Journal of Multivariate Analysis 130 (2014) 176–193

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Goodness of fit tests for linear mixed models

Min Tang [a,*], Eric V. Slud [b,c], Ruth M. Pfeiffer [d,*]

[a] Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA
[b] Department of Mathematics, University of Maryland, College Park, MD, USA
[c] Center for Statistical Research & Methodology, US Census Bureau, USA
[d] National Cancer Institute, Division of Cancer Epidemiology and Genetics, Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 13 August 2013
Available online 5 April 2014

AMS subject classifications:
62J12
62F05
62J20

Keywords:
Asymptotic efficiency
Information matrix
Maximum likelihood estimators
Method of moments
Model fit
Random effects

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.

Published by Elsevier Inc.

1. Introduction

The linear mixed model (LMM) [15] 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 model building, and approaches are also available for comparing non-nested models [3,5]. 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 [20,26] 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 [10] 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 [25] and Crainiceanu and Ruppert [4]. Jiang [8] and Ritz [22] assessed the distributional assumptions for the random effects in LMMs. Claeskens and Hart [2] proposed tests for normality of the random effects and/or error terms. Lombardía and Sperlich [11] introduced a test

* Corresponding authors.
E-mail addresses: amethyst1984tm@gmail.com (M. Tang), pfeiffer@mail.nih.gov (R.M. Pfeiffer).

http://dx.doi.org/10.1016/j.jmva.2014.03.012
0047-259X/Published by Elsevier Inc.

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 [17] proposed 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 [24] 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 [24] 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,

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

where $\mathbf{Y}_{N\times 1}$ is the vector of observations; $\mathbf{X}_{N\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ 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; $\boldsymbol{\beta}$ 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 $\boldsymbol{\alpha}_r$, an $m_r \times 1$ random vector, for $r = 1, \ldots, R$. The random effects $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_R$ are i.i.d. and independent of the error term $\boldsymbol{\varepsilon}$. In the next section we assume that the components α_{kr} of $\boldsymbol{\alpha}_r$ and $\boldsymbol{\varepsilon}$ 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 $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi})$ be the parameters of model (1), where $\boldsymbol{\psi} = (\sigma_\epsilon^2, \sigma_1^2, \ldots, \sigma_R^2)$ 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,

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, m,\ j = 1, \ldots, n_i, \tag{2}$$

where, using a slightly different notation, the $1 \times p$ vector $\mathbf{x}_{ij}^T = (1, x_{ij1}, \ldots, x_{ij(p-1)})$ denotes covariates for the jth observation within the ith cluster. The first entry in $\mathbf{x}_{ij}$ is set to be 1 to accommodate an intercept term in the model. We let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$ denote the vector of observations for the ith 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 $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. Then under model (2), $\mathbf{Y}$ is also normal, $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with a block diagonal covariance matrix $\mathbf{V}$, where each of the m $n_i \times n_i$ blocks $\mathbf{V}_i$, $i = 1, \ldots, m$, has entries $\sigma_\alpha^2 + \sigma_\epsilon^2$ on the diagonal and entries σ_α^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 $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}) = (\boldsymbol{\beta}, \sigma_\alpha^2, \sigma_\epsilon^2)$ are estimated by maximum likelihood (MLE). Here $\mathbf{X}$ is considered to be fixed.

Under Assumptions1.1–1.6 stated in Theorem 1, consistency and asymptotic normality of the MLE $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}})$ follow from [16], i.e. $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(0, \mathbf{J}^{-1})$, where $\mathbf{J}$ denotes the limiting Fisher information matrix. Wand [30] showed that under model (2), $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\psi}}$ are asymptotically uncorrelated and thus

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