



Published in final edited form as:

*J Multivar Anal.* 2014 September ; 130: 176–193. doi:10.1016/j.jmva.2014.03.012.

## Goodness of Fit Tests for Linear Mixed Models

Min Tang<sup>1,\*</sup>, Eric V. Slud<sup>2</sup>, and Ruth M. Pfeiffer<sup>3,\*</sup>

<sup>1</sup>Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA

<sup>2</sup>Department of Mathematics, University of Maryland, College Park, MD, USA

<sup>3</sup>National Cancer Institute, Division of Cancer Epidemiology and Genetics, Bethesda, MD, USA

### Abstract

Linear mixed models (LMMs) are widely used for regression analysis of data that are assumed to be clustered or correlated. Assessing model fit is important for valid inference but to date no confirmatory tests are available to assess the adequacy of the fixed effects part of LMMs against general alternatives. We therefore propose a class of goodness-of-fit tests for the mean structure of LMMs. Our test statistic is a quadratic form of the difference 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 when model parameters are estimated by maximum likelihood or by least squares and method of moments, and study its power under local alternatives both analytically and in simulations. Data on repeated measurements of thyroglobulin from individuals exposed to the accident at the Chernobyl power plant in 1986 are used to illustrate the proposed test.

### Keywords

asymptotic efficiency; information matrix; maximum likelihood estimators; method of moments; model fit; random effects

## 1. Introduction

The linear mixed model (LMM) (McCulloch and Searle, 2001) extends the linear model by including random effects in addition to the usual fixed effects in the linear predictors. By incorporating random effects LMMs can accommodate clustered or correlated data. Developments in model fitting algorithms and their implementations in statistical packages (e.g. *lme* in *R*; PROC Mixed in SAS 9.2; SAS Institute, Cary, NC) have greatly facilitated the applications of LMMs.

Two important steps in modeling are selecting a model and checking its fit. Often model selection is done by comparing nested models, via likelihood ratio or score tests, as part of

\*Corresponding authors. amethyst1984tm@gmail.com; pfeiffer@mail.nih.gov.

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

model building, and approaches are also available for comparing non-nested models (Cox, 1961; Godfrey, 1988). Variables are often selected for inclusion into a model if their p-value obtained from a Wald test meets some significance criterion. AIC, BIC and other model selection principles (Rao and Wu, 1989; Shao, 1997) also focus on selection of covariates. Once a model is selected, its fit should be assessed. For fixed effects models this is done by checking residuals and formal goodness of fit tests, such as score or Wald tests or likelihood ratio tests based on nested models. Khuri, Mathew and Sinha (1998) discussed likelihood ratio testing for fixed effects within LMMs. The literature for assessing the fit of LMMs against general alternatives is limited, and is mostly concerned with specification of the random effect distributions. Likelihood ratio testing for the presence of random effects in LMMs has been discussed by Self and Liang (1987) and Crainiceanu and Ruppert (2004). Jiang (2001) and Ritz (2004) assessed the distributional assumptions for the random effects in LMMs. Claeskens and Hart (2009) proposed tests for normality of the random effects and/or error terms. Lombardía and Sperlich (2008) introduced a test for the hypothesis of a linear fixed effect part in a generalized linear mixed model against the alternative of a semiparametric fixed effect part. Pan and Lin (2005) propose checking the adequacy of 2-level generalized linear mixed models based on the maximum absolute partial sums of residuals over a scalar projection of covariates. Their approach allows for assessing overall model fit as well as the functional form of individual components of the fixed effects part. However, to date there is no general easily computable test for checking the fit of the fixed-effect part of a model against unspecified alternatives, including omitted covariates or interaction terms or misspecifications of the functional form of covariates. Such a test is needed as a model-building tool.

Examination of the residuals of a model is a standard way to judge the quality of model fit. This can be done in many different ways. One useful way is to classify the response into mutually exclusive events defined in terms of the covariates and then assess for each category the deviation of the observed values and the expected values under the model. For survival data, Schoenfeld (1980) presented a class of omnibus chi-squared goodness of fit tests for the proportional hazards regression model, based on the observed minus the expected values of the covariates at each failure time. In this article, we adopt the idea of Schoenfeld (1980) and develop a goodness of fit test for the mean structure of LMMs by comparing the observed and expected values computed from the model within cells of a partition of the covariate space.

The rest of the paper is organized as follows. In Section 2 we present the linear mixed model, introduce the goodness of fit test statistic, and derive its asymptotic properties, including its theoretical power under local alternatives. We first assume that the random effects components and the error term are normally distributed and parameters are estimated by maximum likelihood (Section 2.2). We then relax the assumption of normality and only require finite higher order moments for the random effect and the error term and estimate parameters using least squares and method of moments (Section 2.3). We study the power of the test in simulations in Section 3, present a data example in Section 4 and close with a discussion in Section 5.

## 2. Goodness of fit test statistic for linear mixed models

### 2.1. The linear mixed model

We consider the linear mixed model (LMM) with additive random effects,

$$\mathbf{Y} = \mathbf{X}\beta + \sum_{r=1}^R \mathbf{Z}_r \alpha_r + \varepsilon, \quad (1)$$

where  $\mathbf{Y}_{N \times 1}$  is the vector of observations;  $\mathbf{X}_{N \times p} = (\mathbf{x}_1^T, \dots, \mathbf{x}_N^T)$  is the design matrix for the fixed effects part of the model, where  $\mathbf{x}_i$  denotes the  $p \times 1$  covariate vector for individual  $i$ ;  $\beta$  is a  $p \times 1$  vector of unknown fixed effects parameters;  $\mathbf{Z}_r$  is the known  $N \times m_r$  design matrix for the random effect  $\alpha_r$ , an  $m_r \times 1$  random vector, for  $r = 1, \dots, R$ . The random effects  $\alpha_1, \dots, \alpha_R$  are i.i.d. and independent of the error term  $\varepsilon$ . In the next section we assume that the components  $\alpha_{kr}$  of  $\alpha_r$  and  $\varepsilon$  are normally distributed. Within the LMM with a single random effect, we later require no distributional assumptions on the random effect and the error terms, but only the finiteness of their  $4 + \delta$  moments for some  $\delta > 0$ . We let  $\theta = (\beta, \psi)$  be the parameters of model (1), where  $\psi = (\sigma_\varepsilon^2, \sigma_1^2, \dots, \sigma_R^2)$  is the vector of all variance components.

An important special case of model (1) is the 2-level LMM, that includes only a single random effect,

$$y_{ij} = \mathbf{x}_{ij}^T \beta + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n_i, \quad (2)$$

where, using a slightly different notation, the  $1 \times p$  vector  $\mathbf{x}_{ij}^T = (1, x_{ij1}, \dots, x_{ij(p-1)})$  denotes covariates for the  $j$ th observation within the  $i$ th cluster. The first entry in  $\mathbf{x}_{ij}$  is set to be 1 to accommodate an intercept term in the model. We let  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$  denote the vector of observations for the  $i$ th cluster. The normally distributed cluster specific random effects  $\alpha_i \sim N(0, \sigma_\alpha^2)$  are assumed to be independent of the error terms  $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ . Then under model (2),  $\mathbf{Y}$  is also normal,  $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{V})$  with a block diagonal covariance matrix  $\mathbf{V}$ , where each of the  $m n_i \times n_i$  blocks  $V_i$ ,  $i = 1, \dots, m$ , has entries  $\sigma_\alpha^2 + \sigma_\varepsilon^2$  on the diagonal and entries  $\sigma_\alpha^2$  elsewhere. Throughout this paper, we regard models (1) and (2) as conditional specifications of the distribution of  $\mathbf{Y}$  given  $\mathbf{X}$ .

### 2.2. Test statistic and its asymptotic behavior when parameters are estimated by maximum likelihood

**2.2.1. LMM with a single random effect**—We first discuss the 2-level LMM in (2) when both the random effect and the error term are normally distributed and derive our test statistic for the setting where the model parameters  $\theta = (\beta, \psi) = (\beta, \sigma_\alpha^2, \sigma_\varepsilon^2)$  are estimated by maximum likelihood (MLE). Here  $\mathbf{X}$  is considered to be fixed.

Under Assumptions 1.1-1.6 stated below in Theorem 1, consistency and asymptotic normality of the MLE  $\hat{\theta} = (\hat{\beta}, \hat{\psi})$  follow from Miller (1977), i.e.  $\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, \mathbf{J}^{-1})$ , where  $\mathbf{J}$  denotes the limiting Fisher information matrix. Wand (2007) showed that under model (2),  $\hat{\beta}$  and  $\hat{\psi}$  are asymptotically uncorrelated and thus

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_{\beta\beta} & 0 \\ 0 & \mathbf{M} \end{bmatrix}, \quad (3)$$

where

$$\mathbf{J}_{\beta\beta} = \lim_{N \rightarrow \infty} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} / N. \quad (4)$$

We assume that  $\mathbf{J}_{\beta\beta}$  is positive definite (Assumption 1.5, Theorem 1).

To test the goodness of fit of the mean structure of the LMM (2), we first divide the covariate space into  $L$  disjoint regions  $E_1, \dots, E_L$ . These regions are based on categorizations of single covariates or of composites of multiple covariates that may or may not be included in the current model. For example, for a single continuous covariate with support on the interval  $(a, b)$ , the cells  $E$  could be defined by  $E_l = (c_l, c_{l+1}]$ ,  $l = 1, \dots, L - 1$  where  $a = c_1 < c_2 < \dots < c_{L-1} < c_L = b$ . For a categorical (discrete) variable  $X$  that takes the values  $c_l$  for  $l = 1, \dots, L$ , the partition can be defined by  $E_l = \{X = c_l\}$ . We compute the observed and expected sums in each region  $E_l$  as

$$f_l = \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} y_{ij}, \quad (5)$$

$$e_l(\beta) = \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} E(y_{ij}) = \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} \mathbf{x}_{ij}^T \beta, \quad (6)$$

where  $I$  denotes the indicator function. When the cell partition is based on covariates not included in the model (2), then we let  $\mathbf{x}_{ij}$  denote the vector of all available covariates and  $\mathbf{x}_{ij}^*$  the covariates used in the model, and use  $e_l(\beta^*) = \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} (\mathbf{x}_{ij}^*)^T \beta^*$ , where  $\beta^*$  corresponds to the coefficients of  $\mathbf{x}_{ij}^*$ . However, for notational simplicity we employ the expressions (5) and (6) throughout.

Letting  $\mathbf{f} = (f_1, \dots, f_L)$  and  $\mathbf{e}(\beta) = (e_1(\beta), \dots, e_L(\beta))$ , the observed minus the expected vector is

$$\mathbf{f} - \mathbf{e}(\beta_0) = \begin{pmatrix} \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_1\}} (y_{ij} - \mathbf{x}_{ij}^T \beta_0) \\ \vdots \\ \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_1\}} (y_{ij} - \mathbf{x}_{ij}^T \beta_0) \end{pmatrix}. \quad (7)$$

Since the true parameter vector  $\beta_0$  is unknown, to create a test statistic we must replace it by a consistent asymptotically normal estimator, the MLE  $\hat{\beta}$  as in (7) within Section 2.2 and the generalized least squares estimator (15) in Section 2.3. We further make Assumption 1.7 which, together with Assumptions 1.4 and 1.5, ensures the existence of components of the limiting variance covariance matrix for the test statistic.

**Theorem 1.** *We make the following assumptions:*

*Assumption 1.1.* The true parameter point  $\theta_0 = (\beta_0, \psi_0)$  is an interior point of  $\Theta = \left( \mathcal{R}^p, (\mathcal{R}^+)^{R+1} \right)$ .

*Assumption 1.2.*  $\alpha_i \tilde{N} \left( 0, \sigma_\alpha^2 \right)$  and  $\epsilon_{ij} \tilde{N} \left( 0, \sigma_\epsilon^2 \right)$ .

*Assumption 1.3.*  $\mathbf{X}$  is fixed and has full rank.

*Assumption 1.4.*  $\lim_{\kappa \rightarrow \infty} \limsup_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m I_{\{n_i^2 \geq \kappa\}} n_i^2 = 0$ .

*Assumption 1.5.*  $\mathbf{J}_{\beta\beta} = \lim_{N \rightarrow \infty} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} / N$  exists and is positive definite.

*Assumption 1.6.* The  $2 \times 2$  matrix  $\mathbf{M}$  with elements defined below exists and is positive definite;

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

where  $\mathbf{G}_0 = \mathbf{I}$  is the  $N \times N$  identity matrix and  $\mathbf{G}_1$  is the block-diagonal matrix with  $m$  blocks and each block is an  $n_i \times n_i$  matrix of all 1s. After some algebra,

$$\begin{aligned} [M]_{00} &= \lim_{N \rightarrow \infty} \frac{1}{2N} \text{tr} \left( \mathbf{V}^{-2} \right) = \lim_{N \rightarrow \infty} \frac{1}{2N} \sum_{i=1}^m \left( \frac{n_i - 1}{\sigma_\epsilon^4} + \frac{1}{(\sigma_\epsilon^2 + n_i \sigma_\alpha^2)^2} \right), \\ [M]_{01} &= [M]_{10} = \lim_{N \rightarrow \infty} \frac{1}{2N} \text{tr} \left( \mathbf{V}^{-2} \mathbf{1}_N^{\otimes 2} \right) = \lim_{N \rightarrow \infty} \frac{1}{2N} \sum_{i=1}^m \frac{n_i}{(\sigma_\epsilon^2 + n_i \sigma_\alpha^2)^2}, \\ [M]_{11} &= \lim_{N \rightarrow \infty} \frac{1}{2N} \text{tr} \left( \mathbf{V}^{-1} \mathbf{G}_1 \mathbf{V}^{-1} \mathbf{G}_1 \right) = \lim_{N \rightarrow \infty} \frac{1}{2N} \sum_{i=1}^m \frac{n_i^2}{(\sigma_\epsilon^2 + n_i \sigma_\alpha^2)^2} \end{aligned}$$

It is easy to see that matrix  $\mathbf{M}$  is the average of nonnegative definite matrices. Under Assumption 1.4,  $\mathbf{M}$  is positive definite if and only if  $\liminf_{m \rightarrow \infty} \sum_{i=1}^m n_i / m > 1$ . Thus the main restriction in Assumption 1.6 is the requirement that  $\mathbf{M}$  exists.

Assumption 1.7. For any cell partition  $E_1, \dots, E_L$  of the covariate space,

$$\lim_{N \rightarrow \infty} \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} \mathbf{x}_{ij}^T / N \text{ exists for } l = 1, \dots, L.$$

For model (2), under Assumptions 1.1-1.7, as  $N \rightarrow \infty$ ,

$$\sqrt{N} \begin{pmatrix} \{\mathbf{f} - e(\beta_0)\} / N \\ \hat{\beta} - \beta_0 \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, \mathbf{D}\mathbf{V}\mathbf{D}^T),$$

where

$$\mathbf{D} = \begin{pmatrix} N^{-1/2} \begin{bmatrix} I_{\{x_{11} \in E_1\}} & \cdots & I_{\{x_{m1} \in E_1\}} \\ \vdots \\ I_{\{x_{11} \in E_L\}} & \cdots & I_{\{x_{m1} \in E_L\}} \end{bmatrix} \\ N^{-1/2} \mathbf{J}_{\beta\beta}^{-1} \mathbf{X}^T \mathbf{V}^{-1} \end{pmatrix}_{(L+p) \times N},$$

and

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

The off-diagonal and diagonal elements of  $\mathbf{H}$ ,  $H_{lk}$  and  $H_{ll}$ , are

$$H_{lk} = \sigma_\alpha^2 \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^m \left\{ \left( \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} \right) \left( \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_k\}} \right) \right\}, \quad (8)$$

$$H_{ll} = \sigma_\epsilon^2 \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} + \sigma_\alpha^2 \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^m \left( \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} \right)^2, \quad (9)$$

and

$$\Lambda = \begin{pmatrix} \Lambda_1^T \\ \vdots \\ \Lambda_L^T \end{pmatrix}_{L \times p} = \lim_{N \rightarrow \infty} \frac{1}{N} \begin{pmatrix} \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_1\}} \mathbf{x}_{ij}^T \\ \vdots \\ \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_L\}} \mathbf{x}_{ij}^T \end{pmatrix}. \quad (10)$$

The proof of Theorem 1 is given in Appendix A.

**Corollary 1.** Consistent estimators for the quantities given in (8), (9), (10) and (4) are

$$\begin{aligned} \hat{H}_{lk} &= \hat{\sigma}_\alpha^2 \frac{1}{N} \sum_{i=1}^m \left\{ \left( \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} \right) \left( \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_k\}} \right) \right\}^2 \\ \hat{H}_{ll} &= \sigma_\epsilon^2 \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} + \sigma_\alpha^2 \frac{1}{N} \sum_{i=1}^m \left( \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} \right)^2, \\ \hat{\Lambda}_l^T &= \sum_{i=1}^m \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} x_{ij}^T / N, \quad \hat{\mathbf{J}}_{\beta\beta} = \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X} / N. \end{aligned}$$

**Corollary 2.** For model (2), under Assumptions 1.1-1.7 in Theorem 1, as  $N \rightarrow \infty$ ,  $\{\mathbf{f} - \mathbf{e}(\hat{\beta})\} / \sqrt{N} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Sigma)$ , where  $\Sigma = \mathbf{H} - \Lambda \mathbf{J}_{\beta\beta}^{-1} \Lambda^T$  is an  $L \times L$  matrix that is consistently estimated by  $\hat{\Sigma} = \hat{\mathbf{H}} - \hat{\Lambda} \hat{\mathbf{J}}_{\beta\beta}^{-1} \hat{\Lambda}^T$  based on Corollary 1.

The proof of Corollary 2 is given in Appendix B.

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

$$T = \frac{1}{N} \{\mathbf{f} - \mathbf{e}(\hat{\beta})\}^T \hat{\Sigma}_\zeta^- \{\mathbf{f} - \mathbf{e}(\hat{\beta})\}, \quad (11)$$

where  $\hat{\Sigma}_\zeta^-$  denotes the Moore-Penrose generalized-inverse (also called ‘pseudo-inverse’) of a slight modification of the consistent variance estimator  $\hat{\Sigma}$ . We define  $\hat{\Sigma}_\zeta^-$  in the following paragraph. Under the null hypothesis that model (2) is the true model,  $T$  has an asymptotic central  $\chi_r^2$  distribution, where with probability converging to 1 as  $N \rightarrow \infty$ ,  $r = \text{rank}(\hat{\Sigma}_\zeta) = \text{rank}(\Sigma)$ . This result applies to the modification  $\hat{\Sigma}_\zeta$  of any consistent estimator  $\hat{\Sigma}$  of the  $\Sigma$  matrix, such as restricted maximum likelihood (REML) estimators.

The modification  $\hat{\Sigma}_\zeta$  of variance estimates we describe next applies in several places in this paper. The issue is always to avoid numerical instabilities and asymptotic distributional anomalies due to rank differences between consistent nonnegative-definite variance estimators  $\hat{\Sigma}$  and the true asymptotic variance  $\Sigma$ . Assume in what follows that there exists a known threshold  $\zeta$  (say,  $10^{-4}$ ) smaller than all non-zero singular values of  $\Sigma$ . Since  $\Sigma$  is nonnegative definite, all singular values are nonnegative. Denoting the spectral

decomposition of  $\hat{\Sigma}$  as  $\sum_{k=1}^q c_{kN} \mathbf{v}_{kN} \mathbf{v}_{kN}^T$ , where  $\{\mathbf{v}_{kN}\}_{k=1}^q$  is an orthonormal basis of eigenvectors of  $\hat{\Sigma}$  and  $c_{kN}$  are the corresponding eigenvalues, the modification  $\hat{\Sigma}_\zeta$  of  $\hat{\Sigma}$  is defined as  $\hat{\Sigma}_\zeta = \sum_{k=1}^N I_{[c_{kN} > \zeta]} c_{kN} \mathbf{v}_{kN} \mathbf{v}_{kN}^T$ , and its Moore-Penrose pseudo-inverse is then  $\hat{\Sigma}_\zeta^- = \sum_{k=1}^q I_{[c_{kN} > \zeta]} (c_{kN})^{-1} \mathbf{v}_{kN} \mathbf{v}_{kN}^T$ . We prove in Proposition 1, Appendix D, that  $\hat{\Sigma}_\zeta$  is coordinate-free and unique, that  $\text{rank}(\hat{\Sigma}_\zeta) = r$  with probability converging to 1, and that the asymptotic distribution of  $T$  in (11) is  $\chi_r^2$  and would persist if the Moore-Penrose

choice of pseudo-inverse were replaced by any other generalized-inverse of  $\hat{\Sigma}_\zeta$  as defined in Rao (1973, Sec. 1.b). In addition, the assumption of a fixed known threshold  $\xi$  can be replaced by allowing  $\xi = \xi_N$  to depend non-randomly on  $N$  and converge to 0 sufficiently slowly. (The rate would depend on the specific consistent estimator  $\hat{\Sigma}_\zeta$ .)

**2.2.2. LMM with multilevel additive random effects**—We now consider the general LMM with multilevel additive random effects given in equation (1) with a fixed covariate matrix  $\mathbf{X}$ .

Again, the covariate space, comprised of covariates not all of which need be included in the model, is divided into  $L$  disjoint regions  $E_1, \dots, E_L$ , and for  $l = 1, 2, \dots, L$ , we define the observed and expected vectors  $\mathbf{f} = (f_1, \dots, f_L)$  and  $\mathbf{e}(\beta) = (e_1(\beta), \dots, e_L(\beta))$  as

$f_l = \sum_{k=1}^N I_{\{x_k \in E_l\}} y_k$ , and  $e_l(\beta) = \sum_{k=1}^N I_{\{x_k \in E_l\}} E(y_k) = \sum_{k=1}^N I_{\{x_k \in E_l\}} \mathbf{x}_k^T \beta$ . The impact of the choice of  $L$  is discussed further in the simulation section (Section 3).

Conditions in Miller (1977) that ensure the consistency and asymptotic normality of the MLE of  $\theta = (\beta, \psi)$  are given in the Supplementary Material S1 (assumptions A.1-A.7). We additionally make assumptions A.8-A.9 (Supplementary Material S1) to ensure the existence of large-sample averages involving  $\mathbf{x}_k$  and  $I_{[x_k \in E_j]}$ .

**Theorem 2.** For model (1), under Assumptions A.1-A.9 given in the Supplementary Material S1,

$$T = \{f - e(\hat{\beta})\}^T \hat{\Sigma}_\zeta^- \{f - e(\hat{\beta})\} / N \xrightarrow{\mathcal{D}} \chi_r^2, \quad (12)$$

as  $N \rightarrow \infty$ , where  $\hat{\beta}$  is the MLE of  $\beta$ ,  $\hat{\Sigma}_\zeta$  is a consistent estimator of  $\Sigma = \mathbf{H} - \Lambda \mathbf{J}_{\beta\beta}^{-1} \Lambda^T$ ,  $\hat{\Sigma}_\zeta^-$  is the modification of  $\hat{\Sigma}_\zeta$  as defined in the last paragraph of Section 2.2.1,  $\hat{\Sigma}_\zeta^-$  denotes the Moore-Penrose pseudoinverse of  $\hat{\Sigma}_\zeta$ , and for  $r \equiv \text{rank}(\Sigma)$ ,  $P(\text{rank}(\hat{\Sigma}_\zeta) = r) \rightarrow 1$ , as  $N \rightarrow \infty$ . Here  $\mathbf{H} = \lim_{N \rightarrow \infty} \mathbf{F} \mathbf{V} \mathbf{F}^T$ , with

$$\mathbf{F} = \frac{1}{\sqrt{N}} \begin{pmatrix} I_{\{x_1 \in E_1\}} & \cdots & I_{\{x_N \in E_1\}} \\ \vdots & & \vdots \\ I_{\{x_1 \in E_L\}} & \cdots & I_{\{x_N \in E_L\}} \end{pmatrix} \quad (13)$$

and the  $l$ -th row of  $\Lambda$  given by

$$\Lambda_l^T = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N I_{\{x_k \in E_l\}} \mathbf{x}_k^T \quad (14)$$



The proof of Theorem 2 is similar to that of Theorem 1 and is given in Supplementary Material S2.

**2.3. 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), but now only require that  $E(\alpha_i) = E(\epsilon_{ij}) = 0, Var(\alpha_i) = \sigma_\alpha^2, Var(\epsilon_{ij}) = \sigma_\epsilon^2$ , and that there is a  $\delta > 0$ , such that  $E(\alpha_i^{4+\delta}) < \infty$  and  $E(\epsilon_{ij}^{4+\delta}) < \infty$ , instead of assuming normality of  $\alpha_i$  and  $\epsilon_{ij}$ . To compensate for the weaker distributional assumptions, we assume for simplicity in probability limit theorems that the covariate vectors  $\mathbf{x}_{ij}$  are random, that  $\{(\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}), n_i\}$  are i.i.d. and that  $E(n_i^2) < \infty$  and  $\|E(\mathbf{x}_i^T \mathbf{x}_i)\| < \infty$ , where  $\mathbf{x}_i$  denotes the  $n_i \times p$  matrix of covariates for the  $i$ th cluster.

We estimate  $\beta$  by the generalized least squares estimator

$$\tilde{\beta} = (\mathbf{X}^T \tilde{\mathbf{V}}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \tilde{\mathbf{V}}^{-1} \mathbf{Y}) = (\mathbf{X}^T \tilde{\mathbf{V}}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \tilde{\mathbf{V}}^{-1} \mathbf{Y}) + o_p(1), \quad \text{as } N \rightarrow \infty, \quad (15)$$

where  $\mathbf{V}$  depends on the variance components  $\psi = (\sigma_\alpha^2, \sigma_\epsilon^2)$ . They are estimated by the method of moments by equating the right-hand sides of

$$E \left\{ \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \right\} = \sum_{i=1}^m \sum_{j=1}^{n_i} (\mathbf{x}_{ij}^T \beta - \bar{\mathbf{x}}_i^T \beta)^2 + (N - m) \sigma_\epsilon^2 \quad (16)$$

$$E \left\{ \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}_{..})^2 \right\} = \sum_{i=1}^m n_i (\bar{\mathbf{x}}_i^T \beta - \bar{\mathbf{x}}_{..}^T \beta)^2 + \left( N - \frac{1}{N} \sum_{i=1}^m n_i^2 \right) \sigma_\alpha^2 + (m - 1) \sigma_\epsilon^2 \quad (17)$$

respectively with their estimates based on the sum of squares within groups (SSW) and the sum of squares between groups (SSB) in the analysis of variance, given by

$$SSW = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad \text{and} \quad SSB = \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}_{..})^2 = \sum_{i=1}^m n_i \bar{y}_i^2 - N \bar{y}_{..}^2$$

The notation  $\bar{y}_i$  stands for  $\bar{y}_i = \sum_j y_{ij} / n_i$  and  $\bar{y}_{..} = \sum_i \sum_j y_{ij} / N$ , and similar averages apply to the covariates  $\mathbf{x}$ . Because different clusters are independent, and

$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2, n_i \bar{y}_i^2$ , and  $\sum_{j=1}^{n_i} y_{ij}$  have finite second moments,  $SSW/m$  and  $SSB/m$  satisfy the law of large numbers. The estimating equations (15), (16) and (17) can be solved

iteratively for  $\theta = (\beta, \sigma_\alpha^2, \sigma_\epsilon^2)$  and yield consistent estimates  $\tilde{\theta}$  (Richardson and Welsh, 1994; Jiang, 1996).

To obtain the test statistic, we again compute the observed and expected values in each of the  $L$  cells of the covariate space as in (5) and (6).

**Theorem 3.** For the LMM (2), under Assumptions 3.1-3.3 in Appendix D, as  $N \rightarrow \infty$ ,  $\{\mathbf{f} - e(\tilde{\beta})\} / \sqrt{N} \xrightarrow{\mathcal{D}} N(0, \Sigma)$ , where  $\Sigma = \mathbf{H} - \mathbf{\Lambda} \mathbf{J}_{\beta\beta}^{-1} \mathbf{\Lambda}^T$ . Thus  $T = \{\mathbf{f} - e(\tilde{\beta})\}^T \hat{\Sigma}_\zeta^{-1} \{\mathbf{f} - e(\tilde{\beta})\} / N \xrightarrow{\mathcal{D}} \chi_k^2$ , where  $\hat{\Sigma}$  is a consistent estimator of  $\Sigma$ ,  $\hat{\Sigma}_\zeta^{-1}$  is the modification of  $\hat{\Sigma}$  as defined in the last paragraph of the section 2.2.1,  $\hat{\Sigma}_\zeta^{-1}$  denotes the Moore-Penrose pseudoinverse of  $\hat{\Sigma}_\zeta$ , and  $P(\text{rank}(\hat{\Sigma}_\zeta) = \text{rank}(\Sigma) = r) \rightarrow 1$   $N \rightarrow \infty$ .

As  $\mathbf{X}$  is a matrix of random variables in this section, Assumption 3.3 ensures that Assumptions 1.4-1.7 hold, which are needed when  $\mathbf{X}$  are assumed to be fixed in Section 2.1. The matrices  $\mathbf{H}$ ,  $\mathbf{\Lambda}$  and  $\mathbf{J}_{\beta\beta}$  are the same as for the two-level LMM (2) when parameters are estimated by maximum likelihood under the assumption of normality of  $\alpha_j$  and  $\epsilon_{ij}$  and are specified in (8), (9), (10) and (4). The proof of the theorem is given in Appendix E.

**Remark 1.** Theorem 1 and Theorem 3 are still valid when empirical quantiles instead of fixed cut-offs are used to define cell partition, is we assume that the empirical quantiles of coordinates of  $\mathbf{x}_j$  converge to unique limits or under Assumption 3.3.

**2.4. Power of the test**

For the multi-level LMM (1), we derive the theoretical power under local, and more specifically under contiguous alternatives for the test in (12) for the situation where some covariates that influence the outcome  $y$  are omitted from model (1). This case also covers omitted interactions of covariates or omitted higher order terms and is thus practically relevant.

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 to fit model (1), with  $p^* < p$ . The null hypothesis is  $H_0: \theta_N = \theta_0$ . We assess the power of  $T$  under the alternative

$$H_1: \theta_N = \theta_0 + a / \sqrt{N}, \quad (18)$$

with  $\theta_0 = (\beta_0, \psi_0)$ , where several components of  $\beta_0$  are 0. The vector  $\beta_0^*$  corresponding to  $\mathbf{X}^*$ . Here  $a / \sqrt{N}$  is the vector difference between the parameter values under the alternative hypothesis and the parameter values under the null hypothesis.

Based on the derivation for Theorem 2, we have that under  $H_0$ ,  $\{\mathbf{f} - e(\hat{\beta}^*)\} / \sqrt{N} \xrightarrow{\mathcal{D}} N(0, \Sigma^*)$ . By applying Le Cam’s third lemma (see Appendix F for details), we find that under the alternative hypothesis  $H_1$  in (18),

$$\{\mathbf{f} - \mathbf{e}(\hat{\beta}^*)\} / \sqrt{N} \xrightarrow{\mathcal{D}} N(\boldsymbol{\tau}, \boldsymbol{\Sigma}^*), \quad (19)$$

where

$$\boldsymbol{\tau} = \lim_{N \rightarrow \infty} \left[ \boldsymbol{\Lambda} - \boldsymbol{\Lambda}^* \{(\mathbf{X}^*)^T \mathbf{V}^{-1} \mathbf{X}^*\}^{-1} \{(\mathbf{X}^*)^T \mathbf{V}^{-1} \mathbf{X}\} \right] \mathbf{a}, \quad (20)$$

with  $\boldsymbol{\Lambda}$  given by expression (14).  $\boldsymbol{\Lambda}^*$  is computed using  $\mathbf{X}^*$  in place of  $\mathbf{X}$  in (14).

Thus under  $H_1$ ,  $T^*$  has a limiting noncentral  $\chi^2$  distribution

$$T^* = \frac{1}{N} \{\mathbf{f} - \mathbf{e}(\hat{\beta}^*)\}^T (\hat{\boldsymbol{\Sigma}}_\zeta^*)^{-1} \{\mathbf{f} - \mathbf{e}(\hat{\beta}^*)\} \xrightarrow{\mathcal{D}} \chi_r^2(\lambda), \quad (21)$$

where  $r = \text{rank}(\hat{\boldsymbol{\Sigma}}_\zeta^*)$  and the non centrality parameter is  $\lambda = \boldsymbol{\tau}^T (\hat{\boldsymbol{\Sigma}}_\zeta^*)^{-1} \boldsymbol{\tau}$ . For a given type  $I$  error level  $\alpha$ , the power is thus  $P(T^* > \chi_{r,\alpha}^2)$ , where  $\chi_{r,\alpha}^2$  is the  $1 - \alpha$  quantile of the central  $\chi_r^2$  distribution and  $P$  denotes the non central  $\chi_r^2(\lambda)$  distribution. In the computation of the power we use the Moore-Penrose inverse of a modification  $\hat{\boldsymbol{\Sigma}}_\zeta^*$  as in (11) of a consistent estimator  $\hat{\boldsymbol{\Sigma}}^*$  in (20).

As an illustration, we show the asymptotic power to detect lack of fit for an omitted covariate for the two-level LMM (2). We assume  $\mathbf{Y} \sim N(\mathbf{X}^T \boldsymbol{\beta}, \mathbf{V})$ , where  $\mathbf{X} = (1, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$  and  $\mathbf{V}$  is the block diagonal covariance matrix. The  $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_j$  are i.i.d. and drawn from a multivariate normal distribution

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \rho_{13} \\ 0 & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} \right), \quad (22)$$

and  $\mathbf{x}_{ij}$  and  $n_j$  are independent. In fitting the model, we omit  $\mathbf{x}_3$  leading to  $\mathbf{X}^* = (1, \mathbf{x}_1, \mathbf{x}_2)$  and  $\mathbf{a} = (0, 0, 0, \boldsymbol{\beta}_3)$  in (20). For this setting,  $\boldsymbol{\tau}$  in (20) and  $\boldsymbol{\Sigma}^*$  in (19) can be computed explicitly as a functions of the moments of  $\mathbf{X}$  and  $n_j$  (section 2.2.3., Tang, 2010). We study the impact of the magnitude of the variance components  $\sigma_\alpha^2$  and  $\sigma_\epsilon^2$  and the correlations  $\rho_{13}$  and  $\rho_{23}$  in (22) 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 theoretical quantiles of the omitted covariate  $x_3$  with  $L = 8$  cells. For  $\rho_{13} = 0.5$  and  $\rho_{23} = 0.6$ , Figure 1 (left panel) plots the theoretical power against  $\beta_3 / (\sigma_\alpha^2 + \sigma_\epsilon^2)^{1/2}$  for three choices of  $(\sigma_\alpha^2, \sigma_\epsilon^2)$  all corresponding to the same overall variance  $\sigma_\alpha^2 + \sigma_\epsilon^2$  and varying  $\beta_3$  on the x-axis. For any fixed pair of

$(\sigma_\alpha^2, \sigma_\epsilon^2)$ , the power of the test, not surprisingly, increases as a function of  $\beta_3$ , the coefficient of the omitted covariate  $x_3$ . This can also be seen by taking a first order Taylor expansion of the theoretical power formula around  $\lambda = 0$ , as the power for  $\lambda$  close to zero depends linearly on  $\lambda = \boldsymbol{\tau}^T(\boldsymbol{\Sigma}^*)^{-1}\boldsymbol{\tau}$ , which is a function of  $\beta_3^2$ . Figure 1 (left panel) shows that for any fixed  $\beta_3$  the power increases when the random effect  $\sigma_\alpha^2$  decreases compared to the error term  $\sigma_\epsilon^2$ . Figure 1 (right panel) plots the power for  $\sigma_\alpha^2=1, \sigma_\epsilon^2=0.25$  for different choices of  $(\rho_{13}, \rho_{23})$ . The power increases as  $\rho_{13}^2 + \rho_{23}^2$  decreases. For  $\rho_{13} = 0$  and  $\rho_{23} = 0$ , that is, when  $x_3$  is uncorrelated with  $x_1$  and  $x_2$ , the power is not affected by the individual values of  $\sigma_\alpha^2$  and  $\sigma_\epsilon^2$ , but only depends on the sum  $\sigma_\alpha^2 + \sigma_\epsilon^2$ . The theoretical formulas for power under contiguous-alternatives given here will generally be close to the actual power only for very large sample sizes. However, numerical studies presented in the next section show that these formulas often also agree with empirical power in samples of moderate size ( $m = 50, N = 150$  to  $N = 200$ ).

### 3. Simulations to assess power and robustness of the test statistic

For a given number of clusters  $m$ , we first generated the cluster sizes  $n_i$  from a uniform distribution on  $\{2, 3, 4, 5\}$  for  $i = 1, \dots, m$  and then drew  $N = \sum_{i=1}^m n_i$  independent covariates  $\mathbf{x}_{ij}$  for all simulations presented below. We present scenarios for which we believe our test would be practically most relevant: models with omitted main effects (Scenario I), omitted interaction terms and main effects (Scenarios II and III) and misspecified functional forms of a covariate (Scenario IV). We covered a range of effect sizes to provide a fair assessment of the performance of our test.

#### 3.1. Main effects only (Scenario I)

Here  $\mathbf{x}_{ij} = (x_{1ij}, x_{2ij}, x_{3ij})$  were drawn from the multivariate normal distribution given in (22). Given  $\mathbf{X} = (\mathbf{1}, x_1, x_2, x_3)$ ,  $\boldsymbol{\beta}$ ,  $\sigma_\alpha = 1$  and  $\sigma_\epsilon = 0.5$ , we generated  $\mathbf{Y}$  from a multivariate normal distribution,  $\mathbf{Y} \sim N(\mathbf{X}^T \boldsymbol{\beta}, \mathbf{V})$ .

**3.1.1. Size and power of T**—To check the size of the test for various choices of cell partition based on  $\mathbf{X}$ , we let  $\rho_{13} = \rho_{23} = 0$  in (22),  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, 1, 1)$ ,  $m = 500$  and fit model (2) with all covariates  $\mathbf{X}$  in the model. Table 1 gives the number of cells and the covariates that are the basis of the cell partition in the first column. Cell partitions in the computation of the test statistic  $T$  were based on empirical quantiles of the respective components of  $\mathbf{X}$ . For all cell partitions in Table 1 the empirical sizes were close to the nominal  $\alpha$  levels of 0.05 and 0.1.

To assess the power of the test, we generated data from model (2) that includes all three covariates but then omitted  $x_3$  in fitting the model to the data. We set  $(\rho_{13}, \rho_{23}) = (0.5, 0.6)$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, 1, 0.25)$ . We then generated  $K = 2000$  datasets for a given  $\mathbf{X}$ . We repeated this data generation process for  $D = 1000$  independently drawn design matrices  $\mathbf{X}$ . For a given  $\mathbf{X}$ , we computed the theoretical power of  $T^*$  in (21) based on the asymptotic  $\chi^2$  distribution with the true values of  $\sigma_\alpha^2$  and  $\sigma_\epsilon^2$ . The empirical moments for  $\mathbf{X}$  were used in the

calculation of the non-centrality parameter  $\lambda$ . For a given  $\mathbf{X}$  and each generated  $\mathbf{Y}$ , we calculated the estimated theoretical power based on the asymptotic  $\chi^2$  distribution with the variance components estimated based on the given  $\mathbf{Y}$  and empirical moments of  $\mathbf{X}$  in (20). We then repeated the calculation of the estimated theoretical power for each of the  $K = 2000$  generated  $\mathbf{Y}$  and by taking the average, we obtained a mean estimated theoretical power for that given  $\mathbf{X}$ . For each given design matrix  $\mathbf{X}$ , we also calculated the empirical power based on  $K = 1000$  iterations on  $\mathbf{Y}$ . For  $m = 20$  clusters the cell partition was based on the empirical quantiles of the omitted  $x_3$  with  $L = 8$  cells to avoid empty cells, but for  $m = 50$  or  $500$ , we used theoretical quantiles of  $x_3$  as cell boundaries for computational ease.

The mean theoretical power (“Theo.Pow.”), the mean estimated theoretical power (“Theo.Pow.hat”) and the empirical power (“Empi.Pow.n”) agreed very well, even when  $m$  is small (Table 2). However, only for  $m = 500$  was there adequate power to detect lack of fit when  $\beta_3 = 0.25$ , which is 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.

### 3.1.2. Robustness of T with respect to error and random effects distributions

—In Table 2 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  $\epsilon$  from a  $t$  distribution with  $k = 3$  or  $5$  degrees of freedom (d.f.) instead of from a  $N(0, \sigma_\epsilon^2)$ . We rescaled the variance of  $\epsilon$  so that the noise had the same variance as in the normal case. The power of the test under a  $t$ -distribution was virtually the same as with a normal error distribution indicating that our test is robust to symmetric violations of normality. For example, for  $m = 50$  with  $\beta_3 = 0.8$ , the power was 0.83 for the normal error distribution and for  $t$ -distributions with 3 and 5 d.f. (Table 2). We also used the same misspecification for the random effects distribution, and observed very similar results (Table 2). We chose the  $t$  distribution because it is symmetric but has heavier tails than the normal distribution and it satisfies the conditions given in Section 2.3 on the existence of moments of the random effects and errors.

**3.1.3. Impact of choice of the cell partition on power**—As is true for Pearson’s chi-squared test, the choice of cell partition strongly impacts the performance of our goodness of fit test. To illustrate the impact of the cell partition on the power of our test we generated  $y$  from a model with  $E(y) = 1 + x_1 + x_2 + 0.15x_3$ , with  $\sigma_\alpha = 1$  and  $\sigma_\epsilon = 0.5$ , but then omitted  $x_3$  in the subsequent model fitting. We studied cell 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  $x_3$ , all based on empirical quartiles of the covariates. Table 3 shows that a lack of fit is detectable by our test statistic only when the cell partition involves the omitted covariate  $x_3$ , and power decreases as correlations ( $\rho_{13}$ ,  $\rho_{23}$ ) between the covariates increase.

## 3.2. Normally distributed covariates with an omitted interaction term (Scenario II)

We generated  $y$  from a linear model with  $E(y) = 1 + x_1 + x_2 + \beta_3 x_1 x_2$ , where  $x_1$  and  $x_2$  are independent and  $x_i \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2$ . We first let  $\beta_3 = 0.2$  and set  $\sigma_\alpha = 1$ ,  $\sigma_\epsilon = 0.5$ . We

then fit model (2) without  $x_3 = x_1 x_2$ . The test had adequate power only when the cell partition is based on empirical quantiles of  $x_1$  and  $x_2$ , or on the omitted interaction term  $x_3 = x_1 x_2$ , but not if the cell partition was based on either  $x_1$  or  $x_2$  alone, for  $\rho_{12} = 0$  and  $\rho_{12} = .3$  (Table 1, Web Supplementary Material). Figure 2 shows the power of the test as a function of the number of cells computed based on quantiles of the omitted covariate  $x_3$  for various values of  $\mu_i = E(X_i)$ ,  $i = 1, 2$ . The power was higher for smaller absolute values of  $\mu_1$  and  $\mu_2$  and was largest for  $L = 11$  cells for  $\mu_1 = 2$  and  $\mu_2 = 1$  and  $L = 7$  cells for  $\mu_1 = 1$  and  $\mu_2 = 0.5$ .

Figure 3 plots the theoretical power against  $\beta_3 / (\sigma_\alpha^2 + \sigma_\epsilon^2)^{1/2}$  for three choices of  $(\sigma_\alpha^2, \sigma_\epsilon^2)$  corresponding to the same overall variance  $\sigma_\alpha^2 + \sigma_\epsilon^2$  and varying  $\beta_3$  on the x-axis when the cell partition was based on  $x_3$  with  $L = 8$  cells using fixed cell boundaries. Our conclusions are consistent with those in Section 3.1. For any fixed pair of  $(\sigma_\alpha^2, \sigma_\epsilon^2)$ , the power of the test increased as a function of  $\beta_3$ , the coefficient of the omitted covariate  $x_3$ . For any fixed  $\beta_3$ , the power increased when the random effect  $\sigma_\alpha^2$  decreases compared to the error term  $\sigma_\epsilon^2$ .

### 3.3. Omitted main effect and interaction term (Scenario III)

Here we generated  $y$  from a LMM that includes three covariates  $x_1$ ,  $x_2$  and  $x_3$  through  $E(y) = 1 + x_1 + x_2 + 0.05x_3 + 0.1x_1x_3$ , with  $\sigma_\alpha = 1$  and  $\sigma_\epsilon = 0.5$ . We then omitted  $x_3$  and any interactions with  $x_3$  from the model fitting and investigate the power of our test in working model M1 with covariates  $x_1$ ,  $x_2$ , and working model M2 with  $x_1$ ,  $x_2$  and  $x_1^2$ . The cell partition for M1 and M2 was based on empirical quantiles of  $x_1$  with  $L = 8$  cells. We simulated  $x_3 \sim N(0, \sigma^2)$  with  $\sigma = 1.5$ , and let  $x_1 = e^{x_3}$ . The covariate  $x_2 \tilde{\chi}_1^2$  was generated independently of  $x_1$  and  $x_3$ .

Under this setting, with  $m = 500$  clusters, our test had power 1 to detect lack of fit of the working model M1. The Wald test for inclusion of the quadratic term  $x_1^2$  in model M2 had power 1. As  $x_3$  is not available, a Wald test cannot be applied for any term related to  $x_3$ . We thus would select model M2 (with covariates  $x_1$ ,  $x_2$ ,  $x_1^2$ ) and a Wald-type test is not able to further assist in testing model inadequacy. However, our proposed test had power of 0.938 to detect lack of fit of M2, under the cell partition based on  $x_1$ , a covariate in M2.

This example, in which the omitted covariate is not available in the dataset but a correlated variable is, shows the usefulness of our test in addition to Wald type tests. However, our test has reasonable power against omitted-covariate alternatives only when the partitions of the covariate space are based on variables that are correlated with the omitted covariates.

### 3.4. Misspecified functional form of a covariate (Scenario IV)

We generated  $y$  from a LMM with  $E(y) = 1 + x_1 + x_2 + 0.1x_3^2$  with  $\sigma_\alpha = 1$  and  $\sigma_\epsilon = .5$ , and studied the power of our test when instead of  $x_3^2$  only  $x_3$  is used in the working model.  $\mathbf{x} = (x_1, x_2, x_3)$  was generated from the multivariate normal distribution given in (22) with  $(\rho_{13}, \rho_{23}) = (.5, .6)$ . We used empirical quantiles of  $x_3$  to define  $L = 8$  cells for  $m = 500$  or 50 clusters.

When  $m = 500$ , our test had approximately 87% power in detecting model inadequacy (Table 4). The Wald test to assess the significance of  $x_3$  however had of only approximately 6% power. Thus based on Wald test,  $x_3$  would not be included in the model and therefore it is unlikely that the higher order term  $x_3^2$  would be considered.

We also investigated the impact of symmetric misspecification of the error and random effect distribution in this scenario. When the t-distribution was used for the error term (or for the random effect term) results are similar to those for the normally distributed error term (or random effect term) (Table 4). Thus in this scenario, our test was robust to symmetric misspecification of the error or random effect distribution, similar to Scenario I.

**3.5. Remarks**

The primary purpose of the goodness of fit tests studied in this paper is to assess the quality of the fixed-effect part of the mean response in the presence of a mixed-effect variance structure. Yet it is well known that there is ambiguity in Gaussian linear models as to which terms contribute to the fixed-effect predictors and which terms to the variance. To be specific, we consider the model

$$Y_{ij} = \beta_0 + \beta_1^T \mathbf{X}^* + \gamma X_3 + \alpha_i + \epsilon_{ij} \quad (23)$$

where  $\mathbf{X}^* = (X_1, X_2)$ ,  $(\mathbf{X}^*, X_3)$  are jointly normally distributed with means 0, and  $\tilde{\alpha} \sim \mathcal{N}(0, \sigma_\alpha^2)$ , and the random error  $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ . By grouping the  $\gamma X_3$  term together with the error  $\epsilon$ , we see that model (23) is equivalent to the model

$$Y = \beta_0 + \beta_1^{*T} \mathbf{X}^* + \alpha_i + \epsilon^* \quad (24)$$

where  $\beta_1^*$  and  $\epsilon^*$  are defined in terms of  $E(X_3|\mathbf{X}^*) = M^T \mathbf{X}^*$  and  $V(X_3|\mathbf{X}^*) = \sigma_R^2$  by  $\epsilon^* = \epsilon + \gamma(X_3 - M^T \mathbf{X}^*)$ ,  $\beta_1^* = \beta_1 + M\gamma$ , and  $V(\epsilon^*) = \sigma_\epsilon^2 + \gamma^2 \sigma_R^2$ . This argument shows that the portion of a normal linear model describing  $E(Y|\mathcal{D})$  is not uniquely determined, where  $D$  denotes the data-vector of covariates, that is  $\mathcal{D} = (\mathbf{X}^*, X_3)$  in (23), and  $\mathcal{D} = \mathbf{X}^*$  in (24). However, since our goodness of fit tests for adequacy of the mean structure are considered conditional on  $D$ , and are specified in terms of covariate-defined cells, these two models (23) and (24) are in fact distinguishable if cells under (23) are taken to depend non-trivially on the omitted covariate  $X_3$ .

This argument also highlights the lack of power for the test in the setting of main effects (Scenarios I and III) with an omitted covariate when the cell partition was not based on the omitted covariate or a transformation of it, or for an omitted interaction term, when the cell partition is based on only one of the variables that define the interaction (Scenario II). When cell partitions are based on only on  $\mathbf{X}^*$  no lack of fit in the mean structure can be detected, as it is correctly specified with respect to  $\mathbf{X}^*$ .

## 4. Data example

On April 26, 1986, an accident at the Chernobyl power plant in 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 materials contaminated the territory. Radioisotopes of iodine, e.g. I-131, are concentrated in the thyroid gland. Belarusians exposed to the accident were enrolled in a cohort study to evaluate the relationship between I-131 doses and thyroid cancer risk (Stezhko et al, 2004). Investigators were also interested in studying iodine deficiency in this population, as it impacts I-131 absorption.

We therefore evaluated the relationship between levels of serum thyroglobulin ( $TG$ ), a marker of iodine deficiency, and variables that might reflect or impact dietary iodine intake, including age at the time of exam, age at the 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 (yes/no), presence of goiter (yes/no) and presence of any thyroid abnormality (yes/no).

We used data on  $m = 933$  men from four of the five study regions, who had complete covariate information, whose  $ATG$  and  $TSH$  levels were measured by a luminescence assay, and who had  $TG > 80$  (to exclude those with thyroid disease). Among these men, 404 had a single  $TG$  measurement, 484 had two, 42 three and 3 four  $TG$  measurements during follow-up, resulting in  $N = 1510$  observations.  $\log(TG)$  was normally distributed (Anderson-Darling test p-value  $p=0.09$ ).

We fit various models using Proc GLIMMIX, SAS 9.2. Model 1 included all the variables mentioned above, with the exception of presence of nodules, and an interaction term of  $ATG$  levels with presence of any thyroid abnormality that was marginally significant (Wald test p-value  $p = 0.054$ ) and had a log-likelihood value of  $-1625.2$ . The random effect variance estimate was  $\hat{\sigma}_\alpha^2 = 0.29$  and the error variance estimate was  $\hat{\sigma}_\epsilon^2 = 0.25$ . Model 2 had no interaction term, but included presence of nodules and resulted in a log-likelihood of  $-1621.3$ . The variance component estimates were similar to model 1,  $\hat{\sigma}_\alpha^2 = 0.29$  and  $\hat{\sigma}_\epsilon^2 = 0.26$ . However, as models 1 and 2 are not nested, we could not compare them using a likelihood ratio test.

To assess the fit of both models, each person in the dataset was assigned to one of the  $L = 8$  cells defined by the quartiles of  $ATG$  and the response “yes” or “no” to the question “presence of any thyroid abnormality”. There was no indication of lack of fit for either model, with  $p = 0.32$  and  $p = 0.40$  for models 1 and 2 respectively. We also calculated the test statistic for a second cell partition with  $L = 4$  cells defined by “presence of nodules” (yes/no) and “presence of goiter” (yes/no), with  $p = 0.19$  and  $p = 0.70$  for models 1 and 2 respectively. These results suggested that both models fit the data adequately. Thus omitting the interaction term of the variable “presence of any thyroid abnormality” with  $ATG$  levels does not affect the fit to the data.



## 5. 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 the mean structure of a linear mixed model, with cell partitions based on covariates. We described the asymptotic properties of the test when parameters are estimated and developed its theoretical power under local alternatives. We assessed factors that affect the power, the impact of choice of cell partitions on the test as well as the robustness of the test with respect to error distribution and distribution of random effects in simulations. When a specific covariate associated with outcome is omitted, such as an interaction term or a covariate correlated with terms already in the model, cell partitions based on the omitted covariate result in adequate power of the test. In our simulations we studied models involving only a few covariates. In such cases, Wald testing and likelihood-based model building tools could undoubtedly be used instead. In practical settings our test would be recommended when there are many potential predictors that should in fact not appear in the model. In such circumstances, many nonlinear terms involving omitted variables would not be Wald-tested. We also found that the estimated theoretical power calculated using Le Cam's third lemma was reliable at least when the number of clusters  $m$  was 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 robust to symmetric violations of the normality assumption of the error distribution as well as the violation of normality of the random effects distribution.

This goodness of fit test can be used to test the statistical adequacy of the fixed effects part of a finally selected LMM. It should not be used if one wants to test if a specific covariate should be included in the model, as standard tests such as the Wald test have better power for that purpose (e.g. Scenario I, Table 2). However, when a covariate is missing from the dataset, our test can detect model inadequacy when the cell partition is based on an existing covariate in the working model, which is correlated with the omitted covariate, while no Wald-type test can be applied (Scenario III). Also, the Wald test did not have power to select a variable that entered the mean model only through a quadratic term (Scenario IV), while our test clearly showed lack of fit of the finally selected model with respect to cells defined by that variable. This is particularly important in the situation when many predictors are available, and testing all possible higher order terms or interactions is not practical. In addition, investigators might not consider the inclusion of a higher order term for a variable that has no main effect. We have shown using simple examples that our proposed test has good power to detect many sorts of model inadequacies, not all of which would be tested exhaustively by other methods.

To implement the test one only needs 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 must modify the estimated variance matrix  $\hat{\Sigma}$ , projecting its eigenspace corresponding to extremely small eigenvalues to  $\mathbf{0}$ , to ensure the correct degrees of freedom for the test statistic.

Pan and Lin (2005) developed methods for checking the adequacy of generalized linear mixed models by comparing the cumulative sums of residuals over covariates or predicted values. Our proposed test has additional flexibility in defining cells based on multiple covariates, the test statistic follows a known distribution and is thus easily computed, and we present a broader class of LMMs.

Our goodness of fit test examines multiple features of the data, corresponding to residuals within each covariate cell and bears some relation to the multiaspect framework by Pesarin and Salmaso (2010), Salmaso and Solari (2005) and Marozzi (2007). Future work could attempt to adapt their permutational approaches for several populations to the goodness of fit test in a single population.

Notably, the cell partition used for our test is based on covariates, not on the response variable as for standard Pearson  $\chi^2$  statistic. In future research we plan to further investigate the choice of covariate-based cells partition on the performance of our proposed test. A related issue is sparse cells. Our asymptotic results were derived letting the sample size go to infinity for a fixed cell partition and thus asymptotically cells are not sparse. However, in a real dataset the issue of sparse cells could arise. Maydeu-Olivares and Joe (2005, 2006) and Cagnone (2012) studied the impact of sparse cells when assessing the goodness of fit of latent variable models. For use with heavily cross-classified and sparse covariate-space cell decompositions the limited-information approach of Maydeu-Olivares and Joe (2005, 2006) could be used in our setting and will be part of future investigations. Other possible extensions include derivation of the distribution of the test statistic for random components with heavy tails, for example, under symmetric  $\alpha$ -stable distributional assumption for the errors and random effects. However, these extensions of the mixed model theory presented in our paper are technically difficult and we are not aware of any related results in the literature.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgment

We thank the investigators of the ‘US-Belarusian Study of Thyroid Cancer and Other Thyroid Diseases Following the Chernobyl Accident’ (National Cancer Institute, Columbia University, U.S.A. and Republican Research Center of Radiation Medicine and Human Ecology, Belarus) for providing the data and Jincao Wu for help with computations. We also thank the reviewers for helpful comments and suggestions. This work is part of M. Tang’s Ph.D. thesis done at the University of Maryland.

## Appendix

### Appendix A: Proof of Theorem 1

Let  $\mathbf{J}$  be the limit of the sample information matrix per observation given in (3). The consistency of the MLE  $\hat{\theta}$  in model (2) follows from Miller (1977). By Taylor series expansion of the score function  $\mathbf{S}(\boldsymbol{\theta}) = \nabla \log L(\boldsymbol{\theta})$ , where  $L(\boldsymbol{\theta})$  denotes the likelihood function,

$$\sqrt{N} (\hat{\theta} - \theta_0) \approx \left\{ -\frac{1}{N} \frac{\partial S(\theta_0)}{\partial \theta} \right\}^{-1} \frac{1}{\sqrt{N}} S(\theta_0) \approx \mathbf{J}^{-1} \frac{1}{\sqrt{N}} S(\theta_0). \quad (25)$$

As the Fisher information (3) is block diagonal,  $\mathbf{J}^{-1} = \begin{bmatrix} \mathbf{J}_{\beta\beta}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^{-1} \end{bmatrix}$ . Under Assumption 1.2  $\mathbf{Y} - \mathbf{X}\beta \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ , and the score functions for  $\beta$ , i.e. the first  $p$  components of  $S(\theta)$ , are  $S_\beta(\theta) = \mathbf{X}^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta)$ . By extracting the first  $p$  components of (25), with  $A \approx B$  denoting  $A - B \xrightarrow{P} \mathbf{0}$ , we have

$$\sqrt{N} (\hat{\beta} - \beta_0) \approx \mathbf{J}_{\beta\beta}^{-1} S_\beta(\theta_0) / \sqrt{N} = \mathbf{J}_{\beta\beta}^{-1} \mathbf{X}^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta_0) / \sqrt{N}.$$

Thus,

$$\begin{aligned} \sqrt{N} \begin{pmatrix} \{(\mathbf{f} - \mathbf{e}(\beta_0))\} / N \\ \hat{\beta} - \beta_0 \end{pmatrix} &\approx \begin{pmatrix} N^{-1/2} \begin{bmatrix} I_{\{x_{11} \in E_1\}} & \cdots & I_{\{x_{mm} \in E_1\}} \\ \vdots \\ I_{\{x_{11} \in E_L\}} & \cdots & I_{\{x_{mm} \in E_L\}} \end{bmatrix} \\ N^{-1/2} \begin{bmatrix} I_{\{x_{11} \in E_L\}} & \cdots & I_{\{x_{mm} \in E_L\}} \\ N^{-1/2} \mathbf{J}_{\beta\beta}^{-1} \mathbf{X}^T \mathbf{V}^{-1} \end{bmatrix} \end{pmatrix} (\mathbf{Y} - \mathbf{X}\beta_0 - \mathbf{0}) \\ &= \mathbf{D}_{(L+p) \times N} (\mathbf{Y} - \mathbf{X}\beta_0), \end{aligned}$$

which is a linear combination of Gaussian random variables.

Under Assumptions 1.4, 1.5, 1.7, which ensure the existence of components of the covariance matrix of the test statistic, we get as  $N \rightarrow \infty$ ,

$$\sqrt{N} \begin{pmatrix} \{(\mathbf{f} - \mathbf{e}(\beta_0))\} / N \\ \hat{\beta} - \beta_0 \end{pmatrix} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{D}\mathbf{V}\mathbf{D}^T). \quad (26)$$

## Appendix B: Proof of Corollary 2

Under asymptotic normality of  $\sqrt{N} (\hat{\beta} - \beta_0)$ ,

$$\begin{aligned} &\frac{1}{\sqrt{N}} \{ \mathbf{f} - \mathbf{e}(\hat{\beta}) \} = \frac{1}{\sqrt{N}} \{ \mathbf{f} - \mathbf{e}(\beta_0) \} + \frac{1}{\sqrt{N}} \{ \mathbf{e}(\beta_0) - \mathbf{e}(\hat{\beta}) \} \\ &\approx \frac{1}{\sqrt{N}} \{ \mathbf{f} - \mathbf{e}(\beta_0) \} = \frac{1}{\sqrt{N}} \nabla \mathbf{e}(\beta_0) \{ \hat{\beta} - \beta_0 \} \approx \frac{1}{\sqrt{N}} \{ \mathbf{f} - \mathbf{e}(\beta_0) \} - \Lambda \sqrt{N} \{ \hat{\beta} - \beta_0 \} \\ &= (\mathbf{I} - \Lambda) \sqrt{N} \begin{pmatrix} \{ \mathbf{f} - \mathbf{e}(\beta_0) \} / N \\ \hat{\beta} - \beta_0 \end{pmatrix}. \end{aligned}$$

Since  $N^{-1/2} \{ \mathbf{f} - \mathbf{e}(\hat{\beta}) \}$  is a linear combination of components of the left hand side of (26),  $N^{-1/2} \{ \mathbf{f} - \mathbf{e}(\hat{\beta}) \} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Sigma)$ , with  $\Sigma = \mathbf{H} - \Lambda \mathbf{J}_{\beta\beta}^{-1} \Lambda^T$ .

## Appendix C: Proposition 1

**Proposition 1.** Suppose that a sequence  $Z_N$  of random  $q$ -vectors is asymptotically distributed as  $\mathcal{N}(0, \Sigma_0)$ , where  $\text{rank}(\Sigma_0) = r - 1$  and there exists a known  $\xi > 0$  smaller than the minimum positive eigenvalue of  $\Sigma_0$ , and that  $\hat{\Sigma}$  is a consistent covariance-matrix-valued estimator of  $\Sigma_0$ .

Let the spectral decomposition of  $\hat{\Sigma}$  be given by

$$\hat{\Sigma} = \sum_{k=1}^q c_{kN} \mathbf{v}_{kN} \mathbf{v}_{kN}^T$$

where  $c_{kN}$  are the eigen values and  $\{\mathbf{v}_{kN}\}_{k=1}^q$  form an orthonormal eigenbasis determined from  $\hat{\Sigma}$ . Define

$$\hat{\Sigma}_\zeta = \sum_{k=1}^q c_{kN} I_{[c_{kN} > \zeta]} \mathbf{v}_{kN} \mathbf{v}_{kN}^T \quad \text{and} \quad \hat{\Sigma}_\zeta^- \equiv \sum_{k=1}^q I_{[c_{kN} > \zeta]} (1/c_{kN}) \mathbf{v}_{kN} \mathbf{v}_{kN}^T$$

and let  $\tilde{\Sigma}^-$  be any other generalized inverse of  $\hat{\Sigma}_\zeta$ , i.e. any matrix such that  $\hat{\Sigma}_\zeta \tilde{\Sigma}^- \hat{\Sigma}_\zeta = \hat{\Sigma}_\zeta$ . Then

- i.  $P(\text{rank}(\hat{\Sigma}_\zeta) = \text{rank}(\Sigma_0)) \rightarrow 1$  and  $\hat{\Sigma}_\zeta \xrightarrow{P} \Sigma_0$ .
- ii.  $Z_N^T \hat{\Sigma}_\zeta^- Z_N \xrightarrow{\mathcal{D}} \chi_r^2$  and  $Z_N^T \tilde{\Sigma}^- Z_N \xrightarrow{\mathcal{D}} \chi_r^2$  as  $N \rightarrow \infty$ .

**Proof of Proposition.** Note first that while the eigenvectors  $\mathbf{v}_{kN}$  are not necessarily uniquely determined if any eigenvalues have multiplicity greater than 1, the eigenspaces spanned by  $\{\mathbf{v}_{kN}: 1 \leq k \leq q, c_{kN} > s\}$  are uniquely and measurably determined from  $\hat{\Sigma}$  for each real  $s > 0$ . Therefore  $\{c_{kN}\}_{k=1}^q$  and all of the random variance matrices  $\hat{\Sigma}_\zeta, \hat{\Sigma}_\zeta^-$  are well-defined, coordinate-free and measurably defined from  $\hat{\Sigma}$ .

Without loss of generality, let the eigenvalues  $c_{kN}$  of  $\hat{\Sigma}$  be indexed in nondecreasing order. Since the  $k$ 'th smallest eigenvalue is a continuous function on the set of  $q \times q$  symmetric nonnegative definite matrices (Golub and van Loan 1983, pp. 18-19), it follows from the convergence  $\hat{\Sigma} - \Sigma_0 \xrightarrow{P} 0$ , that for arbitrarily small  $\epsilon \in (0, \xi)$ , the event

$$A_N(\epsilon) \equiv \left[ c_{q-r,N} \leq \epsilon, \quad c_{q-r+1,N} > \zeta, \quad \sup_{\mathbf{x}: \|\mathbf{x}\|=1} \|(\hat{\Sigma} - \Sigma_0)\mathbf{x}\| \geq \epsilon \right]$$

has probability converging to 1 as  $N \rightarrow \infty$ . This implies that on  $A_N(\epsilon)$ , the range space of  $\hat{\Sigma}_\zeta$  is exactly the span of the eigenvectors  $\mathbf{v}_{kN}$  with  $k \leq q - r + 1$ , and therefore that  $\text{rank}(\hat{\Sigma}_\zeta) = r$  on the event  $A_N(\epsilon)$ . Moreover, on the event  $A_N(\epsilon)$ , for all  $\mathbf{x} \in \mathbf{R}^q$  with  $\|\mathbf{x}\| = 1$ ,

$$\begin{aligned} \|(\hat{\Sigma}_\zeta - \Sigma_0) \mathbf{x}\| &\leq \|(\hat{\Sigma} - \Sigma_0) \mathbf{x}\| + \|(\hat{\Sigma}_\zeta - \hat{\Sigma}) \mathbf{x}\| \\ &\leq \epsilon + \left\| \sum_{k=1}^{q-r} c_{kN} \mathbf{v}_{kN} \left( \mathbf{x}^T \mathbf{v}_{kN} \right) \right\| \leq \epsilon + \epsilon \end{aligned}$$

since  $\max\{|c_{kN}|: k = 1, \dots, q-r\} \leq \epsilon$  and  $\sum_{k=1}^q (\mathbf{x}^T \mathbf{v}_{kN})^2 = \|\mathbf{x}\|^2 = 1$ . This shows the matrix sup-norm of  $\hat{\Sigma}_\zeta - \Sigma_0$  converges in probability to 0 as  $N \rightarrow \infty$ , completing the proof of (i).

By (i), the asymptotic distribution of  $Z_N$  is the same as  $\hat{\Sigma}_\zeta^{1/2} W$ , where  $W \sim \mathcal{N}(0, \mathbf{I}_{q \times q})$  is independent of  $Z_N$  and the matrix square-root is the symmetric square-root equal to  $\sum_{k=1}^q c_{kN}^{1/2} \mathbf{I}_{c_{kN} > \zeta} \mathbf{v}_{kN} \mathbf{v}_{kN}^T$ . Therefore, by the continuous mapping theorem, the asymptotic distribution of  $Z_N^T \hat{\Sigma}_\zeta^- Z_N$  is the same as the distribution of  $W^T (\hat{\Sigma})^{1/2} \hat{\Sigma}_\zeta^- (\hat{\Sigma})^{1/2} W$ , which is  $\chi_r^2$  since  $(\hat{\Sigma}_\zeta^-)^{1/2} \hat{\Sigma}_\zeta^- (\hat{\Sigma}_\zeta^-)^{1/2}$  is symmetric and idempotent with trace  $r$ . The only feature of  $\hat{\Sigma}_\zeta^-$  that has been used in this proof is the generalized-inverse property  $\hat{\Sigma}_\zeta \hat{\Sigma}_\zeta^- \hat{\Sigma}_\zeta = \hat{\Sigma}_\zeta$  shared by  $\tilde{\Sigma}^-$ . This fact about generalized inverses, which completes the proof of assertion (ii), was previously proved in detail by Rao (1973, 1b.5.(viii), 3b.4.(vii) or 3b.5.(iv)).

### Appendix D: Assumptions for Theorem 3

For the rest of the Appendices we employ the notation  $v^{\otimes 2} = v v^T$  for any vector  $v$ .

**Assumption 3.1.** The true parameter point  $\theta_0 = (\beta_0, \psi_0)$  is an interior point of  $\Theta = (\mathcal{R}^p, (\mathcal{R}^+)^{\mathbf{R}+1})$ .

**Assumption 3.2.**  $E(\alpha_i) = E(\epsilon_{ij}) = 0$ ,  $Var(\alpha_i) = \sigma_\alpha^2$ ,  $Var(\epsilon_{ij}) = \sigma_\epsilon^2$  and there is a  $\delta > 0$ , such that  $E(\alpha_i^{4+\delta}) < \infty$  and  $E(\epsilon_{ij}^{4+\delta}) < \infty$ .

**Assumption 3.3.**  $\mathbf{X}$  is a matrix of random variables,  $(x_j, n_j)$  are i.i.d. with  $\|E(\mathbf{x}_i^T \mathbf{x}_i)\| < \infty$ ,  $E(\mathbf{x}_1^T \mathbf{x}_1)$  being positive definite, and  $E(n_i^2) < \infty$ .

### Appendix E: Proof of Theorem 3

The following Lemma is used in proving Theorem 3.

**Lemma 1.** Let  $\{u_{in}: n = 1, 1 \leq i \leq n\}$  be a triangular array of i.i.d. random variables within each row (i.e., across  $i$ ) with mean 0 and finite variance  $\sigma_w^2$ , and that these variables are independent of the random array  $\{c_{in}: n = 1, 1 \leq i \leq n\}$  which satisfies, as  $n \rightarrow \infty$ , (a)  $\max_{1 \leq i \leq n} |c_{in}| \rightarrow 0$  and (b)  $\sum_{i=1}^n c_{in}^2 \rightarrow \kappa$  in probability, where  $\kappa \in (0, \infty)$ . Then  $\sum_{i=1}^n c_{in} u_{in} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \kappa)$  as  $n \rightarrow \infty$ .

**Proof of Lemma 1:**  $\left\{ \sum_{i=1}^k c_{in} u_{in} \right\}_{k=1}^n$  is a martingale with respect to the filtration  $\mathcal{F}_{kn} = \sigma(\{c_{in}, u_{in} : 1 \leq i \leq k\})$  and the Lemma follows directly from the Martingale Central Limit Theorem (Hall and Heyde, 1980).

**Proof of Theorem 3:** Let the  $n \times p$  covariate matrix for the  $i$ -th cluster be

$$\mathbf{x}_i = (\mathbf{x}_{i1}^T, \dots, \mathbf{x}_{in_i}^T). \text{ Then}$$

$$\sqrt{N}(\tilde{\beta} - \beta_0) = \sqrt{N}(\mathbf{X}^T \tilde{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \tilde{\mathbf{V}}^{-1} (\mathbf{Y} - \mathbf{X} \beta_0) \approx \left( \frac{\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}}{N} \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^m \mathbf{x}_i^T \mathbf{V}_i^{-1} (y_i - \mathbf{x}_i \beta_0).$$

Then

$$\begin{aligned} \left\{ \mathbf{f} - \mathbf{e}(\tilde{\beta}) \right\} N &= \left\{ \mathbf{f} - \mathbf{e}(\beta_0) \right\} / \sqrt{N} + \left\{ \mathbf{e}(\beta_0) - \mathbf{e}(\tilde{\beta}) \right\} / \sqrt{N} \\ &\approx \left\{ \mathbf{f} - \mathbf{e}(\beta_0) \right\} / \sqrt{N} - \nabla \mathbf{e}(\beta_0) (\tilde{\beta} - \beta_0) / \sqrt{N} \\ &\approx \sum_{i=1}^m \frac{1}{\sqrt{N}} \left\{ \begin{pmatrix} z_{i1} \\ \vdots \\ z_{iL} \end{pmatrix} - \frac{\nabla \mathbf{e}(\beta_0)}{N} \left( \frac{\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}}{N} \right)^{-1} \mathbf{x}_i^T \mathbf{V}_i^{-1} (y_i - \mathbf{x}_i \beta_0) \right\}, \end{aligned}$$

with  $z_{il} = \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} (y_{ij} - E(y_{ij})) = \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_l\}} (y_{ij} - \mathbf{x}_{ij} \beta_0)$ ,  $i = 1, \dots, m$ ,  $l = 1,$

$\dots, L$ . Let  $\tilde{\mathbf{\Lambda}} = N^{-1} \nabla \mathbf{e}(\beta_0) \xrightarrow{P} \mathbf{\Lambda}$ ,  $\tilde{\mathbf{J}}_{\beta\beta} = N^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \xrightarrow{P} \mathbf{J}_{\beta\beta}$ . We next show that

$(\mathbf{f} - \mathbf{e}(\tilde{\beta})) / \sqrt{N}$  has a limiting Gaussian distribution by using the multivariate Central Limit Theorem. For any constant vector  $\mathbf{C} = (C_1, \dots, C_L)^T$ , since the inverse of  $\mathbf{V}_i$  is

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

$$\begin{aligned} \mathbf{C}^T N^{-1/2} \left\{ \mathbf{f} - \mathbf{e}(\tilde{\beta}) \right\} &\approx \sum_{i=1}^m \frac{1}{\sqrt{N}} \left\{ \sum_{l=1}^L C_l z_{il} - \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{J}}_{\beta\beta}^{-1} \mathbf{x}_i^T \mathbf{V}_i^{-1} (y_i - \mathbf{x}_i \beta_0) \right\} \\ &= \sum_{i=1}^m \left[ \frac{1}{\sqrt{N}} \sum_{l=1}^{n_i} \left\{ \sum_{l=1}^L C_l I_{\{x_{ij} \in E_l\}} - \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{J}}_{\beta\beta}^{-1} \left( \frac{1}{\sigma_\epsilon^2} x_{ij} - \frac{n_i \sigma_\alpha^2}{\sigma_\epsilon^2 (\sigma_\epsilon^2 + n_i \sigma_\alpha^2)} \bar{x}_i \right) \right\} \right] \alpha_i \\ &+ \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{1}{\sqrt{N}} \left\{ \sum_{l=1}^L C_l I_{\{x_{ij} \in E_l\}} - \mathbf{C}^T \tilde{\mathbf{\Lambda}} \tilde{\mathbf{J}}_{\beta\beta}^{-1} \left( \frac{1}{\sigma_\epsilon^2} x_{ij} - \frac{n_i \sigma_\alpha^2}{\sigma_\epsilon^2 (\sigma_\epsilon^2 + n_i \sigma_\alpha^2)} \bar{x}_i \right) \right\} \epsilon_{ij} \\ &= \sum_{i=1}^m c_{i,n_i} \alpha_i + \sum_{s=1}^N w_s \epsilon_s, \end{aligned}$$

where the double index  $(i, j)$  is placed in one-to-one correspondence with the single index  $s$ .

Because  $\{\alpha_i\}_{i=1}^m$  and  $\{\epsilon_s\}_{s=1}^N$  are i.i.d and satisfy conditions (a) and (b) of Lemma 1, the above sums have limiting normal distributions as  $m \rightarrow \infty$ . Because  $\alpha_i$  and  $\epsilon_{ij}$  are

independent  $\sum_{i=1}^m c_{i,n_i} \alpha_i$  and  $\sum_{s=1}^N w_s \epsilon_s$  are conditionally independent given  $(x_j, n_j)$ . As the two sums are jointly normal and asymptotically uncorrelated they are asymptotically

independent and the limiting distribution of  $\mathbf{C}^T N^{-1/2} (\mathbf{f} - \mathbf{e}(\tilde{\beta}))$  is normal. Moreover, for any constant vector  $\mathbf{C}$ , its limiting variance is of the form  $\mathbf{C}^T \Sigma \mathbf{C}$  with the same fixed  $\Sigma$ ,

$$\Sigma = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^m \text{Var} \begin{pmatrix} \sum_{j=1}^m I_{\{x_{ij} \in E_i\}} (y_{ij} - x_{ij} \beta_0) \\ \vdots \\ \sum_{j=1}^{n_i} I_{\{x_{ij} \in E_L\}} (y_{ij} - x_{ij} \beta_0) \end{pmatrix} - \lim_{N \rightarrow \infty} \left[ \frac{\nabla \mathbf{e}(\beta_0)}{N} \right] \left[ \frac{\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}}{N} \right]^{-1} \left[ \frac{\nabla \mathbf{e}(\beta_0)}{N} \right]^T = \mathbf{H} - \Lambda \mathbf{J}_{\beta\beta}^{-1} \Lambda^T.$$

Therefore,  $N^{-1/2} \{\mathbf{f} - \mathbf{e}(\tilde{\beta})\} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Sigma)$ , and

$T = \{\mathbf{f} - \mathbf{e}(\tilde{\beta})\}^T \Sigma^{-1} (\mathbf{f} - \mathbf{e}(\tilde{\beta})) / N \xrightarrow{\mathcal{D}} \chi_r^2$ , where  $r = \text{rank}(\Sigma)$ . We replace  $\Sigma$  with  $\hat{\Sigma}_\zeta$ , the reconstructed estimated variance matrix defined as immediately following equation (11) by means of the singular value decomposition applied to any consistent estimator  $\hat{\Sigma}$  of  $\Sigma$ . One such consistent estimator of  $\Sigma$  is to replace all parameters in  $\Sigma$  with least squares and method of moments estimators. Based on Proposition 1 in Appendix C,  $\text{rank}(\hat{\Sigma}_\zeta) = \text{rank}(\Sigma)$  for large  $N$ . Thus

$$T = \{\mathbf{f} - \mathbf{e}(\tilde{\beta})\}^T \hat{\Sigma}_\zeta^{-1} \{\mathbf{f} - \mathbf{e}(\tilde{\beta})\} / N \xrightarrow{\mathcal{D}} \chi_r^2.$$

## Appendix F: Derivation of the power of the test

We derive the power of the test for LMM (1) under contiguous alternatives, based on Le Cam's third lemma (Van der Vaart, 2000).

**Lemma 2. (Le Cam's third lemma)** Let  $P_N$  and  $Q_N$  be two measures on a measurable space, corresponding to a null distribution under investigation, and an alternative hypothesis respectively. Suppose  $W_N$  is a real valued statistic for every  $N$ . If

$$\left( W_N, \log \frac{dQ_N}{dP_N} \right) \xrightarrow{\mathcal{D}_N} N_{L+1} \left( \left[ \begin{array}{c} \mu \\ -\sigma/2 \end{array} \right], \left[ \begin{array}{cc} \Sigma & \tau \\ \tau^T & \sigma^2 \end{array} \right] \right), \quad (27)$$

then  $W_N \xrightarrow{\mathcal{D}_N} N_L(\mu + \tau, \Sigma)$ .

Let  $H_0: \theta_N = \theta_0$ , and  $H_1: \theta_N = \theta_0 + a / \sqrt{N}$ , where  $a$  is a constant vector and  $\theta_N \rightarrow \theta_0$ , as  $n \rightarrow \infty$ . By Taylor expansion, under Theorem 5.21 in van der Vaart (2000),

$$\begin{aligned} \log \frac{dQ_N}{dP_N} &= \log \frac{\text{Likelihood}(\theta_N; \mathbf{Y}, \mathbf{X})}{\text{Likelihood}(\theta_0; \mathbf{Y}, \mathbf{X})} = \log \frac{L(\theta_N)}{L(\theta_0)} \\ &\approx (\nabla \log(L(\theta_0)))^T \frac{\mathbf{a}}{\sqrt{N}} + \frac{1}{2} \frac{\mathbf{a}^T}{\sqrt{N}} (\nabla^{\otimes 2} \log(L(\theta_0))) \frac{\mathbf{a}}{\sqrt{N}} \approx (S_N(\theta_0))^T \frac{\mathbf{a}}{\sqrt{N}} - \frac{1}{2} \mathbf{a}^T \mathbf{J}(\theta_0) \mathbf{a}, \end{aligned}$$

where

$$\begin{aligned}
 S_N(\theta_0) &= \nabla \log(L(\theta_0)) \\
 &= \begin{bmatrix} \mathbf{X}^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta_0) \\ -\frac{1}{2} \text{tr} \left( \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \sigma_a^2} \right) + \frac{1}{2} (\mathbf{Y} - \mathbf{X}\beta_0)^T \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \sigma_a^2} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta_0) \\ -\frac{1}{2} \text{tr} (\mathbf{V}^{-1}) + \frac{1}{2} (\mathbf{Y} - \mathbf{X}\beta_0)^T \mathbf{V}^{-1} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta_0) \end{bmatrix} \quad (28)
 \end{aligned}$$

and the limit of the sample Fisher information per observation is

$$\mathbf{J}(\theta_0) = \lim_{N \rightarrow \infty} -\nabla^{\otimes 2} \log(L(\theta_0)) / N = \lim_{N \rightarrow \infty} \text{Var} \left( \mathbf{S}_N(\theta_0) / \sqrt{N} \right).$$

Thus

$$\log \frac{dQ_N}{dP_N} \xrightarrow{\mathcal{P}_N} N \left( \frac{1}{2} \mathbf{a}^T \mathbf{J}(\theta_0) \mathbf{a}, \mathbf{a}^T \mathbf{J}(\theta_0) \mathbf{a} \right).$$

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 estimate the coefficient  $\beta^*$  corresponding to  $\mathbf{X}^*$ . The sum over the expected values under the model,  $\mathbf{e}(\cdot)$ , in (6) has  $\mathbf{R}^p$  as its domain. Let  $\mathbf{e}^*(\cdot)$  denote the sum over the expected values under the reduced model, computed using  $\mathbf{X}_{N \times p}^*$  instead of  $\mathbf{X}_{N \times p}$  with domain  $\mathbf{R}^{p^*}$ . Let  $W_N = (\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)) / \sqrt{N}$  be the first vector component of (27). Under the null hypothesis  $P_N$ ,  $W_N \rightarrow N(0, \Sigma^*)$ , based on Corollary 2.

Next, we compute the variance-covariance matrix  $\Sigma$  in (27), which is equivalent to the variance-covariance matrix of  $\mathbf{a}^T \mathbf{S}_N(\theta_0) / \sqrt{N}$  and  $(\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)) / \sqrt{N}$ .

$$\begin{aligned}
 & \frac{a}{\sqrt{N}} (\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)) \approx \frac{1}{\sqrt{N}} (\mathbf{f} - \mathbf{e}^*(\beta_0^*)) - \frac{1}{\sqrt{N}} \nabla \mathbf{e}^*(\beta_0^*) (\hat{\beta}^* - \beta_0^*) \\
 \approx & \frac{1}{\sqrt{N}} (\mathbf{f} - \mathbf{e}^*(\beta_0^*)) - \Lambda^* \sqrt{N} (\hat{\beta}^* - \beta_0^*) \\
 \approx & \frac{1}{\sqrt{N}} (\mathbf{f} - \mathbf{e}^*(\beta_0^*)) - \Lambda^* (\mathbf{J}_{\beta\beta}^*)^{-1} (\mathbf{X}^*)^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}^* \beta_0^*) / \sqrt{N} = \frac{1}{\sqrt{N}} (\mathbf{A} - \mathbf{B}) (\mathbf{Y} - \mathbf{X}^* \beta_0^*),
 \end{aligned}$$

where  $\mathbf{J}_{\beta\beta}^*$  denotes the information matrix corresponding to  $\beta^*$ , and

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

Thus,



$$\begin{aligned} \text{Cov} \left( \frac{\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)}{\sqrt{N}}, \log \frac{dQ_N}{dP_N} \right) &= \text{Cov} \left( \frac{\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)}{\sqrt{N}}, \frac{\mathbf{a}^T \mathbf{S}_N(\theta_0)}{\sqrt{N}} \right) \\ &= \frac{1}{N} \text{Cov} \left( \mathbf{f} - \mathbf{e}^*(\hat{\beta}^*), \mathbf{a}_1^T \mathbf{S}_\beta + a_2 \mathbf{S}_{\sigma_\alpha^2} + a_3 \mathbf{S}_{\sigma_\epsilon^2} \right). \end{aligned} \quad (29)$$

Under equation (28), since both  $\text{tr} \left( \mathbf{V}^{-1} \left( \partial \mathbf{V} / \partial \sigma_\alpha^2 \right) \right)$  and  $\text{tr}(\mathbf{V}^{-1})$  are scalars, we have

$$\begin{aligned} &\text{Cov} \left( \mathbf{f} - \mathbf{e}^*(\hat{\beta}^*), \mathbf{S}_{\sigma_\alpha^2} \right) \\ &= \text{Cov} \left( (\mathbf{A} - \mathbf{B}) (\mathbf{Y} - \mathbf{X}^* \beta_0^*), \frac{1}{2} (\mathbf{Y} - \mathbf{X} \beta_0)^T \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \sigma_\alpha^2} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X} \beta_0) \right) \\ &= (\mathbf{A} - \mathbf{B}) \text{Cov} \left( \mathbf{Y} - \mathbf{X}^* \beta_0^*, \frac{1}{2} (\mathbf{Y} - \mathbf{X} \beta_0)^T \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \sigma_\alpha^2} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}^* \beta_0^*) \right) = 0. \end{aligned}$$

Similarly, we get  $\text{Cov} \left( \mathbf{f} - \mathbf{e}^*(\hat{\beta}^*), \mathbf{S}_{\sigma_\epsilon^2} \right) = 0$ . Therefore (29) becomes

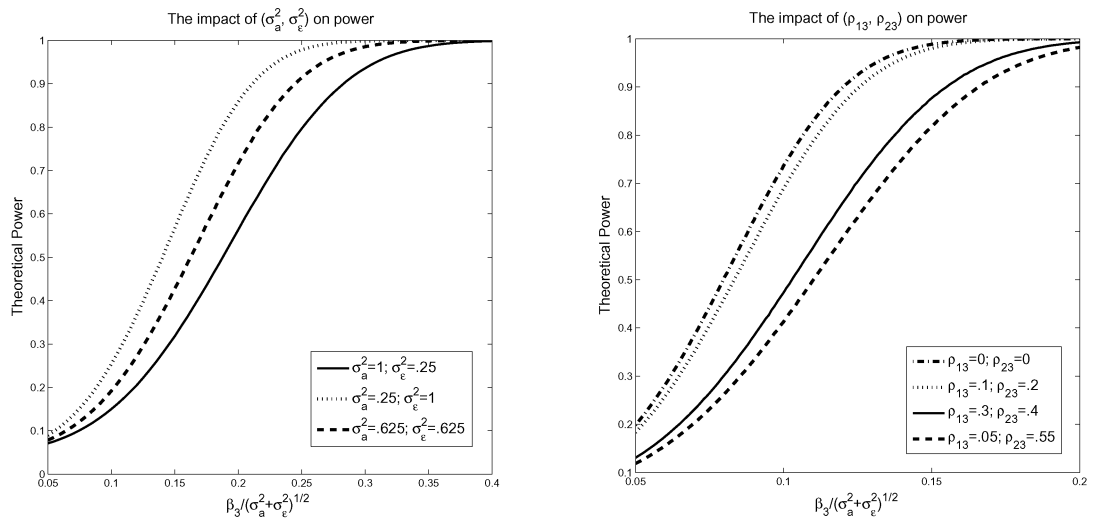
$$\begin{aligned} \text{Cov} \left( \frac{\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)}{\sqrt{N}}, \log \frac{dQ_n}{dP_n} \right) &= \frac{1}{N} \text{Cov} \left( \mathbf{f} - \mathbf{e}^*(\hat{\beta}^*), \mathbf{a}_1^T \mathbf{S}_\beta \right) = \frac{1}{N} (\mathbf{A} - \mathbf{B}) \text{Var}(\mathbf{Y}) \mathbf{V}^{-1} \mathbf{X} a_1 \\ &= \left\{ \Lambda - \frac{1}{N} \Lambda^* (\mathbf{J}_{\beta\beta}^*)^{-1} [(\mathbf{X}^*)^T \mathbf{V}^{-1} \mathbf{X}] \right\}_{a_1} = \left\{ \Lambda - \Lambda^* [(\mathbf{X}^*)^T \mathbf{V}^{-1} (\mathbf{X}^*)]^{-1} [(\mathbf{X}^*)^T \mathbf{V}^{-1} \mathbf{X}] \right\}_{a_1}. \end{aligned}$$

Since both  $\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)$  and  $\mathbf{a}_1^T \mathbf{S}_\beta$  can be written as a matrix multiplied by the same normal vector  $\mathbf{Y} - \mathbf{X} \beta_0 = \mathbf{Y} - \mathbf{X}^* \beta_0^*$ , we obtain asymptotic joint normality of  $\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)$  and  $\mathbf{a}_1^T \mathbf{S}_\beta$ . Because  $\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)$  is asymptotically uncorrelated with both  $\mathbf{S}_{\sigma_\alpha^2}$  and  $\mathbf{S}_{\sigma_\epsilon^2}$  as shown in the above,  $\mathbf{f} - \mathbf{e}^*(\hat{\beta}^*)$  and  $\mathbf{a}_1^T \mathbf{S}_N(\theta_0)$  are also asymptotically jointly normal.

## References

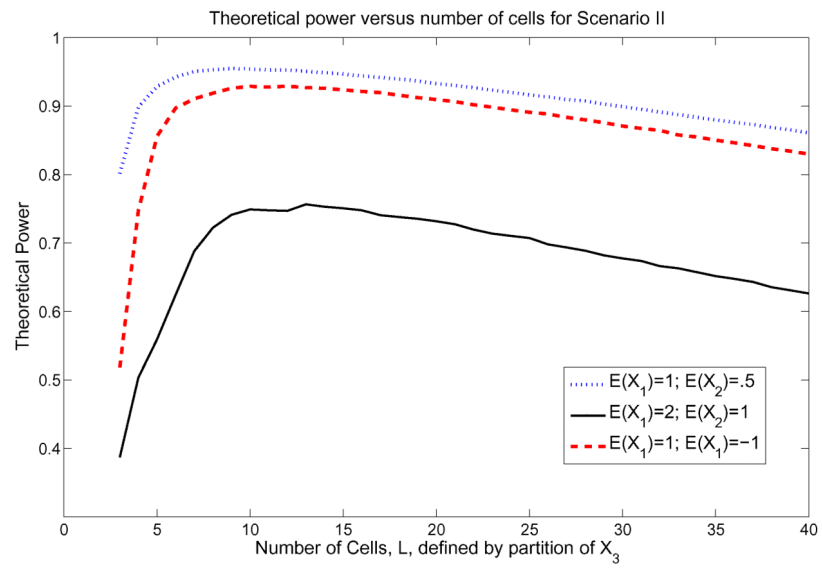
- Cagnone S. A note on goodness of fit test in latent variable models with categorical variables. *Commun. Stat. A Theor.* 2012; 41:2983–2990.
- Claeskens G, Hart JD. Goodness-of-fit tests in mixed models. *Test.* 2009; 10:1100–1120.
- Cox, DR. Tests of separate families of hypotheses; *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*; 1961. p. 105–123.
- Crainiceanu CM, Ruppert D. Likelihood ratio tests in linear mixed models with one variance component. *J. R. Statist. Soc. B.* 2004; 66:165–185. D.
- Godfrey, LG. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches* (Econometric Society Monographs). Cambridge University Press; Cambridge, United Kingdom: 1988.
- Golub, GH., Van Loan, CF. *Matrix Computations*. The Johns Hopkins Studies in Mathematical Sciences, The Johns Hopkins University Press; Baltimore and London: 1983.
- Hall, P., Heyde, CC. *Martingale Limit Theory and its Applications*. Academic Press; New York: 1980.
- Jiang J. REML Estimation: Asymptotic Behavior and Related Topics. *Ann. Stat.* 1996; 24:255–286.
- Jiang J. Goodness-of-fit tests for mixed model diagnostics. *Ann. Stat.* 2001; 4:1137–1164.
- Khuri, AI., Mathew, T., Sinha, BK. *Statistical Tests for Mixed Linear Models*. John Wiley & Sons; New York: 1998.

- Lombardía MJ, Sperlich S. Semiparametric inference in generalized mixed effects models. *J. R. Statist. Soc. B.* 2008; 70:913–930.
- Marozzi M. Multivariate tri-aspect non-parametric testing. *J. Nonparametr. Stat.* 2007; 19:269–282.
- Marozzi M. A modified Cucconi test for location and scale change alternatives. *Colomb. J. Statist.* 2012; 35:369–382. M.
- Maydeu-Olivares A, Joe H. Limited- and full-information estimation and goodness-of-fit testing in 2n contingency tables: A unified framework. *J. Am. Stat. Assoc.* 2005; 100:1009–1020.
- Maydeu-Olivares A, Joe H. Limited information goodness-of-fit testing in multidimensional contingency tables: a unified framework. *Psychometrika.* 2006; 71:713–732.
- McCulloch, CE., Searle, SR. *Generalized, Linear, and Mixed Models.* John Wiley & Sons; New York: 2001.
- Miller JJ. Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Ann. Stat.* 1977; 5:746–762.
- Pan Z, Lin DY. Goodness-of-fit methods for generalized linear mixed models. *Biometrics.* 2005; 61:1000–1009. [PubMed: 16401273]
- Pesarin, F., Fortunato, L. *Permutation Tests for Complex Data: Theory, Applications and Software.* John Wiley & Sons; New York: 2010.
- Rao, CR. *Linear statistical inference and its applications.* second Edition. John Wiley & Sons; New York: 1973.
- Rao CR, Wu Y. A strongly consistent procedure for model selection in a regression problem. *Biometrika.* 1989; 76(2):369–374.
- Richardson AM, Welsh AH. Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *Austral. J. Statist.* 1994; 36:31–43.
- Ritz C. Goodness-of-fit tests for mixed models. *Board of the Foundation of Scand. J. Stat.* 2004; 31:443–458.
- Salmaso L, Solari A. Multiple aspect testing for case-control designs. *Metrika.* 2005; 62:331–340.
- Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika.* 1980; 67:145–153.
- Self SG, Liang K. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Amer. Statist. Assoc.* 1987; 82:605–610.
- Shao J. An asymptotic theory for linear model selection. *Stat. Sinica.* 1997; 7:221–264.
- Stezhko VA, Buglova EE, Danilova LI, et al. A cohort study of thyroid cancer and other thyroid diseases after the Chernobyl accident: Objectives, design and methods. *Radiat. Res.* 2004; 161:481–492. [PubMed: 15038762]
- Tang, M. Ph.D Thesis: Goodness of fit test for generalized linear mixed models. Department of Mathematics, University of Maryland; College Park: 2010.
- van der Vaart, AW. *Asymptotic Statistics.* Cambridge University Press; Cambridge, United Kingdom: 2000.
- Wand MP. Fisher information for generalized linear mixed models. *J. Multivariate. Anal.* 2007; 98:1412–1416.

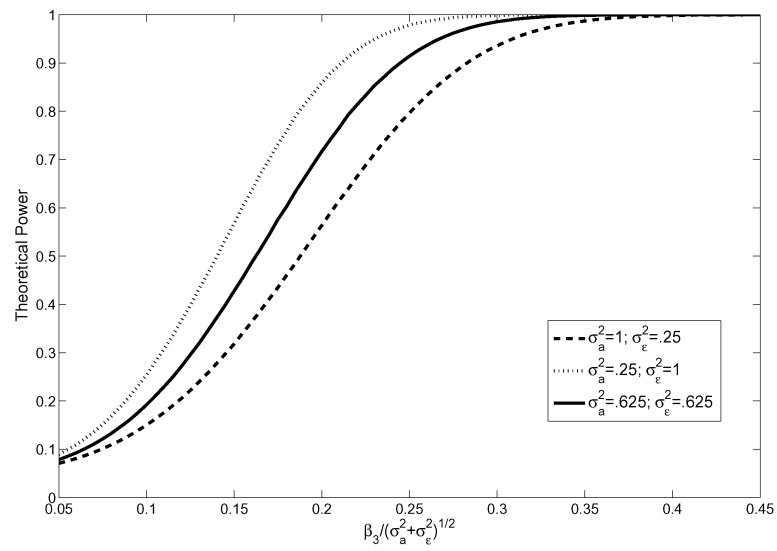


**Figure 1.**

Left: Theoretical power as a function of  $(\sigma_\alpha^2, \sigma_\epsilon^2)$ ; Right: Power as a function of  $(\rho_{13}, \rho_{23})$ .



**Figure 2.**  
The impact of number of cells for cell partition on theoretical power



**Figure 3.**

The impact of  $(\sigma_\alpha^2, \sigma_\epsilon^2)$  on theoretical power,  $\rho_{12} = 0$  (Scenario II)

**Table 1**

Empirical size of the test under different cell partitions (Scenario I).  $m = 500$ ,  $E(N) = 1750$ ,  $\beta_3 = 1$ ,  $\rho_{13} = \rho_{23} = 0$ ,  $\sigma_a = 1$ ,  $\sigma_e = .5$ ,  $K = 2000$ .  $L$  denotes the number of cells for the test statistic.

$L$	$\alpha$	Emp. Size	$\alpha$	Emp. Size
8 ( $x_1$ )	0.05	0.052	0.1	0.103
3×4 ( $x_1, x_2$ )	0.05	0.053	0.1	0.108
5×4 ( $x_1, x_3$ )	0.05	0.045	0.1	0.094
6×7 ( $x_2, x_3$ )	0.05	0.047	0.1	0.096

**Table 2**

Power and robustness study (Scenario I) with  $L = 8, K = 2000, D = 1000, (\rho_{13}, \rho_{23}) = (.5, .6), \sigma_a = 1, \sigma_e = .5$ . Standard deviation (std.dev.) relates to variation across the randomly generated 500 covariate matrices X

Power	<u>m = 500, EN = 1750</u>		<u>m = 50, EN = 175</u>		<u>m = 20, EN = 70</u>	
	mean	std.dev.	mean	std.dev.	mean	std.dev.
$\beta_3 = .25$						
Theo.Pow.	0.800	0.040	0.120	0.023	0.086	0.018
Theo.Pow.hat	0.799	0.039	0.125	0.023	0.090	0.019
Empi.Pow.n	0.799	0.037	0.111	0.022	0.063	0.017
Misspecification of the error term distribution						
Empi.Pow. $t_3$	0.798	0.036	0.112	0.023	0.062	0.019
Empi.Pow. $t_5$	0.799	0.036	0.111	0.022	0.063	0.018
Misspecification of the random intercept distribution						
Empi.Pow. $t_3$	0.817	0.033	0.132	0.027	0.076	0.021
Empi.Pow. $t_5$	0.799	0.036	0.116	0.023	0.067	0.018
$\beta_3 = .8$						
Theo.Pow.	1	0	0.847	0.104	0.541	0.189
Theo.Pow.hat	1	0	0.821	0.096	0.512	0.151
Empi.Pow.n	1	0	0.820	0.102	0.444	0.161
Misspecification of the error term distribution						
Empi.Pow. $t_3$	1	0	0.821	0.102	0.449	0.164
Empi.Pow. $t_5$	1	0	0.821	0.102	0.444	0.162
Misspecification of the random intercept distribution						
Empi.Pow. $t_3$	1	0	0.852	0.074	0.545	0.155
Empi.Pow. $t_5$	1	0	0.824	0.094	0.472	0.159

**Table 3**

Impact of cell partition on empirical power when covariate  $x_3$  with  $\beta_3 = .15$  is omitted from model fitting (Scenario I) with  $m = 500$ ,  $\sigma_a = 1$ ,  $\sigma_e = 0.5$ ,  $K = 2000$ .

Cell Variables	$\rho_{13} = 0, \rho_{23} = 0$		$\rho_{13} = 0.2, \rho_{23} = 0.3$		$\rho_{13} = 0.4, \rho_{23} = 0.5$	
	L=12	L=42	L=12	L=42	L=12	L=42
$x_1$	0.056	0.060	0.046	0.044	0.045	0.044
$x_2$	0.055	0.046	0.048	0.048	0.051	0.050
$x_3$	0.985	0.871	0.936	0.748	0.630	0.367
$x_1, x_2$	0.054	0.050	0.052	0.046	0.051	0.048
$x_1, x_3$	0.968	0.821	0.896	0.732	0.578	0.382
$x_2, x_3$	0.962	0.843	0.913	0.752	0.642	0.435

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**Table 4**

Power and robustness study (Scenario IV).  $L = 8$ ,  $K = 2000$ ,  $D = 1000$ ,  $(\rho_{13}, \rho_{23}) = (.5, .6)$ ,  $\beta_3 = 0.1$ ,  $\sigma_a = 1$ ,  $\sigma_e = .5$ . Note: The standard deviation (std.dev.) relates to variation across the randomly generated 1000 covariate matrices  $\mathbf{X}$ . Each number was obtained as the mean over 1000 simulated covariate matrices  $\mathbf{X}$  and for each generated  $\mathbf{X}$ ,  $K = 2000$  iterations were used to simulate the response vector  $\mathbf{Y}$ .

	<b>m= 500, EN = 1750</b>		<b>m = 50, EN = 175</b>	
	<b>Mean Power</b>	<b>std.dev.</b>	<b>Mean Power</b>	<b>std.dev.</b>
<b>Cell Variable <math>x_3</math></b>	0.873	0.035	0.122	0.028
<b>Wald Test on <math>x_3</math></b>	0.059	0.024	0.065	0.026
<b>Misspecification of the error term distribution</b>				
<b><math>\varepsilon_{ij}</math> simulated from <math>t_3</math></b>				
<b>Cell Variable <math>x_3</math></b>	0.873	0.035	0.123	0.030
<b>Wald Test on <math>x_3</math></b>	0.059	0.024	0.065	0.026
<b><math>\varepsilon_{ij}</math> simulated from <math>t_5</math></b>				
<b>Cell Variable <math>x_3</math></b>	0.873	0.035	0.122	0.029
<b>Wald Test on <math>x_3</math></b>	0.060	0.025	0.066	0.030
<b>Misspecification of the random effect distribution</b>				
<b><math>\alpha_i</math> simulated from <math>t_3</math></b>				
<b>Cell Variable <math>x_3</math></b>	0.878	0.032	0.139	0.033
<b>Wald Test on <math>x_3</math></b>	0.061	0.027	0.067	0.031
<b><math>\alpha_i</math> simulated from <math>t_5</math></b>				
<b>Cell Variable <math>x_3</math></b>	0.872	0.035	0.126	0.030
<b>Wald Test on <math>x_3</math></b>	0.061	0.026	0.067	0.031