Journal of Multivariate Analysis 161 (2017) 123–140

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Assessing skewness, kurtosis and normality in linear mixed models

Alexandra Soberón [a,*], Winfried Stute [b]

[a] Departamento de Economía, Universidad de Cantabria, Avenida Los Castros s/n, E-39005 Santander, Spain
[b] Mathematical Institute, University of Giessen, Arndtstr. 2, D-35392 Giessen, Germany

ARTICLE INFO

Article history:
Received 7 October 2016
Available online 8 August 2017

AMS subject classifications:
62F03
62F05
62H15
62J10

Keywords:
Kurtosis
Linear mixed model
Longitudinal data
Moment estimator
Normality
Skewness

ABSTRACT

Linear mixed models provide a useful tool to fit continuous longitudinal data, with the random effects and error term commonly assumed to have normal distributions. However, this restrictive assumption can result in a lack of robustness and needs to be tested. In this paper, we propose tests for skewness, kurtosis, and normality based on generalized least squares (GLS) residuals. To do it, estimating higher order moments is necessary and an alternative estimation procedure is developed. Compared to other procedures in the literature, our approach provides a closed form expression even for the third and fourth order moments. In addition, no further distributional assumptions on either random effects or error terms are needed to show the consistency of the proposed estimators and tests statistics. Their finite-sample performance is examined in a Monte Carlo study and the methodology is used to examine changes in the life expectancy as well as maternal and infant mortality rate of a sample of OECD countries.

1. Introduction

Linear mixed models are often used to study intra-group correlation patterns. They have attracted considerable attention, e.g., in biomedical, social and economic sciences. For mathematical tractability, a common assumption in these models is that the random effects and the error terms are normally distributed. When this is the case, it is well known that maximum likelihood estimation (MLE) and restricted maximum likelihood estimation (RMLE) perform quite well; see, e.g., [2,5,6]. However, conclusions derived under these restrictive assumptions may not be robust to departures from Gaussianity, especially when data show multimodality and skewness; see, e.g., [16].

The impact of misspecification in the random effects distribution has been extensively investigated in the literature; see, e.g., [9,10,15,16]. However, there seems to be no general consensus about its effect and the proposed alternatives are restricted to specific mixed models. Formal tests to detect mixture distributions in the random effects have also been proposed. For example, [7] provides a goodness-of-fit test for both random effects and error terms, but the statistic does not have an exact χ^2 distribution and it is a bit cumbersome to implement in practice. More recently, [11] proposed to check the normality of the random effects using gradient functions.

Therefore, it is of practical interest to develop a simple test that enables to check the normal distribution assumption of both random effects and error terms. This is not an easy task, however, because in linear mixed models the lack of Gaussianity can arise in more than one component of the regression error. The identification of which component is causing the departure

* Corresponding author.
E-mail address: alexandra.soberon@unican.es (A. Soberón).

http://dx.doi.org/10.1016/j.jmva.2017.07.010
0047-259X/

from normality is then crucial. This paper is concerned with the efficient estimation and testing of linear mixed models when standard distributional assumptions such as normality or symmetry cannot be justified. Specifically, some very simple and intuitive tests to detect departures from normality in the form of skewness and/or kurtosis are proposed based on moment conditions, for which estimators of the higher order moments are necessary.

Note that despite the great importance of higher order moments for statistical inference, there are few references in the literature that define and study estimators for higher than second order moments. Two examples are [3,8] but the methods they advocate exhibit some weaknesses: the first is not easily extended to multivariate problems, and the second is not valid if symmetry fails in practice. To overcome this situation, [13] developed an alternative method that provides consistent estimators of higher order moments. However, their technique is somewhat problematic for the third and fourth moments. In these cases there are several choices of estimating equations that can provide consistent estimators, so finding efficient ones becomes a critical issue.

Thus, the aim of this paper is twofold: (i) to develop a new approach leading directly to efficient estimators of the higher order moments of the error term, thereby solving the efficiency issue in [13]; (ii) to propose tests for detecting departures from normality for both random effects and error terms, based on a proposal made by [4] in the context of longitudinal models. The method proposed here has the following promising features. First, it results in closed form expressions for the estimators of the higher order moments, including the third and the fourth order moment. Second, using generalized least squares (GLS) residuals, no further distributional assumptions on either random effects or error terms are needed to show the consistency of these estimators. Third, the proposed tests enable one to identify departures from normality in the form of skewness and/or kurtosis of each component of the regression error, jointly or separately. Fourth, the estimators of the higher order moments allow one to study distributional properties of the estimators through their Fourier transform. To the best of our knowledge, all these results are original. For methodological reasons, we restrict our attention to the linear mixed model, but the underlying ideas are applicable to handle non-linear, semi-parametric or nonparametric models as well.

Finally, in order to illustrate the feasibility and possible gains of the proposed method, a Monte Carlo study is conducted to assess the finite-sample performance of our proposed estimators and test statistics. As a concrete example, an empirical study based on data from the Organization for Economic Co-operation and Development (OECD) is carried out to measure and assess the importance of various factors determining two commonly accepted measures of health care outcomes: life expectancy and infant mortality.

The rest of the article is organized as follows. In Section 2, we introduce the linear mixed model and describe the estimation method. In Section 3, the corresponding asymptotic properties are studied. In Section 4, we derive some tests to detect departures from normality in the form of skewness or kurtosis, and we study their asymptotic properties. Section 5 contains some simulation results and an empirical application to illustrate the usefulness of the method. Section 6 presents our main conclusions. All proofs are collected in the Appendix.

2. Statistical model and estimation procedure

Assume that data are available from a linear mixed model of the form

$$y_{ij} = \alpha + x_{ij}^{\top}\beta + b_i + \epsilon_{ij}, \tag{1}$$

where $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, \ell_i\}$ denote the group index and the measurements within this group, respectively, ℓ_i is the sample size within group i, and n is the number of subjects. Also, y_{ij} is the response variable corresponding to the jth observation in the ith group, while x_{ij} is a $p \times 1$ vector of covariates. Further, the relation between x_{ij} and y_{ij} described in Eq. (1) contains an intercept parameter α, a $p \times 1$ fixed effect parameter vector β, and some random effects b_i, all of which are unknown. All these quantities are also perturbed by random errors ϵ_{ij}. Throughout this paper, it is assumed that ϵ_{ij} and b_i are independent and identically distributed (iid), and ϵ_{ij} is independent of all b_i and x_{ij} for all i and j. Without loss of generality, we further assume that the expectations of the random effects and errors are zero so that β is identifiable. Otherwise, unknown nonzero expectations can be incorporated in the intercept.

For each $k \in \{2, \ldots, 8\}$, let $\gamma_b^k = \mathrm{E}(b_i^k)$ and $\gamma_\epsilon^k = \mathrm{E}(\epsilon_{ij}^k)$ be the kth moments of the unobserved random effects and the idiosyncratic errors, respectively. This paper is concerned with the efficient estimation and testing of these unknown terms. Since these estimators are based on a suitable combination of the GLS residuals of (1), in this section we first obtain the estimators of the parameters of interest, β and α, and later we focus on the estimation of γ_b^k and γ_ϵ^k.

In all our developments, we are unwilling to impose any condition on the statistical relationship between b_i and the covariates of the model. Rewriting the regression model (1) in vectorial form, we have

$$y_i = \iota_{\ell_i}\alpha + x_i\beta + v_i, \quad v_i = \iota_{\ell_i} b_i + \epsilon_i, \tag{2}$$

where for any $i \in \{1, \ldots, n\}$, v_i is a composed error term, $y_i = (y_{i1}, \ldots, y_{i\ell_i})^{\top}$, $v_i = (v_{i1}, \ldots, v_{i\ell_i})^{\top}$, and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{i\ell_i})^{\top}$ are ℓ_i-dimensional vectors for subject i, $x_i = (x_{i1}, \ldots, x_{i\ell_i})^{\top}$ is an $\ell_i \times p$ dimensional matrix, and ι_{ℓ_i} is a unitary vector of length ℓ_i.

In order to obtain consistent estimators for β, a standard solution is to sweep out b_i by the $\ell_i \times \ell_i$ idempotent transformation matrix $Q_{\ell_i} = I_{\ell_i} - \iota_{\ell_i}(\iota_{\ell_i}^{\top}\iota_{\ell_i})^{-1}\iota_{\ell_i}^{\top}$, leading, for each $i \in \{1, \ldots, n\}$, to

$$Q_{\ell_i} y_i = Q_{\ell_i} x_i \beta + Q_{\ell_i}\epsilon_i, \tag{3}$$

where I_{ℓ_i} is a $\ell_i \times \ell_i$ identity matrix.