Testing transformations for the linear mixed model

Matthew J. Gurka\textsuperscript{a, *}, Lloyd J. Edwards\textsuperscript{b}, Leena Nylander-French\textsuperscript{c}

\textsuperscript{a}Department of Public Health Sciences, Division of Biostatistics and Epidemiology, University of Virginia School of Medicine, P.O. Box 800717, Charlottesville, VA 22908-0717, USA
\textsuperscript{b}Department of Biostatistics, CB #7420, University of North Carolina, 3105H McGavran-Greenberg, Chapel Hill, NC 27599-7420, USA
\textsuperscript{c}Department of Environmental Sciences and Engineering, CB #7431, University of North Carolina, 4114G McGavran-Greenberg, Chapel Hill, NC 27599-7431, USA

Received 9 February 2006; received in revised form 12 May 2006; accepted 13 May 2006
Available online 27 June 2006

Abstract

Transformation of the response of a linear model is a popular method in practice when attempting to satisfy the assumptions of the model. Environmental research routinely uses log-transformations due to the nature of the observed data. The choice of the transformation is often made based upon previous experience or on the comparison of models with different transformed responses. Often a transformation parameter is estimated when fitting a model to a set of data. However, in practice interpretability becomes an issue, as it is only desired to know if a particular transformation is appropriate. Thus, inference tools for a hypothesized value of the transformation, such as the log-transformation in environmental exposure models, have their merit. An examination of hypothesis tests of the transformation parameter in the general linear mixed model will be beneficial due to its practical applications, particularly for areas of environmental research. The effect of outliers on inference about the transformation parameter is also studied.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Linear mixed model; Transformation; Inference; Outliers; Longitudinal data; Environmental research

1. Introduction

1.1. Motivation

The general linear mixed model is a powerful and flexible tool for representing continuous longitudinal data, and consequently it is becoming a standard procedure in many areas of biometric research. As is the case for univariate linear models with independently and identically distributed errors, it is quite common in practice to transform the response in order to meet the Gaussian assumptions of the error components in the mixed model. Environmental research is an example of a field that routinely uses transformations in both univariate linear models and mixed models (e.g., Sahl et al., 1994; Bracken and Patterson, 1996). Transforming the response, often by the natural logarithm, is common for environmental data in which the exposure variable tends to be positively skewed.

Estimation of models with a parametric transformation has been studied extensively for univariate linear models and has begun to receive attention in linear mixed models (Gurka et al., 2006; Lipsitz et al., 2000). However, inference techniques for the transformation parameter developed for the univariate model have not been explicitly extended to...
the mixed model. Such tests are useful in practice when for interpretability reasons, a particular transformation is chosen rather than the exact value of the estimated transformation (e.g., a natural log-transformation). The increasing popularity of linear mixed models, especially for longitudinal data analysis, provides the motivation for the study of inference about the transformation when applied to the linear mixed model. The environmental data analysis described below is just one example of the use of transformations in linear mixed models for longitudinal data. Specifically, we wish to first examine the value of likelihood-based hypothesis tests, such as a Wald test, score test, and likelihood ratio test (LRT), in this context. Furthermore, we will examine the influence of outliers on any tests developed for the transformed mixed model, either at the subject or observation level.

1.2. Literature review

Estimation using a parametric transformation in univariate linear models has been studied extensively (e.g., Box and Cox, 1964; Bickel and Doksum, 1981). Though debated from a methodological standpoint (e.g., Hinkley and Runger, 1984), the use of transformations in univariate models is commonplace in practice. Gurka et al. (2006) presented an extensive methodological treatment for the estimation of the Box–Cox transformation when applied to the general linear mixed model. The authors were able to demonstrate that a single transformation parameter for a linear mixed model of repeated measurements was sufficient in targeting the normality of both sources of randomness associated with the model. An estimation method and techniques for the improved approximation of the variance of the fixed effect estimators in the mixed model were also proposed.

Inference on the transformation parameter in the Box–Cox family of transformations has been an interest in univariate linear models due to its utility in determining whether a specific transformation value is appropriate for a given data set and model. Box and Cox (1964) proposed a LRT for testing their proposed transformation parameter. Andrews (1971) introduced an exact test of significance, which is claimed to be more robust, while Atkinson (1973) proposed a score statistic with better power properties than Andrews’ test. Carroll (1980) proposed a test that serves as a sort of compromise to the tests of Andrews and Atkinson. Atkinson and Lawrance (1989) reviewed and compared other tests of the transformation in the univariate linear model through simulations.

The likelihood-based approach to fitting linear mixed models is particularly sensitive to outliers in the data. Diagnostics developed for the univariate linear model (e.g., Cook and Weisberg, 1982) do not have a simple extension to the linear mixed model due to the multiple definitions of what constitutes an outlier in the mixed model setting. In the context of longitudinal data, an independent sampling unit (subject) could be an outlier compared to the other subjects in the data set, or a particular observation within a subject could exhibit an unusually large influence on the model fit. Christensen et al. (1992) extended Cook’s distance for measuring influence to the fixed effects of the mixed model, conditional on the variance parameters of the model. Verbeke and Molenberghs (2000) extended the notion of local influence first described by Cook (1986) to the mixed model when using the maximum likelihood approach for estimation.

The influence of outliers on estimation and inference of the transformation parameter when applied to the univariate linear model has also been examined. Cook and Wang (1983) proposed case-deletion diagnostics for outliers that affect the estimation of the transformation parameter. Atkinson (1982) provided another method for identifying influential cases in this setting based on the score statistic (Atkinson, 1973) for the hypothesis that the transformation parameter equals a certain value. In discussing inference techniques for the transformation in the linear mixed model, it becomes apparent that the impact of outlying observations needs to be studied.

1.3. Motivating example: magnetic-field exposure data

Given that exposure data tend to be positively skewed, environmental scientists often transform the response by taking the natural logarithm (e.g., Sahl et al., 1994; Bracken and Patterson, 1996). McCurdy et al. (2001) examined low-frequency magnetic fields among working women and homemakers. Based on experience with similar data and after a simple examination of the distribution of the collected exposure data, they fit a linear mixed model of log-transformed magnetic-field exposure measured repeatedly on working women as a function of exposure time and occupation group. For the purposes of this research, a simpler model was studied, including only two occupational groups (Homemaker, \( n = 69 \), and Other, \( n = 130 \)):

\[
\log(y_{ij}) = \beta_0 + \beta_1 x_i + \beta_2 t_{ij} + b_i + e_{ij}.
\]
The model expresses $y_{ij}$, the time-weighted average magnetic-field exposure for woman $i$ at measurement $j$, as a function of an occupational group ($x_i = 0$ if woman $i$ is a homemaker; $x_i = 1$ if she has another occupation) and the percentage of time woman $i$ at measurement $j$ was within three feet of a potential magnetic-field source, $t_{ij}$. The random subject effect and the within-subject error term are assumed to independently follow normal distributions, $b_i \sim N(0, \sigma_b^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$, respectively; $i \in \{1, \ldots, 199\}$ and $j \in \{1, 2\}$.

The positive response $y_{ij}$ is actually transformed using the form proposed by Box and Cox (1964) with a scalar transformation parameter $\lambda$:

$$y_{ij}^{(\lambda)} = \begin{cases} 
\left(\frac{y_{ij}^2 - 1}{\lambda}\right) / \lambda, & \lambda \neq 0, \\
\log(y_{ij}), & \lambda = 0.
\end{cases}$$

Even though it is commonplace within this area of study to transform the exposure response by the natural logarithm, one should examine the validity of such a transformation when fitting a linear mixed model. For this example, the hypothesis of interest is $H_0: \lambda = 0$. Using the estimation methods proposed in Gurka et al. (2006), $\hat{\lambda} = -0.16$ for this particular model of the data. This value provides the most “valid” model of form (1) with respect to the distributional assumptions of $b_i$ and $e_{ij}$. However, interpretability of this exact estimate in practice is an issue, and it would be preferred to determine the validity of the log-transformation for this model of the data. Simple examination of the distribution of the raw exposure data is not sufficient in determining the validity of the log-transformation for the model of interest.

2. An overview of the Box-Cox transformation for the linear mixed model

Combining the methods used in transforming response variables in ordinary linear regression with the theory of the general linear mixed model results in a slightly modified linear mixed model with a transformed response for subject $i \in \{1, \ldots, m\}$:

$$h(y_i) = X_i \beta + Z_i b_i + e_i,$$

in which $y_i$ is a $n_i \times 1$ vector of observations on the $i$th subject, $X_i$ is a $n_i \times p$ known, constant design matrix for the $i$th subject with rank $p$, and $\beta$ is a $p \times 1$ vector of unknown, constant population parameters. $Z_i$ is a $n_i \times q$ known, constant design matrix for the $i$th subject with rank $q$ corresponding to $b_i$, a $q \times 1$ vector of unknown, random individual-specific parameters, and $e_i$ is a $n_i \times 1$ vector of random within-subject error terms. Additionally, let $e_i = Z_i b_i + e_i$ be the “total” error term of the model (3).

For model (3), we make the following distributional assumptions: $b_i \sim \mathcal{N}(0, D)$, and $e_i \sim \mathcal{N}(0, R_i)$ independent of $b_i$. The covariance matrices $D$ and $R_i$ are characterized by unique parameters contained in the $k \times 1$ vector $\theta$. We assume a conditionally independent model; i.e., $R_i = \sigma^2 I_{n_i}$. Thus, the total variance for the response vector in (3) is $\nabla \{h(y_i)\} = \nabla \{e_i\} = \Sigma = Z_i D Z_i + \sigma^2 I_{n_i}$, where $\nabla \{\cdot\}$ is the variance–covariance operator.

Since the model in this context represents a single positive outcome measured repeatedly, the transformation of the response vector $y_i$, $h(y_i)$, will be of form (2) proposed by Box and Cox (1964), for all subjects $i \in \{1, \ldots, m\}$ and measurement occasions $j \in \{1, \ldots, n_i\}$. However, in order to reliably estimate $\lambda$ using existing computational procedures, one should replace (2) with a “scaled” transformation (Box and Cox, 1964):

$$h(y_{ij}) = w_{ij}^{(\lambda)} = \begin{cases} 
\left(\frac{y_{ij}^2 - 1}{\lambda}\right) / \lambda, & \lambda \neq 0, \\
\tilde{y} \log(y_{ij}), & \lambda = 0.
\end{cases}$$

This leads to the model

$$h(y_i) = w_i^{(\lambda)} = X_i \beta * + Z_i b_{si} + e_{si},$$

in which $\tilde{y} = \left(\prod_{i=1}^{m} \prod_{j=1}^{n_i} y_{ij}\right)^{1/N}$ is the geometric mean of the observations pooled across the $m$ subjects and $N = \sum_{i=1}^{m} n_i$ is the total number of observations. The assumptions of model (5) remain consistent with those of model (3). Conditional on the geometric mean, the Jacobian of (4) is equal to one.
Restricted, or residual, maximum likelihood estimation (REML) of mixed models is recommended when interest lies in accurate estimators of the variance components of the mixed model (Verbeke and Molenberghs, 2000). Since REML estimation is quickly becoming the standard method of estimation of the parameters of the linear mixed model, REML estimation will be used for the parameters of model (5). The residual log-likelihood with respect to the original response \( y_i \), ignoring constant terms, is that of a standard general linear mixed model:

\[
L_R (w, \lambda | \theta_e) = -\frac{1}{2} \sum_{i=1}^{m} \log \left| \Sigma_{ei} \right| - \frac{1}{2} \log \left| \sum_{i=1}^{m} X_i' \Sigma_{ei}^{-1} X_i \right| - \frac{1}{2} \sum_{i=1}^{m} \left( w_i^{(\lambda)} - X_i \hat{\beta}_e \right)' \Sigma_{ei}^{-1} \left( w_i^{(\lambda)} - X_i \hat{\beta}_e \right),
\]

where

\[
\hat{\beta}_e = \hat{\beta}_e(\lambda, \theta_e) = \left( \sum_{i=1}^{m} X_i' \Sigma_{ei}^{-1} X_i \right)^{-1} \sum_{i=1}^{m} X_i' \Sigma_{ei}^{-1} w_i^{(\lambda)}. \]

In Gurka et al. (2006), obtaining reliable estimates of all of the parameters of the transformed mixed model was of interest, and thus for interpretability reasons the model of the transformed response \( y_i^{(\lambda)} \) was of primary concern. Since the present focus is on inference techniques for \( \lambda \), the proposed methods will be based on the scaled model (5).

3. Inference about the transformation in the linear mixed model

The primary concern will be the development of inference tools for the hypothesis

\[
H_0 : \lambda = \lambda_0.
\]

Tests of this hypothesis are important in practice when one simply wishes to determine whether or not a certain transformation is valid given the data and model. An ideal first test of the hypothesis \( H_0 : \lambda = 1 \) would help determine whether or not a transformation for a particular model is even necessary. A test on the validity of the log-transformation of the response in fitting a linear mixed model, i.e. testing \( H_0 : \lambda = 0 \), would prove to be very useful, as demonstrated by the discussed magnetic-field exposure data.

Tests developed for the parameters of the mixed model, especially \( \beta_e \), are approximate due to the fact that the covariance parameter vector \( \theta_e \) must be estimated. Consequently, when extending inference techniques to the transformed mixed model, we will assess the approximate distribution theory within the context of inference on the transformation parameter.

3.1. A Wald test

The estimation techniques for \( \lambda \) and \( \varphi' \left( \hat{\lambda} \right) \) developed in Gurka et al. (2006) can be employed to derive a Wald test for hypothesis (7). For the Wald test, one could estimate \( \hat{\lambda} \) using the simple grid search method detailed in Gurka et al. (2006). Then, one can estimate the variance of \( \hat{\lambda} \) from the inverse of the observed information matrix of \( (\beta_e, \lambda) \) evaluated at \( (\hat{\beta}_e, \hat{\lambda}) \), treating the estimate of \( \theta_e \) as fixed and known (Gurka et al., 2006). See the Appendix for details.

Using the REML estimate of \( \lambda, \hat{\lambda} \), and the variance estimate of \( \hat{\lambda} \) that results from inverting the observed information matrix, the resulting Wald statistic for hypothesis (7) is

\[
T_W = \frac{(\hat{\lambda} - \lambda_0)^2}{\varphi' \left( \hat{\lambda} \right)}. \tag{8}
\]

Asymptotically under the null hypothesis, \( T_W \sim \chi^2(1) \). The appropriateness of this distributional assumption when \( \theta_e \) is estimated must be assessed. For the general linear mixed model without a transformation parameter, methods to account for the variability introduced by estimating \( \theta_e \) when making inferences about \( \beta \) have been proposed (Kenward and Roger, 1997). The estimated variance of \( \hat{\beta}_e \) is underestimated when ignoring the variability in estimating \( \theta_e \), and hence statistics with approximate \( t \) and \( F \) distributions have been recommended (Verbeke and Molenberghs, 2000, pp. 56–57).

The downward bias of \( \hat{\varphi}' \left( \hat{\beta}_e \right) \) is not readily evident in \( \hat{\varphi}' \left( \hat{\lambda} \right) \), which has been shown to be a very good approximation of the true variance of \( \lambda \) (Gurka et al., 2006). Monte Carlo standard error estimates from Gurka et al. (2006) demonstrate
the reliability of inference about \( \lambda \) for large samples. However, the suitability of the chi-squared distribution of \( T_W \) for smaller sample sizes needs to be assessed, both in terms of the number of subjects and the number of observations per subject.

3.2. A score test

In a similar fashion to the Wald test, one can develop a score test for hypothesis (7) using the observed information matrix of \((\hat{\beta}_*, \lambda)\) evaluated at \( \lambda = \lambda_0 \). The score test, though, will have computational advantages over the Wald test due to the fact that \( \hat{\lambda} \) does not need to be estimated. This could be of great concern when fitting complex linear mixed models that are slow to converge, even before attempting to estimate \( \lambda \). Therefore, it might serve as a relatively “quick” test useful for data analysts who first simply wish to see if a transformation is needed. The score test for (7) then follows:

\[
T_S = \left[ -\sum_{i=1}^{m} \left[ \frac{\partial w_i^{(\lambda)}}{\partial \lambda} \right] \Sigma_{ai}^{-1} \left( w_i^{(\lambda)} - X_i \hat{\beta}_* \right) \right]^2 \hat{\beta}_* = \hat{\beta}_{s0}, \lambda = \lambda_0
\]

(9)

where \( \hat{\beta}_{s0} \) is the estimate of \( \beta_* \) when \( \lambda = \lambda_0 \), and \( \gamma'(\lambda) \) is computed using methods presented in the Appendix. \( T_S \) can be computed using the REML estimates of the model parameters. When \( \theta_s \) is unknown, its estimate is substituted, and \( \Sigma_{ai} \) is used in (9). Similar to the Wald test, when \( \theta_s \) is assumed known, \( T_S \sim \chi^2(1) \) under the null hypothesis asymptotically. This approximation also needs to be evaluated for various sample sizes and for the case when \( \theta_s \) is unknown.

3.3. A Likelihood ratio test

A LRT for hypothesis (7) can be developed, but only if model (5) is fitted using ML estimation. REML estimation is based on maximizing the likelihood function of a set of error contrasts of the response of model (5). The error contrasts of the model fitted under \( H_0 \) will differ from those of the general model fitted with no restrictions. Hence, \( T_L \) is invalid under REML because the two residual log-likelihoods using different mean structures can no longer be compared (Gurka, 2006). The LRT statistic can be written in the following form:

\[
T_L = -2 \left\{ L \left( w, \lambda_0 | \theta_{s0} \right) - L \left( w, \hat{\lambda} | \theta_s \right) \right\},
\]

(10)

where \( L \left( w, \lambda_0 | \theta_{s0} \right) \) is the maximum log-likelihood of the fitted model under the null hypothesis (7), and \( L \left( w, \hat{\lambda} | \theta_s \right) \) is the maximum log-likelihood of the fitted model with no restrictions on the parameter space. The log-likelihood to be maximized is as follows, ignoring constants:

\[
L \left( w, \lambda | \theta_s \right) = -\frac{1}{2} \sum_{i=1}^{m} \log |\Sigma_{ai}| - \frac{1}{2} \sum_{i=1}^{m} \left( w_i^{(\lambda)} - X_i \hat{\beta}_* \right) \Sigma_{ai}^{-1} \left( w_i^{(\lambda)} - X_i \hat{\beta}_* \right).
\]

(11)

Again, when \( \theta_s \) is unknown, its estimate is substituted, and \( \Sigma_{ai} \) is used in (11). Like the Wald statistic, the LRT statistic is computationally intensive. The ML estimate of \( \lambda \) must be estimated and the corresponding model fitted, in addition to the fact that the model under \( H_0 \) must also be fitted. Under some regularity conditions, and when \( \theta_s \) is assumed known, \( T_L \sim \chi^2(1) \) asymptotically under \( H_0 \).

4. Exploration of magnetic-field exposure data

The linear mixed model of interest for the magnetic-field exposure data is as follows, with \( i \in \{1, \ldots, 199\}; j \in \{1, 2\}:

\[
y_{ij}^{(\lambda)} = \beta_0 + \beta_1 x_i + \beta_2 t_{ij} + b_i + e_{ij}.
\]

(12)

The present focus is on making inference on \( \lambda \); specifically, the hypothesis

\[
H_0 : \lambda = 0
\]

(13)
will be tested to determine whether or not the model of interest (1) is valid. Under the null hypothesis, model (12) reduces to model (1). Tests of hypothesis (13) are based on fitting the model of the scaled response, $w_{ij}^{(2)}$, in order to take advantage of existing estimation procedures.

The first step in making a decision regarding (13) is to implement the estimation procedure outlined in Gurka et al. (2006) to determine $\hat{\lambda}$ using REML estimation, as the Wald test depends on this value. For this particular model and data set, $\hat{\lambda} = -0.16$ with a standard error estimate of 0.06. It is first appropriate to assess the normality assumption of the random intercept and within-unit error term of the fitted mixed model. Fig. 1 displays the normal quantile–quantile
Table 1: Test statistics of the hypothesis $H_0: \lambda = 0$ for model (12) fitted to the original magnetic-field exposure data, with and without the one influential subject.

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>Included outlying subject</th>
<th>Removed outlying subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Value</td>
<td>$p$-value</td>
</tr>
<tr>
<td>Wald ($T_W$)</td>
<td>8.56</td>
<td>0.003</td>
</tr>
<tr>
<td>Score ($T_S$)</td>
<td>7.90</td>
<td>0.005</td>
</tr>
<tr>
<td>Likelihood ratio ($T_L$)$^a$</td>
<td>9.09</td>
<td>0.003</td>
</tr>
</tbody>
</table>

$^a$The likelihood ratio test ($T_L$) was computed from ML estimates.

(Q–Q) plots on the resulting residuals, $\hat{b}_t$ and $\hat{e}_{ij}$, using the fit of model (12) of the log-transformed response, as well as the residuals from the fit of the untransformed model. The Q–Q plots in Fig. 1 supply evidence that indeed the log-transformation is appropriate to achieve approximate normality of the error terms of the mixed model fit from this data.

Closer examination of the Q–Q plots for $\hat{b}_t$ and $\hat{e}_{ij}$ show at least one influential subject that might impact the inference techniques and the resulting conclusions. The subject in question can easily be identified in the Q–Q plot of $\hat{b}_t$ and $\hat{e}_{ij}$ for the untransformed model. The random intercept estimate for the subject (ID #30) is $\hat{b}_{30} = 14.8$, nearly five times greater in magnitude than the next most influential intercept estimate, and $\hat{e}_{30,1} = 19.4$, almost four times greater in magnitude than the next most outlying within-subject residual. Though its influence is diminished in the transformed models, one still must address whether or not this subject affected the estimation of $\lambda$ and resulting inference. The measurements of the subject’s exposure in question were in fact real, and hence the investigators were unwilling to eliminate this subject from the data set for final analysis.

Table 1 displays the computed Wald and score test statistics using REML, and the LRT statistic using ML, for model (12) fit to the magnetic-field exposure data and their corresponding $p$-values for their assumed chi-squared distributions. The Wald and score test statistics using ML were also computed; the values were very similar to those using REML and thus are not displayed. Based on these results, one would reject the null-hypothesis and claim the log-transformation is invalid. However, this contradicts most of what has been done in this topic area, and the conclusion does not correspond with the normality assessment of the Q–Q plots of Fig. 1. Keeping in mind the potential effect of the one outlying subject on the test statistics for hypothesis (13), the tests were computed on the data set with this influential subject’s observations removed; results are also displayed in Table 1. Without the two observations from the one subject, the transformation estimate from the fitted model is $\hat{\lambda} = -0.09$. The standard error estimate of $\hat{\lambda}$ is again equal to 0.06. For all three tests, one now fails to reject the null hypothesis that the log-transformation is the appropriate transformation for model (12). This indicates that estimation of $\lambda$ and the resulting likelihood-based hypothesis tests are very sensitive to outliers. The extent of which the resulting sensitivity is due to subject-level or observation-level outliers within the context of repeated measures data remains in question. The magnitude of the sensitivity for the inference techniques will be examined and compared through simulations for both outlying subject-level data as well as influential observations within subjects.

5. Simulations

Two motivations arose for performing simulation analyses within this research. First, it was of interest to compare the size and power of the proposed hypothesis tests, $T_W$, $T_S$, and $T_L$ for both large and small samples (Goal 1). The second interest was to compare the sensitivity of the tests to outliers (Goal 2). Data sets were generated from the following model:

$$\log(y_{ij}) = y^{(0)}_{ij} = 5 + 2.0x_i + 1.0t_{ij} + b_i + e_{ij}.$$  \hspace{1cm} (14)

The model contains an intercept parameter with the first cohort as the referent occupational group, a parameter representing the increment due to the effect of the second occupational cohort, and a common time parameter. In order to assure $y_{ij} > 0$, the intercept was set equal to 5. Only a random intercept was included in the model, $b_i$, independent of the within-subject error term $e_{ij}$. The distributions of the two random components of the model depended on the
The specified number of subjects (Table 2) of the log-transformed model (17) for a given number of subjects \( m \) and observations per subject \( n_i \). The resulting size and power estimates for the Wald and score tests using REML estimation and for the LRT using ML estimation are reported in Table 2. Corresponding results for the Wald and score tests under ML were very similar to those under REML and are not displayed. Examination of means and variances of the test statistics under the null hypothesis (not shown) allows us to conclude that all three tests are in fact nearly equivalent for large samples and exhibit a chi-square distribution with one degree of freedom under the null hypothesis. The Wald test seems to best exhibit a chi-square distribution with one degree of freedom when the sample size is not as large. The question of “how large is sufficiently large” must be addressed within the context of estimation and inference of the transformation parameter \( \lambda \) when applied to mixed models for longitudinal data. In the case of the simulations generated from the log-transformed model, the simulations performed for Goal 1 demonstrate that the Wald test performs fairly well in the presence of a small number of independent subjects (\( m \)). Even the score test performs reasonably in this setting. Unlike the instance in which one is making inference about the fixed effects \( \beta \), in which a larger number of subjects contribute more information than a larger number of observations per subject (\( n_i \)), each observation for each subject is contributing equal information for the estimation and inference of \( \lambda \). Even though this is a repeated measures setting in which observations are correlated, one can see the value of more observations per subject in making better inferences about \( \lambda \), since \( \lambda \) is simply providing information on the optimum scale of the observations for the particular model in question. The power of all three tests is less than desirable when \( n_i \leq 4 \), no matter the value of \( m \), when trying to detect a difference in magnitude of 0.10 between the hypothesized value of \( \lambda \) and its true value. However, when larger magnitudes are of interest, as is usually the case in practice, power is more than sufficient for even modest sample sizes.

### Table 2

Power and size under REML of the Wald and score tests, and under ML of the likelihood ratio test, of \( H_0 : \lambda = \lambda_0 \), estimated from 10,000 replications of the log-transformed model (17) for a given number of subjects \( m \) and observations per subject \( n_i \).

<table>
<thead>
<tr>
<th>( m )</th>
<th>( n_i )</th>
<th>( T_W )</th>
<th>( T_S )</th>
<th>( T_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_0 = 0 )</td>
<td>( \lambda_0 = 0.1 )</td>
<td>( \lambda_0 = 0.25 )</td>
<td>( \lambda_0 = 0.5 )</td>
<td>( \lambda_0 = 0 )</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.06</td>
<td>0.15</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>0.23</td>
<td>0.82</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.05</td>
<td>0.32</td>
<td>0.93</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>0.05</td>
<td>0.25</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>0.04</td>
<td>0.42</td>
<td>0.99</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.05</td>
<td>0.58</td>
<td>&gt; 0.99</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td>0.05</td>
<td>0.46</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>0.73</td>
<td>&gt; 0.99</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.05</td>
<td>0.87</td>
<td>&gt; 0.99</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>60</td>
<td>2</td>
<td>0.05</td>
<td>0.63</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>0.88</td>
<td>&gt; 0.99</td>
<td>&gt; 0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.05</td>
<td>0.96</td>
<td>&gt; 0.99</td>
<td>&gt; 0.99</td>
</tr>
</tbody>
</table>

Nominal type I error (when \( \lambda_0 = 0 \)) is \( \alpha = 0.05 \).
This phenomenon also explains the outlying subject in the analysis of the magnetic-field data, as this subject also are considered outliers relative to the other observations from that particular subject, inference on significant amount of information for making inferences on support the conclusions from the Goal 1 simulations, that each observation from each subject contributes a significant amount of information for making inferences on.

### Table 3

| Size of the tests of $H_0 : \lambda = 0$, estimated from 10,000 replications of the log-transformed model (17) in the presence of outliers (distribution specified); number of subjects $m = 40$ and observations per subject $n_i = 4$. |
|---|---|---|
| **Distribution of $b_i$** | **Distribution of $e_{ij}$** | **Estimated size of the test** |
| $b_i \sim N(0, 0.5)$ | $e_{ij} \sim N(0, 0.5)$ | $T_W$ | $T_S$ | $T_L^a$ |
| $b_i \sim [0.99, N(0, 0.5) + 0.01, N(0, 5)]$ | $e_{ij} \sim N(0, 0.5)$ | 0.05 | 0.04 | 0.05 |
| $b_i \sim [0.97, N(0, 0.5) + 0.03, N(0, 5)]$ | $e_{ij} \sim N(0, 0.5)$ | 0.06 | 0.05 | 0.06 |
| $b_i \sim [0.95, N(0, 0.5) + 0.05, N(0, 5)]$ | $e_{ij} \sim N(0, 0.5)$ | 0.07 | 0.06 | 0.07 |
| $b_i \sim [0.99, N(0, 0.5) + 0.01, N(0, 10)]$ | $e_{ij} \sim N(0, 0.5)$ | 0.08 | 0.07 | 0.09 |
| $b_i \sim [0.97, N(0, 0.5) + 0.03, N(0, 10)]$ | $e_{ij} \sim N(0, 0.5)$ | 0.08 | 0.06 | 0.08 |
| $b_i \sim [0.95, N(0, 0.5) + 0.05, N(0, 10)]$ | $e_{ij} \sim N(0, 0.5)$ | 0.13 | 0.10 | 0.13 |
| $b_i \sim N(0, 0.5)$ | $e_{ij} \sim [0.99, N(0, 0.5) + 0.01, N(0, 5)]$ | 0.15 | 0.12 | 0.16 |
| $b_i \sim N(0, 0.5)$ | $e_{ij} \sim [0.97, N(0, 0.5) + 0.03, N(0, 5)]$ | 0.13 | 0.12 | 0.13 |
| $b_i \sim N(0, 0.5)$ | $e_{ij} \sim [0.95, N(0, 0.5) + 0.05, N(0, 5)]$ | 0.23 | 0.22 | 0.24 |
| $b_i \sim N(0, 0.5)$ | $e_{ij} \sim [0.99, N(0, 0.5) + 0.01, N(0, 5)]$ | 0.29 | 0.27 | 0.29 |

Nominal type I error is $\alpha = 0.05$.

$^a$The likelihood ratio test ($T_L$) was computed under ML.

### 5.2. Goal 2: effect of outliers on performance of test statistics

A simulation study was also performed in order to obtain a better understanding of the impact of outliers on inference about $\lambda$ for the transformed mixed model. Again, the log-transformed true model (14) was examined in this simulation study. In a similar fashion to Goal 1, the test statistics were evaluated for each of 10,000 replicates in testing the hypothesis $H_0 : \lambda = 0$; $m = 40$ and $n_i = 4$. However, the distributions of the two random components of model (14) were altered to generate a certain percentage of both subject and observation level outliers within each generated data set. The size (when testing $H_0 : \lambda = 0$) was compared for the Wald and score tests using REML estimation and for the LRT using ML estimation when the data are generated from model (14); results for each of the varying specified distributions are displayed in Table 3. Results for the Wald and score tests using ML estimation again were nearly identical to those using REML estimation and are not displayed.

As hypothesized and demonstrated through the magnetic-field exposure data, the simulation results show that even a small percentage of outliers at either the subject or observation level greatly impact the performance of the three examined likelihood-based inference techniques. The size of the two tests of $H_0 : \lambda = 0$ grows dramatically as this proportion of influential observations increases, the magnitude of such an increment much greater than that when subject-level outliers are included. The three tests perform very poorly when observation-level outliers are present. This phenomenon also explains the outlying subject in the analysis of the magnetic-field data, as this subject also had very large observation-level residuals after fitting the model. The results of the Goal 2 simulations demonstrate that outliers on both levels affect estimation and inference of $\lambda$ in the transformed mixed model. One could presume that outliers on the observation-level have more of an impact than those on the subject level. This notion seems to support the conclusions from the Goal 1 simulations, that each observation from each subject contributes a significant amount of information for making inferences on $\lambda$. When only a small number of observations per subject are considered outliers relative to the other observations from that particular subject, inference on $\lambda$ can be affected dramatically.

### 6. Conclusions and discussion

Likelihood-based inference techniques have been proposed and discussed for testing the hypothesis of $H_0 : \lambda = \lambda_0$, where $\lambda$ is the Box–Cox transformation parameter applied to a linear mixed model. The proposed tests serve as somewhat reliable tools when one simply wishes to perform a simple test of whether or not a certain value of a transformation is needed when fitting a mixed model to a set of data, particularly when the data set is sufficiently large. Given the asymptotic equivalency of the presented tests, one must attempt to discern which test is best in a small sample setting or in the presence of outlying subjects or observations. In concluding which inference technique is “best”, it is suggested
that the Wald test under REML should be calculated if computational time is not an issue. Given the speed of modern computers and efficiency of statistical programs, computational constraints should not arise too often except in the case of a very complex model whose fit is slow to converge.

Computational issues may exist when utilizing the proposed inference techniques based on large sample likelihood theory when the log-transformation is most appropriate. When $\lambda = 0$, the computation of the second derivative becomes problematic. In this situation, a simple solution was taken, substituting with a value very close to 0 to avoid division by 0. Thus, in the context of inference on $\lambda$, alternative variance estimation techniques or inference tools may need to be proposed that are valid for all true and hypothesized values of $\lambda$.

The examined tests are sensitive to outliers at both levels of longitudinal data, although they are particularly sensitive to outliers present within subjects. As noted earlier in explaining the magnetic-field data outlier, investigators, particularly in environmental research, wish to include outlying subjects in the final data analysis if in fact their observations are real, since they contribute information for the assessment of the relationship(s) of interest. One must take this into account when using any of the discussed tests. Diagnosing which subjects or observations are outliers and determining the effect of these influential observations on the resulting test statistic and corresponding conclusions is recommended. In the case of the magnetic-field exposure data, the inclusion of the one influential subject resulted in rejection of the log-transformation as a valid transformation for the model, a transformation that is commonplace within this field of research. However, when removing the one subject, the log-transformation is concluded to be valid. One could then utilize this transformation for the final model fit to the data including the influential subject.

The issue of outliers and their impact on inference of $\lambda$ is only addressed partially within this research. Careful examination of the data, particularly through graphical techniques, will provide information as to the presence of outliers within this context. The proposed tests can then be computed with and without the candidate outliers, and decisions can be made accordingly to produce the best, most interpretable model for the given data. However, it is apparent that better inference techniques that are more robust to influential observations could be useful. Extending Andrews’ “exact” test (Andrews, 1971), a test with little power in the linear univariate model, might prove to be beneficial. Another possible extension could be performed on the more robust test (Carroll, 1980). Robust estimation under REML for the linear mixed model was proposed by Richardson and Welsh (1995), and could also be the basis for inference about the transformation parameter in this setting.

Some might argue that inference techniques for the transformation parameter are not of much importance. One could simply estimate the parameter using techniques described in Gurka et al. (2006), and then decide on the best transformation. However, it is not evident in many applied settings whether or not a certain preferred value of $\lambda$ in terms of interpretability is valid simply by examining the estimate of $\lambda$. Often data analysts are not interested in a precise value of $\lambda$ and only want to know if a certain value of $\lambda$ is suitable. The repeated use of diagnostic tools for various values of $\lambda$, such as normal Q–Q plots on the model residuals, might provide sufficient information as to whether or not a particular value of $\lambda$ is valid. Again, though, subjective methods are being relied upon in determining the validity of the transformation. The proposed tests provide objective ways of deciding which value of $\lambda$ to use in the final model while keeping in mind the issue of interpretability. We do not claim that these tests are perfect or necessarily appropriate in certain cases, but they provide a simple way of allowing an investigator to decide on the best scale of the response to use when fitting a linear mixed model.

Acknowledgment

This research was supported in part by National Institute of Environmental Health Sciences (NIEHS) Grants T32-ES07018 and P30-ES10126.

Appendix

The observed information matrix described in Section 3 used in estimating the variance of $\hat{\lambda}$ and the fixed effects estimators in the scaled model is as follows:

$$
\mathcal{A} \left( \hat{\beta}_s, \hat{\lambda} \right) = \begin{bmatrix} \sum_{i=1}^{m} X_i' \sum_{u=1}^{U} X_i & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix},
$$
where

\[ J_{12} = - \sum_{i=1}^{m} \Sigma_{x_i}^{-1} \left( \frac{\partial w_i^{(\lambda)}}{\partial \lambda} \right) \right|_{\lambda=\hat{\lambda}}, \]

\[ J_{22} = \left\{ \sum_{i=1}^{m} \left( \frac{\partial w_i^{(\lambda)}}{\partial \lambda} \right) \right\}^{-1} \left( \frac{\partial w_i^{(\lambda)}}{\partial \lambda} \right) + \sum_{i=1}^{m} \left( \frac{\partial w_i^{(\lambda)}}{\partial \lambda} \right) \right\}^{-1} \left( \Sigma_{x_i}^{-1} \left( w_i^{(\lambda)} - X \hat{\beta}_s \right) \right), \]

\[ \frac{\partial w_i^{(\lambda)}}{\partial \lambda} = \lambda^{-1} \hat{y}^{(\lambda)} \hat{s}_i^{(\lambda)} - \left[ \lambda^{-1} + \log (\hat{y}) \right] w_i^{(\lambda)}, \]

\[ \frac{\partial^2 w_i^{(\lambda)}}{\partial \lambda^2} = \lambda^{-1} \hat{y}^{(\lambda)} \hat{t}_i^{(\lambda)} - 2 \left[ \lambda^{-1} + \log (\hat{y}) \right] \hat{s}_i^{(\lambda)} + \left[ \lambda^{-1} + \log (\hat{y}) \right]^2 + \lambda^{-2} w_i^{(\lambda)}, \]

and \( \hat{s}_i^{(\lambda)} \) and \( \hat{t}_i^{(\lambda)} \) are \( n_i \times 1 \) vectors such that the \( j \)th element is \( y_{ij}^{\hat{\lambda}} \log (y_{ij}) \) and \( y_{ij}^{\hat{\lambda}} (\log (y_{ij}))^2 \), respectively. Inverting the observed information matrix for \( (\hat{\beta}_s, \hat{\lambda}) \) given above leads to estimates of \( \gamma^{-1}(\hat{\beta}_s), \gamma^{-1}(\hat{\lambda}) \), and \( \text{Cov}(\hat{\beta}_s, \hat{\lambda}) \).

References