $0 < \zeta < \lambda$. *If we represent* $\hat{\mathbf{\Sigma}}_N^0$ *in terms of an orthonormal eigenbasis by* $\hat{\mathbf{\Sigma}}_N^0 = \sum_{k=1}^{q} c_k v_{kN} v_{kN}^T$ *and define the random matrix* $\hat{\mathbf{\Sigma}}_N = \sum_{k=1}^{q} c_k \, I_{[c_k > \zeta]} \, v_{kN} v_{kN}^T$, *then* $P(rank\ (\hat{\mathbf{\Sigma}}_N) = rank\ (\mathbf{\Sigma})) \to 1$.

Proof. Let $\mathcal{V}$ and $\mathcal{V}^{\perp}$ be the range and null space of $\mathbf{\Sigma}$ respectively. Because $\hat{\mathbf{\Sigma}}_N^0 \xrightarrow{P} \mathbf{\Sigma}$, we have that $\forall\ \mathbf{w} \in \mathcal{V}^{\perp}$ with $\mathbf{w} \neq \mathbf{0}$ and $\forall\ \mathbf{v} \in \mathcal{V}$ with $\mathbf{v} \neq \mathbf{0}$,

$$\hat{\mathbf{\Sigma}}_N^0 \mathbf{w} \to \mathbf{0}, \quad (\hat{\mathbf{\Sigma}}_N^0 - \mathbf{\Sigma})\mathbf{v} \xrightarrow{P} \mathbf{0}.$$

By definition of the matrix $\hat{\mathbf{\Sigma}}_N$,

$$\left\{\mathbf{w} :\ \|\hat{\mathbf{\Sigma}}_N^0 \mathbf{w}\| \leq \zeta \|\mathbf{w}\|\right\} \ \subset\ \left\{\mathbf{w} :\ \hat{\mathbf{\Sigma}}_N \mathbf{w} = \mathbf{0}\right\}, \text{ and } \|\hat{\mathbf{\Sigma}}_N \mathbf{v}\| \ >\ \zeta \|\mathbf{v}\|.$$

Thus $P(\hat{\mathbf{\Sigma}}_N \mathbf{w} = \mathbf{0}) \to 1$ and $P(\hat{\mathbf{\Sigma}}_N \mathbf{v} \neq \mathbf{0}) \to 1$. Overall, $\forall\ \mathbf{w} \in \mathcal{V}^{\perp}$ with $\mathbf{w} \neq \mathbf{0}$ and $\forall\ \mathbf{v} \in \mathcal{V}$ with $\mathbf{v} \neq \mathbf{0}$,

$$P(\mathbf{w} \in \text{range of } \hat{\mathbf{\Sigma}}_N,\ \mathbf{v} \in \text{null space of } \hat{\mathbf{\Sigma}}_N) \to 1.$$

Therefore, $P\left(\text{rank } (\hat{\mathbf{\Sigma}}_N) = \text{rank } (\mathbf{\Sigma})\right) \to 1$, as $N \to \infty$. □

Bibliography

[1] Agresti, A. *Categorical Data Analysis*, 2nd ed. New York: Wiley, 2002.

[2] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100-1120.

[3] Bickel, P. J. and Doksum, K. A. *Mathematical Statistics*, 2nd ed. Prentice Hall; 2006.

[4] Christiansen, C. L. and Morris, C. N. (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Assocation*. **92**, pp. 618-632.

[5] Claeskens, G. and Hart, J. D. (2009). Goodness-of-fit tests in mixed models. *TEST*. **18**, pp. 213-239.

[6] Cox, D. R. (1961). Tests of separate families of hypotheses. Proc. Fourth Berkeley Symp. on Math. Statist. and Prob., Vol. **1**, 105-123.

[7] Crainiceanu, C. M. and Ruppert D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Statist. Soc. B* **66**, 165-185.

[8] Dorgan, J. F. , Baer, D. J. , Albert, P. S. , Judd, J. T. , Brown, E. D. , Corle, D. K. , et al. (2001) Serum hormones and the alcohol-breast cancer association in postmenopausal women. *J Natl Cancer Inst* **93**, pp. 710-5.

[9] Godfrey, L. G. *Misspecification Tests in Econometrics*. Cambridge, 1988.

[10] Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., and Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis*, Vol. **51**, pp, 5142-5154.

[11] Jiang, J. (1998b). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, **93**, pp. 720-729.

[12] Jiang, J. (1999). Conditional inference about generalized linear mixed models. *The Annals of Statistics*. **27**, pp. 1974-2007.

[13] Jiang, J. (2001). Goodness-of-fit tests for mixed model diagnostics. *The Annals of Statistics* **4**, 1137-1164.

[14] Jiang, J. and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test*. Vol. **15**, No. **1**, pp. 1-96.

[15] Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*. Wiley, New York.

[16] Khuri, A. I., Mathew, T. and Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models*. New York: Wiley.

[17] Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*. Vol. 38, No. **4**, pp. 963-974.

[18] Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *Annals of Statistics*. Vol. **17**, No, **2**, pp. 624-642.

[19] Lee, Y., Nelder, J. A., and Pawitan, Y.. *Generalized Linear Models with Random Effects Unified Analysis via H-likelihood*. Chapman Hall/CRC, 2006.

[20] Lin, D. Y., Wei, L. J. and Ying, Z. (2002) Model-checking techniques based on cumulative residuals. *Biometrics*, Vol **58**, No. **1**, pp. 1-12.

[21] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd ed.

[22] McCulloch, C. E. and Searle, S. R. *Generalized, Linear, and Mixed Models*. Wiley, 2001.

[23] Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics*. **5**, 746-762.

[24] Pan, Zhiying and Lin, D. Y. (2005). Goodness-of-Fit Methods for Generalized Linear Mixed Models. *Biometrics*. **61**, 1000-1009.

[25] Park, T. and Lee, S.-Y. (2004). Model diagnostic plots for repeated measures data. *Biometrical Journal*, **46**, pp. 441-452.

[26] Prasad, N. G. N. and Rao, J. N. K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statitical Association*. Vol. **85**, No. **409**, pp. 163-171.

[27] Rao, C. R. and Wu, Y. (1989) A strongly consistent procedure for model selection in a regression problem. *Biometrika*, Vol. **76**, No. **2**, pp. 369-374.

[28] Rao, C. R., Wu, Y., Konishi, S. and Mukerjee, R. (2001) On model selection. *IMS Lecture Notes - Monograph Series*, Vol. **38**, pp. 1-64.

[29] Ritz, C. (2004). Goodness-of-fit tests for mixed models. *Board of the Foundation of the Scandinavian Journal of Statistics* **31**, 443-458.

[30] Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67**, 145-153.

[31] Self, S. G. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605-610.

[32] Shao, J. (1997) An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221-264.

[33] Slud, E. and Kedem, B. (1994). Partial likelihood analysis of logistic regression and autoregression. *Statistica Sinica* **4**, 89-106.

[34] Stezhko, VA, Buglova, EE, Danilova, LI, et al (2004). A cohort study of thyroid cancer and other thyroid diseases after the Chornobyl accident: Objectives, design and methods. *RADIATION RESEARCH* Volume **161**, Issue **4**, 481-492.

[35] Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **67**, Issue **1**, pp. 250-251.

[36] Van der Vaart, A. W. , *Asymptotic Statistics*. Cambridge, 2000.

[37] Wand, M. P. (2007). Fisher information for generalised linear mixed models. *Journal of Multivariate Analysis*. **98**, 1412-1416.

[38] Weiss, R. E. *Modelling Longitudinal data*. Springer, 2005.