USE OF R^2 STATISTICS FOR ASSESSING GOODNESS-OF-FIT AND MODEL SELECTION IN THE LINEAR MIXED MODEL FOR LONGITUDINAL DATA

Jean Guilmond Orelien

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor in Public Health in the Department of Biostatistics (School of Public Health)

Chapel Hill
2007

Approved by:

Dr. Lloyd J. Edwards

Dr. Keith Muller

Dr. Paul Stewart

Dr. Bahjat Qaqish

Dr. Victor Schoenbach

ABSTRACT

JEAN G. ORELIEN: Use Of Pseudo-R^2 For Assessing Goodness-Of-Fit and Model Selection in The Linear Mixed Model for Longitudinal Data
(Under the direction of Dr. Lloyd Edwards)

In the Linear Mixed Model (LMM), several R^2 statistics have been proposed for assessing goodness-of-fit. However, the performance of these statistics has not been demonstrated. In this dissertation research, first we show that many of the R^2 statistics that have been proposed in the statistical literature are not appropriate to assess adequacy of the fixed effect terms because they are unable to detect when important covariates are missing from the model. A distinction is made between R^2 statistics that can be classified as marginal and those that can be classified as conditional. We show through simulations that only marginal R^2 statistics are appropriate for assessing the adequacy of the fixed effects in the LMM. To remedy the shortcoming of R^2 statistics that have been proposed, we introduce new R^2 statistics that measure the extent to which the model at hand is better than a null model and statistics that measure how much of the variation in the outcome is explained by the model at hand assuming that the model is adequate. Results from simulations show that our proposed R^2 statistics perform well in assessing adequacy of model fit for the fixed effects or selection of the fixed effects covariates. For selecting the random effects, our proposed R^2 statistics are able to distinguish between a model that includes a time covariate and one that doesn't (such as a model with only a random intercept). However, these statistics were unable to discriminate between a full and a reduced model in the random effects that both included a time covariate such as a full model with an intercept, linear and quadratic component for time and a reduced model with an intercept and linear component for time. We found that even when the true model of the random effects involves variables (polynomial components) beyond the linear term, the reduced model with an intercept and a linear term for time may be as good as the full model.

ACKNOWLEDGEMENTS

In the course of completing this dissertation, I am forever indebted to many family members, friends, colleagues and faculty in the school of Public Health at UNC. I am very thankful for my wife whose support has made it possible to put the time toward work and the pursuit of a doctorate degree knowing that she had everything under control on the "home front". To my kids (Katina, Vladimir, Vijay and Valexa), I thank you for accepting that your dad has not spent as much time as he wanted with you while completing this degree and growing a small business. To my mom who wanted her son to be a doctor, this one is for you. I'm also thankful for the support of my brother-in-law. To my brother Jonas, I am blessed to have you as a brother. Brother, this degree is proof that you too can reach high to achieve your dreams. Don't ever give up on your dreams. To the rest of my family members, I thank you so much for the confidence you have in me. I thank all my friends who have understood that because of my busy schedule, I did not stay in touch as much as I should.

I cannot forget to mention some colleagues who have been very helpful in providing encouragement for the pursuit of this degree or in working around my schedule. Rich Cohn, my former supervisor at Constella Group (formerly Analytical Sciences, Inc.) encouraged me to pursue a doctorate in Statistics and he was also very accommodating. Rich, you taught me a lot about how to treat other people. Steven Yevich, one of the first clients of SciMetrika (the firm I founded), also provided encouragement and was very understanding of my schedule.

Last but not least, it was a privilege to learn from world class faculty at UNC. I’d like to thank members of my dissertation committee and in particular the chair Dr. Lloyd Edwards. Dr. Edwards, your class on Linear Mixed Model was the inspiration for this dissertation. A special thank you to Dr. Shoenback who has taught me everything I know about epidemiology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Literature Review

1.1 Introduction

Few tools are available for assessing goodness-of-fit (GOF) in the linear mixed model. Traditional statistics such as the likelihood ratio test (LRT), the Akaike Information Criterion (AIC) introduced by Akaike (1974), or the Bayesian Information Criterion (BIC) by Shwarz (1978) require that two models be fitted to the data. For the LRT, the two models must be nested. The AIC and BIC can be used when the two models are not nested, though a non-nested mean structure violates the assumptions used to originally derive the AIC. However, in comparing the values of AIC or BIC, it is not clear what magnitude of difference constitutes a meaningful or significant one.

Recently, other statistics have been proposed in the statistical literature for assessing goodness-of-fit in linear mixed models. It is not clear which, if any, of these statistics performs best or what their limitations are. The purpose of this literature review is to evaluate tools that have been recently proposed for assessing GOF in linear mixed models. Tools that have been proposed for other classes of models that can be applied to linear mixed models are also examined. First, we focus on tools for assessing the adequacy of a given model. Second, we investigate tools for assessing the covariance structure of a model assuming that the fixed effect function is properly defined. Third, we look at tools that have been developed for GOF in the Generalized Linear Multivariate Model (GLMM). These tools are of interest because every GLMM can be expressed as a linear mixed model.

1.2 The general linear mixed model

Assume the following linear mixed model (Harville 1977, Laird and Ware 1982):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \qquad (1)$$

where $i \in$ {1, 2, ..., n} is the index for the independent sampling units (ISU) and

$\mathbf{y}_i$ is an $n_i \times 1$ vector of observations from the i^{th} independent sampling unit (subject),

$\mathbf{X}_i$ denotes an $n_i \times p$ fixed effects design matrix for the i^{th} subject,

$\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown, constant, fixed effect parameters,

$\mathbf{Z}_i$ denotes an $n_i \times q$ random effects design matrix for the i^{th} subject,

$\mathbf{b}_i$ is a $q \times 1$ vector of unobservable random effects for the i^{th} subject, and

$\mathbf{e}_i$ denotes an $n_i \times 1$ vector of unobservable within-subject error terms.

It is also assumed that $\mathbf{b}_i$ has a multivariate normal distribution $N_q(\mathbf{0}, \mathbf{G})$ independent of $\mathbf{e}_i$, which has a multivariate distribution $N_{n_i}(\mathbf{0}, \mathbf{R}_i)$.

$$E\begin{bmatrix}\mathbf{b}_i\\ \mathbf{e}_i\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \text{ and } V\begin{bmatrix}\mathbf{b}_i\\ \mathbf{e}_i\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_i\end{bmatrix},$$

where $\mathbf{G}$ is a $q \times q$ unknown covariance matrix for the random effects and $\mathbf{R}_i$ is an $n_i \times n_i$ unknown covariance matrix for the within-subject error terms. With these assumptions, we have $\boldsymbol{\Sigma}_i = V(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$. In many applications, $\mathbf{R}_i$ is taken to be $\sigma^2\mathbf{I}_{n_i}$, known as the conditional independence assumption for the error term (Laird and Ware 1982).

By stacking the vectors of responses and associated matrices, the mixed model can also be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}$$

where

$\mathbf{y} = (\mathbf{y}_1' \mid \mathbf{y}_2' \mid . . . |\mathbf{y}_n')'$ is $N \times 1$ and $N = \sum_{i=1}^{n} n_i$; $\mathbf{X} = (\mathbf{X}_1' \mid \mathbf{X}_2' \mid . . . |\mathbf{X}_n')'$ is $N \times p$;

$\mathbf{Z} = Diag(\mathbf{Z}_1, \mathbf{Z}_2, . . ., \mathbf{Z}_n)$ is $N \times nq$; and $\mathbf{b} = (\mathbf{b}_1' \mid \mathbf{b}_2' \mid . . . |\mathbf{b}_n')'$ is $nq \times 1$ and

$\mathbf{e} = (\mathbf{e}_1' \mid \mathbf{e}_2' \mid . . . |\mathbf{e}_n')'$ is $N \times 1$. The distributional assumptions are that $\mathbf{b} \sim N_{nq}(\mathbf{0}, \mathbf{G} \otimes \mathbf{I}_n)$

independent of $\mathbf{e} \sim N_N(\mathbf{0}, \mathbf{R})$, $\mathbf{R} = Diag(\mathbf{R}_1, \mathbf{R}_2, . . ., \mathbf{R}_n)$ is $N \times N$. Also,

$\mathbf{\Sigma} = V(\mathbf{y}) = Diag(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, ..., \mathbf{\Sigma}_n)$.

A brief overview of approaches to parameter estimation for the model in (1) is given by Ware (1985). The use of maximum likelihood (ML) and restricted maximum likelihood (REML) approaches for linear mixed models was first discussed by Harville (1977). Laird and Ware (1982) proposed a Bayesian approach to estimation and the use of the EM algorithm for both the Bayesian approach and the ML approach. Detailed formulae for computing ML and REML estimates using the EM algorithm with suggestions on how to speed convergence are given in Laird, Lange, and Stram (1987).

1.3 The Concordance Correlation Coefficient

Lin (1989) proposed the concordance correlation coefficient (CCC) as a way to evaluate reproducibility between two sets of measurements, as in the case where there is a “gold standard” assay or instrumentation and the intent is to measure whether a new assay can reproduce the results from the gold standard assay or instrumentation. If the new assay is successful, then the plot of the new assay’s results versus that of the gold standard should fall along the 45 degree or equality line.

Consider *n* pairs of *independent* measurements $(x_i,\ y_i)$ and for $i \neq j,$ the pairs (x_i, y_i) and (x_j, y_j) are independent with $E(x_i) = \mu_x,\ E(y_i) = \mu_y,\ V(x_i) = \sigma_x^2,\ V(y_i) = \sigma_y^2,$ and $\operatorname{cov}(x_i, y_i) = \sigma_{xy}$. The CCC is denoted by ρ_c where:

$$\rho_c = 1 - \frac{E\left[\left(x_i - y_j\right)^2\right]}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

and its estimate is given in the equation below by

$$\hat{\rho}_c = \frac{2S_{12}}{S_1^2 + S_2^2 + (\overline{x} - \overline{y})^2} \qquad (2)$$

where $\overline{x} = \sum_{i=1}^{n} x_i$, $\overline{y} = \sum_{i=1}^{n} y_i$, $S_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$, $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, and

$$S_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Lin (1992) showed how to estimate sample sizes for computing the CCC. A discussion of other methods for assessing agreement is found in Lin, Hedayat, Sinha, and Yang (2002). Muller and Buttner (1994) provide a discussion of the different intraclass correlation coefficient (ICC) statistics used to assess agreement between measurements and how to make an appropriate choice.

1.3.1 Generalization of the CCC

Chinchilli, Martel, Kumanyika, and Lloyd (1996) proposed a weighted concordance correlation coefficient for repeated measures designs. For paired observations (such as arise between observed and predicted values from longitudinal data), each vector of observations from the pairs of measurements are separately modeled with a random-coefficient growth curve model. One potential problem with this approach is the fact that there may be a limited number of variables available to use as covariates in modeling one measurement as a function of the other. This approach would be impractical in the context of using the CCC as a goodness-of-fit statistic for linear mixed models. It is not clear how one would choose the covariates to model the predicted and observed values.

Vonesh, Chinchilli, and Pu (1996) proposed that an unweighted CCC denoted r_c be used to assess goodness-of-fit for the generalized nonlinear mixed effect models. For these models, they formulate the CCC as follows:

$$r_c = 1 - \frac{\sum_{i=1}^{n} (\mathbf{y}_i - \widehat{\mathbf{y}}_i)'(\mathbf{y}_i - \widehat{\mathbf{y}}_i)}{\left(\sum_{i=1}^{n} (\mathbf{y}_i - \bar{y}\mathbf{1}_i)'(\mathbf{y}_i - \bar{y}\mathbf{1}_i)\right) + \left(\sum_{i=1}^{n} (\widehat{\mathbf{y}}_i - \hat{y}\mathbf{1}_i)'(\hat{\mathbf{y}}_{\mathbf{i}} - \hat{y}\mathbf{1}_i)\right) + N(\bar{y} - \hat{y})^2} \qquad (3)$$

where *n* is the number of independent sampling units (or subjects), $\mathbf{y}_i$ is the vector of observed values for the *i*th subject, $\widehat{\mathbf{y}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i$ is the vector of predicted values for the *i*th subject, $\hat{y}$ is the grand average of the predicted values, $\bar{y}$ is the grand average of the observed values, and *N* is the total number of observations.

The r_c measures the percent agreement between observed and predicted values. A value of 1 corresponds to perfect agreement, and values close to 0 correspond to a lack of fit. Note that in their paper, Vonesh et al. (1996) erroneously gave the range for r_c as being between -1 and 1.

Barnhart and Williamson (2001) proposed using GEE to model the CCC. Their contribution is to adjust the estimate of the CCC through modeling with variables that are potential predictors. The results of their simulation showed that confidence intervals for the CCC can be wide for moderate samples. For making comparisons (such as comparing two raters for data that involve multiple measurements), their test is liberal, resulting in higher rates of rejection of the null hypothesis than the stated nominal type I error rate. The liberality of the test could be an artifact of using a distribution free approach to derive estimates. When only an intercept is included in the model, the estimates from Barnhart and Williamson (2001) and Lin (1989) are the same.

Remarking that the CCC is based on the squared function of distance, King and Chinchilli (2001) proposed a generalized CCC for both continuous and categorical data. The authors based their formula for the CCC on convex functions of distance. For categorical data, their class of estimators has similarities with the kappa and weighted kappa statistics. The choice of a particular distance function is akin to choosing a set of weights to estimate a weighted kappa. Their extended version of the CCC can be used in situations where there is an interest in estimating agreement for more than two raters or assays. Barnhart, Haber, and Song (2002) proposed an overall CCC for assessing interobserver variability when there are more than two observers. It turns out that the overall CCC that they proposed is equivalent to the generalized CCC of King and Chinchilli (2001) when the squared distance function is used.

1.3.2 Objections to the use of the CCC

Atkinson and Nevill (1997) object to the use of the CCC and other correlation methods to compare measurements. Their primary argument is that correlation methods such as the CCC are highly sensitive to sample heterogeneity (the fact that with a varied sample, larger values can be obtained) and can lead to erroneous conclusions. Nickerson (1997) remarked that although Lin (1989) objected to the use of intraclass correlation coefficients for assessing reproducibility, the CCC is nearly identical to a subset of coefficients in that group. Liao and Lewis (2000) urge caution when using correlation coefficients and advocate the need for an improved correlation coefficient.

1.3.3 The CCC as a goodness-of-fit statistic

The recommendation for use of the CCC is based on the observation that predicted and observed values are similar to a set of measurements from two instrumentations: a gold standard (observed values) and another set of measurements (predicted values). However, for models such as generalized linear mixed models in which observations from the same subject are correlated, Vonesh et al. (1996) did not take into consideration the fact that the assumptions underlying the CCC, as outlined by Lin (1989), are not applicable. Three assumptions of the CCC are that (a) the two sets of measurements come from a bivariate normal distribution, (b) the two sets of measurements have equal variances, and (c) each pair of measurements for an individual observation are independent of all other pairs. While Lin (1989) showed that the CCC is robust to deviation from normality, the fact that observations from a generalized linear mixed model are correlated raises questions about the use of the CCC for such models. Given the issues of the underlying assumptions of the CCC being violated for correlated models, simulations would be desirable to ensure that the CCC is a suitable goodness-of-fit statistic in such models. Other issues not addressed by Vonesh et al. (1996) include transformations and models in the class of generalized non-linear mixed models, such as logistic regression, for which the observed values are not continuous.

Zheng (2000) recommended the use of r_c, R_1^2 (that he denotes the proportional reduction in entropy measure) and the proportional reduction in deviance measure. However, he did not provide any simulation or analytical results in support of this recommendation. An example is given by Zheng (2000) where he analyzed data on growth measurement published by Pothoff and Roy (1964), with age, gender, and their interaction as potential predictors. The analysis of the data showed that high values of the r_c were obtained from any model that included age as an explanatory variable. Even when other statistically significant terms are removed from the model, the value of the r_c remained relatively unchanged at 0.99. The fact that his computations yielded high values of the r_c even when important terms were removed was not a cause of concern. This was a demonstration to him that the r_c and the other three statistical measures that he proposed could discriminate between "statistical significance" and "practical importance" (Zheng, 2000). The possibility that there could be a problem with the r_c and the other statistics he proposed was not explored. This should have been a consideration in light of the remarks by Atkinson and Nevill (1997) that the r_c, like other intraclass correlation measures, is sensitive to sample heterogeneity. It should be noted that if indeed there is an issue with the inability of r_c to discriminate when other significant variables are missing from the model, then the other statistics proposed by Zheng (2000) are likely to suffer the same deficiency. Although values for these statistics were not as high as the r_c, when significant terms were removed, they too exhibited little change.

1.4 Pseudo-R^2 Measures in Generalized Linear and Nonlinear Models

Various R^2 statistics have been proposed in generalized linear models. Some of these statistics are specific to a subclass of models and would not be applicable to the generalized linear mixed model. For example, various statistics have been proposed for logistic regression and specific members of the exponential family such as Gamma or Poisson distributions. In section 3.1, we focus on statistics that have been proposed specifically for generalized linear mixed models. Statistics that have been proposed for other generalized linear models and that can be applied to linear mixed models are discussed in section 3.2.

1.4.1 Pseudo-*R*^{2} statistics for linear mixed models

Besides the r_c, Vonesh and Chinchilli (1997) proposed a statistic that we denote R_1^2 for assessing GOF in a generalized linear mixed model.

$$R_1^2 = 1 - \frac{\sum_{i=1}^{n_i} (\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^{n_i} (\mathbf{y}_i - \bar{y}\mathbf{1}_i)'(\mathbf{y}_i - \bar{y}\mathbf{1}_i)} \tag{4}$$

This statistic is the counterpart to the traditional R^2 in linear models and as such lends itself to ease of interpretation. However, R_1^2 does not explicitly take into account the random components of the model and no simulation results were offered.

Zheng (2000) proposed two other pseudo-R^2 measures besides the CCC for linear mixed models. The first one, denoted D_{rand} is the same statistic as R_1^2. However, it is referred to by Zheng (2000) as the proportional reduction in deviance. Although D_{rand} is similar to R_1^2, expression for it is given in (5) as it will be useful to simplify the expression for the other statistic proposed by Zheng (2000).

$$D_{rand} = 1 - \frac{\sum_{i=1}^{n} d_i(\mathbf{y}_{i,} \hat{\mathbf{y}}_i)}{\sum_{i=1}^{n} d_i(\mathbf{y}_{i,} \bar{y}\mathbf{1}_i)}, \tag{5}$$

where $i \in$ {1, 2, ..., n} is the index for the ISU and $\bar{y}$ is the grand average of the observed values. Let $L(\boldsymbol{\mu}, \sigma; \mathbf{y})$ denote the joint log-likelihood given the predictors and random effects, where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$. The numerator in equation 5, $\sum_{i=1}^{n} d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -2(L(\hat{\mathbf{y}}_i, \sigma; \mathbf{y}_i) - L(\mathbf{y}_{i,} \sigma; \mathbf{y}_i))$ is defined as the deviance under the model at hand and the denominator $\sum_{i=1}^{n} d_i(\mathbf{y}_i, \bar{y}\mathbf{1}_i) = -2(L(\bar{y}\mathbf{1}_i, \sigma; \mathbf{y}_i) - L(\mathbf{y}_{i,} \sigma; \mathbf{y}_i))$ is defined as the deviance under the null model.

Another statistic proposed by Zheng (2000) is P_{rand}, the proportional reduction in penalized quasi-likelihood (PQL).

$$P_{rand} = 1 - \frac{\sum_{i=1}^{n} d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i)/(2\hat{\sigma}) + \hat{\mathbf{b}}'(\hat{\mathbf{G}} \otimes \mathbf{I}_n)^{-1}\hat{\mathbf{b}}/2}{\sum_{i=1}^{n} d_i(\mathbf{y}_i, \bar{y}\mathbf{1}_i)} \tag{6}$$

where $\sum_{i=1}^{n} d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i)/(2\hat{\sigma}) + \hat{\mathbf{b}}'(\hat{\mathbf{G}} \otimes \mathbf{I}_n)^{-1}\hat{\mathbf{b}}/2$ is defined as the negative of the PQL, $\hat{\mathbf{b}}$ is the estimated vector of random effect parameters for all subjects, and $\hat{\mathbf{G}}$ is the estimated covariance for the random effect parameters.

PQL measures the proportional reduction in the log-likelihood from the model at hand compared to a null model. It takes values between 0 and 1, with values close to 0 indicating a lack of fit and/or large random effects and values of 1 indicating a perfect fit and/or a small random effect. It should be noted that PQL is an attempt similar to the statistics AIC (Akaike 1974) and BIC (Shwarz 1978) to account for additional covariates in the model with a "penalty" term. PQL, as opposed to AIC and BIC, is a pseudo-R^2, with values for lack of fit and perfect fit. However, with these statistics the suitability of a "penalty term" over others needs to be demonstrated.

Xu (2003) proposed two statistics, in addition to the traditional R^2 (R_1^2) of Vonesh and Chinchilli (1997), for explaining the variations in a linear model: a statistic denoted r^2 that measures the proportion of explained variation and a statistic denoted ρ^2 that measures the proportion of explained randomness. The statistic r^2 is derived from the fact that for the model in (1), the variability in the dependent variable $\mathbf{y}_i$ that is not explained by the covariates (both fixed and random) is $\mathrm{V}(\mathbf{y}_i | \mathbf{X},\ \mathbf{b}_i) = \mathrm{V}(\mathbf{e}_i) = \sigma^2 \mathbf{I}_{n_i}$ and the total variance of $\mathbf{y}_i$ under a "null" model that assumes that the covariates have no effect is $V(\mathbf{y}_i) = \sigma_0{}^2 \mathbf{I}_{n_i}$. The statistic r^2 is given by

$$r^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0{}^2} \text{ and estimates } \Omega^2 = 1 - \frac{\mathrm{V}(y_{ij} \,|\, \mathbf{X},\ \mathbf{b})}{\mathrm{V}(y_{ij})},$$

where $j \in \{1, 2, \ldots, n_i\}$ so that y_{ij} is the j^{th} element of $\mathbf{y}_i$, $\hat{\sigma}^2$ is the estimate of σ^2 for the model at hand, and $\hat{\sigma}_0{}^2$ is the estimator for the residual variance of a "null" model. The null model could take the following form:

$$\mathbf{y}_i = \mathbf{1}_i \beta_0 + \mathbf{1}_i b_{0i} + \mathbf{u}_i \qquad (7)$$

where β_0 is an unknown fixed coefficient, b_{0i} is an unknown random coefficient that has a normal distribution with mean 0, and $\mathbf{u}_i$ is the unobservable within-subject random error term for the model (that is, equation 7 represents a model with fixed and random effect intercepts).

The null model could also take the following form:

$$\mathbf{y}_i = \mathbf{1}_i \beta_{00} + \mathbf{u}_{0i} \tag{8}$$

where β_{00} is an unknown fixed coefficient and $\mathbf{u}_{0i}$ is the unobservable within-subject random error term for the model (a model with a fixed effect intercept and no random intercept).

Explained randomness was first introduced by Kent (1983). Xu (2003) defines the randomness of a random variable Y as a monotonic transformation of its entropy, $\exp[-2I(\theta)]$, where $I(\theta) = E[\log p(y;\theta)]$ is the expected log-likelihood. Under the linear mixed model in (1), residual randomness is defined as $D(y_{ij} \mid \mathbf{X},\mathbf{b}) = \exp\{-2E[\log p(y_{ij} \mid \mathbf{X},\mathbf{b})]\}$. The proportion of explained randomness is then given as:

$$\rho^2 = 1 - \frac{D(y_{ij} \mid \mathbf{X},\mathbf{b})}{D(y_{ij} \mid \mathbf{b}_\mathbf{0}^*)},$$

where $D(y_{ij} \mid \mathbf{b}_\mathbf{0}^*)$ is the expected log-likelihood of a null model [such as in (7) and (8)]. For the model in (1) and the null model in (7), it can be shown that an estimator of ρ^2 is:

$$\hat{\rho}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \exp\left(\frac{RSS}{N\hat{\sigma}^2} - \frac{RSS_0}{N\hat{\sigma}_0{}^2} \right) \tag{9}$$

where RSS is the residual sum of squares for the model in (1) and RSS_0 is the residual sum of squares under model (7). The statistic ρ^2 takes values between 0 and 1. In the absence of random effect terms from the model, it can be shown that ρ^2 is equal to the R^2 of traditional linear models.

Xu (2003) conducted a limited simulation to assess the ability of the statistics r^2, ρ^2, and R^2 to estimate the predictability of covariates in the model where predictability is defined in terms of values of $\Omega^2 = 1 - \frac{\mathrm{V}(y_{ij} \,|\, \mathbf{X},\ \mathbf{b})}{\mathrm{V}(y_{ij})}$. He conducted a simulation with 100 replicates for two cases: (a) 50 clusters with 5 observations per cluster and (b) 10 clusters with 25 observations per cluster. Data were simulated for different values of the "strength" of the fixed and random effects terms. For each set of simulated data, values of Ω^2 could be computed exactly. The results of the simulations show that r^2, ρ^2, and R_1^2 tend to give reasonable estimates of Ω^2. For large clusters, the three statistics yield almost similar results. With smaller clusters, ρ^2 and R_1^2 tend to overestimate Ω^2.

It should be noted that there are several limitations to the simulation results proposed by Xu (2003). Besides the small number of replications (100), the simulations did not address the ability of the statistics to discriminate overfitting or the effect of excluding significant covariates from the model. That is, it would be useful to ascertain how the statistics vary when there is overfitting (overestimation of Ω^2 would be expected) or how they vary when important covariates are excluded (underestimation of Ω^2 would be expected provided that the model does not include additional explanatory variables that exhibit a spurious relationship with the outcome). Another issue with the simulations is that they did not include sufficient variation of the fixed effect and random effect terms. Specifically, it would be desirable to determine how well these statistics estimate Ω^2 when (a) the fixed effects account for a small proportion of the variability in the outcome relative to the random effects and (b) the fixed effects account for a large proportion of the variability in the outcome relative to the random effects.

Some of the pseudo-R^2 statistics in this section attempt to quantify "explained residual variation" such as r_c and R_1^2 as opposed to measures of "explained risk" (e.g., ρ^2 or r^2).

The difference between explained residual variation and explained risk was described by Korn and Simon (1991). According to these authors, "explained risk" is "a way of quantifying how much better predictions are when using the covariates compared to when not using them." On the other hand, "explained residual variation" defined as the proportional decrease in residual variation incorporates the "explained risk" and GOF ("applicability of the model to the data").

1.4.2 Marginal versus conditional

Vonesh et al (1996) and Vonesh and Chinchilli (1997) discussed the concept of conditional versus marginal R^2. For r_c and R_1^2, when the computations of the predicted values in the formula of these statistics involve the random effects ($\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i$), they are referred to by Vonesh et al (1996) and Vonesh and Chinchilli (1997) as conditional R^2. On the other hand, when the computation of the predicted values in the formula of these statistics involve only the fixed effect components ($\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}$), these statistics are referred to as marginal R^2. While the concept of conditional versus marginal R^2 was introduced for r_c and r_c, this concept could be applied to other statistics such as r_c and R_2^2. Vonesh et al (1996) and Vonesh and Chinchilli (1997) noted that the marginal version of their statistics might be undervalued in that these statistics don't account for the random effects. For assessing the combined effects of the fixed and random effects, the authors recommend that the conditional version of their statistics be used. In that sense, the conditional R^2 can be seen as an omnibus goodness-of-fit statistic.

1.4.3 Other approaches for GOF for generalized nonlinear models

Various approaches have been proposed in the statistical literature on pseudo-R^2 for generalized linear models or a subclass of generalized linear models, such as logistic, Poisson, and survival models. A good review of pseudo-R^2 measures for logistic regression is given in DeMaris (2002). Similar measures for Poisson models are given in Cameron and Windmeijer (1996), Mittlbock and Waldhor (2000), and Heinzl and Mittlbock (2003). For survival models, pseudo-R^2 measures were proposed by Graf and Schumacher (1995), Schemper and Stare (1996), Xu and O'Quigley (1999), Schemper (2000), Henderson, Jones, and Stare (2001), and O'Quigley and Xu (2001). For GEE models, no equivalent pseudo-R^2 was uncovered in our literature search. However, GOF tests have been proposed by Barnhart and Williamson (1998), Horton et al. (1999), and Pan (2001) for binary outcomes and Pan (2002) for any GEE model.

Many of the approaches described in the previous section for linear mixed models were first proposed either for generalized linear models as a whole or for a subclass of generalized linear models. In particular, the equivalent of D_{rand} for generalized linear models was first proposed by Cameron and Windmeijer (1996) and then by Zheng and Agresti (2000). Similarly, the measure of explained randomness was first proposed for survival models by Kent and O'Quigley (1988) and Xu and O'Quigley (1999).

1.5 Adequacy of the Covariance Structure

1.5.1 Graphical methods

The graphical methods for selecting the covariance matrix of a linear mixed model that have been proposed can be divided into those that can be used as exploratory tools and those that can be used as diagnostic tools. We will restrict our attention in this review to diagnostic tools for determining the adequacy of the covariance matrix once a model has been fitted. For a review of graphical exploratory analysis techniques helpful in selecting a parsimonious covariance structure for fitting the model in (1), the reader is referred to Diggle, Liang, and Zeger (1994), Dawson, Gennings, and Carter (1997), Zimmerman (2000), and Pourahmadi (2002). In addition to exploratory analysis consisting of plotting the observations of each subject versus time, Weiss and Lazaro (1992) proposed plotting the residuals in a similar manner. This is an “omnibus” type of goodness-of-fit similar to the plotting of residuals versus predicted values to ascertain the adequacy of the model and to detect outliers. However, Weiss and Lazaro (1992) recognized that their graphical approach does not address the issue of the adequacy of the covariance structure and that additional graphics are needed.

Grady and Helms (1995) proposed graphical plots that can be used as model selection tools but also as a way to assess the adequacy of the covariance structure. The basic approach derived from their paper would consist of fitting a cell means model to the data with an unstructured covariance structure. The estimated covariance structure would then be plotted to ascertain which covariance structure best fits the data. The plot they suggested consists of plotting actual values of (covariance or correlations) as a function of lag time between measures.” The trend in the graph would then be suggestive of the covariance structure to use in the model. This approach is in essence a model selection tool similar to those cited earlier (Dawson et al., 1997). The approach by Grady and Helms (1995) can also be considered a goodness-of-fit tool, because the authors proposed that the estimated covariance matrix from a fitted model could be compared to the covariance structure of the cell means model by looking visually at the graphical plots. The closeness of the covariance of the fitted model to that of the cell means model would be an indication of a good fit.

1.5.2 Analytical methods

Two analytical approaches were proposed by Vonesh et al. (1996) to assess the adequacy of the covariance structure in generalized nonlinear mixed effect models. One of these tools is an R^2 type statistic similar to r_c denoted the variance-covariance concordance correlation, $r(\hat{\omega})$. This statistic measures the distance, scaled to 1, between the estimated covariance matrix of $\boldsymbol{\beta}$ from the model at hand and the "sandwich" covariance matrix estimator proposed by Liang and Zeger (1986). Vonesh et al. (1996) argued that since the covariance of $\boldsymbol{\beta}$ based on the "sandwich" estimator and the one based on the assumed structure of $\boldsymbol{\Sigma}_i$ (the covariance matrix of $\mathbf{y}_i$) will converge to the same limit if the assumption is correct, it makes sense to compare the goodness of fit of $\hat{\boldsymbol{\Sigma}}_i$ to $\mathrm{V}(\mathbf{y}_i)$ by comparing how close the two estimators of the covariance of $\boldsymbol{\beta}$ are to each other. The other tool proposed by Vonesh et al. (1996) is a pseudo-likelihood ratio test (PLRT) to determine if the estimated covariance matrix is significantly different from the robust covariance estimator.

The statistic for $r(\hat{\omega})$ is given by:

$$r(\hat{\omega}) = 1 - \frac{\| (\hat{\boldsymbol{\omega}} \,\text{-}\, \mathbf{h}) \|^2}{\| \hat{\boldsymbol{\omega}} \|^2 + \| \mathbf{h} \|^2} = \frac{\| (\hat{\boldsymbol{\omega}} \,\text{-}\, \mathbf{h}) \|^2}{\| \hat{\boldsymbol{\omega}} \|^2 + p} \,, \qquad (10)$$

where:

$\widehat{\boldsymbol{\omega}} = \text{vech}\left\{\widehat{\boldsymbol{\Gamma}}^{\text{-1/2}}\widehat{\boldsymbol{\Gamma}}_{\mathbf{R}}\widehat{\boldsymbol{\Gamma}}^{\text{-1/2}}\right\}$. $\widehat{\boldsymbol{\Gamma}}$ is the covariance estimate of $\boldsymbol{\beta}$ under the assumed covariance structure and $\widehat{\boldsymbol{\Gamma}}_{\mathbf{R}}$ is the robust estimator of the covariance of $\boldsymbol{\beta}$, $\mathbf{h} = vech(\mathbf{I}_{\mathbf{p}})$, where p=dim($\boldsymbol{\beta}$).

Vonesh and Chinchilli (1997) noticed that while $r(\widehat{\omega})$ could be useful for detecting gross differences between the assumed covariance structure of $\mathbf{y}_i$ and its true value, “it is less useful to detect moderate but important suggestions.” The authors suggest that the pseudo-likelihood ratio test they have developed be used instead. That test statistic is given by:

$$\widehat{\lambda} = n\left\{\ln|\widehat{\boldsymbol{\Gamma}}| - \ln|\widehat{\boldsymbol{\Gamma}}_{\mathbf{R}}| + \text{trace}\left(\widehat{\boldsymbol{\Gamma}}_{\mathbf{R}}\widehat{\boldsymbol{\Gamma}}^{\text{-1}}\right) - p\right\} \tag{11}$$

Under the null hypothesis of Ho: $\text{var}(\mathbf{y}_i) = \boldsymbol{\Sigma}_i$, the test has an approximate chi-square distribution with $p(p+1)/2$ degrees of freedom. It should be noted that the degrees of freedom are based on an asymptotic likelihood ratio test and is the difference between the number of parameters that would have to be estimated assuming an unstructured covariance structure and the number of parameters assuming a more parsimonious model. So, the degrees of freedom for the test would be expected to be less than $p(p+1)/2$. Results from a limited simulation (only 400 replicates) of this test were provided. The simulation results showed that the pseudo-likelihood ratio test performs well when $\mathbf{y}_i$ has a multivariate normal distribution. The test may not be valid if $\mathbf{y}_i$ follows a moderately (such as a multivariate T distribution) to heavily skewed distribution. Besides the small number of replications for the simulation, the results were based on a 2 × 2 covariance structure, which may affect generalizibility of the results.

1.6 GOF in the GLMM

In developing other GOF tools for linear mixed models, one approach is to review similar GOF in other classes of models. Such a review will help determine if new GOF in the linear mixed models can be patterned after the GOF from other classes of models. GLMM is one of the likeliest candidates because every GLMM can be expressed as a linear mixed model and because of the existence of many GOF statistics for GLMMs.

1.6.1 Relationship between the GLMM and the linear mixed model

Assume a GLMM:

$$\mathbf{Y} = \mathbf{X\Xi} + \mathbf{E} \tag{12}$$

where:

$\mathbf{Y}$ $(N\ x\ p)$ is an array of N random row matrices, { $\mathbf{Y}_\mathbf{i}$ }, each $1\,x\ p$ and mutually independent, $\mathcal{E}\mathbf{Y} = \mathbf{X\Xi}$,

$\mathbf{X}$ is the N x q matrix for the dependent variables which is assumed to be known and fixed,

$\mathbf{\Xi}$ is the q x q matrix of regression coefficients that are unknown, unknowable and fixed,

$V(\mathbf{Y}') = \mathbf{I} \otimes \mathbf{\Sigma}$ (block-diagonal covariance matrix), and

$V(\mathbf{Y}_i') = \mathbf{\Sigma}$.

Note that the model in (12) can be written as a linear mixed model:

$$\left(\mathbf{Y}_i'\right)_{p\,\mathrm{x}\,1} = \begin{bmatrix} \mathbf{X}_i & \mathbf{O}_{1xq} & \ldots & \mathbf{O}_{1xq} \\ \mathbf{O}_{1xq} & \mathbf{X}_i & \ldots & \mathbf{O}_{1xq} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{O}_{1xq} & \mathbf{O}_{1xq} & \ldots & \mathbf{X}_i \end{bmatrix}_{p\,\mathrm{x}\,(pq)} \left[\mathrm{vec}\left(\mathbf{\Xi}\right)\right]_{(pq)\,\mathrm{x}\,1} + \left(\mathbf{E}_\mathbf{i}\right)_{p\,\mathrm{x}\,1}$$

where $\mathbf{X}_i$ corresponds to the i^{th} row of $\mathbf{X}$.

It should be pointed out that the resulting linear mixed model has the following characteristics:

- Every “subject” has the same number of observations; that is, there are no missing data within subjects.
- Observations within “subjects” are correlated but the “subjects” are mutually independent.

Hence, because of the limitations above, while every GLMM can be written as a linear mixed model, the inverse is not necessarily true.

1.6.2 GOF statistics in the GLMM

A review of GOF statistics for the GLMM is provided by Cramer and Nicewander (1979). The authors discussed seven statistics that can be used for assessing GOF. Only four of the seven statistics are commonly used (refer for example to Huberty 1994 or Tatsuoka and Lohnes 1988). These four statistics are Roy’s largest root (RLR), Pillai-Bartley trace (PBT), Hotelling-Lawley trace (HLT), and Wilks (W). To give expressions for these statistics, we first need to define the Wilks lambda criterion (Wilks, 1932) commonly denoted as Λ.

Let $\mathbf{E}_{p\times p}$ be the error sum of squares and cross products (SSCP) matrix, that is, $\mathbf{E}_{p\times p} = (\mathbf{Y} - \widehat{\mathbf{Y}})'_{p\times N}(\mathbf{Y} - \widehat{\mathbf{Y}})_{N\times p}$, and let $\mathbf{H}$ be the hypothesis matrix (in the context of GOF, this is equivalent to testing that there is no effect or no relationship between the outcomes and the predictor variables, so that $\mathbf{H}_{p\times p}$ corresponds to the between sum of squares matrix). Thus, we have $\mathbf{H}_{p\times p} = [\widehat{\mathbf{Y}}'_{p\times N}\widehat{\mathbf{Y}}_{N\times p} - p\,\overline{\mathbf{y}}_{p\times 1}\overline{\mathbf{y}}'_{1\times p}]$, where the i^{th} element of $\overline{\mathbf{y}}$ is the average of the elements of $\mathbf{Y_i}$, the i^{th} row of $\mathbf{Y}$. Let $\mathbf{T = H + E}$ (that is, $\mathbf{T}$ is the total sum of squares and cross products matrix about the mean).

The Wilks lambda criterion is defined as

$$\Lambda = \frac{|\mathbf{H}|}{|\mathbf{T}|}$$

PB is defined as the trace of $\mathbf{HT}^{-1}$, RLR is the largest eigenvalue of $\mathbf{HT}^{-1}$, and HLT is the trace of $\mathbf{HE}^{-1}$. The statistic W takes different forms depending on the author. Cramer and Nicewander (1979) define W as $1-\Lambda^{1/s}$, Huberty (1994) defines W as $1-\Lambda$, and Tatsuoka and Lohnes (1988) use W = Λ. In the remainder of this document, we will use $W = 1-\Lambda^{1/s}$.

Cramer and Nicewander (1979) show that these four statistics are functions of the canonical correlation. Given two set of variables $\mathbf{Y}$ and $\mathbf{X}$, a measure of multivariate association between the two sets of variables can be obtained by considering the linear combination of the $\mathbf{Y}$ variables that has the greatest multiple correlation with the $\mathbf{X}$ variables. It can be shown that the coefficients for the $\mathbf{X}$ variables that maximize the matrix of correlations between the two sets of variables are the right eigenvectors of a matrix $\mathbf{M_x}$ ($\mathbf{M_x}$ is a function of the matrices of correlations among the $\mathbf{X}$ and $\mathbf{Y}$ variables and the matrices of correlations between the two sets of variables). The eigenvalues of that matrix are referred to as the squared canonical correlations. For a more detail exposition on canonical correlations and the GLMM, the reader is referred to Muller (1982) or Gittens (1979).

The eigenvalues of $\mathbf{HT}^{-1}$ are generalizations of squared canonical correlations (Muller and Peterson, 1984). Olson (1976) and Muller and Peterson (1984) showed simple mathematical relationships between the eigenvalues of $\mathbf{HT}^{-1}$, $\mathbf{HE}^{-1}$, and $\mathbf{ET}^{-1}$. That is, all four tests are functions of the canonical correlations. Muller (1982) remarked that with these GOF statistics, one is implicitly measuring the canonical correlations (the strength of the relationship). An equivalent interpretation pointed out by Cramer and Nicewander (1979) that may be more accessible to the lay user is that these statistics measure the strength of the association between the $\mathbf{Y}$ and the $\mathbf{X}$.

Using the eigenvalues of $\mathbf{HT}^{-1}$ which are the generalized canonical correlations, the test statistics are given by

$RLR = \hat{\rho}_1^2$ (i.e., largest eigenvalue or largest squared generalized canonical correlations)

$$PB = \sum_{k=1}^{s} \hat{\rho}_k^2$$

$$HLT = \sum_{k=1}^{s} \hat{\rho}_k^2 / (1 - \hat{\rho}_k^2)$$

$$\hat{W} = 1 - \left(\prod_{k=1}^{s} (1 - \hat{\rho}_k^2) \right)^{1/s}$$

where s = min(p, q)

Muller and Peterson (1984) provide a useful review of the distributions of these GOF statistics. All of these except RLR have an approximate F distribution. Because of its lack of sensitivity and poor power in most instances, most authors suggest that RLR should be avoided (e.g., Olson 1976, Muller 1982, and Tatsuoka and Lohnes 1988). There is no consensus in the statistical literature on which of the other three statistics is to be preferred. Olson recommended the use of PB because of its robustness and power. However, Schatzoff (1966) found in simulations that PB tended to perform poorly and gave preference to either HLT or W. Most authors agree that with a large sample, tests based on the three statistics (PB, HLT, and W) are asymptotically equivalent. Given the lack of consensus, one may want to consider the advice offered by Tatsuoka and Lohnes (1988) to examine the conclusions from the four statistics and, in case they differ, to make the final decision based on the consequences of making a type I versus a type II error.

1.7 Conclusion

In this literature review, I demonstrate that there are few tools that have been validated for assessing goodness-of-fit for linear mixed effect models. Analytical or simulation results on the performance of the statistics or tests that have been proposed to assess goodness-of-fit have been lacking or not presented at all. There is a need to evaluate how well these statistics perform and to characterize any limitations or conditions under which they might not be recommended. If the performance of these tools is less than satisfactory, other tools need to be developed. In developing these tools, one may want to discriminate between tools that are measures of residual variation versus those that are measures of explained risk. The interpretation is different depending on whether the GOF tool measures residual variation or explained risk. It may be that a GOF tool from one class (measures of residual variation versus measures of explained risk) performs better than that from another. A possible approach to developing new tools for GOF in the linear mixed models is to look at similar statistics in other classes of models, such as GLMM.

References

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transaction on Automatic Control*, AC-19, 716-723.

Atkinson, G., and Nevill, A. (1997), "Comment on the Use of Concordance Correlation to Assess the Agreement Between Two Variables," *Biometrics*, 53, 775-777.

Barnhart, H. X., Haber, M., and Song, J. L. (2002), "Overall Concordance Correlation Coefficient for Evaluating Agreement among Multiple Observers," *Biometrics*, 58, 1020-1027.

Barnhart, H. X., and Williamson, J. M. (1998), "Goodness-of-Fit Tests for GEE Modeling with Binary Responses," *Biometrics*, 54, 720-729.

-----(2001), "Modeling Concordance Correlation via GEE to Evaluate Reproducibility," *Biometrics*, 57, 931-940.

Cameron, A. C., and Windmeijer, F. A. G. (1996), "R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization," *Journal of Business & Economic Statistics*, 14, 209-220.

Chinchilli, V. M., Martel, J. K., Kumanyika, S., and Lloyd, T. (1996), "A Weighted Concordance Correlation Coefficient for Repeated Measurement Designs," *Biometrics*, 52, 341-353.

Cramer, E. M., and Nicewander, A. W. (1979), "Some Symmetric, Invariant Measures of Multivariate Association," *Psychometrika*, 44, 43-53.

Dawson, K. S., Gennings, C., and Carter, W. H. (1997), "Two Graphical Techniques Useful in Detecting Correlation Structure in Repeated Measures Data," *American Statistician*, 51, 275-283.

DeMaris, A. (2002), "Explained Variance in Logistic Regression---A Monte Carlo Study of Proposed Measures," *Sociological Methods & Research*, 31, 27-74.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, New York, NY: Oxford University Press.

Gittens, R. (1979), "Ecological Applications of Canonical Analysis," in *Multivariate Methods in Ecological Work*, ed. L. Orloci and C. R. Rao, Fairland: International Co-op Publication.

Grady, J. J. and Helms, R. W. (1995), "Model Selection Techniques for the Covariance-Matrix for Incomplete Longitudinal Data," *Statistics in Medicine*, 14, 1397-1416.

Graf, E. and Schumacher, M. (1995), "An Investigation on Measures of Explained

Variation in Survival Analysis," *Statistician*, 44, 497-507.

Harville D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320-340.

Heinzl, H., and Mittlbock, M. (2003), "Pseudo R-Squared Measures for Poisson Regression Models with Over- or Underdispersion," *Computational Statistics & Data Analysis*, 44, 253-271.

Henderson, R., Jones, M., and Stare, J. (2001), "Accuracy of Point Predictions in Survival Analysis," *Statistics in Medicine*, 20, 3083-3096.

Horton, N. J., Bebchuk, J. D., Jones, C. L., Lipsitz, S. R., Catalano, P. J., Zahner, G. E. P., and Fitzmaurice, G. M. (1999), "Goodness-of-Fit for GEE: An Example with Mental Health Service Utilization," *Statistics in Medicine*, 18, 213-222.

Huberty, C. J. (1994), *Applied Discriminant Analysis*, New York, NY: John Wiley.

Kent, J. T. (1983), "Information Gain and a General Measure of Correlation," *Biometrika*, 70, 163-174.

Kent, J. T., and O'Quigley, J. (1988), "Measure of Dependence for Censored Survival Data," *Biometrika*, 75, 525-534.

King, T. S., and Chinchilli, V. M. (2001), "A Generalized Concordance Correlation Coefficient for Continuous and Categorical Data," *Statistics in Medicine*, 20, 2131-2147.

Korn, E. L., and Simon, R. (1991), "Explained Residual Variation, Explained Risk, and Goodness of Fit," *American Statistician*, 45, 201-206.

Laird, N., Lange, N., and Stram, D. (1987), "Maximum Likelihood Computations with Repeated Measurements: Applications of the EM Algorithm," *Journal of the American Statistical Association*, 97-105.

Laird, N., and Ware, J. H. (1982), "Random Effect Models for Longitudinal Data," *Biometrics*, 38, 963-974.

Liang K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13-22.

Liao, J. J. Z., and Lewis, J. W. (2000), "A Note on Concordance Correlation Coefficient," *Pda Journal of Pharmaceutical Science and Technology*, 54, 23-26.

Lin, L., Hedayat, A. S., Sinha, B., and Yang, M. (2002), "Statistical Methods in Assessing

Agreement: Models, Issues, and Tools," *Journal of the American Statistical Association*, 97, 257-270.

Lin, L. I. (1989), "A Concordance Correlation-Coefficient to Evaluate Reproducibility," *Biometrics*, 45, 255-268.

Lin, L. I. K. (1992), "Assay Validation Using the Concordance Correlation-Coefficient," *Biometrics*, 48, 599-604.

Mittlbock, M. and Waldhor, T. (2000), "Adjustments for R-2-measures for Poisson Regression Models," *Computational Statistics & Data Analysis*, 34, 461-472.

Muller, K. E. (1982), "Understanding Canonical Correlation through the General Linear Model and Principal Components," *The American Statistician*, 36, 342-354.

Muller, K. E., and Peterson, B. L. (1984), "Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis," *Computational Statistics and Data Analysis*, 2, 143-158.

Muller, R., and Buttner, P. (1994), "A Critical Discussion of Intraclass Correlation-Coefficients," *Statistics in Medicine*, 13, 2465-2476.

Nickerson, C. A. E. (1997), "A note on "A concordance correlation coefficient to evaluate reproducibility"," *Biometrics*, 53, 1503-1507.

Olson, C. L. (1976), "On Choosing a Test Statistic in Multivariate Analysis of Variance," *Psychological Bulletin*, 83, 579-586.

O'Quigley, J., and Xu, R. H. (2001), "Explained Variation in Proportional Hazards Regression," in *Handbook of Statistics in Clinical Oncology*, New York, NY: Marcel Dekker.

Pan, W. (2001), "Model selection in estimating equations," *Biometrics*, 57, 529-534.

-----(2001), "Akaike's information criterion in generalized estimating equations," *Biometrics*, 57, 120-125.

Pan, W. (2002), "Application of Conditional Moment Tests to Model Checking for Generalized Linear Models," *Biostatistics*, 3, 267-276.

Pothoff, R. F., and Roy, S. N. (1964), "A generalized multivariate analysis of variance model useful especially for growth curve problems," *Biometrika*, 51, 313-326.

Pourahmadi, M. (2002), "Graphical Diagnostics for Modeling Unstructured Covariance Matrices," *International Statistical Review*, 70, 395-417.

Schatzoff, M. (1966), "Sensitivity Comparisons Among Tests of the General Linear Hypothesis," *Journal of the American Statistical Association*, 61, 415-435.

Schemper, M. (2000), "Predictive Accuracy and Explained Variation in Cox Regression," *Biometrics*, 56, 249-255.

Schemper, M., and Stare, J. (1996), "Explained Variation in Survival Analysis," *Statistics in Medicine*, 15, 1999-2012.

Shwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.

Tatsuoka, M. M., and Lohnes, P. R. (1988), *Multivariate Analysis: Techniques for Educational and Psychological Research*, New York, NY: MacMillian.

Vonesh, E. F., and Chinchilli, V. M. (1997), *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, New York, NY: Marcel Dekker.

Vonesh, E. F., Chinchilli, V. P., and Pu, K. W. (1996), "Goodness-of-Fit in Generalized Nonlinear Mixed-Effects Models," *Biometrics*, 52, 572-587.

Ware, J. H. (1985), "Linear Models for the Analysis of Longitudinal Studies," *The American Statistician*, 39, 95-101.

Weiss, R. E., and Lazaro, C. G. (1992), "Residual Plots for Repeated Measures," *Statistics in Medicine*, 11, 115-124.

Wilks, S. S. (1932), "Certain Generalizations in the Analysis of Variance," *Biometrika*, 24, 241-251.

Xu, R. H. (2003), "Measuring Explained Variation in Linear Mixed Effects Models," *Statistics in Medicine*, 22, 3527-3541.

Xu, R. H., and O'Quigley, J. (1999), "R^2 Type of Measure of Dependence for Proportional Hazards Models," *Statistics in Medicine*, 12, 83-107.

Zheng, B. Y. (2000), "Summarizing the Goodness of Fit of Generalized Linear Models for Longitudinal Data," *Statistics in Medicine*, 19, 1265-1275.

Zheng, B. Y., and Agresti, A. (2000), "Summarizing the Predictive Power of a Generalized Linear Model," *Statistics in Medicine*, 19, 1771-1781.

Zimmerman, D. L. (2000), "Viewing the Correlation Structure of Longitudinal Data through a PRISM," *American Statistician*, 54, 310-318.

2 Fixed Effect Variable Selection in Linear Mixed Models Using R^2 Statistics

Abstract

In the linear mixed model (LMM), several R^2 statistics have been proposed for assessing the goodness-of-fit of fixed effects. However, the performance of these statistics has not been fully demonstrated either analytically or through simulations. We report results of simulations to asses the ability of these statistics to select the most parsimonious model. R^2 statistics from a full model were compared to other models in which fixed effect covariates were removed. The full model was also compared to an overfitted model that included additional covariates not linked to the outcome. All models compared involved the same random effects. In this paper, we show that R^2 statistics that involve the residuals are unable to adequately discriminate between the correct model and one from which important fixed-effects covariates are omitted if the computation of the predicted values for the residuals included the random effects (referred to as conditional R^2 statistics). However, if the random effects are excluded from the computation of the predicted values that lead to the residuals, these R^2 statistics (referred to as marginal R^2 statistics) are able to select the most parsimonious model. Other R^2 statistics that have been proposed by Xu [25] performed poorly in that there was little variation in the value of these statistics from a full model to a reduced model.

2.1 Introduction

Few diagnostic tools are available for assessing the adequacy of linear mixed models (LMMs). Three statistics that are often used and are available in most statistical software are the Akaike's information criterion (AIC), the Bayesian information criterion (BIC), and the likelihood ratio test (LRT). Unlike the R^2 of traditional linear regression, these statistics cannot be used to ascertain the extent to which the proposed model can explain variation in the outcome. Their use is limited to the comparison of several models fitted to the same data. In the case of the LRT, the models to be compared must be nested, whereas for the AIC or the BIC, it is not clear what constitutes a significant difference. Also, for comparing models with different fixed effects the use of AIC or BIC may be inappropriate when restricted maximum likelihood (REML) has been used for estimation (Verbeke and Molenbergh, 2000) [20]. Similarly, Whelham and Thompson [24] noted that the log-likelihood ratio test may not be valid under REML for comparing 2 models with different fixed effect terms. Hence, statistics similar to the R^2 of traditional linear regression are needed to answer questions such as (a) how much better is it to use the model at hand compared to another model, and (b) how much of the variation in the outcome can be explained by the model at hand or by a subset of the covariates.

Kvalseth [8] proposed eight criteria for evaluating R^2 statistics:

1. R^2 should have reasonable interpretation and utility as a GOF measure.
2. R^2 should be independent of the units of measurement.
3. The potential range of values should be well defined with endpoints corresponding to perfect and complete lack of fit.

4. R^2 should be sufficiently general to be applicable to any type of model, whether the covariates are random or nonrandom and regardless of the statistical properties of the model.
5. R^2 should not be confined to any specific model-fitting technique.
6. R^2 should be such that values for different models fitted to the same data set are directly comparable.
7. Relative values of R^2 ought to be generally compatible with those derived from other acceptable measures of fit.
8. Positive and negative residuals should be weighted equally.

Cameron and Windmeijer [2] proposed four more criteria:

- R^2 does not decrease as regressors are added (without degree-of-freedom correction)
- R^2 based on residual sum of squares coincides with R^2 based on explained sum of squares
- There is a correspondence between R^2 and a significance test on all slope parameters and between changes in R^2 as regressors are added and significance tests
- R^2 has an interpretation in terms of information content of the data

Recently, several R^2 statistics having interpretation and properties similar to the traditional R^2 have been proposed for assessing the goodness-of-fit (GOF) of fixed effect covariates in the LMM. The purpose of this paper is to evaluate the performance of R^2 statistics for the LMM in selecting the most parsimonious model. That is, we are focusing primarily on the ability of these statistics to discriminate between a fully specified model and one from which important fixed-effects covariates are missing. We believe that a desired property of an R^2 is that it should decrease in value when important covariates are removed from the model. The decrease in value should be proportional to how much the variation in the outcome depends on or can be explained by the variables that have been removed. In comparing two models, one must take into account the fact that the LMM consists of two sub-models: the fixed-effect covariates and random-effect covariates. Because the impact of the misspecification of random-effect covariates on fixed effects is not clear, we are restricting this evaluation to cases where the models to be compared have the same random-effect covariates and the same covariance structure. This is not a major limitation because in most cases the analyst is primarily interested in assessing the effects of the fixed-effect terms.

We note, however, that if focus is on the covariance, greater effort must be exerted in achieving a good model for the covariance. To compare covariance structures, it is usually assumed that the mean structure has been correctly specified. Covariance model selection techniques that require the assumption include the LRT (Jennrich and Schluchter [6]; Schaalje et al. [16]; Grady and Helms [4]), information criteria (AIC and BIC), and predictive approaches such as PRESS [12]). As a first step to assessing the performance of R^2 statistics, we focus our attention on fixed effects. As a result, the performance of R^2 statistics for changing covariance structures is beyond the scope of this paper. However, we consider the topic as an active area of future research.

There are many applications such as in clinical trials or epidemiological studies where repeated measurements are taken on a subject and the interest of the researcher is to ascertain the effect over time of the explanatory variables (such as treatment and individual patient characteristics that may impact treatment) on the outcome. R^2 statistics discussed in this paper could be used to aid in selecting the most parsimonious model, particularly when an effect such as an interaction term may be statistically significant but in actuality does not contribute much in explaining variance of the outcome relative to other variables in the model.

For example, Potthoff and Roy [15] presented data for an orthodontic study that involved 27 children, 16 boys and 11 girls. For each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was measured at ages 8, 10, 12, and 14 years with complete data for each child. The objectives of the study were to determine whether, on the average over time (age), distances are larger for boys than for girls and whether, on the average over time, the rate of change of the distance is similar for boys and girls (see Section VI). For this study, using the linear mixed model we find that the effect of the age-by-gender interaction is statistically significant ($p < 0.03$). The interpretation of a significant age-by-gender interaction is that the rate of change of the distance with respect to age is statistically different between boys and girls. However, as shown in Table 5, the interaction term's proportionate reduction in residual variance is practically negligible. As a result, though the rate of change of the distance between boys and girls is statistically different, the effect contributes so little to explaining the proportionate reduction in residual variance that the interaction effect can possibly be excluded. In this example, excluding the interaction effect based on R^2 clearly has an impact on the results of the study.

After defining and giving notations for the LMM in Section 2.2, we review the R^2 statistics and give formulas for them in Section 2.3. We describe data generation techniques for our simulation in Section 2.4. Results from our simulation are presented in Section 2.5. An example of the use of the statistics evaluated are given in Section 2.6. A discussion of these results follows in Section 2.7. We end the paper with concluding remarks in Section 2.8.

2.2 The Linear Mixed Model

Assume the following linear mixed model (Harville [5]; Laird and Ware [9]):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i\,, \qquad (1)$$

where $i \in$ {1, 2, ..., n} is the index for the independent sampling units (ISU),

$\mathbf{y}_i$ is an $n_i \times 1$ vector of observations from the i^{th} independent sampling unit (subject),

$\mathbf{X}_i$ denotes an $n_i \times p$ fixed-effects design matrix for the i^{th} subject,

$\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown, constant, fixed-effect parameters,

$\mathbf{Z}_i$ denotes an $n_i \times q$ random-effects design matrix for the i^{th} subject,

$\mathbf{b}_i$ is a $q \times 1$ vector of unobservable random effects for the i^{th} subject, and

$\mathbf{e}_i$ denotes an $n_i \times 1$ vector of unobservable within-subject error terms.

It is also assumed that $\mathbf{b}_i$ has a multivariate normal distribution $N_q(\mathbf{0}, \mathbf{G})$ independent of $\mathbf{e}_i$, which has a multivariate distribution $N_{n_i}(\mathbf{0}, \mathbf{R}_i)$.

$$E\begin{bmatrix}\mathbf{b}_i \\ \mathbf{e}_i\end{bmatrix} = \begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix} \text{ and } V\begin{bmatrix}\mathbf{b}_i \\ \mathbf{e}_i\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_i\end{bmatrix},$$

where $\mathbf{G}$ is a $q \times q$ unknown covariance matrix for the random effects and $\mathbf{R}_i$ is an $n_i \times n_i$ unknown covariance matrix for the within-subject error terms for the i^{th} subject. With these assumptions, for the i^{th} subject we have $\boldsymbol{\Sigma}_i = V(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$. In many applications,

$\mathbf{R}_i$ is taken to be $\sigma^2\mathbf{I}_{n_i}$, known as the conditional independence assumption for the error term [10].

By stacking the vectors of responses and associated matrices, the mixed model can also be expressed as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zb} + \mathbf{e},$$

where $\mathbf{y} = (\mathbf{y}'_1, \mathbf{y}'_2, . . ., \mathbf{y}'_n)'$ is $N \times 1$,

$N = \sum_{i=1}^{n} n_i$, $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2, . . ., \mathbf{X}_n)'$ is $N \times p$, $\mathbf{Z} = Diag(\mathbf{Z}_1, \mathbf{Z}_2, . . ., \mathbf{Z}_n)$ is $N \times nq$,

$\mathbf{b} = (\mathbf{b}'_1, \mathbf{b}'_2, . . ., \mathbf{b}'_n)'$ is $nq \times 1$, and $\mathbf{e} = (\mathbf{e}'_1, \mathbf{e}'_2, . . ., \mathbf{e}'_n)'$ is $N \times 1$. The distributional assumptions are that $\mathbf{b} \sim N_{nq}(\mathbf{0}, \mathbf{G} \otimes \mathbf{I}_n)$ independent of $\mathbf{e} \sim N_N(\mathbf{0}, \mathbf{R})$,

$\mathbf{R} = Diag(\mathbf{R}_1, \mathbf{R}_2, . . ., \mathbf{R}_n)$ is $N \times N$. Also, $\mathbf{\Sigma} = V(\mathbf{y}) = Diag(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, ..., \mathbf{\Sigma}_n)$ is $N \times N$.

A brief overview of approaches to parameter estimation for the model in (1) is given by Ware [23]. The use of maximum likelihood (ML) and restricted maximum likelihood (REML) approaches for linear mixed models was first discussed by Harville [5]. Laird and Ware [10] proposed a Bayesian approach to estimation and the use of the expectation-maximum (EM) algorithm for both the Bayesian approach and the ML approach. Detailed formulas for computing ML and REML estimates using the EM algorithm with suggestions on how to speed convergence are given in Laird, Lange, and Stram [9].

2.3 Proposed R^2 Statistics in the LMM

Vonesh, Chinchilli, and Pu [22] proposed that an unweighted concordance correlation coefficient (CCC), denoted r_c, be used to assess goodness-of-fit for generalized nonlinear mixed-effect models. For these models, they formulate the CCC as

$$r_c = 1 - \frac{\sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\left(\sum_{i=1}^{n} (\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i}) \right) + \left(\sum_{i=1}^{n} (\hat{\mathbf{y}}_i - \hat{y}\mathbf{1}_{n_i})'(\hat{\mathbf{y}}_{\mathbf{i}} - \hat{y}\mathbf{1}_{n_i}) \right) + N(\bar{y} - \hat{y})^2}, \qquad (2)$$

where *n* is the number of independent sampling units (or subjects),

n_i is the number of observations for subject i

$\mathbf{y}_i$ is the vector of observed values for the *i*th subject,

$\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i$ is the vector of predicted values for the *i*th subject,

$\hat{y}$ is the grand average of the predicted values,

$\bar{y}$ is the grand average of the observed values,

N is the total number of observations, and

$\mathbf{1}_{n_i}$ is an n_i x 1 vector of 1's.

It should be noted that the CCC was first introduced by Lin [11] as a way to evaluate reproducibility between two sets of measurements, as in the case where there is a "gold standard" assay or instrumentation and the intent is to measure whether a new assay can reproduce the results from the gold standard assay or instrumentation. If the new assay is successful, then the plot of the new assay's results versus that of the gold standard should fall along the 45 degree or equality line. Hence, an interpretation for r_c is that it measures the degree of agreement between the observed and estimated values.

Besides the r_c, Vonesh and Chinchilli [21] proposed a statistic that we denote R_1^2 for assessing GOF in a generalized linear mixed model. Assuming $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$ then

$$R_1^2 = 1 - \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})} \tag{3}$$

This statistic is the counterpart to the traditional R^2 in linear models and as such lends itself to ease of interpretation. However, R_1^2 does not explicitly take into account the random components of the model.

In addition to the CCC and R_1^2, Zheng [26] proposed P_{rand}, the proportional reduction in penalized quasi-likelihood (PQL).

$$P_{rand} = 1 - \frac{\sum_{i=1}^{n} d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i)/(2\hat{\sigma}) + \hat{\mathbf{b}}'(\hat{\mathbf{G}} \otimes \mathbf{I}_n)^{-1}\hat{\mathbf{b}}/2}{\sum_{i=1}^{n} d_i(\mathbf{y}_i, \bar{y}\mathbf{1}_{n_i})/(2\hat{\sigma})}, \tag{4}$$

Where $d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i) = \sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)$ is the deviance (McCullagh and Nelder [13]),

$\sum_{i=1}^{n} d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i)/(2\hat{\sigma}) + \hat{\mathbf{b}}'(\hat{\mathbf{G}} \otimes \mathbf{I}_n)^{-1}\hat{\mathbf{b}}/2$ is defined as the negative of the PQL,

$\hat{\mathbf{b}}$ is the estimated vector of random-effect parameters for all subjects, and

$\hat{\mathbf{G}}$ is the estimated covariance for the random-effect parameters. Zheng [26] also showed that P_{rand} can be expressed as

$$P_{rand} = 1 - \frac{(1/2\hat{\sigma})\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i) + \hat{\mathbf{b}}'(\hat{\mathbf{G}} \otimes \mathbf{I}_n)^{-1}\hat{\mathbf{b}}/2}{(1/2\hat{\sigma})\sum_{i=1}^{n}(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})} \tag{5}$$

Zheng [26] interprets PQL as a measure of the proportional reduction in the log-likelihood from the model at hand compared to a null model that consists of only a fixed-effect intercept. It takes values between 0 and 1, with values close to 0 indicating a lack of fit or large random effect and a value of 1 indicating a perfect fit or small random effect. Using the expression in (5), PQL can be seen as an attempt similar to the statistics AIC [2] and BIC [17] to account for additional covariates in the model with a "penalty" term. However, the desirability of a penalty term over others needs to be demonstrated.

Xu [25] proposed three statistics for explaining the variations in a linear model: two statistics that measure the proportion of explained variation (which we denote Ω^2 and R_2^2) and another, denoted ρ^2, that measures the proportion of explained randomness. The statistic $\hat{\Omega}^2$ as proposed by Xu [25] is given by

$\hat{\Omega}^2 = 1 - \dfrac{\hat{\sigma}^2}{\hat{\sigma}_0{}^2}$ and is meant to estimate $\Omega^2 = 1 - \dfrac{\mathrm{V}(y_{ij} \mid \mathbf{X},\ \mathbf{b})}{\mathrm{V}(y_{ij} \text{ under a null model})}$,

where $j \in \{1, 2, \ldots, n_i\}$ so that y_{ij} is the j^{th} element of $\mathbf{y}_i$, $\hat{\sigma}^2$ is the estimate of σ^2 for the model at hand, and $\hat{\sigma}_0{}^2$ is the estimator for the residual variance of a "null" model. The null model could take the following form:

$$\mathbf{y}_i = \mathbf{1}_{n_i} \beta_0 + \mathbf{1}_{n_i} b_{0i} + \mathbf{u}_i, \qquad (6)$$

where β_0 is an unknown fixed parameter,

b_{0i} is an unknown random coefficient that has a normal distribution with mean 0, and

$\mathbf{u}_i$ is the unobservable within-subject random error term for the model (i.e., equation 6 represents a model with fixed- and random-effect intercepts). We define $V(\mathbf{u}_i) = \sigma_0{}^2 \mathbf{I}_{n_i}$ and $V(b_{0i}) = \tau_{00}^2$.

The null model could also take the following form:

$$\mathbf{y}_i = \mathbf{1}_{n_i}\beta_{00} + \mathbf{u}_{0i}, \tag{7}$$

where β_{00} is an unknown fixed coefficient and $\mathbf{u}_{0i}$ is the unobservable within-subject random-error term for the model (a model with a fixed-effect intercept and no random effects).

The second statistic suggested by Xu [25], R_2^2, is given by

$$R_2^2 = 1 - \left(\frac{RSS}{RSS_0}\right), \tag{8}$$

where RSS is the residual sum of squares for the model in (1) and RSS_0 is the residual sum of squares under the model in (6).

Explained randomness was first introduced by Kent [7]. Xu [25] defines the randomness of a random variable Y as a monotonic transformation of its entropy, $\exp[-2I(\theta)]$, where $I(\theta) = E[\log p(y;\theta)]$ is the expected log-likelihood. Under the linear mixed model in (1), residual randomness is defined as $D(y_{ij} \mid \mathbf{X},\mathbf{b}) = \exp\{-2E[\log p(y_{ij} \mid \mathbf{X},\mathbf{b})]\}$. The proportion of explained randomness is then given as

$$\rho^2 = 1 - \frac{D(y_{ij} \mid \mathbf{X},\mathbf{b})}{D(y_{ij} \mid \mathbf{b}_0^*)},$$

where $D(y_{ij} \mid \mathbf{b}_0^*)$ is the expected log-likelihood of a null model [such as in (6) and (7)]. For the model in (1) and the null model in (6), Xu [25] defines the estimator of ρ^2 as

$$\hat{\rho}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\exp\left(\frac{RSS}{N\hat{\sigma}^2} - \frac{RSS_0}{N\hat{\sigma}_0{}^2}\right), \tag{9}$$

where RSS is the residual sum of squares for the model in (1) and RSS_0 is the residual sum of squares under model (6). The statistic ρ^2 takes values between 0 and 1. In the absence of random effect terms from the model, it can be shown that ρ^2 is equal to the R^2 of traditional linear models. When the null model in (7) as opposed to the one in (6) is used, we will denote Ω^2, R_2^2 and ρ^2, respectively, by Ω_0^2, $R_{0_2}^2$ and ρ_0^2, in which case $\mathrm{R}_{0_2}^2$ is the same as the R_1^2 of Vonesh and Chinchilli [21]. Throughout the remainder of this paper, we will assume null model (6) as opposed to null model (7). The rationale for using model (6) is that in the absence of any covariates in the context of the LMM, one would prefer it over model (7) to reduce the data.

Xu [25] conducted a limited simulation to assess the ability of the statistics $\hat{\Omega}^2$, R_2^2, ρ^2, $\hat{\Omega}_0^2$, R_1^2 (again same as $R_{0_2}^2$) and ρ_0^2 to estimate the predictability of covariates in the model where he defined predictability in terms of values of

$\Omega^2 = 1 - \dfrac{\sigma^2}{\mathrm{V}(y_{ij} \text{ under a null model})}$. Xu [25] conducted a simulation with 100 replicates for

two cases: (a) 50 clusters with 5 observations per cluster and (b) 10 clusters with 25 observations per cluster. Data were simulated for different values of the “strength” of the fixed- and random-effects terms. For each set of simulated data, values of Ω^2 could be computed exactly. The results of the simulations show that r^2, $\hat{\rho}^2$, and R_1^2 tend to give reasonable estimates of Ω^2. For large clusters, the three statistics yield almost similar results. With smaller clusters, $\hat{\rho}^2$ and R_1^2 tend to overestimate Ω^2.

It should be noted that there are several limitations to the simulation results proposed by Xu [25]. Besides the small number of replications (100), the simulations did not address the ability of the statistics to discriminate overfitting or the effect of excluding significant covariates from the model. That is, it would be useful to ascertain how the statistics vary when there is overfitting or how they vary when important covariates are excluded.

2.3.1 Conditional versus marginal R^2 Statistics

Vonesh et al [22] and Vonesh and Chinchilli [21] introduced the concept of conditional versus marginal R^2. For r_c and R_1^2, when the computations of the predicted values in the formula of these statistics involve the random effects ($\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i$), they are referred to by Vonesh et al [22] and Vonesh and Chinchilli [21] as conditional R^2. On the other hand, when the computation of the predicted values in the formula of these statistics involve only the fixed effect components ($\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}$), these statistics are referred to as marginal R^2. While the concept of conditional versus marginal R^2 was introduced for r_c and R_1^2, this concept could be applied to other statistics such as P_{rand} and R_2^2 introduced respectively by Zheng [26] and Xu [25]. It can be shown that if only the fixed effects are used in the computations of SSR and SSR_0 for R_2^2, the same results as the marginal R_1^2 would be obtained. Note that Vonesh et al [22] and Vonesh and Chinchilli [21] recommend that the conditional version of their statistics (r_c and R_1^2) be used because they can account for the combined effects of both fixed and random effects. In that sense, the conditional versions of r_c and R_1^2 can be viewed as overall GOF statistics that can give a measure of the adequacy of both the random and fixed effects. On the other hand, the marginal version should be seen as measuring only the effects of the fixed effects.

Table 2.1 summarizes the statistics that are reviewed in this paper.

2.4 Data Generation Techniques for Simulation Study

The simulation study is based on theoretical rather than actual data in order to vary the range of parameters to assess how each of the statistics performs under various conditions. We simulated a longitudinal continuous outcome with three fixed-effect terms: a time covariate (age) and two dichotomous variables (gender and treatment). The random-effect covariates consisted of a random intercept term and a linear term for age. We generated data assuming that the covariance of the random effects was unstructured. The within-subject error covariance was assumed to be $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$ in all simulated data sets. A total of three data sets were generated. The data sets differed in their values of within-subject correlation (we wanted the correlation within subject to be respectively around 0.1, 0.5, and 0.8). This was accomplished by changing the values of σ^2 (12, 45, and 250, respectively). In all three data sets, there were 64 subjects and 6 observations per subject. We give below the values of other parameters used in the simulation:

$\mathbf{X} = [\mathbf{1}_6, (5,\ 6,\ 7,\ 7.25,\ 7.5,\ 7.75)', \mathbf{1}_6 I_1, \mathbf{1}_6 I_2]$; $\mathbf{Z} = [\mathbf{1}_6, (5, 6, 7, 7.25, 7.5, 7.75)']$ where $\mathbf{1}_6$ is a 6×1 vector of ones, I_1 and I_2 are indicator function taking values 1 or 0. We took $\boldsymbol{\beta} = \begin{pmatrix} 10 \\ 6 \\ 11 \\ 11 \end{pmatrix}$,

and $\mathbf{G} = \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}$.

Parameter values for two of the fixed effects (gender and treatment) were chosen to be the same so that a lack of change in the value of the GOF statistics from the removal of one of these variables would not be attributed to the fact that the variable being removed accounted for little in the variation of the outcome (compared to other variables). For the covariance matrix of the random effects, parameter values were chosen so that the correlation between the random intercept and age would be 0.5. With these parameter values, we achieved values of Ω^2 between 0.14 and 0.77. Ω^2 was computed using formula from Xu [25] as $\Omega^2 = 1 - \frac{\sigma^2}{(\beta_{\text{age}}^2 + \tau_1^2)\,\text{var}(age) + \sigma^2}$.

2.5 Results of the Simulation

Linear mixed models were fitted for each of the 10,000 samples within each of the three simulated data sets described in section IV. The models consisted of a full model with age, gender, and treatment as fixed effects, an overfitted model with 2 extraneous variables not related to the outcome, a reduced model with the variable for treatment removed (reduced model 1), and a second reduced model with both variables for gender and treatment removed (reduced model 2). For all models, the random-effect covariates were the same, consisting of an intercept and age. An unstructured covariance structure was assumed for the random effects (same as the simulated data). The within-subject error variance was assumed to be fixed and constant ($\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$ for each subject). The data were generated and analyzed with SAS v9.1 (SAS Institute, Cary, North Carolina) using restricted maximum likelihood (REML) estimation. For each sample within each data set, the R^2 statistics described in Section 2.3 were computed. Samples in which the Hessian matrix or the covariance matrix of the random effects were not positive definite were removed.

Tables 2.2 gives the mean, minimum, and maximum values obtained in each of the three data sets for conditional R^2 and the R^2 statistics proposed by Xu [25]. The range for all the statistics are between 0 and 1, although when the Hessian matrix or the covariance matrix of the random effects was not positive definite, negative results (not shown) were obtained. Values of the statistics proposed by Xu [25] were close to values of Ω^2 as theorized by the author. For all the R^2 statistics, no noticeable change was observed between the overfitted model and the full model (a desirable property). Unfortunately, in comparing the full to the reduced models, no noticeable change was observed in the mean value of the R^2 statistics. In fact, it appears that each of the R^2 statistics in Table 2.2 remained constant from the full to the reduced models.

In Table 2.3, we computed marginal R^2 for each data set. No change was noticed between the overfitted models and the full models. Unlike the conditional R^2, there was a noticeable decrease in the value of the marginal R^2 statistic when important covariates are removed from the model with the size of the decrease being larger when two variables are removed as opposed to one. Lower values of the statistics were obtained with higher within subject correlation. Hence, a low value of the statistic can be due to a misspecified model or high within subject correlation.

To understand better the values in Table 2.2 (particularly why the conditional R^2 statistics are unable to discriminate between the "true" model and a misspecified one), we computed intermediate results (the numerator and denominator) that formed these statistics (we refer to misspecification of the fixed effects here as omitting cross-sectional or baseline variables). We computed the average sum of squares of the residuals and average of $\hat{\sigma}^2$ from each of the models in Table 2.2. For all three data sets, in the results (not shown) little difference was observed between the full, the overfitted and reduced models in the average sum of squares of the residuals or average values of $\hat{\sigma}^2$. All the conditional R^2 statistics and the statistics proposed by Xu [25] take the form of $\left(1-\frac{\text{Numerator}}{\text{Denominator}}\right)$, with the denominator being constant for a given data set. As we have shown that on average the numerator does not change when the fixed effects are misspecified, this explains why in Tables 2.2, there is little change from a full model to a reduced model in the values of the conditional R^2 statistics and the R^2 statistics proposed by Xu [25]. Other results from our simulation that might be unexpected are worth reporting. We investigated the fixed-parameter estimates when the fixed effects are misspecified. Values of the fixed-parameter estimates, except for the intercept, appear to be robust to misspecification of the fixed effects for these simulations.

2.6 Data Example

We analyzed data on dental fissures first reported by Pothoff and Roy [15]. The data consist of measurements of continuous outcome (the distance in millimeters between the pituary and the pterygomaxillary fissure) measured at ages 8, 10, 12, and 14 on 11 boys and 16 girls. These data have been analyzed in the context of the LMM by various authors, including Zheng [26] who used it to compute CCC and P_{rand}.

The data were analyzed with PROC MIXED in the SAS System using restricted maximum likelihood (REML). Several nested models were fitted by removing fixed effect terms. The full model consisted of age, gender, and age-by-gender interactions as fixed effects. The random effects were an intercept term and age. Table 2.4 gives our results for conditional R^2 and the statistics proposed by Xu [25]. Our results for r_c, R_1^2 and P_{rand} are the same as those published by Zheng [26]. Note that there is little variation from one reduced model to the next for all of the GOF statistics in Table 2.4. In Table 2.5, we give values of marginal R^2 for the data. The marginal R^2 statistics show that the model with age and gender is better than the model with age alone. Also, the values of the marginal R^2 statistics are much lower (less than 0.6 for marginal r_c and less than 0.5 for marginal R_1^2 and P_{rand}) compared to the conditional R^2 statistics suggesting that the fit of the model may be inadequate.

2.7 Discussion

These results show that the most common forms of R^2 statistics that have been proposed as GOF measures in the LMM do not perform adequately because they are unable to discriminate when important covariates are missing from the model for the examples considered. The simulations also explain why these statistics are not appropriate measures of GOF. For the conditional version of r_c, it is inappropriate to use a measure of agreement between observed and predicted values because in the LMM these predicted values are robust to misspecification of the fixed-effect function (cross-sectional or baseline variables omitted). The robustness of the predicted values also explains why other conditional R^2 based on them, such as P_{rand} and R_1^2, are inappropriate. As for $\hat{\Omega}^2$ that uses estimates of σ^2 in its numerator, we have shown that it is also inappropriate because of the robustness of $\hat{\sigma}^2$ to misspecification of the fixed-effect function. Since σ^2 is the within-subject variability—a population parameter that should have a fixed value—even if it were to change from one model to the other, this would only indicate bias in the estimate.

We obtained satisfactory results for the marginal counterparts of 3 of the statistics that we have reviewed (r_c, P_{rand} and R_1^2) in that a) they were able to differentiate between the full model and one in which important covariates were removed and b) the values of these statistics showed little change between the full model and an overfitted one. Hence, these statistics could be used in selecting the most parsimonious model for the fixed effect covariates. Because there was almost no difference in the values of the statistics P_{rand} and R_1^2, we are recommending that the analyst considers computing only one of these statistics. In choosing between P_{rand} and R_1^2 our preference would be for R_1^2 because it is easier to compute and interpret in that it is a straightforward extension of the traditional R^2. Also, because the only difference between P_{rand} and R_1^2 is in a penalty term to correct for additional variables in the model, one may question the adequacy of that penalty term. It should also be pointed out that low values of the marginal statistics might be an indication that there is high within or between subject variability in the data. Hence, in the case where low values are obtained the analyst should consider comparing the values of the within subject variance, between subject variance or a combination of the two to the overall variance. Large within subject or between subject variability has serious implications for users of the data. This indicates that even if additional variables (related to the outcome) are included in the model, the values of the R^2 statistics would not increase substantially.

A major result of our simulation is supported by Verbeke and Fieuws [18] who found that estimates of σ^2 were robust to misspecification of cross-sectional or baseline fixed effects. This result explains why the R^2 proposed by Xu [25] do not perform adequately in determining the most parsimonious model (although they appear to be estimating a population parameter). Another result of our simulation, the fact that parameter estimate for a covariate is robust to misspecification of the fixed effects for well-defined models was confirmed by Verbeke et al. [19] and Verbeke and Fieuws [18]. Hence, while a limitation of our results is that they are based on simulations, key findings of these simulations have been demonstrated analytically or confirmed by other independent simulations.

2.8 Conclusion

We have shown through simulations that conditional R^2 and similar statistics proposed by Xu [25] are inadequate in comparing two linear mixed models with the same random effects but different fixed effects. The inadequacy of these R^2 statistics revealed by our simulations put into question their usefulness as a GOF tool for any mixed model. Consequently, we suggest that they should not be used in assessing GOF in the LMM. On the other hand, marginal R^2 statistics were useful in identifying the most parsimonious model. However, it is unclear that marginal R^2 statistics will be useful if the random effects are misspecified. Future studies should investigate the development of other R^2 statistics that can select the most parsimonious model.

References

1. Akaike, H., 1974. A new look at the statistical model identification. IEEE Transaction on Automatic Control, AC-19, 716-723.

2. Cameron, A.C. and Windmeijer, F.A.G., 1996. R-Squared measures for count data regression models with applications to health-care utilization. Journal of Business & Economic Statistics, 14(2) 209-220.

3. Cnaan, A., Laird, N.M., and Slasor, P., 1997. Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. Statistics in Medicine, 16(20) 2349-2380.

4. Grady, J. J., and Helms, R. W. (1995). Model Selection Techniques For The Covariance Matrix For Incomplete Longitudinal Data. Statistics in Medicine, 14, 1397-1416.

5. Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72(358) 320-340.

6. Jennrich, R. I., and Schluchter, M. D. (1986). Unbalanced Repeated-Measures Models With Structured Covariance Matrices. Biometrics, 42, 805-820.

7. Kent, J.T., 1983. Information gain and a general measure of correlation. Biometrika, 70(1) 163-174.

8. Kvalseth, T.O., 1985. Cautionary note about R^2. The American Statistician, 39(4) 279-285.

9. Laird, N., Lange, N., and Stram, D., 1987. Maximum likelihood computations with repeated measurements: applications of the EM algorithm. Journal of the American Statistical Association, 82(397) 97-105.

10. Laird, N. and Ware, J.H., 1982. Random effect models for longitudinal data. Biometrics, 38(4) 963-974.

11. Lin, L.I., 1989. A concordance correlation-coefficient to evaluate reproducibility. Biometrics, 45(1) 255-268.

12. Liu, H., Weiss, R. E., Jennrich, R. I., and Wenger, N. S. (1999). PRESS Model Selection in Repeated Measures Data. Computational Statistics and Data Analysis, 30, 169-184.

13. McCullagh P. and Nelder J.A. (1989). Generalized Linear Models. 1989, CRC Press, Boca Raton, FL, 1989, 33-34

14. Morrell, C.H., Pearson, J.D., and Brant, L.J., 1997. Linear transformations of linear mixed-effects models. The American Statistician, 51(4) 338-343.

15. Pothoff, R.F. and Roy, S.N., 1964. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51 313-326.

16. Schaalje, B., Zhang, J., Pantula, S. G., and Pollock, K. H. (1991). Analysis of Repeated-Measurements Data From Randomized Block Experiments. Biometrics, 47, 813-824.

17. Shwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics, 6(2) 461-464.

18. Verbeke G. and Fieuws S. The effect of miss-specified baseline characteristics on inference for longitudinal trends in linear mixed models. Biostatistics, advanced publication on line March 23, 2007.

19. Verbeke, G. et al., 2006. A comparison of procedures to correct for base-line differences in the analysis of continuous longitudinal data: a case-study. Journal of the Royal Statistical Society, Series C Applied Statistics, 55(1) 93-102.

20. Verbeke, G. and Molenberghs, G., 2000. Linear Mixed Models For Longitudinal Data. Springer-Verlag, New York.

21. Vonesh, E.F., Chinchilli, V.M., 1997. Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker, New York, 1997, 419-424.

22. Vonesh, E.F., Chinchilli, V.M., Pu, K.W., 1996. Goodness-of-fit in generalized nonlinear mixed-effects models. Biometrics, 52 572-587.

23. Ware, J.H., 1985. Linear models for the analysis of longitudinal studies. The American Statistician, 39(2) 95-101.

24. Welham, SJ, Thompson, R. A Likelihood Ratio Test for Fixed Model Terms Using Residual Maximum Likelihood. Journal of the Royal Statistical Society, Series B. 1997; 59 (3) 701-714

25. Xu, R.H., 2003. Measuring explained variation in linear mixed effects models. Statistics in Medicine, 22(22) 3527-3541.

26. Zheng, B.Y., 2000. Summarizing the goodness of fit of generalized linear models for longitudinal data. Statistics in Medicine, 19(10) 1265-1275.

Table 2.1 Summary of statistics reviewed

Statistic	Formula	Author(s)
r_c **(also CCC)**	$r_c = 1 - \dfrac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\left(\sum_{i=1}^{n}(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})\right) + \left(\sum_{i=1}^{n}(\hat{\mathbf{y}}_i - \hat{y}\mathbf{1}_{n_i})'(\hat{\mathbf{y}}_{\mathbf{i}} - \hat{y}\mathbf{1}_{n_i})\right) + N(\bar{y} - \hat{y})^2}$ (1)	**Vonesh et al. [22]**
R_1^2	$R_1^2 = 1 - \dfrac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})}$ (1)	**Vonesh and Chinchilli [21]**
P_{rand}	$P_{rand} = 1 - \dfrac{\sum_{i=1}^{n} d_i(\mathbf{y}_i, \hat{\mathbf{y}}_i)/(2\hat{\sigma}) + \hat{\mathbf{b}}'(\hat{\mathbf{G}} \otimes \mathbf{I}_n)^{-1}\hat{\mathbf{b}}/2}{\sum_{i=1}^{n} d_i(\mathbf{y}_i, \bar{y}\mathbf{1}_{n_i})/(2\hat{\sigma})}$ (1)	**Zheng [26]**
Ω^2	$\hat{\Omega}^2 = 1 - \dfrac{\hat{\sigma}^2}{\hat{\sigma}_0{}^2}$	**Xu [25]**
R_2^2	$R_2^2 = 1 - \left(\dfrac{RSS}{RSS_0}\right)$ (2)	**Xu [25]**
ρ^2	$\hat{\rho}^2 = 1 - \dfrac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\exp\left(\dfrac{RSS}{N\hat{\sigma}^2} - \dfrac{RSS_0}{N\hat{\sigma}_0{}^2}\right)$	**Xu [25]**

[(1)] For the conditional version of this statistic, $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i$ is used. For the marginal version, $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}$ is used.

[(2)] RSS is the residual sum of squares from the model at hand and RSS_0 is the residual sum of squares from the null mode in (6). Both RSS and RSS_0 are computed using random effects that is $RSS = \sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)$ where $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i$. Note that for the marginal version of this statistic (i.e., using $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}$), the same value as R_1^2 marginal would be obtained.

Table 2.2 Means and ranges for conditional R^2 and R^2 proposed by Xu [25]

			Conditional pseudo-R^2			- R^2 proposed by Xu [25]		
Value of σ^2 in simulated data set	**Model**	**Ω^2 (for full model)[1]**	**Mean of r_c (minimum, maximum) Vonesh et al. [22][2]**	**Mean of R_1^2 (minimum, maximum) Vonesh and Chinchilli [21][2]**	**Mean of P_{rand} (minimum, maximum) Zheng [26][2]**	**Mean of $\hat{\Omega}^2$ (minimum, maximum)**	**Mean of R_2^2 (minimum, maximum) [2]**	**Mean of ρ^2 (minimum, maximum)**
12	Overfitted Model: Age, Gender, Treatment and 2 extraneous variables	0.77	0.97 (0.95, 0.98)	0.94 (0.90, 0.97)	0.94 (0.90, 0.96)	0.78 (0.70, 0.84)	0.79 (0.71, 0.85)	0.79 (0.71, 0.85)
	Full Model: Age, Treatment, Gender		0.97 (0.95, 0.98)	0.94 (0.90, 0.97)	0.94 (0.90, 0.96)	0.78 (0.70, 0.83)	0.79 (0.71, 0.85)	0.79 (0.71, 0.84)
	Reduced Model 1: Age and Treatment		0.97 (0.94, 0.98)	0.94 (0.89, 0.97)	0.94 (0.88, 0.97)	0.78 (0.69, 0.83)	0.79 (0.70, 0.85)	0.79 (0.70, 0.85)
	Reduced Model 2: Age		0.97 (0.94, 0.98)	0.94 (0.89, 0.97)	0.94 (0.89, 0.97)	0.77 (0.69, 0.83)	0.79 (0.69, 0.85)	0.78 (0.69, 0.85)
45	Overfitted Model: Age, Gender, Treatment and 2 extraneous variables	0.47	0.90 (0.83, 0.95)	0.82 (0.72, 0.90)	0.82 (0.72, 0.90)	0.49 (0.35, 0.62)	0.51 (0.36, 0.65)	0.51 (0.36, 0.64)
	Full Model: Age, Treatment, Gender		0.90 (0.83, 0.95)	0.82 (0.72, 0.90)	0.81 (0.72, 0.90)	0.49 (0.36, 0.62)	0.51 (0.36, 0.65)	0.50 (0.36, 0.64)
	Reduced Model 1: Age and Treatment		0.90 (0.82, 0.95)	0.82 (0.71, 0.90)	0.81 (0.71, 0.90)	0.49 (0.35, 0.62)	0.51 (0.35, 0.65)	0.51 (0.35, 0.65)
	Reduced Model 2: Age		0.90 (0.82, 0.95)	0.82 (0.72, 0.90)	0.81 (0.71, 0.90)	0.49 (0.35, 0.62)	0.51 (0.37, 0.65)	0.51 (0.36, 0.65)
250	Overfitted Model: Age, Gender, Treatment and 2 extraneous variables	0.14	0.61 (0.41, 0.79)	0.48 (0.29, 0.68)	0.48 (0.29, 0.67)	0.17 (0.05, 0.35)	0.19 (0.05, 0.39)	0.18 (0.05, 0.39)
	Full Model: Age, Treatment, Gender		0.61 (0.41, 0.78)	0.48 (0.29, 0.67)	0.48 (0.29, 0.66)	0.17 (0.05, 0.35)	0.18 (0.04, 0.39)	0.18 (0.04, 0.39)
	Reduced Model 1: Age and Treatment		0.61 (0.42, 0.78)	0.49 (0.30, 0.67)	0.48 (0.30, 0.67)	0.17 (0.05, 0.35)	0.19 (0.05, 0.41)	0.19 (0.05, 0.40)
	Reduced Model 2: Age		0.61 (0.41, 0.78)	0.49 (0.31, 0.68)	0.49 (0.31, 0.67)	0.17 (0.05, 0.35)	0.20 (0.06, 0.41)	0.20 (0.06, 0.40)

[1] Computed using formula from Xu [25]

[2] Note that for $r_{c,}$ R_1^2, R_2^2 and P_{rand} the numerator was computed using $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_i$

Table 2.3 Means and ranges for marginal R^2

Value of σ^2 in simulated data set	Model	Ω^2 (for full model)[1]	Mean of population-based r_c (minimum, maximum) Vonesh et al. [22] [2]	Mean of population-based R_1^2 (minimum, maximum) Vonesh and Chinchilli [21][2]	Mean of population-based P_{rand} (minimum, maximum) Zheng [26] [2]
12	Overfitted Model: Age, Gender, Treatment and 2 extraneous variables	0.77	0.72 (0.52, 0.86)	0.56 (0.32, 0.75)	0.56 (0.31, 0.75)
	Full Model: Age, Treatment, Gender		0.72 (0.52, 0.86)	0.56 (0.32, 0.75)	0.56 (0.31, 0.75)
	Reduced Model 1: Age and Treatment		0.55 (0.36, 0.74)	0.38 (0.20, 0.59)	0.38 (0.20, 0.58)
	Reduced Model 2: Age		0.33 (0.22, 0.50)	0.20 (0.12, 0.34)	0.19 (0.12, 0.33)
45	Overfitted Model: Age, Gender, Treatment and 2 extraneous variables	0.47	0.64 (0.43, 0.79)	0.47 (0.26, 0.65)	0.47 (0.25, 0.65)
	Full Model: Age, Treatment, Gender		0.64 (0.43, 0.79)	0.47 (0.26, 0.65)	0.47 (0.25, 0.65)
	Reduced Model 1: Age and Treatment		0.48 (0.29, 0.67)	0.32 (0.16, 0.50)	0.31 (0.15, 0.50)
	Reduced Model 2: Age		0.28 (0.17, 0.42)	0.17 (0.09, 0.27)	0.16 (0.09, 0.26)
250	Overfitted Model: Age, Gender, Treatment and 2 extraneous variables	0.14	0.39 (0.18, 0.59)	0.24 (0.10, 0.41)	0.24 (0.10, 0.41)
	Full Model: Age, Treatment, Gender		0.38 (0.18, 0.58)	0.24 (0.10, 0.41)	0.23 (0.10, 0.41)
	Reduced Model 1: Age and Treatment		0.27 (0.07, 0.48)	0.16 (0.04, 0.32)	0.16 (0.03, 0.31)
	Reduced Model 2: Age		0.15 (0.04, 0.30)	0.08 (0.02, 0.18)	0.08 (0.01, 0.17)

[1] Computed using formula from Xu [25]

[2] Note that the numerator for all 3 statistics was computed using $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}$

Table 2.4 Conditional R^2 statistics and R^2 proposed by Xu [25] on the dental data from Pothoff and Roy [15]

Fixed effect term in model	r_c	R_1^2	P_{rand}	$\widehat{\Omega}^2$	R_2^2	ρ^2
Age, Gender, Age-by-Gender Interaction	0.91	0.85	0.83	0.65	0.68	0.68
Age, Gender	0.92	0.86	0.83	0.65	0.70	0.69
Age	0.92	0.86	0.65	0.65	0.70	0.69
Intercept	0.93	0.88	0.85	0.80	0.75	0.73

Note that for $r_{c,}$ R_1^2, R_2^2 and P_{rand} the numerator of these statistics was computed using $\widehat{\mathbf{y}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i$

Table 2.5 Marginal R^2 for dental data of Potthoff and Roy [15]

Model	r_c	R_1^2	P_{rand}
Age, Gender, Age-by-Gender	0.59	0.42	0.40
Age, Gender	0.58	0.41	0.38
Age	0.41	0.26	0.23

Note that the numerator for all 3 statistics was computed using $\widehat{\mathbf{y}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}}$

3 R^2 Statistics as Measures of external and internal consistency in the Linear Mixed Model

Abstract

Several R^2 statistics have been proposed for linear mixed models (LMMs) to assess adequacy of fit. However, Orelien and Edwards [14] showed that many of these statistics performed poorly in that they showed little variation when important variables related to the outcome were missing from the model. It was shown that R^2 statistics that can be classified as marginal are more useful than conditional R^2 statistics in selecting fixed effect covariates. In this chapter, we review the theoretical framework of the different approaches that can be used or have been used for R^2 statistics in the LMM. Limitations of each of these approaches are discussed. We then propose new R^2 statistics based on approaches that have not been considered thus far. Two of the statistics that we propose have the advantage that they can be easily interpreted. One of the statistics measures what we denote as "external consistency" (how well the model performs compared to other competing models, particularly a null model) while the other measures "internal consistency" (how much of the variation in the outcome is explained by the model at hand, assuming that it is the true one). This latter statistic has a corresponding population parameter assuming that the model is fully specified. Simulation results show that these statistics can be used to assess the

goodness of the fit of a model or compare the fixed effects of alternative models. In assessing the ability of the statistics proposed to compare fixed effect covariates of competing models, the comparison is limited to models having the same random effects.

3.1 Introduction

In the LMM, several pseudo-R^2 statistics have been proposed for assessing GOF. Vonesh *et al.* [23] proposed the concordance coefficient correlation coefficient (CCC) — which they denote r_c —to measure the percent agreement between observed and predicted values. Vonesh and Chinchilli [22] proposed in addition to CCC, R_1^2, which is simply R^2 of the traditional linear model. Zheng [26] proposed P_{rand}, which makes adjustments on R_1^2 in a manner similar to the Akaike Information Criteria (AIC) [1] or the Bayesian Information Criteria (BIC) [18]. Xu [25] proposed three statistics—which we denote respectively Ω^2, R_2^2 and ρ—that measure the proportion of variation accounted for by the model. Most of these statistics (r_c, R_1^2, P_{rand} and R_2^2) take the form of $1-\dfrac{\text{Numerator}}{\text{Denominator}}$ where the numerator is the sum of squares of the residuals. Vonesh *et al.* [23] and Vonesh and Chinchilli [22] differentiate between “conditional” and “population” versions of CCC and R_1^2. For conditional versions of these statistics the random effects are included in computing the predicted values as opposed to the population-based R^2 where only the fixed effect components are included. While Vonesh *et al.* [23] and Vonesh and Chinchilli [22] suggested that the population-based versions of the statistics could be used to assess adequacy of the fixed effect components, they indicated that to account for the combined effects of both fixed and random effects, the conditional version should be used. In that sense, the conditional versions of CCC and R_1^2 can be viewed as overall GOF statistics that can give a measure of the adequacy of both the random and fixed effects.

Orelien et al. [14] showed through simulations that only the marginal versions of these statistics could discriminate between two models having the same random effects but in which important fixed effect covariates were missing from one of them. This finding puts into question the usefulness of conditional R^2 statistics to assess model adequacy. However, the use of marginal R^2 statistics is not without problem. The marginal R^2 statistics that have been proposed thus far and were reviewed in our first paper have two limitations—they have no corresponding population parameter (therefore it can be argued that it is not clear exactly what is being measured) and they cannot be used as omnibus GOF to assess the adequacy of the model. These limitations are the motivations behind new R^2 statistics that we are proposing.

While the new R^2 statistics that we are proposing in this chapter could be used in theory as omnibus statistics for assessing adequacy of fit or for comparing any two models whether or not they have the same random effect, our simulations are limited to the cases where the analyst is interested in assessing the adequacy of the fixed effect covariates for one model or comparing (*e.g.*, for the purpose of parsimony) fixed effect covariates of models having the same random effects function. Often, the focus of the analyst is on the fixed effect covariates, not on the random effects function. Cnaan et al [2] indicated that in the LMM, the use of a random intercept and random slope is often sufficient. There are many instances such as in epidemiological studies where repeated measurements are taken on a subject and the interest of investigators is in assessing the effect over time of the explanatory variables (such as treatment and individual patient characteristics that may impact treatment) on the outcome. For example, Cnaan et al. [2] presented data from a clinical trial to compare the efficacy of an experimental drug versus a control. The study included a total of 233 patients from 13 sites, and there were three doses for the experimental drug: low, medium, and high. The endpoint of interest was the Brief Psychiatric Rating Scale (BPRS) at baseline, 1, 2, 3, 4, and 6 weeks. Several competing models in the fixed effects were compared and in all the models that were compared, the random effect function remained the same and consisted of a random intercept and random linear and quadratic time variables with unstructured covariance structure in the random effects. The focus of this chapter is limited to the analysis of longitudinal data similar to that presented by Cnaan et al. [2]. The competing models considered in our simulations have the same random effect (an intercept and a random slope) but different fixed effects.

Another element of the GOF of a LMM which is not examined in our simulations is the ability of the proposed R^2 statistics to determine the adequacy of the covariance structure. To compare covariance structures, it is usually assumed that the mean structure has been correctly specified. Covariance model selection techniques that require the assumption include the LRT (Jennrich and Schluchter [9]; Schaalje et al. [17]; Grady and Helms [6]), information criteria (AIC and BIC), and predictive approaches such as PRESS [13]). Hence, as a first step to assessing the performance of these newly proposed R^2 statistics, in this chapter attention is focused on fixed effects. As a result, the performance of the R^2 statistics for changing covariance structures is beyond the scope of this paper. However, we consider the topic the object of future research. Note that although the interest lies in the fixed effect covariates, the analyst may be unsure about the covariance structure to use. In such a case, one approach could be to use graphical exploratory techniques for selecting the covariance structure such as those proposed by Diggle et al. [5], Dawson et al. [4], Zimmerman [27] and Pourahmadi [16].

The paper is organized as follows: In Section 3.2, we describe notations used for the LMM. Section 3.3 reviews from a theoretical standpoint approaches that have been or can be used for developing R^2 statistics. Four new statistics based on approaches that have not been considered before are proposed in Section 3.4. Data generation techniques and simulation results showing the performance of the statistics proposed are described in Section 3.5. An example is provided in section 3.6. Discussions follow in Section 3.7. We end the chapter with concluding remarks in section 3.8.

3.2 The Linear Mixed Model

The model and data simulation methods are discussed in more details in our earlier paper. The simulation will be based on model (1) below as formulated by Harville [8] and Laird and Ware [11]:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{1}$$

where $i \in \{1, 2, ..., \text{n}\}$ is the index for the independent sampling units (ISU) and

$\mathbf{y}_i$ is an $n_i \times 1$ vector of observations from the i^{th} independent sampling unit (subject),

$\mathbf{X}_i$ denotes an $n_i \times p$ fixed effects design matrix for the i^{th} subject,

$\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown, constant, fixed effect parameters,

$\mathbf{Z}_i$ denotes an $n_i \times q$ random effects design matrix for the i^{th} subject,

$\mathbf{b}_i$ is a $q \times 1$ vector of unobservable random effects for the i^{th} subject, and

$\mathbf{e}_i$ denotes an $n_i \times 1$ vector of unobservable within-subject error terms.

It is also assumed that $\mathbf{b}_i$ has a multivariate normal distribution $N_q(\mathbf{0},\mathbf{G})$ independent of $\mathbf{e}_i$, which has a multivariate distribution $N_{n_i}(\mathbf{0},\mathbf{R}_i)$.

$$E\begin{bmatrix}\mathbf{b}_i\\ \mathbf{e}_i\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \text{ and } V\begin{bmatrix}\mathbf{b}_i\\ \mathbf{e}_i\end{bmatrix}=\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_i\end{bmatrix},$$

where $\mathbf{G}$ is a $q\times q$ unknown covariance matrix for the random effects and $\mathbf{R}_i$ is an $n_i \times n_i$ unknown covariance matrix for the within-subject error terms for the i^{th} subject. With these assumptions, for the i^{th} subject we have $\mathbf{\Sigma}_i = V(\mathbf{y}_i)=\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$. In many applications, $\mathbf{R}_i$ is taken to be $\sigma^2\mathbf{I}_{n_i}$, known as the conditional independence assumption for the error term [11].

3.3 Approaches for Developing R^2 Statistics

In this section, we look at the statistical framework for R^2 in the LMM. We show several categories that can be used to classify existing R^2 statistics and new ones that could be proposed. The advantages and disadvantages of each category are discussed.

3.3.1 R^2 based on comparing the Mahalanobis distance of the model to a null model

A framework for developing R^2 statistics was outlined by Vonesh and Chinchilli [22] and Xu [25] whereby the Mahalanobis distance of the model is compared to the Mahalanobis distance of the null model. This can be represented as:

$$R^2_{\text{distance}} = 1-\frac{\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)'\widehat{\mathbf{\Sigma}}_{nulli}^{-1}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)}{\sum_{i=1}^{n}(\mathbf{y}_i - \widehat{\mathbf{y}}_{nulli})'\widehat{\mathbf{\Sigma}}_{nulli}^{-1}(\mathbf{y}_i - \widehat{\mathbf{y}}_{nulli})} \qquad (2)$$

Where $\hat{\mathbf{y}}_{nulli}$ and $\mathbf{\Sigma}_{nulli}$ are respectively the predicted value for the null model and the covariance matrix for subject i. In the LMM, two types of null model are possible—a model that consists of a fixed effect and random effect intercept (3) or a model that consists of only a fixed effect intercept (4). The first null model can be represented as:

$$\mathbf{y}_i = \mathbf{1}_{n_i}\beta_0 + \mathbf{1}_{n_i} b_{0i} + \mathbf{u}_i, \qquad (3)$$

where β_0 is an unknown fixed parameter, b_{0i} is an unknown random coefficient that has a normal distribution with mean $\mathbf{0}$, and $\mathbf{u}_i$ is the unobservable within-subject random error term for the model (*i.e.*, equation 3 represents a model with fixed and random effect intercepts). We define $V(\mathbf{u}_i) = \sigma_0{}^2\mathbf{I}_{n_i}$ and $V(b_{0i}) = \tau_{00}^2$.

The second null model can be represented as:

$$\mathbf{y}_i = \mathbf{1}_{n_i}\beta_{00} + \mathbf{u}_{0i}, \qquad (4)$$

where β_{00} is an unknown fixed coefficient and $\mathbf{u}_{0i}$ is the unobservable within-subject random-error term for the model (a model with a fixed effect intercept and no random effects) and $V(\mathbf{u}_{0i}) = \sigma_{00}{}^2\mathbf{I}_{n_i}$.

Both R_1^2 of Vonesh et al [23] and Vonesh and Chinchilli [22] and R_2^2 of Xu [25] fall in this category. For R_1^2 given by $R_1^2 = 1 - \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)'(\mathbf{y}_i - \hat{\mathbf{y}}_i)}{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})}$, the null model consists of a fixed effect intercept and no random effects (3) so that for any subject i, $\hat{\mathbf{y}}_{nulli} = \bar{y}\mathbf{1}_{n_i}$ and

$\hat{\mathbf{\Sigma}}_{nulli} = \frac{1}{N-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'_{1\times n_i}(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})_{n_i\times 1}$ which is a constant. For R_2^2 given by

$R_2^2 = 1 - \left(\frac{RSS}{RSS_0} \right)$ where RSS is the residual sum of squares for the model in (1) and RSS_0 is the residual sum of squares under the model in (3) does not fit into this category. However, if we modify it by introducing in the formula, $\widehat{\mathbf{\Sigma}}_{nulli} = \widehat{\mathbf{\Sigma}}_{0i} = \widehat{\tau}_{00}^2 \mathbf{1}'_{n_i \times n_i} \mathbf{1}_{n_i \times n_i} + \widehat{\sigma}_{00}^2 \mathbf{I}_{n_i \times n_i}$ which is the variance for subject i for the null model in (3), a modified version R_2^2, $R_{2(\text{mod})}^2$ is given by

$$R_{2(\text{mod})}^2 = 1 - \frac{\sum_{i=1}^{n} (\mathbf{y}_i - \widehat{\mathbf{y}}_i)' \widehat{\mathbf{\Sigma}}_{0i}^{-1} (\mathbf{y}_i - \widehat{\mathbf{y}}_i)}{\sum_{i=1}^{n} (\mathbf{y}_i - \widehat{\mathbf{y}}_{0i})' \widehat{\mathbf{\Sigma}}_{0i}^{-1} (\mathbf{y}_i - \widehat{\mathbf{y}}_{0i})}$$ where $\widehat{\mathbf{y}}_{0i}$ is the predicted value for the null model in (3). Note that the R^2 of traditional linear models would also fall in this category with the null model consisting of an intercept and thus the predicted value for each observation in that null model being $\overline{y}$ (the average of all observations).

Orelien et al. [14] found that R^2 statistics from this category need to be further classified as either marginal or conditional statistics for the LMM. For conditional R^2 statistics, the computation of the predicted values includes the random effects as opposed to marginal R^2 statistics where only the fixed effect parameter estimates are included in the computation of these predicted values. Results from simulations found that conditional R^2 statistics appeared to be unable to detect when important cross sectional covariates were missing from the model. Another limitation of R^2 based on this framework is that the Mahalanobis distance for the null model in the LMM is not the largest distance for conditional statistics. In traditional linear models, the denominator $\sum_{j=1}^{N} (y_j - \overline{y})' \widehat{\mathbf{\Sigma}}_0 (y_j - \overline{y})$ is the largest distance (where N is the total number of observations and $\widehat{\mathbf{\Sigma}}_0 = \widehat{\sigma}_{00}{}^2$). A clear advantage for these R^2 statistics is that in terms of interpretation, they can be seen as an extension of the R^2 of traditional linear models.

3.3.2 R^2 based on measures of agreement between observed and predicted values

Vonesh et al. [23] proposed the concordance correlation coefficient (CCC) denoted r_c which is given by

$$r_c = 1 - \frac{\sum_{i=1}^{n} (\mathbf{y}_i - \widehat{\mathbf{y}}_i)'(\mathbf{y}_i - \widehat{\mathbf{y}}_i)}{\left(\sum_{i=1}^{n} (\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})'(\mathbf{y}_i - \bar{y}\mathbf{1}_{n_i})\right) + \left(\sum_{i=1}^{n} (\widehat{\mathbf{y}}_i - \hat{y}\mathbf{1}_{n_i})'(\hat{\mathbf{y}}_\mathbf{i} - \hat{y}\mathbf{1}_{n_i})\right) + N(\bar{y} - \hat{y})^2}$$

, where *n* is the number of independent sampling units (or subjects), $\mathbf{y}_i$ is the vector of observed values for the *i*th subject, $\widehat{\mathbf{y}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i$ is the vector of predicted values for the *i*th subject, $\widehat{y}$ is the grand average of the predicted values, $\bar{y}$ is the grand average of the observed values, *N* is the total number of observations, and $\mathbf{1}_{n_i}$ is an n_i x 1 vector of 1's. CCC was first introduced by Lin [12] for comparing the percent agreement between a gold assay and a cheaper one. As such, CCC can be interpreted as measuring the percent agreement between the observed and predicted values. Hence, one could consider an analogous version of r_c to be the squared correlation coefficient between the observed and predicted in traditional linear models (Kvalseth [10]). For categorical data, an analogous statistic is the Kappa coefficient (Cohen [3]).

One of the limitations of using measures of agreement is that it is easy to conceive how two models that are not necessarily a good fit but for which high values of the observed are associated with high predicted values and lower values of the observed are associated with lower predicted values can lead to artificially high values of the statistic. For r_c, Orelien and Edwards [14] showed that the conditional version of the statistic similar to other conditional statistics discussed in section 3.3.1 was unable to detect the absence of important cross-sectional covariates from the model.

3.3.3 R^2 based on comparing the variation explained by the model at hand to that of a null model

Another approach to computing R^2 is to measure the proportion of variation explained by the model. This can be estimated by:

$$\frac{\text{variance explained by the model}}{\text{variance assuming a null model}} =$$

$$1-\frac{\text{variance not explained by the model}}{\text{variance assuming a null model}} \qquad (5)$$

The variance not explained by the model, could be based a) on the component of the variance for an observation or b) the component of the variance for the average of all observations that is not explained by the model. Similarly, one could base the computation for the denominator in (5) accordingly on either the variance for an observation or the variance for the average of all observations. In this chapter, we base all computations for the variance explained by the model at hand or the null model on the variance of the average of all observations.

The framework in (5) could be what Xu [25] try to follow for the statistic $\hat{\Omega}^2$ which he defines as $\hat{\Omega}^2 = 1-\frac{\hat{\sigma}^2}{\hat{\sigma}_0{}^2}$ where $\hat{\sigma}^2$ is the estimated within-subject variance for the model at hand and $\hat{\sigma}_0{}^2$ is the estimated within-subject variance for the model in (3). Orelien and Edwards [14] showed through simulations that problems existed with Ω^2 and that it was therefore not appropriate as an R^2 statistic in the LMM. Specifically, Orelien and Edwards [14] showed that $\hat{\sigma}^2$ is robust to misspecification of the cross-sectional covariates. But we will show that Ω^2 does not fit the framework of (5) in that σ^2 cannot be considered to be the component of the variance not explained by the model and similarly ${\sigma_0}^2$ cannot be considered to be the component of the variance not explained by the null model.

On the other hand, the R^2 of traditional linear model fits the framework in (5). Consider the linear model given by:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \qquad (6)$$

Where $\mathbf{y} = \{y_j\}$, $j = 1, 2, \ldots, N$ is a $N \times 1$ vector of independent observations, $\mathbf{X}$ is an $N \times p$ matrix containing the covariates in the model, β is the $p \times 1$ vector of the unknown parameters and ε is the $N \times 1$ vector of the random errors with $\text{var}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}_N$. One can argue that the variation in $\mathbf{y}$ not explained by the model is the estimate of the random error term $\hat{\sigma}_\varepsilon^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$. Similarly the variance of an observation under null model (4), that is a model with only an intercept term is given by $\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2$. Note that under the model in (4), $\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2$ is the largest variance estimate that can be attained under any model. This is important for the range to be between 0 and 1. In the LMM, it is not clear that the largest maximum variance is always attained under the null model (3). Theoretically, it is thus possible for any R^2 based on the framework discussed in this section to have a lower limit that is less than 0. That is, the range for R^2 in the LMM that follows the framework in this section is $(-\infty, 1)$.

3.3.4 R^2 based on computing the variation explained by the model as a proportion of the variation in the outcome assuming that the fitted model is adequate

An approach which has not been considered either in the LMM or in the traditional linear model is to compute R^2 as a proportion of the variation explained by the model assuming that the model is adequate. One way this can be accomplished is by estimating the component of the variation (of the average of all observations) attributed to the variables in the model as a proportion of the total variance (for the average of all observations)—where total variance is computed assuming that the model at hand is the correct one. For example, in the LMM, the variance of a subject is given by $V(\mathbf{y}_i) = \mathbf{\Sigma}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$ where $\mathbf{R}_i$ is often taken to be $\sigma^2\mathbf{I}_{n_i}$. It can be shown that:

$V(\mathbf{y}_i) = \mathbf{z}_i\mathbf{G}\mathbf{z}_i{'} + \mathbf{R}_i = \tau_0^2\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{n_i\times n_i} + h(\mathbf{z}_i\mathbf{G}\mathbf{z}_i{'}) + \mathbf{R}_i$ where $h(\mathbf{z}_i\mathbf{G}\mathbf{z}_i{'})$ is the expression that remains after taking out τ_0^2 from $\mathbf{z}_i\mathbf{G}\mathbf{z}_i{'}$. and in the case where $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i\times n_i}$, we have

$$V(\mathbf{y}_i) = \tau_0^2\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{n_i\times n_i} + h(\mathbf{z}_i\mathbf{G}\mathbf{z}_i{'}) + \sigma^2\mathbf{I}_{n_i\times n_i} = \begin{pmatrix} \tau_0^2+\sigma^2 & \tau_0^2 & \cdots & \tau_0^2 \\ \tau_0^2 & \tau_0^2+\sigma^2 & \tau_0^2 & \vdots \\ \vdots & \tau_0^2 & \ddots & \tau_0^2 \\ \tau_0^2 & \cdots & \tau_0^2 & \tau_0^2+\sigma^2 \end{pmatrix}_{n_i\times n_i} + h(\mathbf{z}_i\mathbf{G}\mathbf{z}_i{'})$$

That is, in the case where $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i\times n_i}$, the elements in the expression of $\mathrm{var}\left(\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} y_{ij}\right)$ (average of all observations) that does not depend on the variables in the model is:

$$\sum_{i=1}^{n} n_i\sigma^2 + n_i^2\tau_0^2 \qquad (7)$$

When $\mathbf{R}_i \neq \sigma^2 \mathbf{I}_{n_i \times n_i}$, it can be shown that the portion of $\operatorname{var}\left(\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} y_{ij} \right)$ that is not explained by the model is:

$$\sum_{i=1}^{n} \left(n_i^2 \tau_0^2 + \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \mathbf{R}_i [j,k] \right) \tag{8}$$

In the LMM, one could also consider that regardless of the structures of $\mathbf{R}_i$, for the purpose of assessing adequacy of the cross-sectional covariates that the portion of $\operatorname{var}\left(\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} y_{ij} \right)$ that does not depend on the variables in the model is captured only by the variance component for the random intercept. This is supported by findings from Verbeke and Fieuws [20] who demonstrated that the estimate of the within subject variance σ^2 was robust to misspecification of the fixed effects. Hence, we propose that R^2 be derived by assuming that the proportion of variation not accounted for by the model regardless of the structure of $\mathbf{R}_i$ is:

$$\sum_{i=1}^{n} n_i^2 \tau_0^2 \tag{9}$$

The formulae in (7), (8) and (9) use the fact that the variance of the sum of correlated random variables is the sum of the elements of the covariance matrix for these random variables. In the appendix, we show through an example why the expression in (7) is appropriate. Formulae for R^2 statistics based on the framework in this section are given and discussed in more detail in section 3.4. A disadvantage of the R^2 that would fit the framework in this section is that they are not an extension of the traditional R^2 of linear models. Also, one could argue that these types of R^2 measure “internal consistency”, or the extent to which the model indicates that it explains a lot of the variation in the outcome as opposed to whether it does better at explaining the variation in the outcome compared to a trivial model such as an intercept only. The danger could be that a model might be “internally consistent” indicating that it explains a substantial amount of the variation in the outcome and still be inadequate by not being better at explaining the outcome than a trivial model with greater “external consistency”.

3.4 New R^2 Statistics

Four new R^2 statistics are proposed in this section. The statistics in section 3.4.1 are measures of “external consistency” and are based on the framework in Section 3.3.3 (comparing the variation explained by the model at hand to that of a null model) and the ones in section 3.4.2 are measures of “internal consistency” and are based on the framework in Section 3.3.4 (computing the variation explained by the model assuming that the model is adequate). All four statistics are an attempt to measure the proportion of variation explained by the model.

3.4.1 New R^2 statistics based on comparing the variation explained by the model at hand to that of a null model

We propose the statistic *measure of external consistency* (MEC_1) that we define as:

$$MEC_1 = \frac{\text{variance of average obs. explained by the model}}{\text{variance of average obs. assuming a null model}}$$
$$= 1 - \frac{\text{variance of average obs. not explained by the model}}{\text{variance of average obs. assuming a null model}}$$

$$MEC_1 = 1 - \frac{\sum_{i=1}^{n}\left(n_i^2\widehat{\tau}_0^2 + \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\mathbf{R}}_i[j,k]\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\Sigma}_{0i}[j,k]\right)} \qquad \textbf{(10)}$$

Where $\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\mathbf{R}}_i[j,k]$ is the sum of all the elements of the matrix $\widehat{\mathbf{R}}_i$ (for subject i) and $\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\Sigma}_{0i}[j,k]$ is the sum of all the elements of the matrix $\widehat{\Sigma}_{0i}$, the covariance matrix for subject i under null model (3) . Note that the numerator $\sum_{i=1}^{n}\left(n_i^2\widehat{\tau}_0^2 + \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\mathbf{R}}_i[j,k]\right)$ corresponds to the elements of the variation in the outcome not explained by the model and the denominator corresponds to the variation in the outcome assuming null model (3).

The second statistic, MEC_2 is based on assuming that the variation not explained by the model for the purpose of assessing the adequacy of the cross-sectional covariates is given by τ_0^2 :

$$MEC_2 = 1 - \frac{\tau_0^2\sum_{i=1}^{n}\left(n_i^2\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\Sigma}_{0i}[j,k]\right)} \qquad (11)$$

In the case, where $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i\times n_i}$, we propose that:

$$MEC_1 = 1 - \frac{\sum_{i=1}^{n} n_i\sigma^2 + n_i^2\tau_0^2}{\mathrm{N}(\sigma_0^2 + \tau_{00}^2)} \qquad (12)$$

$$MEC_2 = 1 - \frac{\sum_{i=1}^{n} n_i^2\tau_0^2}{\mathrm{N}(\sigma_0^2 + \tau_{00}^2)} \qquad (13)$$

The ranges of both MEC_1 and MEC_2 are between $-\infty$ and 1, with values close to 1 indicating a good fit and values near to or less than 0 indicating a lack of fit. One of the possibilities for negative values of these statistics is that for some data, the null model may lead to lower variance in the outcome compared to the model at hand. As we have indicated in Section 3.3.3, this is one of the drawbacks in working with R^2 that compare the variation explained by the model to that of a null model.

3.4.2 New R^2 statistics based on computing the variation explained by the model as a proportion of the variation in the outcome assuming that the fitted model is adequate

Based on the framework outlined in Section 3.3.4, we are proposing two new R^2 statistics, MIC_1 and MIC_2 (measures of internal consistency), that both measure the proportion of variation explained by the model assuming that the model is adequate. We define MIC_1 as:

$$MIC_1 = 1 - \frac{\text{variation of average obs not explained by the model}}{\text{variance of average obs. assuming model is true}}$$

$$MIC_1 = 1 - \frac{\sum_{i=1}^{n}\left(n_i^2\tau_0^2 + \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\mathbf{R}_i[j,k]\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\Sigma}_i[j,k]\right)} \qquad (14)$$

Where $\widehat{\Sigma}_i = \mathbf{Z}_i\widehat{\mathbf{G}}\mathbf{Z}_i + \widehat{\sigma}^2 I$ and $\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\Sigma}_i[j,k]\right)$ is the sum of all the elements of $\widehat{\Sigma}_i$ and $\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\mathbf{R}}_i[j,k]$ is the sum of all the elements of the matrix $\widehat{\mathbf{R}}_i$. The denominator corresponds to $\text{var}\left(\sum_{i=1}^{n}\sum_{j=1}^{n_i}y_{ij}\right)$, the variance of the sum of all the elements of $\mathbf{y} = \{y_{ij}\}$ with

$i=1$ to n and $j=1,\ ...,\ n_i$. The numerator as shown in section 3.3.4 is part of the denominator and corresponds to the amount of variation not explained by the model. In the case of conditional independence, i.e., $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i \times n_i}$, MIC_1 is given by:

$$MIC_1 = 1 - \frac{\sum_{i=1}^{n} \left(n_i^2 \tau_0^2 + n_i \sigma^2 \right)}{\sum_{i=1}^{i=n} \left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \widehat{\Sigma}_i [j,k] \right)} \tag{15}$$

Under the assumption that the variation not explained by the model for the purpose of assessing the adequacy of the cross-sectional covariates is given by τ_0^2, a counterpart to MEC_2, MIC_2 is defined as:

$$MIC_2 = 1 - \frac{\sum_{i=1}^{n} n_i^2 \widehat{\tau}_0^2}{\sum_{i=1}^{i=n} \left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \widehat{\Sigma}_i [j,k] \right)} = 1 - \frac{\widehat{\tau}_0^2 \sum_{i=1}^{n} n_i^2}{\sum_{i=1}^{i=n} \left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \widehat{\Sigma}_i [j,k] \right)} \tag{16}$$

Note that if $\mathbf{G}$ is not diagonal, negative elements on the off-diagonals of $\mathbf{G}$ could lead to the denominator of (14), (15) or (16) being lower than the numerator. Although, the numerator is part of the denominator, for both MIC_1 and MIC_2 the range may not necessarily be between 0 and 1. If $\mathbf{G}$ is an unstructured covariance matrix with negative values in the off-diagonals, the numerator of these statistics may be larger than the denominator. Hence, the range for MIC_1 and MIC_2 is between 0 and 1 if $\mathbf{G}$ has a diagonal covariance structure and $(-\infty, 1)$ if $\mathbf{G}$ has an unstructured covariance matrix with negative values in the off-diagonals. One could consider using $\sum_{i=1}^{i=n} \left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \mathrm{tr}(\widehat{\Sigma}_i) \right)$ in the denominator so that the range of the statistics is always between 0 and 1. However, we preferred not to use this approach because of problems that would arise with interpretation. Also, the example given in section 3.7 shows that using $\mathrm{tr}(\hat{\sum}_i)$ instead of $\hat{\sum}_i$ could potentially lead to artificially higher values of the statistic. When $\mathrm{tr}(\hat{\sum}_i)$ is used instead of $\hat{\sum}_i$, positive values ranging from 0.36 t0 0.55 are obtained as opposed to the negative values. While in general a finite range might be desirable for an R^2 statistic, in this case if we were to change the formula by using $\mathrm{tr}(\hat{\sum}_i)$ to simply achieve a finite range, we might miss the negative values that would indicate model inadequacy. Table 3.1 summarizes the four statistics that are proposed in this chapter.

3.5 Parameters for the Simulation

We conducted a simulation to assess the performance of the proposed R^2. Details of these simulations were first reported in Orelien and Edwards [14]. For the simulation, using the IML module in SAS, we simulated six sets of data. In each of the data sets, there were 10,000 replications, 64 subjects, and 6 observations per subject. Three of the six data sets assumed a diagonal covariance matrix for the random effects and the other three assumed an unstructured one. The data sets with the same covariance structure for the random effects differed in the values for σ^2 that were 12, 45, and 250. The different values of σ^2 were used to assess how performance of the pseudo-R^2 varied with increased within-subject error which translates into within-subject correlations of about 0.1, 0.5 and 0.8. The parameters used in the analysis were:

$\mathbf{X} = [\mathbf{1}_6, (5,\ 6,\ 7,\ 7.25,\ 7.5,\ 7.75)', \mathbf{1}_6\delta_k, \mathbf{1}_6\delta_l]$; $\mathbf{Z} = [\mathbf{1}_6, (5, 6, 7, 7.25, 7.5, 7.75)']$, where $\mathbf{1}_6$ is a 6×1 vector of ones and $\delta_k, \delta_l = \{0,1\}$. We took $\boldsymbol{\beta} = \begin{pmatrix} 10 \\ 6 \\ 11 \\ 11 \end{pmatrix}$ and $\mathbf{G} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$ for the three data sets with diagonal covariance matrix. For the three data sets with unstructured covariance matrix, we used $\mathbf{G} = \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}$. The population parameter for MIC_1 was computed as $1 - \frac{n_1^2\tau_0^2 + n_1\sigma^2}{\left(\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\Sigma_1[j,k]\right)}$ where $\tau_0^2 = \mathbf{G}[1,1]$, $\sigma^2 = \mathbf{G}[2,2]$ and $\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\Sigma_1[j,k]$ corresponds to the sum of the elements of the matrix $\mathbf{Z}_1\mathbf{G}\mathbf{Z}_1 + \sigma^2 I$. Similarly, the population parameter for MIC_2 was computed as $1 - \frac{n_1^2\tau_0^2}{\left(\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\Sigma_1[j,k]\right)}$. Table 3.2 summarizes the six different data sets that were used in the simulation.

Using the SAS System version 8.2 (SAS Institute, Cary, NC) for each of the six data sets simulated, we fitted the full model, an overfitted model, and two types of reduced models in the fixed effect. For the overfitted model, two randomly generated variables not linked to the outcome were included in the model. Two reduced models were fitted by removing one or two of the binary variables from the fixed effects. The “true” random effect that was simulated was used in each model. That is, the random effects consisted of an intercept and age with either a diagonal covariance matrix or an unstructured covariance matrix. From the models fitted, we computed each of the four statistics discussed in Section 3.4.

In Tables 3.3 and 3.4, we give the values of means and ranges for the proposed R^2 statistics. Table 3.3 is for the data sets with a diagonal covariance matrix and Table 3.4 is for the data sets with an unstructured covariance matrix. For the statistics MEC_1 and MEC_2 negative values are observed for high values of within-subject variance for the data sets with unstructured covariance matrix. Negative values are also observed for the statistics MIC_1 and MIC_2 due to negative values in the off diagonals of the covariance structure for $\hat{\mathbf{G}}$. If $\mathrm{tr}(\hat{\Sigma}_i)$ as opposed to $\hat{\Sigma}_i$ is used in the computations of MIC_1 and MIC_2, values between 0 and 1 that are closer to the population parameters are obtained (refer to Table 3.4). MEC_1 seems to be a better estimate for the population parameter that MIC_1 is supposed to estimate. For the full models and for data with diagonal covariance matrix, estimates of MEC_1 are closer to the population parameter of MIC_1.

Values from the overfitted model do not appear to be different from the values from the full model, indicating that all the statistics are able to discriminate between a model that includes only variables linked to the outcome and one that includes additional variables exhibiting a spurious relationship with the outcome. On the other hand, values from the reduced models appear to be significantly lower than the full model, demonstrating that the statistics can be useful in choosing the most parsimonious model. The more variables that are removed from the full model, the larger the decrease is in the value of the statistics. These R^2 statistics decrease when σ^2 increases, which could be an indication that they are able to account for the fact that less of the variability in the outcome is contributed by the variables in the model when σ^2 is increased. In particular, for $\sigma^2 = 250$, there is little difference in the values for the full and other models, suggesting that all of the models are inadequate given the large within-subject variability. Also, values of the statistics from the data sets with unstructured covariance structure are overall lower compared to the data sets with diagonal covariance matrix structure, which could be explained by the fact that in the data sets with unstructured covariance (where the correlation between observations from the same subject is increased), the model is less efficient.

3.6 Example

We analyzed the data from Pothoff and Roy [15] on dental distance. These data have been analyzed in the context of the LMM by several authors and used as example in [26] in computing pseudo-R^2. We fitted several models to investigate the relationship between dental distance and the explanatory variables age, gender, and age-by-gender interaction in the fixed effects. Age and an intercept were used in the random effects with an unstructured covariance matrix. Values of the new R^2 statistics are given for three models in Table 3.6. For all three models, the values of the measures of external consistency statistics are less than 0 suggesting that these models are not better than a null model. Values of measures of internal consistency variables were also negative.

In Figure 1, we plot the outcome as a function of age. While the figure gives an indication of a possible age and gender effect, there is a lot of within-subject variability and dental distance as a function of age appears to vary considerably from one individual to the next. In Table 3.6, parameter estimates and standard errors show that the magnitude of the effect of the fixed effect variables is small. Although the p-value for age is very small, the 95% confidence interval for the slope is only between 0.28 and 0.68. For the age-by-gender interaction which is also significant, the confidence interval is between 0.04 and 0.26. In contrast, Table 3.7 shows how the between-subject variability of 4.56 dwarfs all other variance components. Also, Table 3.8 shows that most of the variation in the outcome tends to be between subjects. Hence, although marginal-R^2 computed by Orelien and Edwards [14] seems to indicate that the model with age and gender is the most parsimonious model, we believe in light of the new R^2 values and an examination of the fixed parameter estimates and covariance parameters that an equally valid conclusion is that the effect of any of the covariates is small and most of the variation in the data is within and between subjects.

3.7 Discussion

Results from our simulation show that the statistics perform adequately in being able to detect the most parsimonious fixed effects model. When it comes to estimating the proportion of variation explained by the fixed effects in the model, these statistics do not perform as well. Because the range is not well defined in that negative values can be obtained, they may not be considered to be valid R^2 statistics based on the criterion developed by Kvalseth [10] that the range should be well defined. However, because negative values were associated with models with large within-subject variance or models in which important covariates were missing, these negative values do not call into question the usefulness of the statistics. MEC_1 as opposed to MIC_1 and MEC_2 as opposed to MIC_2 gave estimates that were closest to the population parameter for the proportion of variation explained by the model. We believe that this is not a simple coincidence. The denominators of these four statistics (MEC_1, MEC_2, MIC_1 and MIC_2) are all estimates of the average variance of an observation. When $\text{tr}(\hat{\Sigma}_i)$ as opposed to $\hat{\Sigma}_i$ is used in the formula for MIC_1 or MIC_2 in the case where the covariance of the random effects ($\mathbf{G}$) is unstructured, positive values closer to the population parameters are obtained. Hence, substituting $\mathrm{tr}(\hat{\Sigma}_i)$ for $\hat{\Sigma}_i$ should be given consideration in the computation of MIC_1 or MIC_2.

The statistics MEC_1 and MEC_2 depend on the choice of the null model. While we believe that null model (3) is a reasonable choice in the LMM, other choices could be considered by an analyst. In addition to null model (4), another choice of null model for MEC_1 and MEC_2 is to use a model with a fixed effect intercept but with the same covariance structure as the model of interest. The use of such a null model for MEC_1 and MEC_2 could potentially lead to estimates that are closer to the population parameters of MIC_1 and MIC_2 respectively.

Orelien and Edwards [14] showed that many of the statistics that have been proposed (Vonesh et al. [23], Vonesh and Chinchilli [22], Xu [25] and Zheng [26]) in the statistical literature for the LMM were inadequate in that they were unable to distinguish between a full model and one from which important covariates were missing. The statistics that we propose do not suffer from such deficiencies. Also, for many of the statistics that were proposed there is an implicit or explicit assumption of conditional independence such as the ones proposed by Xu [25]. In this chapter, we have proposed formulae for both the general case and the specific case where conditional independence can be assumed. Snijders and Bosker [19] proposed two statistics that were intended to measure the proportion of variation explained by the covariates (although the authors refer to their statistics as measuring the proportion of modeled variance). One of the statistics assesses the proportion of variation explained by the fixed effects and the other explains the proportion of variation due to random effects. While the definition given by Snijders and Bosker [19] for the statistic to assess the proportion of variation $(1-\frac{\mathrm{var}(Y_{ij}-X_{ij}\beta)}{\mathrm{var}(Y_{ij})})$ seems to correspond with our definition of MEC_2, we disagree with the estimates given by the authors. We believe that the estimates that we proposed for the proportion of variation are more complex and reflect the true variance of Y_{ij}.

An analyst might be tempted to use traditional GOF statistics such as the AIC, BIC, AICC, or a likelihood ratio test in assessing the adequacy of the fixed effect components or in choosing the most parsimonious one between two models. However, the use of AIC, BIC, or AICC may be inappropriate when restricted maximum likelihood (REML) has been used in the estimation (Verbeke and Molenbergh, 2000) [21]. Similarly, Whelham and Thompson [14] noted that the log-likelihood ratio test may not be valid under REML. Hence, the new statistics that we propose could be valuable tools for an analyst in assessing model adequacy. Although Gurka (2006) [7] concluded that the AIC, BIC, and AICC could be used under REML for selecting the most parsimonious model in the fixed effects, the author failed to show what magnitude of difference between two models in the values of these statistics constitutes a meaningful difference.

3.8 Conclusion

Based on the results of the simulations, we recommend the use of these newly proposed statistics as they can be useful to the analyst to ascertain whether the cross-sectional covariates are adequate to explain the variation in the data or to select the most parsimonious models among competing ones. Ideally, an adequate model will have high values of external and internal consistency. Based on our simulations, the measures of external consistency and internal consistency may not always agree except in the case where the model is fully specified and the between- and within-subject random errors are relatively small. We recommend that the analyst compute at least one measure of external consistency and one measure of internal consistency. Specifically, we prefer MEC_1 and MIC_1 over their respective counterparts. We believe that in the LMM, a better case can be made for the variance not accounted by the model to be the numerator of MEC_1 or MIC_1 (as opposed to the numerator of MEC_2 or MIC_2). The concept of external and internal measures of consistency that we introduce could be extended to other classes of models beyond the LMM such as non-linear mixed models or any other types of hierarchical models. Future research needs to assess the ability of these statistics to explain adequacy of the random effects.

References

1. Akaike, H., 1974. A new look at the statistical model identification. IEEE Transaction on Automatic Control, AC-19, 716-723.

2. Cnaan, A., Laird, N.M., and Slasor, P., 1997. Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. Statistics in Medicine, 16(20) 2349-2380.

3. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20, 37 -46.

4. Dawson, K. S., Gennings, C., and Carter, W. H. (1997). Two Graphical Techniques Useful in Detecting Correlation Structure in Repeated Measures Data. American Statistician, 51, 275-283.

5. Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). Analysis of Longitudinal Data, New York, NY: Oxford University Press.

6. Grady, J. J., and Helms, R. W. (1995). Model Selection Techniques for The Covariance Matrix For Incomplete Longitudinal Data. Statistics in Medicine, 14, 1397-1416.

7. Gurka, M.J. (2006). Selecting the Best Linear Model under REML. The American Statistician, 60(1) 19-26.

8. Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72(358) 320-340.

9. Jennrich, R. I., and Schluchter, M. D. (1986). Unbalanced Repeated-Measures Models With Structured Covariance Matrices. Biometrics, 42, 805-820.

10. Kvalseth, T.O., 1985. Cautionary note about R^2. The American Statistician, 39(4) 279-285.

11. Laird, N. and Ware, J.H., 1982. Random effect models for longitudinal data. Biometrics, 38(4) 963-974.

12. Lin, L.I., 1989. A concordance correlation-coefficient to evaluate reproducibility. Biometrics, 45(1) 255-268.

13. Liu, H., Weiss, R. E., Jennrich, R. I., and Wenger, N. S. (1999). PRESS Model Selection in Repeated Measures Data. Computational Statistics and Data Analysis, 30, 169-184.

14. Orelien, J. G. and Edwards L.J., 2007. Fixed Effect Variable Selection in Linear Mixed Models Using R^2 Statistics. Computational Statistics and Data Analysis, to be published.

15. Pothoff, R.F. and Roy, S.N., 1964. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51 313-326.

16. Pourahmadi, M. (2002). Graphical Diagnostics for Modeling Unstructured Covariance Matrices. International Statistical Review, 70, 395-417.

17. Schaalje, B., Zhang, J., Pantula, S. G., and Pollock, K. H. (1991). Analysis of Repeated-Measurements Data From Randomized Block Experiments. Biometrics, 47, 813-824.

18. Shwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics, 6(2) 461-464.

19. Snijders T.A. and Bosker R.J., 1994. Modeled Variance in Two-Leveled Models. Sociolgoical Methods and Research, 22 (3) 342-363.

20. Verbeke G. and Fieuws S. The effect of miss-specified baseline characteristics on inference for longitudinal trends in linear mixed models. Biostatistics, advanced publication on line March 23, 2007.

21. Verbeke, G. and Molenberghs, G., 2000. Linear Mixed Models For Longitudinal Data. Springer-Verlag, New York.

22. Vonesh, E.F., Chinchilli, V.M., 1997. Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker, New York, 1997, 419-424.

23. Vonesh, E.F., Chinchilli, V.M., Pu, K.W., 1996. Goodness-of-fit in generalized nonlinear mixed-effects models. Biometrics, 52 572-587.

24. Welham, SJ, Thompson, R. A Likelihood Ratio Test for Fixed Model Terms Using Residual Maximum Likelihood. Journal of the Royal Statistical Society, Series B. 1997; 59 (3) 701-714

25. Xu, R.H., 2003. Measuring explained variation in linear mixed effects models. Statistics in Medicine, 22(22) 3527-3541.

26. Zheng, B.Y., 2000. Summarizing the goodness of fit of generalized linear models for longitudinal data. Statistics in Medicine, 19(10) 1265-1275.

27. Zimmerman, D. L. (2000). Viewing the Correlation Structure of Longitudinal Data through a PRISM. American Statistician, 54, 310-318.

Appendix

We give an example to demonstrate that the elements in the expression of $\operatorname{var}\left(\sum_{i=1}^{n_i}\mathbf{y}_i\right)$ that does not depend on the variables in the model is $\sum_{j=1}^{n_i} n_i\sigma^2 + n_i^2\tau_0^2$ as shown by (7). Assume that $n_i = 3$ and there are 2 variables in the random effect, so that d_i is 2x2. For any subject i, $\operatorname{var}(\mathbf{y}_i)$ can be written as:

$$\operatorname{var}(\mathbf{y}_i) = \begin{pmatrix} 1 & z_{12} \\ 1 & z_{22} \\ 1 & z_{32} \end{pmatrix} \begin{pmatrix} \tau_0^2 & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ z_{12} & z_{22} & z_{32} \end{pmatrix} + \sigma^2\mathbf{I}_{3\times 3} =$$

$$\begin{pmatrix} \tau_0^2 + a_{12}z_{12} & a_{12} + z_{12}a_{22} \\ \tau_0^2 + a_{12}z_{22} & a_{12} + z_{22}a_{22} \\ \tau_0^2 + a_{12}z_{32} & a_{12} + z_{32}a_{22} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ z_{12} & z_{22} & z_{32} \end{pmatrix} + \sigma^2\mathbf{I}_{3\times 3} =$$

$$\begin{pmatrix} \tau_0^2 + 2a_{12}z_{12} + z_{12}^2a_{22} & \tau_0^2 + a_{12}z_{12} + a_{12}z_{22} + z_{22}z_{12}a_{22} & \tau_0^2 + a_{12}z_{12} + a_{12}z_{32} + z_{12}z_{32}a_{22} \\ \tau_0^2 + a_{12}z_{22} + a_{12}z_{12} + z_{12}z_{22}a_{22} & \tau_0^2 + 2a_{12}z_{22} + z_{22}^2a_{22} & \tau_0^2 + a_{12}z_{22} + a_{12}z_{32} + z_{22}z_{32}a_{22} \\ \tau_0^2 + a_{12}z_{12} + a_{12}z_{32} + z_{12}z_{32}a_{22} & \tau_0^2 + a_{12}z_{22} + a_{12}z_{32} + z_{22}z_{32}a_{22} & \tau_0^2 + 2a_{32}z_{12} + z_{32}^2a_{22} \end{pmatrix}$$

$$+\sigma^2\mathbf{I}_{3\times 3} = \tau_0^2 \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} + \sigma^2\mathbf{I}_{3\times 3} +$$

$$\begin{pmatrix} 2a_{12}z_{12} + z_{12}^2a_{22} & a_{12}z_{12} + a_{12}z_{22} + z_{22}z_{12}a_{22} & a_{12}z_{12} + a_{12}z_{32} + z_{12}z_{32}a_{22} \\ a_{12}z_{22} + a_{12}z_{12} + z_{12}z_{22}a_{22} & 2a_{12}z_{22} + z_{22}^2a_{22} & a_{12}z_{22} + a_{12}z_{32} + z_{22}z_{32}a_{22} \\ a_{12}z_{12} + a_{12}z_{32} + z_{12}z_{32}a_{22} & a_{12}z_{22} + a_{12}z_{32} + z_{22}z_{32}a_{22} & 2a_{32}z_{12} + z_{32}^2a_{22} \end{pmatrix}$$

$$=$$

$$\begin{pmatrix} \tau_0^2 + \sigma^2 & \tau_0^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 + \sigma^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 & \tau_0^2 + \sigma^2 \end{pmatrix} +$$

$$\begin{pmatrix} 2a_{12}z_{12} + z_{12}^2a_{22} & a_{12}z_{12} + a_{12}z_{22} + z_{22}z_{12}a_{22} & a_{12}z_{12} + a_{12}z_{32} + z_{12}z_{32}a_{22} \\ a_{12}z_{22} + a_{12}z_{12} + z_{12}z_{22}a_{22} & 2a_{12}z_{22} + z_{22}^2a_{22} & a_{12}z_{22} + a_{12}z_{32} + z_{22}z_{32}a_{22} \\ a_{12}z_{12} + a_{12}z_{32} + z_{12}z_{32}a_{22} & a_{12}z_{22} + a_{12}z_{32} + z_{22}z_{32}a_{22} & 2a_{32}z_{12} + z_{32}^2a_{22} \end{pmatrix}$$

Table 3.1 Summary of the R^2 statistics proposed in this paper

Statistic	**Description**	**General Formula**	**Formula assuming conditional independence**
MEC_1	Measure of external consistency (1)	$1-\frac{\sum_{i=1}^{n}\left(n_i^2\hat{\tau}_0^2+\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\mathbf{R}}_i[j,k]\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_{0i}[j,k]\right)}$	$1-\frac{\sum_{i=1}^{n}\left(n_i^2\hat{\tau}_0^2+n_i\hat{\sigma}^2\right)}{\mathrm{N}(\hat{\sigma}_0^2+\hat{\tau}_{00}^2)}$
MEC_2	Measure of external consistency (2)	$MEC_2=1-\frac{\hat{\tau}_0^2\sum_{i=1}^{n}\left(n_i^2\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_{0i}[j,k]\right)}$	$1-\frac{\hat{\tau}_0^2\sum_{i=1}^{n}\left(n_i^2\right)}{\mathrm{N}(\hat{\sigma}_0^2+\hat{\tau}_{00}^2)}$
MIC_1	Measure of internal consistency (1)	$1-\frac{\sum_{i=1}^{n}\left(n_i^2\hat{\tau}_0^2+\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\mathbf{R}}_i[j,k]\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_i[j,k]\right)}$	$1-\frac{\sum_{i=1}^{n}n_i\hat{\sigma}^2+n_i^2\hat{\tau}_0^2}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_i[j,k]\right)}$
MIC_2	Measure of internal consistency (2)	$1-\frac{\hat{\tau}_0^2\sum_{i=1}^{n}n_i^2}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_i[j,k]\right)}$	$1-\frac{\hat{\tau}_0^2\sum_{i=1}^{n}n_i^2}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_i[j,k]\right)}$

Table 3.2 Description of the data sets used in the simulation

Data Set	Value of σ^2	Proportion of variation explained by the cross-sectional covariates (Population Parameter for MIC_1)	Proportion of variation explained by the cross-sectional covariates (Population parameter for MIC_2)	Type of Covariance Matrix for the random effects
1	12	0.89	0.92	Diagonal, $\mathbf{G} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$
2	45	0.82	0.93	
3	250	0.50	0.96	
4	12	0.92	0.94	Unstructured, $\mathbf{G} = \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}$
5	45	0.86	0.94	
6	250	0.56	0.96	

Table 3.3 Average and Interquartile range for the proposed R^2 using only replicates where Hessian matrix and covariance matrix of random effects are positive definite (data sets with diagonal covariance matrix for the random effects)

σ^2	Model	Average MEC_1 (Q25, Q75)	Average MEC_2 (Q25, Q75)	Average MIC_1 (Q25, Q75)	Average MIC_2 (Q25, Q75)
12	Overfitted Model: Age, Gender, Treatment and 2 variables not related to outcome	0.90 (0.87, 0.95)	0.92 (0.89, 0.97)	0.79 (0.72, 0.89)	0.83 (0.76, 0.93)
12	Full Model: Age, Gender and Treatment	0.90 (0.87, 0.95)	0.92 (0.89, 0.97)	0.79 (0.72, 0.89)	0.83 (0.76, 0.92)
12	Reduced Model 1: Age and Gender	0.67 (0.60, 0.75)	0.69 (0.61, 0.76)	0.55 (0.45, 0.65)	0.58 (0.48, 0.68)
12	Reduced Model 2: Age	0.40 (0.31, 0.49)	0.42 (0.33, 0.50)	0.40 (0.32, 0.49)	0.42 (0.34, 0.50)
45	Overfitted Model: Age, Gender, Treatment and 2 variables not related to outcome	0.80 (0.73, 0.88)	0.86 (0.80, 0.94)	0.58 (0.44, 0.74)	0.71 (0.58, 0.88)
45	Full Model: Age, Gender and Treatment	0.80 (0.73, 0.88)	0.86 (0.80, 0.94)	0.58 (0.44, 0.75)	0.71 (0.58, 0.88)
45	Reduced Model 1: Age and Gender	0.62 (0.51, 0.75)	0.69 (0.57, 0.81)	0.49 (0.34, 0.66)	0.58 (0.42, 0.75)
45	Reduced Model 2: Age	0.40 (0.25, 0.54)	0.46 (0.32, 0.61)	0.40 (0.25, 0.54)	0.46 (0.32, 0.61)
250	Overfitted Model: Age, Gender, Treatment and 2 variables not related to outcome	0.57 (0.49, 0.66)	0.84 (0.76, 0.93)	0.29 (0.16, 0.42)	0.74 (0.61, 0.88)
250	Full Model: Age, Gender and Treatment	0.57 (0.49, 0.66)	0.84 (0.76, 0.93)	0.29 (0.16, 0.43)	0.74 (0.62, 0.88)
250	Reduced Model 1: Age and Gender	0.46 (0.34, 0.60)	0.74 (0.61, 0.87)	0.34 (0.18, 0.50)	0.67 (0.52, 0.84)
250	Reduced Model 2: Age	0.35 (0.18, 0.52)	0.62 (0.46, 0.79)	0.35 (0.18, 0.52)	0.62 (0.46, 0.79)

Table 3.4 Average and Interquartile range for the proposed R^2 using only replicates where Hessian matrix and covariance matrix of random effects are positive definite (data sets with unstructured covariance matrix for the random effects)

σ^2	**Model**	**Average MEC_1 (Q25, Q75)**	**Average MEC_2 (Q25, Q75)**	**Average MIC_1 (Q25, Q75)**	**Average MIC_2 (Q25, Q75)**	**Average MIC_1 (Q25, Q75) using $\text{tr}(\hat{\Sigma}_i)$ in the formula instead of $\hat{\Sigma}_i$**
12	Overfitted Model: Age, Gender, Treatment and 2 variables not related to outcome	0.82 (0.75, 0.91)	0.83 (0.77, 0.92)	0.64 (0.52, 0.82)	0.67 (0.55, 0.85)	0.79 (0.72, 0.89)
12	Full Model: Age, Gender and Treatment	0.82 (0.75, 0.91)	0.83 (0.77, 0.92)	0.64 (0.52, 0.82)	0.67 (0.56, 0.85)	0.79 (0.72, 0.89)
12	Reduced Model 1: Age and Gender	0.63 (0.52, 0.75)	0.64 (0.54, 0.77)	0.51 (0.37, 0.68)	0.53 (0.39, 0.70)	0.55 (0.45, 0.65)
12	Reduced Model 2: Age	0.40 (0.27, 0.56)	0.42 (0.29, 0.58)	0.40 (0.27, 0.56)	0.42 (0.29, 0.58)	0.40 (0.32, 0.49)
45	Overfitted Model: Age, Gender, Treatment and 2 variables not related to outcome	0.39 (0.18, 0.71)	0.45 (0.24, 0.76)	-0.15 (-0.56, 0.45)	-0.05 (-0.46, 0.55)	0.58 (0.44, 0.74)
45	Full Model: Age, Gender and Treatment	0.39 (0.18, 0.71)	0.45 (0.24, 0.76)	-0.15 (-0.55, 0.44)	-0.04 (-0.45, 0.55)	0.58 (0.44, 0.75)
45	Reduced Model 1: Age and Gender	0.22 (-0.04, 0.57)	0.27 (0.02, 0.62)	-0.02 (-0.35, 0.44)	0.05 (-0.28, 0.52)	0.49 (0.34, 0.66)

σ^2	**Model**	**Average** MEC_1 **(Q25, Q75)**	**Average** MEC_2 **(Q25, Q75)**	**Average** MIC_1 **(Q25, Q75)**	**Average** MIC_2 **(Q25, Q75)**	**Average** MIC_1 **(Q25, Q75) using** $\mathrm{tr}(\hat{\Sigma}_i)$ **in the formula instead of** $\hat{\Sigma}_i$
45	Reduced Model 2: Age	0.02 (-0.27, 0.40)	0.07 (-0.22, 0.45)	0.02 (-0.27, 0.40)	0.07 (-0.22, 0.45)	0.40 (0.25, 0.54)
250	Overfitted Model: Age, Gender, Treatment and 2 variables not related to outcome	-1.65 (-2.58, -0.29)	-1.40 (-2.33, -0.04)	-3.22 (-4.64, -1.03)	-2.83 (-4.26, -0.63)	0.29 (0.16, 0.42)
250	Full Model: Age, Gender and Treatment	-1.63 (-2.56, -0.27)	-1.38 (-2.31, -0.02)	-3.19 (-4.60, -1.01)	-2.80 (-4.21, -0.61)	0.29 (0.16, 0.43)
250	Reduced Model 1: Age and Gender	-1.72 (-2.65, -0.33)	-1.47 (-2.39, -0.09)	-2.33 (-3.45, -0.62)	-2.03 (-3.14, -0.33)	0.34 (0.18, 0.50)
250	Reduced Model 2: Age	-1.87 (-2.84, -0.44)	-1.62 (-2.59, -0.19)	-1.87 (-2.84, -0.44)	-1.62 (-2.59, -0.19)	0.35 (0.18, 0.52)

Table 3.5 Pseudo-R^2 for dental data of Potthoff and Roy

Model	MEC_1	MEC_2	MIC_1	MIC_2
Age, Gender, AgexGender	-0.04	0.05	-0.42	-0.3
Age, Gender	-0.55	-0.46	-1.12	-0.99
Age	-0.09	-0.003	-0.09	-0.003

Table 3.6 Parameter estimates for dental data of Potthoff and Roy

Parameter	Estimate	Standard Error	Pvalue
Intercept	17.37	1.18	< 0.000
Age	0.47	0.10	< 0.000
Gender	-1.03	1.54	0.5043
AgexGender	0.30	0.13	0.0224

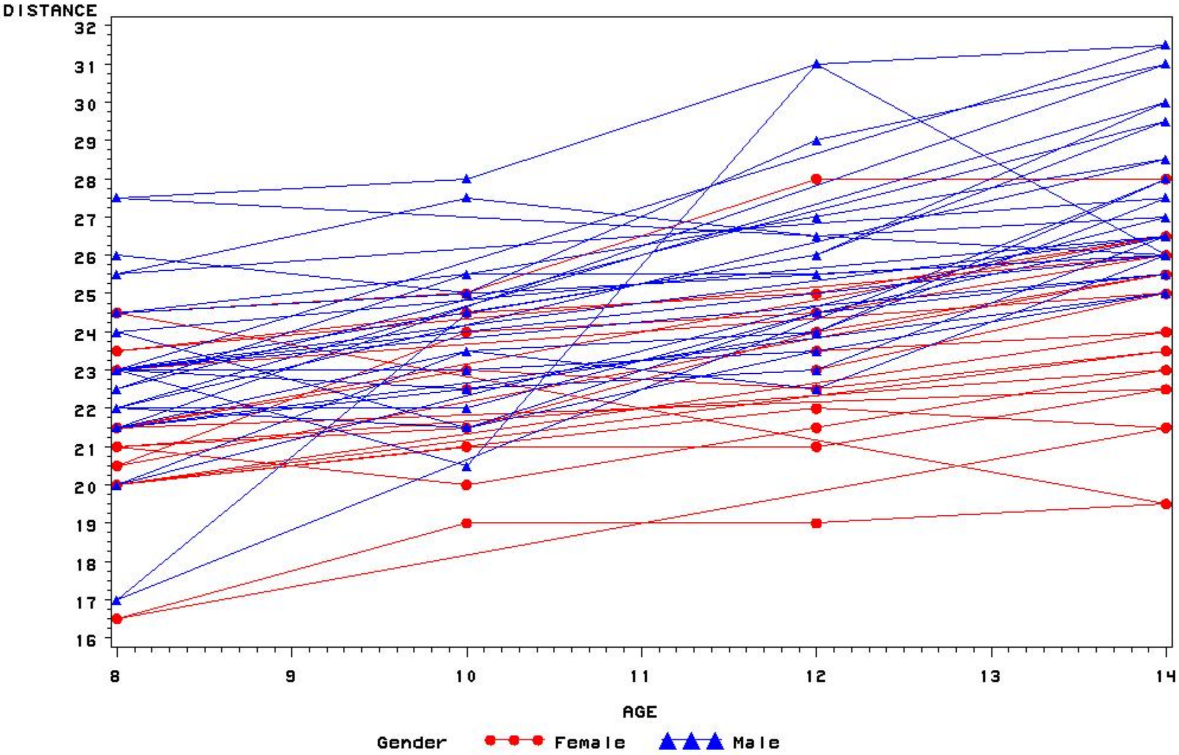
Table 3.7 Covariance Parameter Estimates for dental distance data (full model)

Covariance Parameter	Estimate	Standard Error
UN (1, 1)	4.56	4.67
UN (2, 1)	-0.19	0.38
UN (2, 2)	0.024	0.03
Residual	1.72	0.33

Table 3.8 Distribution of dental distance by age and gender

AGE	Gender	N	Std Dev	Minimum	Maximum
8	**Female**	11	2.12	16.50	24.50
	Male	16	2.45	17.00	27.50
10	**Female**	11	1.90	19.00	25.00
	Male	16	2.14	20.50	28.00
12	**Female**	11	2.36	19.00	28.00
	Male	16	2.65	22.50	31.00
14	**Female**	11	2.44	19.50	28.00
	Male	16	2.09	25.00	31.50

Figure 1 Graph of Distance versus Age



4 Performance of Pseudo-R^2 Statistics in Detecting Misspecification in the Random Effects in Linear Mixed Models

Abstract

Orelien and Edwards [17] showed through simulations that R^2 statistics that can be classified as marginal, that is, the random effects are excluded from the computation of the residuals—as opposed to conditional ones where the random effects are included in the computation of the residuals—are useful in assessing the adequacy of the fixed effects. In Chapter 3, additional R^2 statistics were introduced to assess adequacy of the fixed effect terms. It is obvious that marginal statistics would not be useful in assessing the random effect function, as the values for a marginal statistic would be the same for two models having the same fixed effects but a different random function. On the other hand, in addition to being useful in assessing the adequacy of the fixed effect components, the R^2 statistics introduced in Chapter 3 have the potential to be able to identify adequacy of the random effects function as well. In this chapter we employ simulations to investigate the ability of MEC_1 and MIC_1 (statistics that were introduced in Chapter 3) to detect misspecification in the random effects when the true fixed effects model is known. These statistics—which measure respectively external and internal consistencies—were chosen because they performed best out of the statistics evaluated in Chapter 3. Our simulations are limited to longitudinal studies where the within-subject variance is homogeneous. The results of our simulations show that both

statistics are able to discriminate between models where the random effects contain a time variable and models that do not, such as a model with only a random intercept term. When comparing two competing models that contain a time variable in the random effects such as a reduced model with intercept and linear term for time and a full model that contains a quadratic term in addition to the linear term for time, the statistics in our simulations were unable to detect that the full model was the true one. Additional analysis seems to indicate that in the LMM for longitudinal data even when the true model of the random effects involves variables (polynomial components) beyond the linear term, the reduced model with an intercept and a linear term for time may be as good as the full model.

4.1 Introduction

In the LMM, misspecification of the random effects can lead to biased estimates of the variance of the fixed effect parameters (Heagerty and Kurland [11]). In selecting the random effects for a mixed model, an approach that would come under consideration is the likelihood ratio test (LRT). However, Stram and Lee [19] show that the LRT in this case is a mixture of Chi-square. Lin [15] and Hall and Praestgaard [9] suggested that score tests be used for testing simultaneously that all of the random effects are 0. Albert and Chib [2] proposed an approach for testing whether a single random effect is 0. Chen and Dunson [4] generalized this approach for selecting the best combination of random effects using Gibbs sampling. A major limitation of the method proposed by Chen and Dunson [4] is that it does not lend itself to practical use because of the computations involved. First, software for performing Gibbs sampling is not readily available. Second, depending on the data, convergence may not be achieved. Finally, different values might be obtained depending on the sample size used by an analyst.

Gurka [8] investigated the performance of information criteria such as the AIC [1] and BIC [20] in detecting misspecification of the random effects in the linear mixed model through simulations. There were several limitations to the simulations. The effect of the size of the variance of the random effects was not taken into consideration. In determining whether the correct model was selected, the magnitude of the difference between the information criteria of the models to be compared was not taken into account. Also, in the simulations performed by Gurka [8], the full model in the random effects consisted of an intercept and a linear component for time.

In many population based studies such as clinical trials or medical studies, repeated observations are taken on subjects over time. In most instances, the interest in these studies is to determine the effect of a treatment or an intervention while accounting for the correlation within subjects. For example, Littell et al. [16] gives the example of a clinical trial where the effect of a drug on pulmonary function as measured by FEV1 (forced expiratory volume in 1 second) is investigated. Patients in the study were assigned to 3 treatment groups and measurements of FEV1 were taken at 4 time intervals. Cnaan et al. [3] indicated that often in these types of studies the use of a random intercept and random slope is sufficient to characterize the random effects structure. However, there may be instances where the analyst may judge, based on exploratory analysis or previous experience with the data, that a more complex random effects structure is needed such as one that includes a quadratic term. For longitudinal data analysis, it may be rare to use terms beyond a quadratic component in the random effects. Thus, in terms of competing models for the random effects, in longitudinal data analysis, the considerations are often: i) random intercept, linear effect for time and a quadratic effect for time (full model); ii) random intercept and linear effect for time (reduced model 1); and iii) a random intercept (reduced model 2).

One approach might be to use pseudo-R^2 statistics or other tools that were proposed in Chapter 3 to determine the appropriate random effects structure. But although we have shown that the R^2 statistics discussed in chapter 3 were useful in assessing the adequacy of the fixed effects, their performance in determining the adequacy of the random effects function has not been demonstrated. The advantage of using R^2 statistics over other tools, such as AIC or BIC is that R^2 statistics could provide additional information such as how much of the variation in the outcome is explained by the inclusion or exclusion of a random effect.

There is a direct connection between selection of the random effects and covariance structure for the random effects. For example, choosing between whether the random effects should be a) intercept, linear term for time and a quadratic effect for time (full model) or b) intercept and linear term for time could be translated as to whether the covariance structure for the random effects is of the form $\mathbf{G}_{full} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$ or $\mathbf{G}_{reduced} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$. This is a different question with respect to the covariance structure of the within-subject error term which is often assumed to be of the form $\sigma^2\mathbf{I}$, that is, conditioned on the random effects, the within-subject errors are independent (conditional independence). For assessing covariance of the random error term, there are several issues that need to be taken into considerations, such as cases where the misspecified covariance structure is nested in the true one or convergence issues that may be due to overparameterization. Thus, as in previous chapters, the use of R^2 statistics for assessing the adequacy of the covariance structure of the error term is beyond the scope of our study. For assessing adequacy of the covariance structure, graphical exploratory techniques for selecting the covariance structure, such as those proposed by Diggle et al. [5], Grady and Helms [7] Dawson et al. [5], Zimmerman [26] and Pourahmadi [18] could be used.

This chapter is organized as follows: In Section 4.2, we formulate the LMM and notation. Formulae for the statistics that are reviewed in this chapter are given in Section 4.3. The generation of the simulated data and choice of parameters are discussed in Section 4.4. Results are presented in Section 4.5. An example is given in Section 4.6. We discuss results from the simulation and the example in Section 4.7. Concluding remarks in Section 4.8.

4.2 The Linear Mixed Model

The simulations are based on model (1) below as formulated by Harville [10] and Laird and Ware [14]:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \qquad (1)$$

where $i \in$ {1, 2, ..., n} is the index for the independent sampling units (ISU) and

$\mathbf{y}_i$ is an $n_i \times 1$ vector of observations from the i^{th} independent sampling unit (subject),

$\mathbf{X}_i$ denotes an $n_i \times p$ fixed effects design matrix for the i^{th} subject,

$\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown, constant, fixed effect parameters,

$\mathbf{Z}_i$ denotes an $n_i \times q$ random effects design matrix for the i^{th} subject,

$\mathbf{b}_i$ is a $q \times 1$ vector of unobservable random effects for the i^{th} subject, and

$\mathbf{e}_i$ denotes an $n_i \times 1$ vector of unobservable within-subject error terms.

It is also assumed that $\mathbf{b}_i$ has a multivariate normal distribution $N_q(\mathbf{0},\mathbf{G})$ independent of $\mathbf{e}_i$, which has a multivariate distribution $N_{n_i}(\mathbf{0},\mathbf{R}_i)$.

$E\begin{bmatrix}\mathbf{b}_i\\ \mathbf{e}_i\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$ and $V\begin{bmatrix}\mathbf{b}_i\\ \mathbf{e}_i\end{bmatrix}=\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_i\end{bmatrix}$,

Where $\mathbf{G}$ is a $q\times q$ unknown covariance matrix for the random effects and $\mathbf{R}_i$ is an $n_i\times n_i$ unknown covariance matrix for the within-subject error terms for the i^{th} subject. With these assumptions, for the i^{th} subject we have $\mathbf{\Sigma}_i = V(\mathbf{y}_i)=\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i'+\mathbf{R}_i$. In many applications, $\mathbf{R}_i$ is taken to be $\sigma^2\mathbf{I}_{n_i}$, known as the conditional independence assumption for the within-subject error term [14].

4.3 Pseudo-R^2 Statistics

In this chapter, we are assessing the performance of two statistics that were introduced in chapter 3. In that discussion, a distinction was made between marginal and conditional R^2 statistics. Marginal R^2 statistics are those that involve the residuals and for which the random effect components are excluded in the computation of these residuals. These marginal R^2 statistics are obviously not useful for assessing the adequacy of the random effect terms as their values would not change from a full to a reduced model in the random effects so long as the fixed effect terms are the same. As a result, for assessing the adequacy of the random effects, we are considering only the new statistics that were proposed in Chapter 3. Specifically, we focus on two of the R^2 statistics, MEC_1 and MIC_1, that were shown to have the best performance. Because conditional R^2 statistics proposed by Vonesh et al. [22], Vonesh and Chinchilli [21], Zheng [25] and Xu [24] evaluated in Chapter

2 did not perform well for assessing the adequacy of the fixed effects, they were excluded from consideration in this review.

MEC_1 is given by the formula in (2):

$$MEC_1 = 1 - \frac{\sum_{i=1}^{n}\left(n_i^2\hat{\tau}_0^2 + \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\mathbf{R}}_i[j,k]\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_{0i}[j,k]\right)} \tag{2}$$

Where $\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\mathbf{R}}_i[j,k]$ is the sum of all the elements of the matrix $\hat{\mathbf{R}}_i$ (for subject i) and

$\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_{0i}[j,k]$ is the sum of all the elements of the matrix $\hat{\Sigma}_{0i}$, the covariance matrix for subject i under a null model given by:

$$\mathbf{y}_i = \mathbf{1}_{n_i}\beta_0 + \mathbf{1}_{n_i}b_{0i} + \mathbf{u}_i\,, \tag{3}$$

where β_0 is an unknown fixed parameter, b_{0i} is a random intercept for subject i, and $\mathbf{u}_i$ is the unobservable within-subject random error term for the model (*i.e.*, the null model consists of fixed and random effect intercepts). We define $V(\mathbf{u}_i) = \sigma_0{}^2\mathbf{I}_{n_i}$ and $V(b_{0i}) = \tau_{00}^2$.

The formula for MIC_1 is given by formula (4).

$$MIC_1 = 1 - \frac{\sum_{i=1}^{n}\left(n_i^2\hat{\tau}_0^2 + \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\mathbf{R}}_i[j,k]\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_i[j,k]\right)} \tag{4}$$

where $\hat{\Sigma}_i = \mathbf{Z}_i\hat{\mathbf{G}}\mathbf{Z}_i + \hat{\sigma}^2 I$ and $\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\hat{\Sigma}_i[j,k]\right)$ is the sum of all the elements of $\hat{\Sigma}_i$.

If $\mathbf{R}_i = \sigma^2\mathbf{I}$ then the expressions for MEC_1 and MIC_1, can be simplified as follows:

$$MEC_1 = 1 - \frac{\sum_{i=1}^{n}\left(n_i^2\widehat{\tau}_0^2 + n_i\sigma^2\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\Sigma}_{0i}[j,k]\right)} \qquad \textbf{(5)}$$

$$MIC_1 = 1 - \frac{\sum_{i=1}^{n}\left(n_i^2\widehat{\tau}_0^2 + n_i\sigma^2\right)}{\sum_{i=1}^{i=n}\left(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\widehat{\Sigma}_{i}[j,k]\right)} \qquad (6)$$

The range for both MEC_1 and MIC_1 is $(-\infty, 1)$. For MIC_1, although the numerator is part of the denominator, the statistic can be negative because of negative elements in the off-diagonal elements of $\widehat{\mathbf{R}}_i$.

4.4 Data Generation Techniques

Our simulations were similar to those conducted by Orelien and Edwards [17]. Using the IML module in SAS, we simulated 3 sets of data. In each of the data sets, there were 10,000 replications, 64 subjects, and 6 observations per subject. The random effect for the three data sets consisted of an intercept, random slope and a quadratic term with an unstructured covariance structure. These data sets differed in the values for σ^2 that were 10, 95, and 240. The different values of σ^2 were used to assess how performance of the pseudo-R^2 varied with increased within-subject error, which translates into within-subject correlations of about 0.5, 0.75 and 0.8. The parameters used in the analysis were: $\mathbf{X} = [\mathbf{1}_6, (1,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5)',\ (1,\ 2.25,\ 4,\ 6.25,\ 9,\ 12.25)', \mathbf{1}_6\delta_k, \mathbf{1}_6\delta_l]$;

$\mathbf{Z} = [\mathbf{1}_6, (5, 6, 7, 7.25, 7.5, 7.75)']$, where $\mathbf{1}_6$ is a 6×1 vector of ones and $\delta_k, \delta_l = \{0,1\}$. We took $\boldsymbol{\beta} = \begin{pmatrix} 10 \\ 6 \\ 11 \\ 11 \end{pmatrix}$ and $\mathbf{G} = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 1 & 0.5 \\ 1 & 0.5 & 1.25 \end{pmatrix}$. The population parameter for MIC_1 was computed as $1 - \frac{n_1^2 \tau_0^2 + n_1 \sigma^2}{\left(\sum_{j=1}^{n_1} \sum_{k=1}^{n_1} \Sigma_1[j,k] \right)}$ where $\tau_0^2 = \mathbf{G}[1,1]$, $\sigma^2 = 10$, 95 and 240 and $\sum_{j=1}^{n_1} \sum_{k=1}^{n_1} \Sigma_1[j,k]$ corresponds to the sum of the elements of the matrix $\sum_1 = \mathbf{Z}_1 \mathbf{G} \mathbf{Z}_1 + \sigma^2 I$. Table 1 summarizes the 3 data sets that were used in the simulation.

4.5 Data Analysis and Results

Using the SAS System version 8.2 (SAS Institute, Cary, NC) for each of the simulated data sets, we fitted 3 different models by varying the random effect function. The 3 models for each data set consisted of: a) a full model in the random effects with an intercept, variables for the linear and quadratic components of time; b) a reduced model in the random effects with an intercept and a variable for the linear component of time (reduced model 1); and c) a second reduced model in the random effects with only an intercept term (reduced model 2). For all models, the fixed effects remained the same. The purpose of fitting these models was to ascertain the ability of the R^2 statistics discussed in chapter 3 to identify misspecification in the random effect components.

Results from these analyses are given in Table 2. The results show that overall the most desirable values are obtained for the reduced model that contains the intercept term and the linear term for time (reduced model 1). For the full model, for both MEC_1 and MIC_1, negative values are obtained. Values closest to the population parameter for MIC_1 are obtained for the reduced model with intercept term and age in the random effects (reduced model 1). On the other hand, values for the reduced model with only the random intercept term (reduced model 2) were much lower than the full model. In addition to formula (6), we also computed MIC_1 by substituting $\mathrm{tr}(\hat{\Sigma}_i)$ for $\hat{\Sigma}_i$ (Table 2). Positive values closer to the true population parameter were obtained when $\mathrm{tr}(\hat{\Sigma}_i)$ is used instead $\hat{\Sigma}_i$.

Because the results seem to indicate that reduced model 1 is the best one, we investigated this further by comparing the values and precision of the fixed effect parameters for the full and the reduced models. Table 3 gives average parameter estimates and standard errors for each of the models fit on each data set. There are no discernable differences between full and reduced model with respect to the values of the estimates of the fixed effect parameters; suggesting that estimation of the fixed effects is robust to misspecification of the random effects for this example. Overall, the precision of the estimates is better for the full model, though, the difference between the precision of the parameters for the full model and reduced model 1 is minimal. For the quadratic term for age, slightly better precision is achieved with reduced model 1. In contrast, for the reduced model that consists of only an intercept in the random effect (reduced model 2), there is a remarkable difference in the precision of the estimates compared to that of the other models (full model and reduced model 1). In some instances, the average standard error for the parameter estimates for reduced model 2 is twice as large as that of the other models.

To compare the performance of the 2 statistics versus traditional tools used in assessing GOF, we computed the proportion of times the full versus reduced model 1, the full model versus reduced model 2 and reduced model 1 versus reduced model 2 were identified as being the correct model based on AIC, BIC and LRT. For AIC and BIC, we simply compared values of the statistics within pairs of models. For the LRT, we used the 95% percentile of a Chi-square distribution with the degrees of freedom being the difference in the number of parameters that would need to be estimated for the covariance of the random effects [19]. For example, the number of parameters to be estimated for the covariance matrix of the random effects in the full model is 6, based on $\mathbf{G} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$ and 3 for reduced model 1, based on $\mathbf{G} = \begin{pmatrix} a_{11} \, a_{12} \\ a_{12} \, a_{22} \end{pmatrix}$. Between the full model and reduced model 1, the traditional statistics AIC and BIC do not appear to detect a significant difference except for $\sigma^2 = 10$. Also, for that same value of $\sigma^2 = 10$, the empirical error rate for the LRT is not close to the level $\alpha = 0.05$, which suggests that between two models in the random effects that include a time component, the LRT may be unable to detect a difference even if it exists. For comparing reduced model 2 to the full model or to reduced model 1, the appropriate model is selected over 90% of the time in most cases.

4.6 Example: Schizophrenia Data

Xu [24] presented data from a clinical trial to compare the efficacy of an experimental drug versus a control. The study included a total of 233 patients from 13 sites, and there were three doses for the experimental drug: low, medium, and high. The endpoint of interest was the Brief Psychiatric Rating Scale (BPRS) at baseline, 1, 2, 3, 4, and 6 weeks. Another explanatory variable that was considered was the status of an observation at a patient's last visit, which was defined to be 1 if the patient was discontinued from the study due to lack of therapeutic effect and 0 otherwise. These data were first analyzed by Cnaan et al [3] in the context of a tutorial for the linear mixed model. Six models with fixed effects as described in Table 4.5 were fitted. As was done in the analysis by previous authors, we subtracted 3 weeks (similar to centering) from the time variable prior to the analysis in order to achieve near orthogonality between the linear and quadratic effects. All of the fixed effects models in Tables 5 and 6 were first fitted with a random effect function that consisted of the following components: an intercept, linear effect and quadratic effect for time. Because model 6 for the fixed effects appears to fit the data the best based on values of MEC_1, we fitted those same fixed effects with a reduced random effect model that consisted of an intercept and linear effect for time to obtain model 7. Table 5 gives values for MEC_1 and MIC_1 for the 6 different models in the fixed effects that were originally considered by Cnaan et al. [3]. Table 6 gives values of the fixed effect parameter estimates for models 6 and 7.

According to Table 5, none of the models has good “internal consistency” in that if we assume that the model is correct; the variables don’t explain a significant proportion of the variation in the outcome. At the same time, some of these models appear to have “external consistency” in that they are better at explaining the variation in the data compared to a null model. When the reduced random effect model is used for the best combination of fixed effect variables (model 7 as opposed to model 6), there is a noticeable increase in the value of MEC_1 (i.e., external consistency is improved). Table 6 shows that based on the precision of the parameter estimates for the fixed effects, the reduced model in the random effects is as good as the full model in the random effects.

4.7 Discussion

The results of our simulation show that two statistics MEC_1 and MIC_1 that were proposed in Chapter 3 for assessing internal and external consistency of a model perform satisfactorily in being able to determine misspecification of the random effects structure in the linear mixed model. These statistics fared best in being able to discriminate between a longitudinal model which include the time covariates and a reduced model that contains only a random intercept. In our simulations, the statistics were not able to discriminate between two models that both included time covariates in the random effects but differed in the inclusion of a quadratic term for time and a model with an intercept and linear effect for time. However, we have shown that it is very possible that for the LMM that the model with an intercept and a linear term for time (reduced model 1) in the random effects may be as good as the model with an intercept, linear and quadratic terms for time (full model). Hence, the fact that no differences were detected between the full model and the reduced random effects that omitted the quadratic term may not be an indication that the statistics MEC_1 and MIC_1 are inappropriate for assessing the adequacy of the random effects. Finally, consideration should be given to substituting $\mathrm{tr}(\hat{\Sigma}_i)$ for $\hat{\Sigma}_i$ in the computation of MIC_1 as doing so in our simulations yielded values that were closer to the population parameter (for the reduced model 1).

Our results were similar to those of Gurka [8] for the AIC and BIC for comparing a model with an intercept and a linear term for time and a model with only an intercept term. Gurka [8] showed that the AIC and BIC more than 90% of the time were able to identify the correct model. Heagerty and Kurland [11] found that misspecification of the random effect can lead to biased estimates of the variances of the parameter estimates. Like Gurka [8], the simulations by Heagerty and Kurland [11] were limited to a full model in the random effects that included an intercept and linear term for age. These results also confirm the recommendation by Cnaan et al. [3] that often an intercept and a linear term for time are sufficient for the random effects structure. Cnaan et al. [3] gave no justifications for this recommendation.

4.8 Conclusion

These results confirm that the statistics MEC_1 and MIC_1 can be used to assess the adequacy of the random effects and have similar performance to that of other statistics such as AIC or BIC. Equally important is our finding that misspecification of the random effects can have minimal effect on the fixed parameter estimates or the precision of those estimates when a reduced model with intercept and slope is used even if the true model in the random effects includes a quadratic term. This finding suggests that in longitudinal data analysis involving the LMM, the choice for the random effects can be simplified to deciding whether the linear effect for time needs to be added to the random intercept. Once this choice is made, the analyst can then proceed to focus on determining the subset of fixed effect components that describes best the variation in the outcome. Our results are limited in the fact that we did not investigate the performance of the statistics when the random effects structure was overfitted. This would be useful in the case when the true model is comprised of only a random intercept and a linear term for time is added.

References

1. Akaike, H., 1974. A new look at the statistical model identification. IEEE Transaction on Automatic Control, AC-19, 716-723.

2. Albert, J. and Chib, S., 1997. Bayesian tests and model diagnostics in conditionally independent hierarchical models. Journal of the American Statistical Association 92, 916–925.

3. Cnaan, A., Laird, N.M., and Slasor, P., 1997. Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. Statistics in Medicine, 16(20) 2349-2380.

4. Chen Z. and Dunson D. B., 2003. Random effects selection in linear mixed models. Biometrics, 59, 762-769.

5. Dawson, K. S., Gennings, C., and Carter, W. H. (1997). Two Graphical Techniques Useful in Detecting Correlation Structure in Repeated Measures Data. American Statistician, 51, 275-283.

6. Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). Analysis of Longitudinal Data, New York, NY: Oxford University Press.

7. Grady, J. J., and Helms, R. W., 1995. Model selection techniques for the covariance matrix for incomplete longitudinal data. Statistics in Medicine, 14, 1397-1416.

8. Gurka M. J., 2006. Selecting the best linear mixed model under REML. The American Statistician, 60, 19-26.

9. Hall, D. B. and Praestgaard, J. T., 2001. Order-restricted score tests for homogeneity in generalized linear and nonlinear mixed models. Biometrika 88, 739–751.

10. Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72(358) 320-340.

11. Heagerty P.J. and Kurland B.F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. Biometrika, 88, 973-985.

12. Jennrich, R. I., and Schluchter, M. D. (1986). Unbalanced Repeated-Measures Models With Structured Covariance Matrices. Biometrics, 42, 805-820.

13. Laird, N., Lange, N., and Stram, D., 1987. Maximum likelihood computations with repeated measurements: applications of the EM algorithm. Journal of the American Statistical Association, 82(397) 97-105.

14. Laird, N. and Ware, J.H., 1982. Random effect models for longitudinal data. Biometrics, 38(4) 963-974.

15. Lin, X., 1997. Variance component testing in generalized linear models with random effects. Biometrika 84, 309–326.

16. Littell R. C. et al., 2000. Modeling covariance structures in the analysis of repeated measures data. Statistics in Medicine, 19, 1793-1819.

17. Orelien, J. G. and Edwards L.J., 2007. Fixed Effect Variable Selection in Linear Mixed Models Using R^2 Statistics. Computational Statistics and Data Analysis. In press.

18. Pourahmadi, M. (2002). Graphical Diagnostics for Modeling Unstructured Covariance Matrices. International Statistical Review, 70, 395-417.

19. Stram D.O. and Lee J.W., 1994. Variance components testing in the longitudinal mixed effects model. Biometrics 50, 1171-1177.

20. Shwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics, 6(2) 461-464.

21. Vonesh, E.F., Chinchilli, V.M., 1997. Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker, New York, 1997, 419-424.

22. Vonesh, E.F., Chinchilli, V.M., Pu, K.W., 1996. Goodness-of-fit in generalized nonlinear mixed-effects models. Biometrics, 52 572-587.

23. Ware, J.H., 1985. Linear models for the analysis of longitudinal studies. The American Statistician, 39(2) 95-101.

24. Xu, R.H., 2003. Measuring explained variation in linear mixed effects models. Statistics in Medicine, 22(22) 3527-3541.

25. Zheng, B.Y., 2000. Summarizing the goodness of fit of generalized linear models for longitudinal data. Statistics in Medicine, 19(10) 1265-1275.

26. Zimmerman, D. L. (2000). Viewing the Correlation Structure of Longitudinal Data through a PRISM. American Statistician, 54, 310-318.

Table 4.1 Description of the data sets used in the simulation

Data Set	Value of σ^2	Proportion of variation explained by the cross-sectional covariates (Population Parameter for MEC_1)	Type of Covariance Matrix for the random effects
1	12	0.89	$\mathbf{G} = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 1 & 0.5 \\ 1 & 0.5 & 1.25 \end{pmatrix}$
2	95	0.70	
3	240	0.52	

Table 4.2 Performance of the statistics MEC_1 and MIC_1 in assessing adequacy of random effects

σ^2	Model	MEC_1: Based on comparing to variance of null model (2)	MIC_1: Based on comparing to total variance	MIC_1: Based on comparing to total variance using $\text{tr}(\hat{\Sigma}_i)$ instead of $\hat{\Sigma}_i$
10	Full Model in the random effects: Intercept, age and Agesq	0.74 (0.65, 0.85)	0.71 (0.62, 0.83)	0.87 (0.84, 0.90)
10	Reduced Model 1 in the random effects: Intercept and Age	0.79 (0.75, 0.83)	0.77 (0.72, 0.82)	0.89 (0.87, 0.90)
10	Reduced Model 2 in the random effects: Intercept	0.08 (0.03, 0.13)	-0.00 (0.00, 0.00)	-0.00 (0.00, 0.00)
95	Full Model in the random effects: Intercept, age and Agesq	-1.01 (-1.74, -0.08)	-1.18 (-1.92, -0.17)	0.85 (0.83, 0.88)
95	Reduced Model 1 in the random effects: Intercept and Age	0.51 (0.40, 0.67)	0.47 (0.35, 0.64)	0.77 (0.73, 0.81)
95	Reduced Model 2 in the random effects: Intercept	0.07 (0.02, 0.11)	-0.00 (0.00, 0.00)	-0.00 (0.00, 0.00)
240	Full Model in the random effects: Intercept, age and Agesq	-3.17 (-4.62, -1.25)	-3.45 (-4.87, -1.38)	0.85 (0.83, 0.87)
240	Reduced Model 1 in the random effects: Intercept and Age	0.17 (-0.03, 0.45)	0.12 (-0.10, 0.42)	0.64 (0.59, 0.70)
240	Reduced Model 2 in the random effects: Intercept	0.06 (0.01, 0.10)	-0.00 (0.00, 0.00)	-0.00 (0.00, 0.00)

Table 4.3 Fixed effect parameter estimates and standard errors for each model fitted

σ^2	Model	Average Values for Intercept (Average Standard Error)	Average Values for Age (Average Standard Error)	Average Values for Age Quadratic (Average Standard Error)	Average Values for Treatment (Average Standard Error)	Average Values for Gender (Average Standard Error)
10	Full Model in the random effects: Intercept, age and Agesq	29.98 (1.46)	1.52 (1.27)	1.50 (0.31)	4.00 (0.88)	4.02 (0.88)
10	Reduced Model 1 in the random effects: Intercept and Age	29.98 (1.50)	1.51 (1.40)	1.50 (0.27)	4.00 (0.89)	4.02 (0.89)
10	Reduced Model 2 in the random effects: Intercept	29.99 (3.06)	1.51 (2.27)	1.50 (0.50)	3.98 (2.31)	4.03 (2.31)
95	Full Model in the random effects: Intercept, age and Agesq	29.93 (4.19)	1.52 (3.88)	1.49 (0.86)	4.02 (1.86)	4.02 (1.86)
95	Reduced Model 1 in the random effects: Intercept and Age	30.01 (4.00)	1.47 (3.67)	1.50 (0.79)	4.00 (1.88)	4.02 (1.88)
95	Reduced Model 2 in the random effects: Intercept	30.02 (4.72)	1.49 (4.12)	1.50 (0.90)	3.99 (2.49)	3.98 (2.49)
240	Full Model in the random effects: Intercept, age and Agesq	30.18 (6.54)	1.30 (6.14)	1.54 (1.35)	3.94 (2.48)	4.09 (2.48)
240	Reduced Model 1 in the random effects: Intercept and Age	30.09 (6.19)	1.43 (5.75)	1.51 (1.25)	3.99 (2.50)	4.00 (2.50)
240	Reduced Model 2 in the random effects: Intercept	30.10 (6.65)	1.43 (6.10)	1.51 (1.34)	3.99 (2.78)	3.97 (2.78)

Table 4.4 Using AIC, BIC and LRT for comparing the different models

	AIC, BIC and LRT for comparing full model and reduced model 1			AIC, BIC and LRT for comparing full model and reduced model 2			AIC, BIC and LRT for comparing reduced model 1 and reduced model 2		
σ^2	AIC	BIC	LRT	AIC	BIC	LRT	AIC	BIC	LRT
10	98%	84%	99%	100%	100%	100%	100%	100%	100%
95	29%	6%	41%	100%	100%	100%	100%	100%	100%
240	18%	3%	31%	98%	75%	100%	100%	97%	99%

Table 4.5 Schizophrenia data: measures of external and internal consistency

Model	Fixed effects terms	MEC_1	MIC_1
1	Week, $Week^2$	0.32	-0.24
2	Treatment, Week, $Week^2$	0.34	-0.24
3	Treatment, Week, $Week^2$, $Week \times Treatment$	0.34	-0.23
4	Treatment, Week indicators	0.31	-0.27
5	Treatment, Week, $Week^2$, $Week \times Site$	0.41	-0.29
6	Treatment, Week, $Week^2$, Status	0.54	-0.26
7	Treatment, Week, $Week^2$, Status (reduced model in the random effect)	0.64	0.01

All models above included baseline BPRS and center in the fixed effects. The random effects consisted of an intercept term, a linear and quadratic effect of time for models 1 through 6. For model 7, the random effects consisted of an intercept and a linear effect of time.

Table 4.6 Parameter estimates for the full and reduced model in the random effects

	Parameter estimates for model 6 of Cnaan et al. (1997). The random effect consists of an intercept, linear and quadratic effects for time		Parameter estimates for model 7 of Table 4.5. The random effect consists of an intercept, linear and quadratic effects for time	
Parameters	**Parameter estimate**	**Standard error of the parameter estimates**	**Parameter estimate**	**Standard error of the parameter estimates**
Intercept	26.60	3.53	26.97	3.60
bprs0	0.58	0.06	0.57	0.06
site1	-8.44	3.80	-8.85	3.89
site11	-4.68	3.99	-6.44	4.09
site12	-10.35	3.60	-9.50	3.69
site15	-14.39	4.23	-14.06	4.33
site16	-9.43	3.63	-9.78	3.72
site2	-12.05	3.31	-11.69	3.39
site3	-11.99	2.97	-12.14	3.04
site4	-5.72	4.49	-6.08	4.57
site5	-15.75	4.30	-14.87	4.41
site7	-10.64	3.62	-11.06	3.70
site8	-0.45	3.24	0.23	3.32
site9	-4.14	3.38	-3.94	3.45
status	-16.28	1.07	-16.10	1.11
time	-2.35	0.21	-2.30	0.19
timesq	0.46	0.08	0.49	0.07
trt1	1.14	1.53	0.86	1.57
trt2	-1.01	1.49	-1.33	1.53
trt3	-0.12	1.51	-0.69	1.55

5 Conclusion and Further Research

This dissertation research has contributed to identifying shortcomings in existing R^2 statistics for the LMM, proposed new R^2 statistics, and has demonstrated the suitability of these new statistics in assessing adequacy of both fixed and random effects. However, our research has left some questions on GOF for the LMM unanswered and raised additional ones. In this chapter, we provide some conclusions stemming from the completion of the 3 papers included in the dissertation and discuss opportunities for further research. Our conclusions are in Section 5.1 and areas for future research are discussed in Section 5.2.

5.1 Overall Conclusions

Assumptions

One of the first lessons that must be taken into account is that assumptions that may make sense in the traditional linear models may not be applicable to the LMM. For example, in the traditional linear model, by adding additional covariates in a model one expects the variance of an observation based on the model to be reduced. This may not necessarily be true in the LMM. One of the key findings that might be unexpected is that the conditional residuals would be robust to misspecifications of the cross-sectional covariates. As a result, care must be exercised in using assumptions from traditional linear models or other classes of models to develop R^2 statistics. Based on this dissertation research, it is the humble opinion of the author that extensive simulations are needed to ensure that a proposed R^2 is working as intended—even when appropriate theoretical justifications have been given for the R^2 statistic. Furthermore, because an R^2 statistic has been proven to be adequate for a subclass of models in the LMM, this is not a guarantee that the statistic will have similar performance for all other subclasses of the LMM. This caveat applies to the R^2 that we have proposed as well. Our simulations were limited to instances with longitudinal data and conditional independence. As we will discuss later in this chapter, more simulations are needed to ensure that our results are applicable to other subclasses in the LMM.

Another area for concern in the development of R^2 statistics is that one may need to be careful even when anticipated results are obtained in simulations. For example, our simulations confirm that the statistics proposed by Xu (2003) estimate a population parameter. However, it is clear that this population parameter estimated by the statistics proposed by Xu (2003) is inappropriate for assessing GOF in the LMM.

There is a need for more than one R^2

A second conclusion of this dissertation research is that there is a need for more than one R^2 statistic in the LMM. In paper 2 of this dissertation, we proposed two types of statistics that measure different aspects of a LMM: a) how well the variation of the outcome can be explained by a null model and b) how well the variation of the outcome is explained by the model at hand assuming that it is the true model. Ideally, one would want both types of R^2 statistics to be high, but in the example given in paper 2, one statistic was high and the other low. Because the two types of statistics are measuring different constructs, it is possible for them to be discordant.

While we have shown the need for more than one R^2 for assessing adequacy of the fixed effects, one can see how future research could develop R^2 statistics for assessing adequacy of the random effects or assessing adequacy of the covariance structure. For example, Vonesh et al. (1996) have proposed an R^2 statistic for assessing the covariance of structure that has no relevancy for the adequacy of the fixed effects or the random effects. Hence, the likely scenario is that there is probably more than one R^2 statistic as we have proposed for assessing adequacy of the fixed effect terms, more than one R^2 statistic for assessing adequacy of the random effects and more than one R^2 statistic for assessing the adequacy of the covariance structures.

Simpler Models in the Random Effects or Covariance Structures

In most instances of the LMM, the parameters for the random effects or the parameters for the covariance of these random effects or the error terms are not of primary interest—they can be treated as nuisance parameters—that need to be accounted for so that unbiased estimates of the fixed effect parameters estimates can be obtained. The simulations in paper 3 have indicated that it is possible that simpler models in the random effects might be as efficient as a more complex model in the random effects. In limited simulations of several covariance structures performed, we had difficulty in attaining convergence for the most complex ones such as an unstructured covariance matrix. It is possible that a simpler covariance structure might still lead to unbiased estimates of the fixed effect parameter estimates.

Keeping an "eye on the prize"

As we have stated in this dissertation, in many epidemiological studies with longitudinal data, the primary interests of the researchers are in estimating the fixed effect parameters and their variances. While R^2 statistics provide useful information, they should be complemented with other statistics for determining adequacy of the model. In particular, the analyst should review the values of the fixed effect parameter estimates and their confidence intervals. Small values of the fixed effect parameters (e.g., close to 0), even when they are significant, should be cause for concerns. Wide confidence intervals for one or more fixed effect parameters could also be an indication that the model might be inadequate. The idea is that statistical significance does not mean practical significance. Also, for model selection, the analyst should consider comparing the variance of the fixed effect parameters for choosing the appropriate model.

5.2 Future Research

Use of R^2 in other subclasses of models in the LMM and other classes of models

While the approach that we have proposed seems to work for the types of simulations that we have conducted, it will be worthwhile to conduct additional simulations before extending these results to other subclasses of models. In particular, in our simulations we assumed conditional independence. Also, our study was limited to longitudinal data. A typical feature of models for longitudinal data in the LMM is that the random effects are nested in the fixed effects. There are other models in the LMM where it may not make sense to include all variables that are in the random effects in the fixed effects as well. For example, suppose a study is done in multiple hospitals, the analyst may wish to use hospital in the random effect to account for the clustering (correlation within hospital) without including hospital as a fixed effect. The rationale for doing so is that there is no inherent interest in hospital as a fixed effect. It is simply a nuisance parameter.

Beyond the LMM, there are other classes of models such as nonlinear mixed models (NLM) that the R^2 statistic or approaches we have outlined in Chapter 3 could be useful. Some of the concepts that we introduce such as “external validation” (comparing the model at hand to a null model) and “internal validation” (to compute the variation explained by the model at hand assuming that it is adequate), especially the former, could be applicable to many models. Notice that for any model, a null model consisting of an intercept can always be achieved, including traditional linear models.

Using R^2 for model selection when both fixed and random effects are misspecified

In this dissertation, we have investigated the performance of R^2 for comparing the fixed effects of models having the same random effects or for comparing the random effects for models having the same fixed effects. Comparisons of linear mixed models where both the mean and random effect models are different were not considered and have essentially been ignored in the statistical literature. There are four types of linear mixed models that differ in both the mean and random effects 1) The mean models are different with random effect models the same 2) The random effect models are different with mean models the same. 3) Both the mean and random effect models are different but nested. 4) Both the mean models and random effect models are different but non-nested.

The first type is the same as the problem that we address in chapters 2 and 3 and the second type is similar to the problem we addressed in chapter 4. Situations that arise in types 3 and 4 are beyond the scope of this dissertation. For type 3, one could determine through simulations whether it is a) best to first determine the adequacy of the reduced random effects model and then the fixed effects model or b) first determine the adequacy of the reduced fixed effect model and then the random effect model. Tools developed in this dissertation could be used to come up with such determinations. Additionally, we will assess the performance of the AIC and BIC in selecting such models.

Using R^2 for assessing adequacy of the Covariance Structure

Vonesh and al. (1996) proposed 2 statistics for assessing the adequacy of the covariance structure in the LMM. One of these statistics, the variance-covariance concordance correlation coefficient measures the distance, scaled to 1, between the assumed covariance matrix and the robust covariance matrix ("sandwich estimator") of Liang and Zeger (1986). The other, denoted the pseudo likelihood ratio test (PLRT), is a formal test for detecting significant differences between the 2 covariance matrices. Results from a limited simulation (400 replicates) of the PLRT were presented. No simulation results on the performance of the variance-covariance concordance correlation coefficient were presented.

One approach might be to simulate replicates of multivariate normal data with the following covariance structures: unstructured, Toeplitz, spatial, first order autoregressive, compound symmetry and independence. For each of these covariance structures, one could compute the number of times it is rejected as being inadequate when another covariance-structure in the list is specified (power) and the number of times the test rejects the true covariance structure (type I error). Estimates of the variance-covariance concordance correlation coefficient will be computed to assess its performance when the true covariance structure is specified or misspecified (including the equivalent of a null model for the covariance structure such as complete independence). In performing these simulations, one needs to take into account instances when the covariance structure is misspecified but the misspecified covariance is nested in the true covariance.

The test proposed by Vonesh and Chinchilli (1996) is based on the asymptotic distribution of a statistic that has a chi-square distribution. Another approach in comparing 2 matrices is to use a statistic based on a Wishart distribution. Future research could look into developing and proposing other tests based on the Wishart distribution to assess the adequacy of the covariance structure. The performance of any test proposed would be evaluated through simulations and compared to that of the PLRT.

Another possibility that we have considered is a test based on an approximate F statistic. The test could be constructed as follows. Let $\boldsymbol{\Sigma}_{i(f)}$ be the covariance matrix for the full model and $\boldsymbol{\Sigma}_{i(r)}$ the covariance matrix for a reduced model in the random effects. We define the sum of squares for the full model in the covariance matrix as

$SS(f)=\sum_{i=1}^{n}(\mathbf{y_i}-\mathbf{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{i(f)}^{-1}(\mathbf{y_i}-\mathbf{X}\widehat{\boldsymbol{\beta}})$ and the sum of squares for the full model as

$SS(r)=\sum_{i=1}^{n}(\mathbf{y_i}-\mathbf{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_{i(r)}^{-1}(\mathbf{y_i}-\mathbf{X}\widehat{\boldsymbol{\beta}})$. The approximate F test statistic could then be defined

as $F_{ran}=\dfrac{(n*r_1)SS(r)}{(n*r_2)SS(f)}$ where $r_1=rank\left(\boldsymbol{\Sigma}_{i(f)}\right)$ and $r_2=rank\left(\boldsymbol{\Sigma}_{i(r)}\right)$. There are different

possibilities for the rank of the test that could be investigated. One such possibility is that the degrees of freedom are respectively nr_1 and nr_2 in the numerator and denominator. One could investigate through simulations that under the null hypothesis the full model is not different from the reduced model and that F_{ran} follows an approximate F distribution with degrees of freedom nr_1 and nr_2.