MODEL SELECTION FOR NONNESTED LINEAR MIXED MODELS

Ché L. Smith

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 of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Lloyd J. Edwards

Peggye Dilworth-Anderson

Bahjat Qaqish

Pranab K. Sen

Chirayath Suchindran

ABSTRACT

Ché L. Smith: Model Selection for Nonnested Linear Mixed Models
(Under the direction of Lloyd J. Edwards)

The selection of an adequate and parsimonious model among suitable candidates is an essential aspect of the model-building process. Model selection approaches have been widely studied for the univariate linear model and other models arising from cross-sectional data. As researchers increasingly rely on linear mixed models (LMMs) to characterize longitudinal data, there is a need for improved techniques for selecting among this class of models which requires specification of both fixed and random effects via a mean model and variance-covariance structure. The model selection process for LMMs is further complicated when fixed and/or random effects are nonnested between models. Presently, information criteria such as AIC and BIC dominate model selection criteria used to compare nested and nonnested LMMs. This dissertation explores the development of a hypothesis test to compare nonnested LMMs based on extensions of the work begun by Sir David Cox. Particularly, we address the complex issue of estimating the variance of Cox test statistics through the use of parametric bootstrapping. Various information criteria have been modified for this purpose, but recent investigations have all led to inconclusive results as to which criterion is the best to select among LMMs. We also consider the use of the Extended Information Criterion (EIC) as an improvement on the more commonly used AIC. Application to observed data demonstrates the viability of both the Cox Test and the EIC to select among nonnested LMMs.

To my loving parents:

Patricia Kornegay Smith and Calvin Miles Smith, Jr.;

and to my late grandparents:

Dr. Hobert Kornegay, Jr. & Mrs. Ernestine Price Kornegay,

Dr. Calvin Miles Smith, Sr. & Mrs. Margaret Nixon Smith

ACKNOWLEDGMENTS

I thank God for giving me the courage to see this through; by my own devices, I would have long ago been defeated by real and perceived obstacles. To my advisor, Dr. Lloyd Edwards, thank you for teaching me to always pursue excellence; I pray that your career continues to flourish and that your hard work, innovative approaches to teaching and mentoring, and success in training students do not go unrecognized. I am grateful for my dissertation committee, the Biostatistics department (especially Veronica Stallings and Melissa Hobgood for the tireless support that they provide to students), the Gillings School of Global Public Health, and the Graduate School for supporting my work and giving me this opportunity. I was very fortunate to have several funding sources and work experiences throughout my program. To the Royster Society of Fellows, thank you for acknowledging my academic merit and commitment to scholarship, leadership, and service, and for exposing me to the varied wonderful work that makes UNC such a great campus; many thanks to Randy Davis, Sharon Murray, and others from GlaxoSmithKline, for the generous support during my first year and for the exposure to the pharmaceutical industry; to the Gillings School of Global Public Health for awarding me the Michel Ibrahim Fellowship, and to Felicia Mebane for allowing me to help build the pipeline for talented and diverse students to discover their place in public health; to the NIH Cardiovascular Clinical Trials grant for support in the early years of my program; to IMSD/BBSP (Pat, Ashalla, Josh, and Patrick) for giving me invaluable opportunities to hone my teaching and consulting skills; to Greg Evans, Jeff Carr, Greg Terry, and Sean Simpson at Wake Forest University for the invaluable experience of working on impactful research with a great team; to Henry Frierson, for embracing me in the RES program and helping me get acclimated to Chapel Hill; to Tiffany West, for teaching

me not to fear taking risks; to Vic Schoenbach and the many members of the Minority Student Caucus for recognizing the value of ethnic diversity and for giving me opportunities to serve in many school activities; and to Lisa LaVange and Kant Bangdiwala, for helping me to discover my passion for consulting and that superb written and oral communication are the most valuable skills a biostatistician can have. Richardean Anderson, Annette Dotson, Nagambal Shah, and other educators noticed early on my aptitude and interest in mathematics and took time to nurture me toward this path; Dr. Shah, I especially thank you for being a wonderful mentor and friend, and for teaching me that 'success is a journey'. Thanks to my immediate family (Mom, Dad, Calvin, and Nina) and my extended family for never setting limits on what I could achieve, for setting high expectations without pressuring me, and for offering great examples through their own achievements. I am thankful for true friends who love me beyond any of my academic accomplishments and supportive peers who have offered endless encouragement, examples of discipline through their achievements, accountability, tough love, prayers, and silent support; you all never gave up on me when it seemed everyone else had. To Christian, thanks for being my 'BIOS twin'; we have traveled this journey together and have similar successes and scars that will make for great stories for years to come. Thanks to Gloria Thompson and other UNC staff who were like family; you will never know the impact of your kindness toward me and other students who embark upon this ambitious undertaking. The rest of my village has too many people to name here, but I pray that the rest of my life's work pays forward - times one million - the investment that you *all* have made in me.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Selecting an adequate and parsimonious model is an important, yet often neglected, aspect of data analysis and research. Over the past few decades, the majority of research in model selection has centered around linear regression and other univariate linear models. Any work beyond that class of models has involved the extension of existing model selection methods applied to regression and univariate linear models (e.g., likelihood ratio tests, R^2 coefficient, information criteria) to more complex types of models, with little investigation of the robustness of these methods after extension. Even more rarely have model selection techniques for the linear mixed model been thoroughly examined. Here, an attempt to extend the capability to select among nonnested linear mixed models is made.

Longitudinal studies are often employed to assess changes within individuals and groups over time. One well known example arises from a study by Potthoff and Roy (1964), in which twenty-six children (sixteen girls, eleven boys) were followed every two years for six years, and at each observation the growth of each subject's jaw was measured. In examples like this one, each child has his/her own trajectory of correlated repeated observations, allowing one to study individual variability in growth as well as group averages (e.g, boys vs. girls). The linear mixed model is a useful tool to model longitudinal trends among continuous outcomes following a normal distribution, allowing for both subject-specific trajectories and population-averaged models, as in the above example. As opposed to repeated measures ANOVA, its flexibility lies in the separate modeling of the mean and variance-covariance structures, thus providing many potential 'best' model candidates via various combinations of these two components. Well-developed techniques exist for comparing models with nested mean and/or covariance structures, mostly following from Neyman-Pearson (Neyman and Pearson, 1933) theory. On the other hand, formal approaches to compare models lacking this nested structure (or *nonnested* models) have not been developed.

This review highlights important findings from the literature on model selection for the linear mixed model - with special attention given to the case of nonnested models - and illuminates some opportunities for methodological improvement. In the next section, a detailed specification of the linear mixed model is introduced with specific notation to be used throughout this document. Subsequent sections review the history of model selection techniques for univariate linear models, and their extension to multivariate and correlated data. Special attention is paid to the case of comparing nonnested linear mixed models, which has received little attention. Specifically, two general approaches to model building and selection among nonnested linear mixed models are highlighted - hypothesis testing and the use of information criteria. Particularly for the hypothesis testing approach, ties are made from the econometrics literature which contains applications of selecting among nonnested models of a similar structure to linear mixed models. Finally, the most important findings from the literature review are summarized.

1.2 The Linear Mixed Effects Model

The Linear Mixed Effects Model (referred to hereafter as the *linear mixed model*) is used to analyze multivariate continuous data, particularly longitudinal data. Consider the linear mixed model for repeated measures data specified below (Edwards et al., 2008):

With N independent sampling units (often *persons* in practice), the linear mixed model for person i may be specified as $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathbf{b_i} + \mathbf{e_i}$, where $\boldsymbol{y}_i$ is a $(n_i \times 1)$ vector of correlated observations on person i; in the longitudinal setting, they represent the set of repeated measurements over a specified period of time on a single subject. The $(n_i \times p)$

matrix $\boldsymbol{X}_i$ is the known and fixed design matrix for person i, with full column rank p. $(p \times 1)$ vector $\boldsymbol{\beta}$ is a vector of unknown, constant population parameters, common to all subjects. Matrix $\boldsymbol{Z}_i$, of dimension $(n_i \times q)$, is a known, constant design matrix with rank q for person i corresponding to the $(q \times 1)$ vector of unknown random effects, $\mathbf{b_i}$, while $\mathbf{e_i}$ is an $(n_i \times 1)$ vector of unknown random errors. Random effects are assumed to be independent of random errors, and together they are assumed to follow a normal distribution with mean $\mathbf{0}$ and variance

$$\mathcal{V}\left(\begin{bmatrix} \mathbf{b_i} \\ \mathbf{e_i} \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{bi}}\left(\boldsymbol{\tau}_b\right) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{ei}\left(\boldsymbol{\tau}_e\right) \end{bmatrix}.$$

In the equation above, $\mathcal{V}(\cdot)$ is the covariance operator, while both $\boldsymbol{\Sigma}_{bi}(\tau_b)$ and $\boldsymbol{\Sigma}_{ei}(\tau_e)$ are positive-definite, symmetric covariance matrices. Therefore, $\mathcal{V}(\boldsymbol{y}_i)$ may be expressed as $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i \boldsymbol{\Sigma}_{\boldsymbol{bi}}(\tau_b) \boldsymbol{Z}_i' + \boldsymbol{\Sigma}_{ei}(\tau_e)$. We assume that $\boldsymbol{\Sigma}_i$ can be characterized by a finite set of parameters represented by an $(r \times 1)$ vector $\boldsymbol{\tau}$, which consists of the unique parameters in $\boldsymbol{\tau}_b$ and $\boldsymbol{\tau}_e$.

For simplicity, we will assume here that all subjects have the same number of repeated measurements, so that $n_i = n$ for all subjects, such that$\mathbf{e_i} \sim \mathbf{N_n}(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{e}(\tau_\mathbf{e}))$, where the $(n \times 1)$ vector τ_e collects the n unique parameters of $\boldsymbol{\Sigma}_e(\tau_e)$.

Alternatively, we may specify the linear mixed model for all subjects in a stacked formulation as follows: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}\mathbf{b} + \mathbf{e}$ where $\boldsymbol{y} = (\boldsymbol{y}_1' \ldots \boldsymbol{y}_N')'$; design mnatrix $\boldsymbol{X} = (\boldsymbol{X}_1' \ldots \boldsymbol{X}_N')'$; vector $\boldsymbol{\beta}$ is as specified before; $\boldsymbol{Z} = diag\left(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N\right)$; vector $\mathbf{b}$ is as specified before; and $\mathbf{e} = (\mathbf{e_1'} \ldots \mathbf{e_N'})'$.

Note the dimensions of $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\mathbf{e}$ are $(Nn \times 1)$, $(Nn \times p)$, $(Nn \times q)$, and $(Nn \times 1)$, respectively. Further, $\mathbf{b} \sim \mathcal{N}_\mathbf{q}\left(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{bi}\left(\boldsymbol{\tau}_\mathbf{b}\right) \otimes \mathbf{I_b}\right)$ and $\mathbf{e} \sim \mathcal{N}_\mathbf{Nn}\left(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{e}\right)$. We have $\boldsymbol{\Sigma}_e = diag\left[\boldsymbol{\Sigma}_{e1}\left(\boldsymbol{\tau}_e\right), \ldots, \boldsymbol{\Sigma}_{eN}\left(\boldsymbol{\tau}_e\right)\right]$; thus $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X\beta}, \boldsymbol{\Sigma}\right)$ with $\boldsymbol{\Sigma} = \mathcal{V}(\boldsymbol{y}) = diag\left(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N\right)$.

Hereafter, we adopt this notation to refer to a linear mixed model as a function $\boldsymbol{y}$ of the collection of its complete parameter space $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\tau}')'$ is a $(s \times 1)$ vector, with $s = p + r$.

1.3 Model Selection in the Linear Mixed Model (LMM)

The selection of an adequate model among suitable candidate models is an essential part of the model-building process and has been widely studied for the univariate linear model and other types of models arising from cross-sectional data. Most commonly for linear regression models, researchers employ forward, backward, and stepwise variable selection procedures. Ngo and Brand (2002) discussed how the extension of these procedures to the linear mixed model introduces problems of multiple testing, further complicated by the use of an arbitrary level of significance. Some methods are ad hoc, and have not been well developed or studied. This is especially true for the linear mixed model, which requires selection of both a mean model and a covariance model.

Cheng et al. (2010) recently summarized issues researchers have with building a ’good enough’ linear mixed model, arguing that no model is the ’best’ or ’true’ model. The authors distinguish between the concepts of model choice, model building, and model selection. For example, model choice refers to the decision to use a linear mixed model (as opposed to a different class of model) to characterize data; model building involves using a particular predictor selection strategy; and model selection involves the formulation of a criterion used to determine which model is the relative ’best’ among all candidates. The authors acknowledge that various information criteria have been developed and used to compare models, but they are flawed mostly because they do not provide a formal statistical test to discern models. Further, the authors emphasize that neither information criteria nor likelihood ratio tests (and other F tests) have been well-developed for use among this class of models. They propose the following general five-step strategy for building and selecting an adequate linear mixed model:

1. Specify the maximum model to be considered (both in fixed and random effects).

2. Specify a criterion of goodness of fit of a given model.

3. Specify a predictor selection strategy.

4. Conduct the analysis.

5. Evaluate the reliability of the model chosen.

Cheng et al. emphasize that these steps are also appropriate for building univariate, multivariate, linear, and nonlinear models, along with models for any type of response variable distribution. Their discussion also follows the assumption that the 'best' model is contained, or *nested*, within a maximum model. Of particular interest here is the third item in the above list. It is important to predetermine a strategy that will lead to fitting a select few models with as few predictors as necessary to characterize the outcome variable(s). In most cases, the maximal model rarely ends up being the final model. Ideally, one would implement a well-planned strategy to select a small number of predictors based on the size of the data and knowledge of the subject area. For linear regression, it is straightforward to employ techniques such as forward, backward, or stepwise variable selection. Model selection for the linear mixed model, however, is more complex and requires attention to its two models - the mean structure, and selection of random effects and variance components (Liu and Yang, 2008). In the linear mixed model, it is common to keep one aspect of the model fixed in both models, while using techniques to refine the other model. For example, one may assign a common covariance matrix to both models and compare models with different sets of fixed effects.

Edwards et al. (2008) focus on three general types of model comparisons that may occur when selecting among linear mixed models: first, mean models with common covariance structure (most commonly, nested mean models); second, (nested or nonnested) covariance models with common mean structure; and finally, linear mixed models with different mean and different covariance structures. These categories may be expanded to more specifically outline the types of linear mixed models that may be compared. One may compare:

1) Nested mean models with:
   a. the same covariance structure;
   b. different (but nested) covariance structures; or
   c. nonnested covariance structures, or
2) Nonnested mean models with:
   a. the same covariance structure;
   b. different (but nested) covariance structures; or
   c. nonnested covariance structures.

The first case listed above - comparing nested mean models - is most commonly considered (Diggle et al., 1994; Wolfinger et al., 1993). Selection of a favored model is largely based on extensions of techniques used for univariate linear models, such as likelihood ratio tests, goodness of fit measures, predictive criteria (R^2, $PRESS$, etc.), and information criteria (AIC, BIC, etc.). Existing techniques rely on theories developed by Neyman and Pearson (1933) or Kullback and Leibler (1951). theory. Virtually all recent reviews of variable/model selection in the linear mixed model only consider cases of mean models that are nested (Wang and Schaalje, 2009; Dziak and Li, 2007; Dimova et al., 2011). Here, models are considered *nonnested* if one cannot be obtained as a simple limit of the other. In the following sections, we distinguish in more detail between nested and nonnested models, and describe the history of existing model selection techniques for both cases with particular emphasis on the scarcity of methodology applied to nonnested linear mixed models.

1.3.1 Selecting among Nested Models

The vast majority of model selection techniques were developed to make comparisons between pairs of models that follow a nested structure. Models are considered nested when one model can be obtained by imposing a set of linear constraints on another more general model. In terms of the linear mixed model, one may consider models with nested fixed and/or random effects. Models with nested fixed effects are typically compared under maximum likelihood estimation using likelihood ratio tests, t tests, or F tests. Models with nested covariance structures are compared under restricted maximum likelihood (REML) estimation using χ^2 tests. Wang and Schaalje (2009) reviewed several model selection techniques involving predictive criteria that were originally designed for selecting among ordinary linear models. The authors summarize the history of the adjusted R-squared $\left(R_{adj}{}^2\right)$, the concordance correlation coefficient (CCC_{adj}), and the predicted residual sum of squares $(PRESS)$, noting that though all three were being used for selection of linear mixed models they had not yet been adequately assessed for their ability to select among models. They developed simulations to compare the performance of the predictive criteria against each other, as well as with versions of the most common information criteria (AIC and BIC), and the pseudo F-test. The authors did not determine a single selection criterion that performed consistently better than the others. Moreover, their investigation does not address how well these criteria select among nonnested fixed and/or random effects models. Finally, no assessment was made among large-sample data, which is often the case of interest to many longitudinal data analysts. Other recent investigations that compare information criteria have also led to inconclusive results regarding selecting a 'best' criterion among a group of nested linear mixed models (Dziak and Li, 2007; Shang and Cavanaugh, 2008; Dimova et al., 2011).

1.3.2 Selecting among Nonnested Models

The second case, where models are nonnested, is seldom addressed. There are varied scenarios in which candidate models could be considered nonnested. One may consider comparing models with *nonnested sets of explanatory variables*, where non-linear restrictions to parameter vectors from both models are required to establish a nested structure between competing models. Most commonly seen is the case of alternative, positively correlated, measures of the same concept in nonnested parametric form, where both of which are assumed to influence some outcome (Cole et al., 2005). For example, Dameus et al. (2002) compared models with rival econometric theories to expain the same phenomenon. Additionally, one could compare models with the partially overlapping sets of main effects and with nonested interactions among explanatory variables. Another manifestation of nonnested models of this type occurs in models having additive vs. multiplicative interactions; that is, testing absolute (additive) risk difference vs relative (multipicative) risk (Kalilani and Atashili, 2006). One may also compare models with *nonnested functional forms*; an example of this case is the comparison of linear vs. log-linear functions of a continuous outcome variable. One would use a linear mixed model under the assumption of a Normal distribution for the first function, and a gamma link for the nonlinear model under the assumption of a non-Normal distribution. One could also use a generalized estimating equations (GEE) approach in both, but this case is not explored here.

Here, we focus on the first case, where models have nonnested sets of explanatory variables. That is, one may wish to compare two models for which either the mean models or the covariance models are nonnested. To address comparisons of nonnested linear mixed models, it makes sense to consider approaches used to compare nonnested univariate models. Univariate models do not have random effects, only random errors, eliminating the case of comparing models with nonnested covariance structures; when we refer to nonnested univariate models, we consider models with nonnested explanatory variables. In linear regression, the R^2 statistic is most commonly used to compare nonnested models, though traditionally it is a goodness-of-fit statistic. Edwards et al. (2008) developed an extension of this technique to the linear mixed model to determine an association between repeated measurements and fixed effects. Watnik and Johnson (2002) discuss using a relative efficiency approach to compare nonnested linear regression models and considers globally non-nested models, in which neither design matrix is a subset of the other. Timm et al. (2002) compared nonnested models arising from a meta-analysis of six studies investigating the effect of academic coaching on students' SAT scores. They use a modified Wald test statistic developed in a prior investigation (Timm et al., 2001) which was shown to be more powerful than the hypothesis test developed by Cox (1961). The authors do not consider repeated measures data, but they acknowledge that the use of an information-theoretic approach (using information criteria) as a suitable alternative to their methodology. That approach is considered here as well, and discussed in more detail in a later section.

Pesaran and Weeks (2000) and Szroeter (2007) reviewed methods for comparisons of nonnested econometrics models, including the Cox Test - discussed in more detail in the next section - and encompassing tests. From the behavioral research paradigm, Levy and Hancock (2007) discuss a general hypothesis testing framework and approach that applies to comparing both hierarchical (nested) and non-hierarchical (nonnested) multivariate normal mean and covariance models stemming from the structural equation modeling (SEM) framework. They build on the work proposed by Vuong (1989), who used Kullback-Leibler information theory to contruct tests to compare nonnested models. Vuong proposed directional tests for comparison of nonnested hypotheses under a variety of scenarios under the linear regression framework. This and other approaches have attempted to circumvent having to compare nonnested models by assuming that both models are encompassed by a larger model Mizon and Richard (1986). The authors emphasize the fact that information criteria do not provide a statistical test to compare models, so they cannot elucidate the magnitude of difference between models.

Other notable investigations of comparing nonnested models include Ette (1996), who compared nonnested (or non-hierarchical) nonlinear mixed models arising from longtiduinal observations of pharmacokinetic data; particularly, he compared models with the same covariates, but one having a different parametric form in each model; the two candidate models were thus highly correlated, but not nested by earlier definitions. Ette bootstrapped the log-likelihood difference betweeen candidate models. His investigation illuminates the need for more research related to bootstrapping correlated data and comparing nonnested models arising from longitudinal data. Haskard et al. (2010) discussed ways to compare linear mixed models with nonnested - or *non-stationary* random effects, and used the AIC to compare models. These data were not repeated observations, but rather each vector of observations contained soil samples from nearby locations so that clusters of observations were correlated spatially. In another example using linear mixed models, Morrell et al. (2009) describe how changes in parameterization of time-dependent covariates (mainly, subject's age) can influence results of longitudinal models. The authors compared models with nonnested fixed and random effects, using the AIC and BIC as selection criteria, noting that models with nonnested fixed effects cannot be compared using these criteria under REML estimation.

It is clear that more research is needed on comparing linear mixed models that are nonnested. In the following sections, we review two promising model selection approaches with similar computational challenges that have not yet been applied for selecting among nonnested linear mixed models: hypothesis testing following the theory of Cox (1961) employing parametric bootstrapping to estimate test statistics' variances and distributions, and the Extended Information Criterion (EIC) introduced by Ishiguro et al. (1997) which also employs the use of bootstrapping.

1.4 Testing Separate Hypotheses in the Linear Mixed Model

While recent literature has given much attention to model selection for nonnested regression models and econometrics models, the case of selecting among nonnested linear mixed models remains severely underexplored. Seminal works of Cox beginning in 1961 and 1962 highlight tests of *separate* families of hypotheses, meaning that one hypothesis (or model) cannot be obtained as a simple limit of the other; before his works, ad hoc methods were employed to test separate hypotheses. Since then, Pesaran (1974), Pesaran and Deaton (1978) and several others have also studied this methodology extensively. Parallel approaches have also been developed, employing theories such as the encompassing approach Mizon and Richard (1986), including the J-test (Davidson and MacKinnon, 1993) and Vuong's Test (Vuong, 1989). One major disadvantage of these approaches is that the assumption of an all-encompassing model can become tedious for the linear mixed model, since one must consider both the mean model and covariance structure. Here, we describe Cox's Test of Separate Hypotheses in more detail, providing examples of its implementation, and outlining some considerations for applying this approach to compare separate, or nonnested, linear mixed models.

1.4.1 The Cox Test of Separate Hypotheses

Cox (1961) recognized a need to find a general method for handling testing of separate families of hypotheses, a class of problems which until that time had not received much attention in the literature. Prior to Cox's paper, ad hoc methods had been employed to test nonnested - or *separate* - hypotheses. Two hypotheses are called *separate* if an arbitrary simple hypothesis in one cannot be obtained as a limit of a simple hypothesis in the other. The idea of two hypotheses being separate is a function of both the parameters of the hypotheses and the definitions of the hypotheses. For instance, both the null and alternative hypotheses could have the same number of parameters in their parameter vectors, but the hypotheses are considered separate due to the definitions of the hypotheses. Conventional techniques for comparing nested models, such as Wald's F test, do not apply for the comparisons of nonnested models, as they require restrictions on the models' parameter spaces that do not exist, thus making it more difficult to estimate the distribution of the ratio of likelihood functions. Cox recognized the need to make modifications to recenter the likelihood ratio and standardize a test statistic with an asymptotically normal distribution. Pesaran and Weeks (2000) provide a good summary of the motivation behind Cox's formulation. The literature following Cox's two papers has been extensive, but few well-developed techniques and extensions have been studied or recommended, especially for the linear mixed model. In the following section, Cox's methodology is defined and described in more detail.

Notation and Setup

Let $\boldsymbol{y} = (y_1, y_2, \ldots y_m)$ be an $(m \times 1)$ observed random vector and suppose we are interested in testing the composite null hypothesis, H_1, that the probability density function (pdf) is $f(\boldsymbol{y}, \boldsymbol{\theta}_1)$ against the composite alternative, H_2, that the pdf is $g(\boldsymbol{y}, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are vectors of dimension $(k_1 \times 1)$ and $(k_2 \times 1)$, respectively. That is

$$H_1 : f(\boldsymbol{y}, \boldsymbol{\theta}_1)$$
$$H_2 : g(\boldsymbol{y}, \boldsymbol{\theta}_2)$$

where $\boldsymbol{\theta}_1 \in \Omega_1 \subset R^{k_1 \times 1}$ and $\boldsymbol{\theta}_2 \in \boldsymbol{\Omega}_2 \subset R^{k_2 \times 1}$, where $k_1, k_2 > 1$. Note the following assumptions:

(i) the parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ may be treated as varying continuously even when a component of say $\boldsymbol{\theta}_2$ is the serial number of the observation at which a discontinuity occurs;

(ii) the values of $\boldsymbol{\theta}_1$, or $\boldsymbol{\theta}_2$, are interior to Ω_1, or Ω_2, so that the type of distribution problem discussed by Chernoff (1954) is excluded.

Let $l_1(\hat{\boldsymbol{\theta}}_1)$ be the maximized log-likelihood function of the model proposed under H_1 and $l_2(\hat{\boldsymbol{\theta}}_2)$ be the maximized log-likelihood function under H_2, where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are the maximum likelihood estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. Cox proposed using the following test statistic:

$$T_1 = l_1(\hat{\boldsymbol{\theta}}_1) - l_2(\hat{\boldsymbol{\theta}}_2) - E\left[l_1(\hat{\boldsymbol{\theta}}_1) - l_2(\hat{\boldsymbol{\theta}}_2)\right]_{\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_1}, \tag{1}$$

which compares the observed difference of log-likelihoods with an estimate of the expected difference between log-likelihoods, with expectation taken under H_1. In the expected difference $\boldsymbol{\theta}_1$ is replaced with its maximum likelihood estimate, $\hat{\boldsymbol{\theta}}_1$, under H_1; other unknown paramters are also replaced with estimates under the null. Under the null hypothesis, Cox (1961; 1962) showed the value of T_1 should be nearly zero, while under H_2, T_1 is presumed to be negative. Thus, a large negative value of T_1 leads to the rejection of the null hypothesis, H_1.

Alternatively, a simplified specification of T_1 is given by

$$T_1 = \hat{l}_{12} - E\left[\hat{l}_{12}\right]_{\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_1}$$

where $\hat{l}_{12} = l_1\left(\hat{\boldsymbol{\theta}}_1\right) - l_2\left(\hat{\boldsymbol{\theta}}_2\right)$

Furthermore, a different specification of T_1 is given by

$$T_1 = \hat{l}_{12} - N\left(\text{plim}_{N\to\infty}\frac{\hat{l}_{12}}{N}\right)_{\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_1}.$$

Replacing the expectation in the second term with a probability limit (plim) allows for more flexibility when working with complex expressions, especially those involving products and quotients of random variables. (Dougherty, 2011; **Appendix A**)

A few general remarks are in order (White, 1982):

(i) Usually in likelihood ratio applications, $\boldsymbol{\Omega}_1 \subset \boldsymbol{\Omega}_2$, so that $l_{12} < 0$. Under the assumption of separate families, this inequality - representative of the models having nested structures - may not hold.

(ii) When the components of $\boldsymbol{y}$ are independent, l_{12} is the sum of n independent terms and an application of the central limit theorem will usually prove the asymptotic normality of l_{12}. Approximations to the percentage points of l_{12} can then be obtained under both H_1 and H_2.

(iii) The hypotheses H_1 and H_2 are considered asymmetrically, and are not assumed to be the only possible hypotheses.

(iv) The roles of H_1 and H_2 can be interchanged, yielding corresponding test statistic T_2, where

$$T_2 = l_2(\hat{\boldsymbol{\theta}}_2) - l_1(\hat{\boldsymbol{\theta}}_1) - E\left[l_2(\hat{\boldsymbol{\theta}}_2) - l_1(\hat{\boldsymbol{\theta}}_1)\right]_{\boldsymbol{\theta}_2=\hat{\boldsymbol{\theta}}_2}. \qquad (2)$$

Here, expectation is taken under the new null, H_2, and $\boldsymbol{\theta}_2$ is replaced by its maximum likelihood estimate under H_2, $\hat{\boldsymbol{\theta}}_2$, while all other unknown parameters are replaced by their estimates under H_2.

In order to make inferences using T_1 and/or T_2, one must derive the distribution of these statistics. Cox established that both test statistics have a limiting distribution that is Normal with mean 0 and Cox denoted the variances of T_1 and T_2, $Var(T_1)$ and $Var(T_2)$, respectively, by: $Var(T_1) = V_1(l_{12}) - \mathbf{G_1' I_1^{-1} G_1}$ and $Var(T_2) = V_2(l_{21}) - \mathbf{G_2' I_2^{-1} G_2}$, where $V_1(l_{12})$ is the variance of the log-likelihood ratio taken under the null hypothesis, and $V_2(l_{21})$ is defined correspondingly for the case where the null and alternative hypotheses are interchanged,

$$\mathbf{G_1} = N\frac{\partial}{\partial\boldsymbol{\theta}_1}\operatorname{plim}_{N\to\infty}\frac{\hat{l}_{12}}{N}$$
$$\mathbf{G_2} = N\frac{\partial}{\partial\boldsymbol{\theta}_2}\operatorname{plim}_{N\to\infty}\frac{\hat{l}_{21}}{N}$$

and $\mathbf{I_1}$ and $\mathbf{I_2}$ are the information matrices corresponding to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. Thus, for any pair of hypotheses, two tests may be constructed.

In the following sections, we derive T_1 and T_2 and their corresponding distributions for a pair of linear mixed models with nonnested fixed effects.

Coulibaly and Brorsen (1999) first proposed using a parametric bootstrap to estimate the distribution of test statistics under the corresponding null hypothesis, showing that this technique helped achieve correct test size and higher power, especially in small samples. Dameus et al. (2002), Monfardini (2003), and Godfrey (2007) also establish a need for bootstrapping the distributions of statistics for comparing various types of models applied to econometrics analyses. Huber (1967) discusses the behavior of maximum likelihood estimates under nonstandard conditions, but not much of the literature following Cox's work adequately addresses the distribution of test statistics and their limitations under different variance estimates.

Definition of Separate Families

The hypotheses, H_1 and H_2, are separate in the sense that an arbitrary simple hypothesis H_1 cannot be obtained as a simple limit hypothesis in H_2. That is, $f(y, \theta_1)$ and $g(y, \theta_2)$ represent separate families in that an arbitrary value of θ_1 cannot be approximated arbitrarily closely by $g(y, \theta_2)$.

Existing examples

In his seminal works, Cox presented several scenarios in which one could apply his method; selected examples are described below.

Lognormal vs. Exponential In this example, the null hypothesis assumes a model following a lognormal distribution; that is, $f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{\frac{(\ln y - \mu)^2}{2\sigma^2}}$. The alternative hypothesis assumes an exponential model, given by $g(y) = \theta e^{-\theta y}$. Cox (1961, 1962) showed that

$T_1 = n \log \frac{\hat{\beta}}{\beta_{\hat{\alpha}}}$, where $\beta_{\hat{\alpha}} = e^{\hat{\alpha}_1 + \frac{\alpha_2}{2}}$, and $Var(T_1) = n\left(e^{\hat{\alpha}_2} - 1 - \hat{\alpha}_2 - \frac{1}{2}\hat{\alpha}_2^2\right)$. Not shown here or in other examples, it is trivial to derive the statistic T_2 which assumes $g(y)$ is the null hypothesis, $f(y)$ as the alternative, along with an expression for its variance. **Alternative forms of independent variables** Consider comparing the following sets of models

$$y = a\alpha$$
$$y = b\beta$$

or

$$y = \alpha_1 + \alpha_2 x$$
$$y = \beta_1 + \beta_2 \log x$$

Cox (1961) explained that the two pairs of models above are from a problem considered by Hotelling (1940), and expounded upon in Williams (1959); Cox demonstrated the need for more theoretical exploration for nonnested models of this type.

Alternative forms of the dependent variable

$$E\left(\log y\right) = a\alpha$$
$$E\left(y\right) = a\beta$$

Also from Cox (1961), this example is accompanied by the scenario that in a simple factorial experiment, $E\left(\log y\right) = a\alpha$ has no second order interactions, and the same for the other model; furthermore, the models are nonnested since one model cannot be obtained as a simple limit of the other .

Poisson vs. Geometric distributions

$$\begin{aligned} f_Y(y) &= \frac{e^{-\alpha}\alpha^y}{y!} \quad (y = 0, 1, 2, \ldots) \ ; \\ g_Y(y) &= \frac{\beta^y}{(1+\beta)^{y+1}} \quad (y = 0, 1, 2, \ldots) \end{aligned}$$

Cox (1962) explained that this example, demonstrating the comparison of exact versus asymptotic statistical theory, results in a test statistic $T_1 = -\Sigma \log Y_i! + nl_f\left(\bar{Y}\right)$, where $l_f(\cdot)$ is the log-likelihood function of $f_Y(y)$, which assumes the model $f_Y(y)$ is the null hypothesis and $g_Y(y)$ is the alternative.

Demand analysis models with nonnested functional forms

Dameus et al. (2002) explored comparisons of nonnested demand analysis models (US meat demand), and showed that using a Cox test with a parametric bootstrap was more powerful than using encompassing tests. The authors compared the following two demand analysis models.

First-difference AIDS model:

$$\Delta s_i = \tau_i + \sum_{k=1}^{4} \theta_{ik} D_k + \sum_{j=1}^{4} \gamma_{ij} \Delta \ln(p_j) + \beta_i \left[\Delta \ln(x) - \Delta \ln(P)\right], \quad i = 1, \ldots, 4$$

Rotterdam model:

$$\bar{s}_i \Delta \ln(y_i) = \tau_i + \sum_{k=1}^{4} \theta_{ik} D_k + \sum_{j=1}^{4} \gamma_{ij} \Delta \ln(p_j) + \beta_i \left[\Delta \ln(x) - \sum_{j=1}^{4} \bar{s}_j \Delta \ln(p_j)\right]$$

Multinomial Probit vs. Multinomial Logit Model

Monfardini (2003) compared the multinomial probit vs. multinomial logit model for assessing choice of labor sector among Italian workers.

Multinomial probit model:

$$u_{ij}^{p} = \underline{\mathrm{x}}_i'\alpha_j + \nu_{ij} \quad (\nu_i) = (\nu_{i1}, \ldots, \nu_{iJ-1})' \approx i.i.d.\mathcal{N}\left(\underline{0}, \Sigma\right), j = 1, \ldots, J-1,$$

Multinomial logit model:

$$u_{ij}^{l} = \underline{\mathrm{x}}_i'\delta_j + \eta_{ij} \quad (\eta_i) = (\eta_{i1}, \ldots, \eta_{iJ-1})' \approx i.i.d.Logistic\left(\underline{0}, \Lambda\right), j = 1, \ldots, J-1,$$

To date, this methodology has not been extended to the case of comparing linear mixed models with nonnested fixed and/or random effects. The last two of the extensions outlined below (Pesaran (1974) and Araujo et al. (2005)) provide the most adequate framework for developing test statistics for nonnested linear mixed models. We consider these extensions in more detail in the sections below.

Univariate Linear Regression

Pesaran (1974) derived test statistics to compare two univariate linear regression models with nonnested fixed effects. This section summarizes the formulation. Suppose there are data for N independent subjects; consider the following hypotheses:

$$\begin{aligned} H_1 &: \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{e}_1; \quad \boldsymbol{e}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{1}}^{\mathbf{2}}\mathbf{I_N}) \\ H_2 &: \boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta}_2 + \boldsymbol{e}_2; \quad \boldsymbol{e}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{2}}^{\mathbf{2}}\mathbf{I_N}). \end{aligned}$$

For models in each hypothesis, $\boldsymbol{y}$ is an $(N \times 1)$ vector containing the observed continuous outcome for each of the N subjects. The null hypothesis (H_1) assumes a univariate linear regression model with known $(N \times p)$ design matrix $\boldsymbol{X}$ and random errors distributed normally with mean $\mathbf{0}$ and variance $\sigma_1^2\mathbf{I_N}$, where $\mathbf{I_N}$ is the $(N \times N)$ identity matrix. The alternative hypothesis (H_2) assumes a univariate linear regression model with known $(N \times p)$

design matrix $\boldsymbol{W}$ and random errors also distributed normally with mean $\mathbf{0}$ and variance $\sigma_2^2\mathbf{I_N}$, with $\mathbf{I_N}$ defined as before. Furthermore, assume that $\boldsymbol{X}$ and $\boldsymbol{W}$ are not nested; that is, all of columns of $\boldsymbol{X}$ cannot be obtained from those of $\boldsymbol{W}$ and vice versa. For simplicity, we assume that $\boldsymbol{X}$ and $\boldsymbol{W}$ have the same dimensions. Recall from the discussion of Levy and Hancock (2007) that most often considered is the case of partially overlapping models, where there may be some variables in common between the models but neither design matrix is a subset of the other. We denote the collections of unknown parameters of each model as $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1', \sigma_1^2)'$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2', \sigma_2^2)'$, both vectors having dimension $(p+1 \times 1)$. Pesaran required that the following three limits exist and are finite.

$$\lim_{N\to\infty}\left(\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}\right) = \boldsymbol{\Sigma}_{X'X}$$
$$\lim_{N\to\infty}\left(\frac{1}{N}\boldsymbol{W}'\boldsymbol{W}\right) = \boldsymbol{\Sigma}_{W'W}$$
$$\lim_{N\to\infty}\left(\frac{1}{N}\boldsymbol{X}'\boldsymbol{W}\right) = \boldsymbol{\Sigma}_{X'W}$$

where the matrices $\boldsymbol{\Sigma}_{X'X}$ and $\boldsymbol{\Sigma}_{W'W}$ are nonsingular and $\boldsymbol{\Sigma}_{X'W} \neq \mathbf{0}$. All matrices are of dimension $(N \times N)$.

Formulation of expressions for T_1 and $Var(T_1)$

First, the log-likelihood functions corresponding respectively to the linear models given in hypotheses H_1 and H_2 are defined below:

$$l_1(\boldsymbol{\theta}_1) = -\frac{N}{2}\log\left(2\pi{\sigma_1}^2\right) - \frac{1}{2{\sigma_1}^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_1\right)'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_1\right)$$
$$l_2(\boldsymbol{\theta}_2) = -\frac{N}{2}\log\left(2\pi{\sigma_2}^2\right) - \frac{1}{2{\sigma_2}^2}\left(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}_2\right)'\left(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}_2\right),$$

Defining the log-likelihood functions of H_1 and H_2 as $l_1(\boldsymbol{\theta_1})$ and $l_2(\boldsymbol{\theta_2})$, respectively, and the maximum log-likelihood ratio (or difference in log-likelihood functions) by $\hat{l}_{12} = l_1\left(\hat{\boldsymbol{\theta}}_1\right) - l_2\left(\hat{\boldsymbol{\theta}}_2\right)$, recall the formula for T_1 given by:

$$T_1 = \hat{l}_{12} - N\left[\text{plim}_{N\to\infty}\left(\hat{l}_{12}/N\right)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_1},$$

with *plim* taken under the assumption of H_1. The maximum likelihood estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, given by $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$, respectively, are defined as follows.

First, $\hat{\boldsymbol{\theta}}_1' = \left(\hat{\boldsymbol{\beta}}_1', \hat{\sigma}_1^2\right)$, where

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y},\\ \hat{\sigma}_1^2 &= \frac{\boldsymbol{y}'\left(I_N - M_X\right)\boldsymbol{y}}{N}.\end{aligned}$$

$(N \times N)$ matrix I_N is defined as before, and $(N \times N)$ matrix M_X is given by $M_X = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$.

Next, $\hat{\boldsymbol{\theta}}_2 = \left(\hat{\boldsymbol{\beta}}_2, \hat{\sigma}_2^2\right)$, where

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\boldsymbol{y},\\ \hat{\sigma}_2^2 &= \frac{\boldsymbol{y}'\left(\mathbf{I_N} - \mathbf{M}_{\boldsymbol{W}}\right)\boldsymbol{y}}{N}.\end{aligned}$$

Here, $(N \times N)$ matrix is defined by $\mathbf{M_W} = \boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{\mathbf{-1}}\boldsymbol{W}'$.

Now,

$$l_{12} = \frac{N}{2}\log\left({\sigma_2}^2/{\sigma_1}^2\right) + \frac{1}{2}\left[{\sigma_2}^{-2}\left(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}_2\right)'\left(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}_2\right) - {\sigma_1}^{-2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_1\right)'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_1\right)\right].$$

To compute $\hat{l}_{12}$, we replace the elements of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ with their respective maximum likelihood estimates as defined above.

$$\begin{aligned}
\hat{l}_{12} &= \frac{N}{2}\log\left(\hat{\sigma}_2^2/\hat{\sigma}_1^2\right) + \frac{1}{2}\left[\hat{\sigma}_2^{-2}\left(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)'\left(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right) - \hat{\sigma}_1^{-2}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)\right] \\
&= \frac{N}{2}\log\left(\hat{\sigma}_2^2/\hat{\sigma}_1^2\right) + \frac{1}{2}\Big[\left(\boldsymbol{y}'\left(\mathbf{I} - \mathbf{M_W}\right)\boldsymbol{y}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)'\left(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right) \\
&\quad - \left(\boldsymbol{y}'\left(\mathbf{I} - \mathbf{M_X}\right)\boldsymbol{y}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)\Big] \\
&= \frac{N}{2}\log\left(\hat{\sigma}_2^2/\hat{\sigma}_1^2\right) + \frac{1}{2}\Big[\left(\boldsymbol{y}'\left(\mathbf{I} - \mathbf{M_W}\right)\boldsymbol{y}\right)^{-1}\left(\boldsymbol{y}'\left(\mathbf{I} - \mathbf{M_W}\right)\boldsymbol{y}\right) \\
&\quad - \left(\boldsymbol{y}'\left(\mathbf{I} - \mathbf{M_X}\right)\boldsymbol{y}\right)^{-1}\left(\boldsymbol{y}'\left(\mathbf{I} - \mathbf{M_X}\right)\boldsymbol{y}\right)\Big] \\
&= \frac{N}{2}\log\left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\right)
\end{aligned}$$

Determining the second term of T_1 requires finding the probability limit (plim) of $\hat{l}_{12}$ under the assumption that the null hypothesis is the true model. First, by definition, we know that $\text{plim}_{N\to\infty}\left(\hat{\sigma}_1^2\right) = \sigma_1^2$. Assuming H_1 is the true model, Pesaran (1974) determined the expected value of $\hat{\sigma}_2^2$, denoted by $\hat{\sigma}_{21}^2$ as follows.

$$\hat{\sigma}_{21}^2 = \frac{1}{N}\left(M_W e_1 + M_W\boldsymbol{X}\boldsymbol{\beta}_1\right)'\left(M_W e_1 + M_W\boldsymbol{X}\boldsymbol{\beta}_1\right)$$

Now taking the probability limit of the above quantity, we have

$$\text{plim}_{N\to\infty}\left(\hat{\sigma}_{21}^2\right) = \sigma_1^2 + \boldsymbol{\beta}_1'\lim_{N\to\infty}\left(\frac{1}{N}\boldsymbol{X}'M_W\boldsymbol{X}\right)\boldsymbol{\beta}_1$$

We also have

$$\sigma_{21}^2 = \sigma_1^2 + \boldsymbol{\beta}_1' H\boldsymbol{\beta}_1$$

where

$$H = \mathbf{\Sigma}_{X'X} - \mathbf{\Sigma}_{X'W}\mathbf{\Sigma}_{W'W}^{-1}\mathbf{\Sigma}_{W'X}$$

Now, since under H_1 $\hat{\sigma}_1^2$ is a consistent estimator of σ_1^2, using the above results

$$N\,\text{plim}_{N\to\infty}\left(\frac{\hat{l}_{21}}{N}\right) = \frac{N}{2}\log\left(\frac{\sigma_1^2 + \boldsymbol{\beta}_1' H \boldsymbol{\beta}_1}{\sigma_1^2}\right)$$

Combining the two terms, Pesaran (1974) gave an expression for T_1 as follows:

$$T_1 = \frac{N}{2}\log\left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\right) - \frac{N}{2}\log\left(\frac{\hat{\sigma}_1^2 + \hat{\boldsymbol{\beta}}_1' H \hat{\boldsymbol{\beta}}_1}{\hat{\sigma}_1^2}\right) \tag{3}$$

$$= \frac{N}{2}\log\left(\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2 + \hat{\boldsymbol{\beta}}_1' H \hat{\boldsymbol{\beta}}_1}\right) \tag{4}$$

Distribution of T_1

Pesaran (1974) noted that the distribution of T_1 depends on unknown parameters, since replacing $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ with expressions of their estimates listed above produces a complicated function of the unknown vector $\boldsymbol{\beta}_1$. As a result, the exact distribution of T_1 cannot be derived. The only way to eliminate the dependence on unknown parameters is to compare models that are nested, such that $\mathbf{M_W}\boldsymbol{X} = \mathbf{0}$, which is the case for which a hypothesis test is well-defined.

To obtain an asymptotic variance of T_1, denoted by $\widehat{Var}\left(T_1\right)$, we must derive the two terms according to the formula defined in equation 8.

Pesaran defined the first term as follows:

$$\hat{V}\left(\hat{l}_{12}\right) = \frac{N}{2}\left(\frac{1}{\sigma_{21}^2} - \frac{1}{\sigma_1^2}\right)^2 \sigma_1^4 + \frac{\sigma_1^2}{\sigma_{21}^4}\left(\boldsymbol{X}\boldsymbol{\beta}_1 - \boldsymbol{W}\boldsymbol{\beta}_{21}\right)'\left(\boldsymbol{X}\boldsymbol{\beta}_1 - \boldsymbol{W}\boldsymbol{\beta}_{21}\right)$$

The second term is given by:

$$\frac{1}{N}G_1' \,\mathrm{plim}_{N\rightarrow\infty}\left(NI^{-1}\right)G_1 = \frac{N}{\sigma_{21}^4}\left[\sigma_1^2\boldsymbol{\beta}_1' H \boldsymbol{\Sigma}_{X'X}^{-1} H\boldsymbol{\beta}_1 + \frac{1}{2}\left(\boldsymbol{\beta}_1' H\boldsymbol{\beta}_1\right)^2\right]$$

Combining the two terms and replacing all unknown parameters with their consistent estimates (under H_1), Pesaran's estimate of $\widehat{Var}(T_1)$ is given by

$$\widehat{Var}(T_1) = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_{21}^4}\hat{\boldsymbol{\beta}}_1'\boldsymbol{X}' M_W M_X M_W \boldsymbol{X}\hat{\boldsymbol{\beta}}_1$$

Finally, when H_1 is true, we have that $\frac{T_1}{\left[\widehat{Var}(T_1)\right]^{1/2}}$ approx$\mathcal{N}\left(0,1\right)$

Expressions for T_2 and $Var(T_2)$

When two models are nested, then the choice of a null hypothesis is fairly intuitive; one typically sets the most parsimonious (or *reduced*) model as the null hypothesis. When models are not nested, one must consider that either candidate model can be set as the null hypothesis. In order to implement Cox's test of separate hypotheses, one must consider the case that the second model is the true model. Thus, in this case, one can interchange the models given in each hypothesis and formulate test statistic T_2 and its variance given by $Var\left(T_2\right)$.

$$H_1 : \boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta}_2 + \boldsymbol{e}_2\,;\, \boldsymbol{e}_2 \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\mathbf{2}}^{\mathbf{2}}\mathbf{I_N}\right)$$
$$H_2 : \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{e}_1\,;\, \boldsymbol{e}_1 \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\mathbf{1}}^{\mathbf{2}}\mathbf{I_N}\right)$$

Both models are specified exactly as before, and we now indicate log-likelihood functions from the null and alternative hypotheses as $l_1\left(\boldsymbol{\theta}_2\right)$ and $l_2\left(\boldsymbol{\theta}_1\right)$, respectively. $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are defined as before.

Recall the expression for T_2 given by:

$$T_2 = \hat{l}_{21}\left(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_1\right) - N\left[plim_{N\to\infty}\left(\hat{l}_{21}/N\right)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_2},$$

where $\hat{l}_{21} = l_1\left(\hat{\boldsymbol{\theta}}_2\right) - l_2\left(\hat{\boldsymbol{\theta}}_1\right)$.

Following similarly to the formulation of T_1, it can be shown that an expression for T_2 is given by

$$T_2 = \frac{N}{2}\log\left(\frac{\sigma_1^2}{\sigma_1^2 + \boldsymbol{\beta}_2'\tilde{H}\boldsymbol{\beta}_2}\right),$$

where $\tilde{H} = \boldsymbol{\Sigma}_{W'W} - \boldsymbol{\Sigma}_{W'X}\boldsymbol{\Sigma}_{X'X}^{-1}\boldsymbol{\Sigma}_{X'W}$ with $\boldsymbol{\Sigma}_{W'W}$, $\boldsymbol{\Sigma}_{W'X}$, and $\boldsymbol{\Sigma}_{X'X}$ defined as before.

Also following similarly from previous sections, an asymptotic expression for $Var(T_2)$ is given by

$$\widehat{Var}(T_2) = \frac{\sigma_2^2}{\sigma_{12}^4}\hat{\boldsymbol{\beta}}_2'W'M_XM_WM_XW\hat{\boldsymbol{\beta}}_2$$

Finally, it is also true that $\frac{T_2}{\widehat{Var}(T_2)^{1/2}}$ approx $\sim \mathcal{N}(0,1)$.

1.4.2 Multivariate Linear Regression

Another extension of Cox's methodology for comparing nonnested models was introduced by Araujo et al. (2005). Following from the previous example by Pesaran (1974) and subsequent works involving the same author, the authors developed test statistics to compare nonnested multivariate regression models. Consider the following set of hypotheses:

$$H_1 : \boldsymbol{Y} = \boldsymbol{X}\mathbf{B_1} + \boldsymbol{E_1}$$

$$H_2 : \boldsymbol{Y} = \boldsymbol{W}\mathbf{B_2} + \boldsymbol{E_2}$$

Here, $\boldsymbol{Y}$ is an $N \times k$ matrix of information on k continous outcome variables corresponding to each of N subjects. $\boldsymbol{X}$ and $\boldsymbol{W}$ are, respectively, $N \times p$ and $N \times q$ matrices of regressors;

$\mathbf{B_1}$ and $\mathbf{B_2}$ are, respectively, $p \times k$ and $q \times k$ matrices of parameters. Matrices $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are $N \times k$ matrices whose rows are independent and identically distributed normal random vectors with means equal to zero and $N \times N$ covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively.

So, it follows that $\boldsymbol{E_1} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I_N} \otimes \boldsymbol{\Sigma_1}\right)$ and $\boldsymbol{E_2} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I_N} \otimes \boldsymbol{\Sigma_2}\right)$, and that $\boldsymbol{Y} \sim \mathcal{N}\left(\boldsymbol{X}\mathbf{B_1}, \mathbf{I_N} \otimes \boldsymbol{\Sigma_1}\right)$ under H_1 and $\boldsymbol{Y} \sim \mathcal{N}\left(\boldsymbol{W}\mathbf{B_2}, \mathbf{I_N} \otimes \boldsymbol{\Sigma_2}\right)$ under H_2. That is, $\boldsymbol{Y}$, $\boldsymbol{E}_1$, $\boldsymbol{E}_2$ follow multivariate Normal distributions.

The models represented in each hypothesis are nonnested in that one cannot obtain the columns of $\boldsymbol{X}$ from the columns of $\boldsymbol{W}$, and vice versa. As in Pesaran's example, further assumptions to justify the models being nonnested include the following:

$$\begin{aligned}
\lim_{N\to\infty}\left(\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}\right) &\equiv \boldsymbol{\Sigma}_{X'X} \\
\lim_{N\to\infty}\left(\frac{1}{N}\boldsymbol{W}'\boldsymbol{W}\right) &\equiv \boldsymbol{\Sigma}_{W'W} \\
\lim_{N\to\infty}\left(\frac{1}{N}\boldsymbol{X}'\boldsymbol{W}\right) &\equiv \boldsymbol{\Sigma}_{X'W}.
\end{aligned}$$

Further, it is assumed that $\boldsymbol{\Sigma}_{X'X}$ and $\boldsymbol{\Sigma}_{W'W}$ are nonsingular, and that $\boldsymbol{\Sigma}_{X'W}$ is a non-zero matrix.

We may collect the unknown parameters from each model into $((pk + Nk) \times 1)$ vectors $\boldsymbol{\theta}_1 = (vec\mathbf{B_1}, vec\boldsymbol{\Sigma}_1)'$ and $\boldsymbol{\theta}_2 = (vec\mathbf{B_2}, vec\boldsymbol{\Sigma}_2)'$ corresponding respectively to the models given in H_1 and H_2.

Note that the log-likelihood functions corresponding respectively to the two nonnested models under consideration are given by

$$\begin{aligned}
l_1(\boldsymbol{\theta}_1) &= -\frac{N}{2}\log\left|\Sigma_1^{-1}\right| - \frac{kN}{2}\log(2\pi) - \frac{1}{2}tr\left(\boldsymbol{Y} - \boldsymbol{X}\mathbf{B_1}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\mathbf{B_1}\right)\boldsymbol{\Sigma}_1^{-1} \\
l_2(\boldsymbol{\theta}_2) &= -\frac{N}{2}\log\left|\Sigma_2^{-1}\right| - \frac{kN}{2}\log(2\pi) - \frac{1}{2}tr\left(\boldsymbol{Y} - \boldsymbol{W}\mathbf{B_2}\right)'\left(\boldsymbol{Y} - \boldsymbol{W}\mathbf{B_2}\right)\boldsymbol{\Sigma}_2^{-1}
\end{aligned}$$

Formulation of Expressions for T_1 and $Var(T_1)$ To compute an expression for T_1, as defined in equation 7, Araujo et al. first define $\hat{l}_{12} = \frac{N}{2}\left(\log\left|\hat{\boldsymbol{\Sigma}}_2\right| - \log\left|\hat{\boldsymbol{\Sigma}}_1\right|\right)$, where $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{N}\hat{\boldsymbol{E}}_1^{'}\hat{\boldsymbol{E}}_1$ and $\hat{\boldsymbol{\Sigma}}_2 = \frac{1}{N}\hat{\boldsymbol{E}}_2^{'}\hat{\boldsymbol{E}}_2$. Under H_1, $\hat{\boldsymbol{E}}_1 = M_X\boldsymbol{E}_1$ and $\hat{\boldsymbol{E}}_2 = \mathbf{M_W}\boldsymbol{E_1} + \mathbf{M_W}\boldsymbol{X}\mathbf{B_1}$, where $\mathbf{M_X} = \mathbf{I_N} - \boldsymbol{X}\left(\boldsymbol{X'X}\right)^{\mathbf{-1}}\boldsymbol{X'}$ and $\mathbf{M_W} = \mathbf{I_N} - \boldsymbol{W}\left(\boldsymbol{W'W}\right)^{\mathbf{-1}}\boldsymbol{W'}$.

So it follows that,

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_2 &= \frac{1}{N}\hat{\boldsymbol{E}}_2^{'}\hat{\boldsymbol{E}}_2 = \frac{1}{N}\left(\mathbf{M_W}\boldsymbol{E_1} + \mathbf{M_W}\boldsymbol{X}\mathbf{B_1}\right)'\left(\mathbf{M_W}\boldsymbol{E_1} + \mathbf{M_W}\boldsymbol{X}\mathbf{B_1}\right) \\ &= \frac{1}{N}\left(\boldsymbol{E}_1'\mathbf{M_W}\boldsymbol{E_1} + \mathbf{B_1'}\boldsymbol{X}'\mathbf{M_W}\boldsymbol{E_1} + \boldsymbol{E_1'}\mathbf{M_W}\boldsymbol{X}\mathbf{B_1} + \mathbf{B_1'}\boldsymbol{X}'\mathbf{M_W}\mathbf{X}\mathbf{B_1}\right).\end{aligned}$$

Now, under H_1, the expectation of $\hat{\boldsymbol{\Sigma}}_2$ is given by $\boldsymbol{\Sigma}_{21} = E\left(\hat{\boldsymbol{\Sigma}}_2\right)_{H_1} = \boldsymbol{\Sigma}_1 + \mathbf{B_1'}\bar{\boldsymbol{\Sigma}}\mathbf{B_1}$, where $\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{X'X} - \boldsymbol{\Sigma}_{X'W}\boldsymbol{\Sigma}_{W'W}^{-1}\boldsymbol{\Sigma}_{W'X}$, and as $\hat{\boldsymbol{\Sigma}}_1$ converges to $\boldsymbol{\Sigma}_1$ in probability under H_1, we have

$$N\,\mathrm{plim}_{N\to\infty}\frac{l_{12}\left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right)}{N} = \frac{N}{2}\left(\log\left|\boldsymbol{\Sigma}_1 + \mathbf{B_1'}\bar{\boldsymbol{\Sigma}}\mathbf{B_1}\right| - \log\left|\boldsymbol{\Sigma}_1\right|\right).$$

Putting the two terms together, an expression for T_1 is given below.

$$\begin{aligned}T_1 &= \frac{N}{2}\left(\log\left|\hat{\boldsymbol{\Sigma}}_2\right| - \log\left|\hat{\boldsymbol{\Sigma}}_1\right|\right) - \frac{N}{2}\left(\log\left|\boldsymbol{\Sigma}_1 + \mathbf{B_1'}\bar{\boldsymbol{\Sigma}}\mathbf{B_1}\right| - \log\left|\boldsymbol{\Sigma}_1\right|\right) \\ &= \frac{N}{2}\left(\log\left|\hat{\boldsymbol{\Sigma}}_2\right| - \log\left|\hat{\boldsymbol{\Sigma}}_1 + \frac{\hat{\mathbf{B}}_1'\boldsymbol{X}'\mathbf{M_W}\boldsymbol{X}\hat{\mathbf{B}}_\mathbf{1}}{N}\right|\right).\end{aligned}$$

Note that $\frac{1}{N}\boldsymbol{X}'\mathbf{M_W}\boldsymbol{X}$ is a consistent estimator of $\bar{\boldsymbol{\Sigma}}$.

As in the Pesaran (1974) extension, the distribution of T_1 under H_1 is dependent upon unknown parameters. Only when models are nested - that is, when $\mathbf{M_W}\boldsymbol{X} = \mathbf{0}$ - does this dependence go away.

Araujo et al. (2005) derived the variance of T_1, given by $Var(T_1)$, which also is dependent upon unknown parameters. Their formula is derived by the following series of equations.

$$\begin{aligned}
Var(T_1) = {} & tr\left[\left(\boldsymbol{X}\mathbf{B_1} - \boldsymbol{W}\mathbf{B_{21}}\right)\boldsymbol{\Sigma}_{21}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{21}^{-1}\left(\boldsymbol{X}\mathbf{B_1} - \boldsymbol{W}\mathbf{B_{21}}\right)'\right] \\
& + \frac{N}{2} tr\left[\left(\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\boldsymbol{\Sigma}_1\right]^2 \\
& - \frac{N}{4}\Bigg[\left(2\, vec\, \bar{\boldsymbol{\Sigma}}\mathbf{B_1}\boldsymbol{\Sigma}_{\mathbf{21}}^{\mathbf{-1}}\right)'\left(\boldsymbol{\Sigma}_{X'X}^{-1} \otimes \boldsymbol{\Sigma}_1\right)\left(2\, vec\, \bar{\boldsymbol{\Sigma}}\mathbf{B_1}\boldsymbol{\Sigma}_{\mathbf{21}}^{\mathbf{-1}}\right) \\
& + \left(2\, vech\, \left[\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1}\right] - vech\, diag\left[\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1}\right]\right)'\Delta \\
& \left(2\, vech\, \left[\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1}\right] - vech\, diag\left[\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1}\right]\right)\Bigg],
\end{aligned}$$

where Δ is a complex function of the information matrix of $vech\ \boldsymbol{\Sigma}_1$. More details on these computations are given in the containing paper. Most notably, in this extension to multivariate linear regression models, the formula for $Var(T_1)$ is even more complex than that for the case of comparing univariate linear regression models.

All in all, $T_1\widehat{Var}(T_1)^{-1/2}T_1 approx \sim \mathcal{N}\left(0, 1\right)$.

Expressions for T_2 and $Var(T_2)$

Araujo et al. (2005) did not provide expressions for T_2 and $Var(T_2)$, which are derived similarly to the formulations of T_1 and $Var(T_1)$. Here, we provide expressions for both quantities.

$$T_2 = \frac{N}{2}\left(\log\left|\hat{\boldsymbol{\Sigma}}_1\right| - \log\left|\hat{\boldsymbol{\Sigma}}_2 + \frac{\hat{\mathbf{B}}_2'\boldsymbol{W}'\mathbf{M_X}\boldsymbol{W}\hat{\mathbf{B}}_\mathbf{2}}{N}\right|\right),$$

where $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{W'W} - \boldsymbol{\Sigma}_{W'X}\boldsymbol{\Sigma}_{X'X}^{-1}\boldsymbol{\Sigma}_{X'W}$ and $\frac{1}{N}\boldsymbol{W}'M_X\boldsymbol{W}$ is a consistent estimator of $\tilde{\boldsymbol{\Sigma}}$.

One can find that

$$
\begin{aligned}
Var(T_2) = tr & \left[\left(\boldsymbol{W}\mathbf{B_2} - \boldsymbol{X}\mathbf{B_{12}}\right)\boldsymbol{\Sigma}_{12}^{-1}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_{12}^{-1}\left(\boldsymbol{W}\mathbf{B_2} - \boldsymbol{X}\mathbf{B_{12}}\right)'\right] \\
& + \frac{N}{2} tr\left[\left(\boldsymbol{\Sigma}_{12}^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)\boldsymbol{\Sigma}_2\right]^2 \\
& - \frac{N}{4}\left(2\, vec\, \tilde{\boldsymbol{\Sigma}}\mathbf{B_2}\boldsymbol{\Sigma}_{\mathbf{12}}^{\mathbf{-1}}\right)'\left(\boldsymbol{\Sigma}_{W'W}^{-1} \otimes \boldsymbol{\Sigma}_2\right)\left(2\, vec\, \tilde{\boldsymbol{\Sigma}}\mathbf{B_2}\boldsymbol{\Sigma}_{\mathbf{12}}^{\mathbf{-1}}\right) \\
& + \frac{N}{4}\left(2\, vech\, \left[\boldsymbol{\Sigma}_{12}^{-1} - \boldsymbol{\Sigma}_2^{-1}\right] - vech\, diag\left[\boldsymbol{\Sigma}_{12}^{-1} - \boldsymbol{\Sigma}_2^{-1}\right]\right)'\Gamma \\
& \left(2\, vech\, \left[\boldsymbol{\Sigma}_{12}^{-1} - \boldsymbol{\Sigma}_2^{-1}\right] - vech\, diag\left[\boldsymbol{\Sigma}_{12}^{-1} - \boldsymbol{\Sigma}_2^{-1}\right]\right),
\end{aligned}
$$

where Γ is a complex function of the information matrix of $vech\, \boldsymbol{\Sigma}_2$.

Again, ${T_2\widehat{Var}(T_2)}^{-1/2}T_2 approx \sim \mathcal{N}\left(0, 1\right)$.

All of the previous examples demonstrate a clear need for a similar application to linear mixed models with nonnested functional forms.

1.5 Using Information Criteria to Distinguish Nonnested LMMs

In the linear mixed model, information criteria are most commonly used to compare models with nonnested fixed and/or random effects. As an alternative to a statistical test, the use of information criteria allow one to judge goodness-of-fit of candidate models to data. The premise of comparing models based on information functions dates back to the work of (Kullback and Leibler, 1951), who developed the Kullback-Leibler information (or *divergence, distance, discrepancy*) (Burnham and Anderson (2002)); hereafter, the *K-L information*) function which essentially measures the distance - or loss of information - between a candidate model and an assumed 'true' model. Since we are rarely able to identify the full knowledge of a true model, the development of information criteria overcomes this obstacle by estimating a relative expected K-L information. The formula for most information criteria include a goodness-of-fit term for the candidate model's maximized log likelihood function (usually -2 times the log-likelihood function) along with a penalty term involving the number of parameters estimated by the model to prevent overfitting and/or sample size. The general selection rule is 'smaller is better', meaning that a model having the lowest value of an information criterion - or lowest deviation from an assumed true model - is favored over other candidates. Below, several common information criteria whose theory has been extended to the linear mixed model are reviewed; their advantages and limitations are considered.

1.5.1 Commonly used information criteria

In this section, a brief overview of some information criteria that are commonly applied to linear mixed models are described in more detail. Particular attention is paid to the Akaike Information Criterion (AIC) first proposed in Akaike (1973) and several of its variants. For a more comprehensive review of all available information criteria for the linear mixed model, see Dimova et al. (2011), Greven and Kneib (2010), Wang and Schaalje (2009), Shang and Cavanaugh (2008), Gurka (2006), Pu and Niu (2006), Vaida and Blanchard (2005), Chen and Dunson (2003), Burnham and Anderson (2002), Ngo and Brand (2002), and Cavanaugh (1999).

Akaike Information Criterion (AIC) and variants

Kullback and Leibler (1951) defined the Kullback directed divergence as a measure of the distance between the true model and a candidate model. First proposed in Akaike (1973), the Akaike Information Criterion arises from the Kullback directed divergence. The formula for AIC is given by

$$AIC = -2\,l\,(\boldsymbol{y}\,|\boldsymbol{\theta}\,) + 2p$$

where l is the likelihood function for the observed data $\boldsymbol{y}$ $(N \times 1)$,and $p \times 1$ vector $\boldsymbol{\theta}$ represents the dimension of the unknown parameter space represented by vector $\boldsymbol{\theta}$.

As stated previously, AIC assumes that the candidate model is nested within a larger, 'true' model (Ngo and Brand, 2002). It favors a model with fewer parameters, all else fixed, according to the 'smaller is better' principle. When two models have the same number of parameters, but slightly different forms of a particular independent variable (e.g., continuous vs. categorical age parameterization), using the AIC to select the better model can be misleading.

Several variations of the AIC exist and their formulas are listed below (note that l and p are defined as before):

- $AICc = -2l + 2p\left(\frac{N}{N-p-1}\right)$, also referred to as a 'small-sample AIC (Hurvich and Tsai, 1989)

- $cAIC = -2l_c + 2\left(\hat{\rho} + 1\right)$, where l_c is the conditional log-likelihood function and $\hat{\rho}$ is the estimated effective degrees of freedom (Vaida and Blanchard, 2005; Dimova et al., 2011)

- Variants where the penalty term is estimated using parametric bootstrapping, as considered in Shang and Cavanaugh (2008)

The $AICc$ improves upon the AIC's small smaple bias by correcting for sample size within the penalty term. As a further improvement, Vaida and Blanchard (2005) proposed the Conditional Akaike Information Criterion ($cAIC$) . Liang et al. (2008) highlighted some limitations of Vaida and Blanchard's approach, and proposed a modification to the $cAIC$. Many bootstrap-based variants of the AIC have been developed, and many were shown to outperform the AIC among small samples and under varied estimation approaches; however, many researchers steer clear of this approach because of the assumed computational effort required.

Several investigations have shown that the AIC is inconsistent, particularly among small-sample data (Ngo and Brand, 2002). Gurka (2006) considers the AIC and the variants listed above for comparing fixed effects models; a Monte-Carlo simulation study led to inconclusive results on which criterion performs best. Moreover, Gurka emphasizes certain adjustments required for estimating these criteria under restricted maximum likelihood estimation (REML) and the impact that certain data characteristics can have on each criterion's performance. Most recently, Dimova et al. (2011) compared the *AIC* and some variants in selecting among nested models arising from moderate-sample data. Like other investigations, no one criteria outshone the others; however, a poor performance of the *cAIC* led the authors to recommend that it no longer be used to select linear mixed models. Their investigation omitted Schwarz's Bayesian Information Criterion and any bootstrap variants of the *AIC* such as the Extended Information Criterion (*EIC*), both of which are discussed in more detail below.

Schwarz's Bayesian Information Criterion (*BIC*)

In 1978, Scwharz proposed the Bayesian Information Criterion (*BIC*) as an alternative to the *AIC*. The formula for *BIC* is given by

$$BIC = -2\, l\, (\boldsymbol{y} \mid \boldsymbol{\theta}) + p\, ln\, N$$

where, like before, l represents the log-likelihood function for the $N \times 1$ observed data vector $\boldsymbol{y}$, and $\boldsymbol{\theta}$ is the $p \times 1$ vector of unknown parameters.

Like the *AIC*, the *BIC* penalizes models with many parameters, but this criterion also takes sample size into account penalizing high-dimensional data. Unlike the *AIC*, however, the *BIC* does not assume a true model and thus does not assume a nested structure among models.

Limitations of *AIC* and *BIC*

While the *AIC* and *BIC* are easy to compute, use, and interpret, one major criticism of both is that they are only relative measures of model discrimination; they offer no decision related to statistical significance of differences between candidate models. Additionaly, these and other information criteria will always favor one of the models under consideration, even if none represent the 'best' characterization of the relationship between covariates and the outcome variable. Both criteria have been outperformed by other IC given different types of models.

Preference of one over the other depends on several factors, including the sample size under consideration and the number of unknown parameters in each model. In many cases, the AIC and BIC will select different models.

Burnham and Anderson (2002) offer more discussion of AIC and BIC, and Gurka (2006) examines properties and performance of all of the above criteria under REML in the linear mixed model with inconclusive results as to a 'best' criterion.

A common criticism and major limitation of using an information criterion to select a model is that the absolute value of a criterion for a single model has no direct interpretation. Comparing the values for two models gives no indication of statistical significance in deciding which model is better, and there is no knowledge of the impact of the magnitude of difference between the values of a criterion from two candidate models. Furthermore, the AIC and most of its variants assume a true model within all candidate models are nested; this assumption is not practical for all sets of candidate models.

Li and Wong (2010) analyzed longitudinal data from a clinical trial among scleroderma patients. They considered models with nonnested covariance structures (though no discussion is given to how one distinguishes whether two covariance structures are nested) using the AIC and BIC to compare models. They did acknowledge the existence of variants

In spite of the limitations discussed here, these information criteria remain commonly used to distinguish linear mixed models with nonnested fixed and/or random effects. Many researchers continue to modify the AIC and BIC to account for the limitations listed above. One promising variant of the AIC is described in the following section.

1.5.2 The Extended Information Criterion (*EIC*)

Ishiguro et al. (1997) first proposed the extended information criterion (EIC), an extension of AIC that followed the methodology of Efron (1983); the goal of the EIC is to better estimate the bias of the sample log-likelihood using the bootstrap technique, as compared to the AIC. Like the AIC, the preferred model is ascertained using the 'smaller is better' judging criterion. However, unlike the AIC, the EIC can be used under various methods of parameter estimation (ML, REML, etc.). Further, the EIC does not rely on candidate models being nested.

Notation and Setup The calculation of the EIC is straightforward, but is not readily available for use in statistical software programs. Its general formula is given below.

$$EIC = -2\, l\left(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}\right) + 2\hat{C}^{*} \tag{5}$$

where $\hat{C}^{*}$ is a scalar quantity estimated by

$$\hat{C}^{*} = \frac{1}{B}\sum_{b=1}^{B}\left[l\left(\boldsymbol{y}_{b}{}^{*} \mid \hat{\boldsymbol{\theta}}_{b}^{*}\right) - l\left(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_{b}^{*}\right)\right]$$

where $l\left(\cdot\right)$ represents the log-likelihood function; $\boldsymbol{y}$, like before, is the $N \times 1$ vector containing observed subject responses; $\hat{\boldsymbol{\theta}}$ is the $r \times 1$ vector of unknown parameters corresponding the original data $\boldsymbol{y}$; $\boldsymbol{y}_{b}{}^{*}$ is the b^{th} bootstrap sample of the original data $\boldsymbol{y}$; $\hat{\boldsymbol{\theta}}_{b}^{*}$ is the set of fixed and random effects that maximize the likelihood function of the b^{th} bootstrap data set.

Essentially, $\hat{C}^{*}$ is the averaged (across all B bootstrap samples) difference in the maximum log-likelihood function of the bootstrapped data and the value of a log-likelihood function of the original data with bootstrapped parameter estimates. Among all candidate models, the model with the lowest EIC should be selected.

Implementation of the *EIC*

Pan (1999) compared performance of the *EIC*, *AIC*, and other model selection criteria. Their example data were a small sample with which these criteria were assessed under linear regression, logistic regression, and Cox regression. The author found that the *EIC* outperformed *AIC*, but was not as good as another technique - the BCV (bootstrap cross-validation).

Yafune et al. (2005) used the *EIC* to compare linear mixed models with nested mean and covariance structures under REML estimation, a case which we do not consider since one cannot compare log-likelihood functions of linear mixed models under REML estimation. The authors assessed the performance of the *EIC* using motivating data from two small-sample longitudinal studies. The first study is Potthoff and Roy's dental data Potthoff and Roy (1964), and the authors only used data for the 16 boys in study (observed at 4 equally spaced time points). Second example is data of platelet count for 12 ITP patients (observed at 6 time points). Simulation studies considered only seven models, of which only nested mean model structures were considered. While this paper leverages the use of well-designed longitudinal studies and simulations to assess the performance of the *EIC* and demonstrate its applicability to non-ML estimation procedures, and shows that *EIC* outperforms *AIC* in small sample studies, the authors do not consider large sample studies and do not compare models with nonnested structures.

Shang and Cavanaugh (2008) examined the *EIC* and another bootstrap variant of *AIC* in selecting among mixed models arising from clustered (not necessarily repeated measures) data with small and moderate sample sizes. Their results reinforced the findings of Yafune et al. (2005) that *EIC* outperforms *AIC* in small samples. Among large sample data, the performances of *EIC* and *AIC* in selecting a model were comparable. Again, only nested models were compared.

While the *EIC* was developed to improve the finite-sample bias, it has not been well-studied for its properties in models arising from large-sample data. In addition, its potential to distinguish between nonnested models has not been adequately addressed.

1.6 Summary

The linear mixed model is a very useful tool to characterize longitudinal data; having separately-modeled mean and covariance structures provides great flexibility to build a wide variety of models. However, it also complicates the process of selecting an adequate whole model by requiring correct specification of each structural component. For this reason, most existing model selection techniques focus on the selection of one model (mean or covariance) at a time while holding the other fixed. Techniques to select both mean and covariance models are ad hoc extensions from simpler univariate and multivariate data with uncorrelated continuous outcomes; it has been established that these extensions have not been well studied, nor have they been adequately tested in various model comparison scenarios for the linear mixed model. Particularly, cases where two candidate models are nonnested (in fixed and/or random effects) are rarely considered. Most researchers blindly use information criteria, which are uncalibrated to determine statistically significant differences between models and often lead to inconclusive results. Moreover, hypothesis testing has not been thoroughly leveraged as an option for selecting between nonnested models. Much of the literature on nonnested models is based in econometrics, social science research, and pharmacology; while examples similar to the case of nonnested linear mixed models exist, few attempts have been made to ground or extend these cases to statistical literature. Moreover, discussions of nonnested linear mixed models are usually restricted to nonnested mean models with very little attention paid to the case of nonnested random effects or covariance structures.

We propose two techniques that are incorporated into three methodological papers:

- a hypothesis testing approach

- an information criterion, applied under maximum likelihood estimation.

These two proposed approaches will add to the lack of available model selection techniques for nonnested linear mixed models and advance the discussion and development of more flexible model selection techniques.

CHAPTER 2: A HYPOTHESIS TEST TO SELECT BETWEEN LMMS WITH NONNESTED FIXED EFFECTS

2.1 Introduction

The construction of an adequate statistical model to characterize particular relationships or phenomena is an important step of scientific research. In studies using longitudinal data - particularly repeated measurements - the selection of a linear mixed model, where both a parsimonious mean model and an appropriate covariance structure must be specified, model selection is even more important. Typically, the first step of model selection in the linear mixed model among two candidate models involves assigning a common covariance structure that appropriately characterizes the correlation among repeated measurements to both models and then selecting among nested mean models using techniques such as likelihood ratio tests. When two models are *nested*, that means that one model can be derived by applying a linear restriction to the other model. Once a mean model is chosen, it is kept fixed and an appropriate covariance structure that represents inherent correlation of repeated measurements is then identified. The vast majority of scenarios involve comparing two candidate mean models that are nested, using techniques that are mostly extensions of those used to select among nested univariate linear regression models; these techniques continue to undergo investigations to assess their robustness.

Increasingly and under a wide variety of scenarios, researchers may be interested in comparing nonnested mean models. For instance, econometrics analysts often compare competing indices that measure the same phenomenon to determine which index more appropriately relates observed data to some outcome variable (Dimova et al., 2011; Monfardini, 2003); the indices, though not nested by definition, are often strongly positively correlated precluding their simultaneous inclusion in a single model. The common approach to selecting the more adequate model is that one fits two models, each including only one of the indices, then compares the properties of the models. In many cases, especially for the linear mixed model, the standard suggestion is to compare nonnested models using their information criteria (e.g., AIC, BIC), favoring the model with the smaller information criterion value. However, these criteria have not yet been adequately assessed for their ability to select among a set of nonnested models arising from longitudinal data with continuous outcomes. Furthermore, there does not exist a cadre of well-studied hypothesis tests or other criteria to evaluate nonnested models, so such comparisons have considerable drawbacks.

More relevant to linear mixed models, an important example of comparing nonnested models can be found in the study of obesity as a risk factor for diabetes, hypertension, and other detrimental cardiovascular and metabolic conditions. Researchers have devoted great interest to determining a measure of body fat that best captures the deposition of fat that leads to serious cardiovascular events among aging populations (particularly elderly populations and children). Most commonly used as a proxy for body fat is body mass index (or BMI), which is a function of an individual's weight (in kilograms) and height (in meters). The formula for BMI is given below:

$$BMI = \frac{Weight\,(kg)}{[Height\,(m)]^2}$$

This function produces a range of values with which individuals can be classified as underweight ($BMI < 18kg/m^2$), normal ($18 \leq BMI < 25$), overweight ($25 \leq BMI < 30$), or obese ($BMI \geq 30$). BMI has recently come under much scrutiny for its inability to accurately distinguish detrimental fatty body mass from lean mass, and for consistently misclassifying certain subpopulations as overweight or obese (Lewis et al., 2009; Hojgaard et al., 2008). Other related measures, such as waist circumference (distance around the abdomen, traditionally measured in inches), waist-to-hip ratio (Chan et al., 2003; Cole et al., 2005), and the more recently proposed Body Adiposity Index (BAI) (Bergman et al., 2011), have been purported to more accurately characterize body fat and its association with diabetic and cardiovascular risk.

Mathematically, to assess which (if any) of these alternative measures are better predictors than BMI of some continuous and normally distributed measure of cardiovascular risk among individuals over time, one could build separate models including only one measure at a time since including more than one of these correlated measures in a single model introduces collinearity. All of these measures have been shown to be strongly positively correlated with each other, yet their functional forms are such that separate models each including only one of these measures at a time can be considered nonnested in that one model cannot be obtained as a simple limit or linear restriction of the other. Again, there do not exist robust techniques to compare models of this type that are not nested, but there is great interest in establishing mathematical evidence to favor one measure above the others.

Levy and Hancock (2007) delineated four types of model relations for structural equation models via a framework that can be readily applied to linear mixed models: *completely overlapping models, hierarchically related models, partially overlapping models*, and *nonoverlapping models.* The first two relations represent nested models, in which one model is completely contained in the other. Existing model selection techniques are largely applied to these cases, where likelihood ratio tests can be used to distinguish models. In the discussion of nonnested models, the focus here is on what Levy and Hancock named *partially overlapping models*, which are derived from similar distributions and have some variables in common, but neither model is contained in the other without requiring constraints in both models. Nonoverlapping models are considered to be completely nonnested, where there is no possible set of constraints that could be applied to either model to create a nested structure between models; this case is not addressed here but is worthy of future exploration.

Seminal works by Sir David Cox (1961, 1962) highlighted this case of comparing partially overlapping models, or separate families of (nonnested) hypotheses. Important examples that expound upon Cox's work and help build the case for a need to extend this methodology to linear mixed models are found in, but not limited to, Pesaran (1974), Araujo et al. (2005), Monfardini (2003), and Dameus et al. (2002). To add to the scarce toolbox of available methods to compare and test linear mixed models with nonnested fixed effects, we investigate the development of a set of hypothesis tests based on the work of Cox (1961, 1962) and the extensions to nonnested univariate and multivariate regression models, introduced by Pesaran (1974) and Araujo et al. (2005), respectively. We derive bi-directional test statistics to compare two linear mixed models with nonnested fixed effects, particularly models that could be classified as partially overlapping. The corresponding limiting distributions of the test statistics are also derived, and we show that a closed form for the variance of either test statistic cannot be obtained without making conforming assumptions. In cases where simplifying assumptions made here are not appropriate, we demonstrate that the determinination of the distributions of test statistics can be computed using parametric bootstrapping. Below, we review Cox's methodology and build a foundation for deriving a test of separate hypotheses for linear mixed models with nonnested fixed effects. We use simulation studies to assess the viability of the derived test statistics to select among nonnested fixed effects models, paying particular attention to the issue of distinguishing body fat measures to assess longitudinal risk of cardiovascular events.

2.2 Review of Cox methodology

Recall from **Section 1.4.1** that we covered the background and theoretical overview of Cox's methodology and some notable extensions of his work. Now we introduce notation more relevant to linear models. Let $l_1(\hat{\boldsymbol{\theta}}_1)$ represent the maximized log-likelihood function of the model proposed under H_1 maximized by parameter value $\hat{\boldsymbol{\theta}}_1$ and $l_2(\hat{\boldsymbol{\theta}}_2)$ be the maximized log-likelihood function under H_2, where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are the $(k \times 1)$ maximum likelihood estimates of θ_1 and θ_2, respectively. Cox proposed using the following test statistic,

$$T_1 = l_1(\hat{\boldsymbol{\theta}}_1) - l_2(\hat{\boldsymbol{\theta}}_2) - E\left[l_1(\hat{\boldsymbol{\theta}}_1) - l_2(\hat{\boldsymbol{\theta}}_2)\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_1}, \tag{6}$$

which compares the observed difference of maximized log-likelihoods with an estimate of their expected difference under H_1. In the expected difference $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are replaced with their maximum likelihood estimates under H_1, $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_{21}$, respectively.

Alternatively, T_1 can be expressed as follows:

$$T_1 = \hat{l}_{12} - N\left[\text{plim}_{N\to\infty}\frac{\hat{l}_{12}}{N}\right]_{\boldsymbol{\theta}_1=\hat{\boldsymbol{\theta}}_1} \tag{7}$$

where $\hat{l}_{12} = l_1\left(\hat{\boldsymbol{\theta}}_1\right) - l_2\left(\hat{\boldsymbol{\theta}}_2\right)$; that is, $\hat{l}_{12}$ is the difference in log-likelihood functions or the likelihood ratio between candidate models, and plim refers to the probability limit of the given expression evaluated under H_1. This specification of T_1 differs from the first in its use of the probability limit in lieu of expectation. **Appendix A** delineates the similarities and differences between the probability limit and expectation. Our focus here relies on evaluating probability limits in lieu of taking expectations.

Cox demonstrated that, under the null hypothesis, T_1 is normally distributed with mean 0 and variance given by the equation below:

$$Var(T_1) = V_1\left[l_1(\boldsymbol{\theta}_1) - l_2(\boldsymbol{\theta}_{21})\right] - G_1' I_1^{-1} G_1, \tag{8}$$

where V_1 is the variance of the containing expression evaluated under H_1, $\boldsymbol{\theta}_{21}$ is the plim of $\hat{\boldsymbol{\theta}}_2$ under H_1, $G_1 \equiv N\frac{\partial}{\partial\boldsymbol{\theta}_1}\left[\text{plim}_{N\to\infty}\frac{\hat{l}_{12}}{N}\right]$, and I_1 is the information matrix of $\boldsymbol{\theta}_1$.

Recall that the roles of H_1 and H_2 can be interchanged, yielding corresponding test statistic T_2, where

$$T_2 = \hat{l}_{21} - N\left[\text{plim}_{N\to\infty}\frac{\hat{l}_{21}}{N}\right]_{\boldsymbol{\theta}_2=\hat{\boldsymbol{\theta}}_2} \tag{9}$$

In this case, the probability limit is taken under H_2 and $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_1$ are replaced by their maximum likelihood estimates under H_2, $\hat{\boldsymbol{\theta}}_2$ and $\hat{\boldsymbol{\theta}}_{12}$, respectively.

Asymptotically, $T_1^* = T_1 {Var_1}^{-\frac{1}{2}} T_1$ follows a standard normal distribution under the null hypothesis, H_1.

Decision-Making based on T_1 and T_2

In using two test statistics, it is possible to obtain rejections of both hypotheses (models), as well as non-rejections of both hypotheses. Asymptotically, T_1 has a negative expected value under H_2 and similarly, T_2 has a negative expected value under the alternative hypothesis. Thus, a large negative value of T_1 or T_2 leads to the rejection of the null hypothesis related to each test statistic. The case where both T_1 and T_2 are both significantly positive is inadmissible. From Sawyer (1983), the following table summarizes the decisions that can be made based on the values of T_1 and T_2.

Table 1: Decisions resulting from values of the Cox test statistics

	T_2		
T_1	Sig., $(-)$	Not Sig.	Sig., $(+)$
Sig., $(-)$	Reject H_1 and H_2	Do not reject H_2	Reject H_1 and H_2
Not Sig.	Do not reject H_1	Do not reject H_1 and H_2	Possibly Inadmissible
Sig., $(+)$	Reject H_1 and H_2	Possibly Inadmissible	Inadmissible

Thus, there are nine combinations that may result from computing both T_1 and T_2, resulting in four distinct decisions for a pair of separate hypotheses: rejecting both hypotheses, not rejecting both hypotheses, or rejecting one hypothesis while not rejecting the other. Following Cox's proposed hypothesis testing framework for nonnested models, Pesaran (1974) formulated a test to select among nonnested univariate linear regression models. He found that he could not derive an exact distribution of each test statistic, since each distribution depended on unknown parameters. To overcome this, he derived the variance of each test statistic according to Cox's formula, and in the end replaced unknown parameters with their maximum likelihood estimates, under the given null hypothesis. More than thirty years later, Araujo et al. (2005) carried this work further in developing a similar test for nonnested multivariate linear regression models. The extension to mutivariate models introduced further complexities in deriving the variance of the test statistic, regardless of the model considered as the null hypothesis. Building upon these two extensions, we will derive a hypothesis test for comparing linear mixed models with nonnested fixed effects. We attempt to obtain a closed form, or reasonable estimate, of the variance of each test statistic in order to derive an asymptotic distribution of the test statistics.

2.3 Formulating the Cox Test of Nonnested Hypotheses for LMMs with nonnested fixed effects

Considering the formulations of tests by Pesaran (1974) and Araujo et al. (2005), we propose a similar test of separate hypotheses applicable to the case of comparing two linear mixed models with nonnested fixed effects. One major distinction between the extension by Araujo et al. (2005) and our proposed methodology is that with the linear mixed effects model, we must consider the inherent correlation among the multivariate observations within subjects. From the Araujo et al. extension, it is evident that determining an expression for the variance of the test statistics for linear mixed models may require some conforming assumptions. Below, we present in detail the derivation of bidirectional test statistics T_1 and T_2 and their respective asymptotic variances, denoted by $Var(T_1)$ and $Var(T_2)$; these quantities are used to discuss determination of the asymptotic distribution of T_1 and T_2.

2.3.1 Deriving T_1, computing $Var(T_1)$, and deriving the distribution of T_1

Consider the following hypotheses:

$$H_1 : \boldsymbol{y} = \boldsymbol{X\beta}_1 + \boldsymbol{Z}\mathbf{b_1} + \mathbf{e_1} \tag{10}$$

$$H_2 : \boldsymbol{y} = \boldsymbol{W\beta}_2 + \boldsymbol{Z}\mathbf{b_2} + \mathbf{e_2} \tag{11}$$

alternatively expressed as

$$H_1 : \boldsymbol{y} \sim N\left(\boldsymbol{X\beta}_1, \boldsymbol{\Sigma}_1\right) \tag{12}$$

$$H_2 : \boldsymbol{y} \sim N\left(\boldsymbol{W\beta}_2, \boldsymbol{\Sigma}_2\right) \tag{13}$$

$$\tag{14}$$

The null hypothesis assumes a linear mixed effects model with $(Nn \times p)$ fixed effects design matrix $\boldsymbol{X}$, $(p \times 1)$ fixed effects parameter vector $\boldsymbol{\beta}_1$, $(Nn \times q)$ random effects matrix $\boldsymbol{Z}$ which is a subset of $\boldsymbol{X}$, $(q \times 1)$ random effects parameter vector $\mathbf{b_1} \sim \mathcal{N}_\mathbf{q}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b_1}}\right)$ independent of random errors distributed normally $\boldsymbol{e}_1 \sim \mathcal{N}_{Nn}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{e_1}}\right)$. In the null hypothesis, the variance of $\boldsymbol{y}$ is denoted by $(Nn \times Nn)$ matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{Z\Sigma}_{b_1}\boldsymbol{Z}' + \boldsymbol{\Sigma}_{e_1}$. The alternative hypothesis favors a linear mixed effects model with $(Nn \times p)$ fixed effects matrix $\boldsymbol{W}$, where neither $\boldsymbol{W}$ nor $\boldsymbol{X}$ is assumed to be a subset of the other matrix, although the two matrices may have some common variables and both are assumed to contain the same number of variables. In addition, the $(p \times 1)$ vector of fixed effects parameters in the null hypothesis, $\boldsymbol{\beta}_1$, is assumed to be different from the alternative hypothesis $(p \times 1)$ fixed effects parameter vector, $\boldsymbol{\beta}_2$. Vectors $\mathbf{b_2}$ and $\boldsymbol{e}_2$ are assumed to be independent, and the variance of $\boldsymbol{y}$ under the alternative hypothesis is given by $\boldsymbol{\Sigma}_2 = \boldsymbol{Z\Sigma}_{b_2}\boldsymbol{Z}' + \boldsymbol{\Sigma}_{e_2}$. Again, matrices $\boldsymbol{X}$ and $\boldsymbol{W}$ are not nested; that is, all columns of $\boldsymbol{X}$ cannot be obtained from those of $\boldsymbol{W}$ and vice-versa. However, $\boldsymbol{Z}$ is a subset of $\boldsymbol{W}$. For simplicity, in order to allow $\boldsymbol{Z}$ to be the same in both H_1 and H_2 so that the random effects covariance structures can be specified similarly, and thus nested, we assume that $\boldsymbol{X}$ and $\boldsymbol{W}$ are partially nonnested (Levy and Hancock, 2007).

Recall, the general formula for test statistic T_1 given by equation (6), as well as the alternative formula for T_1 given in equation (7).

In order to derive an expression for T_1 for the specific hypotheses, H_1 and H_2, listed previously, we first note that the log-likelihood functions of models H_1 and H_2 under maximum likelihood (ML) estimation are given according to the multivariate normal distribution by

$$\begin{aligned} l_1(\boldsymbol{\theta}_1) &= -\frac{Nn}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_1| - \frac{1}{2}tr\,(\boldsymbol{y}-\boldsymbol{X\beta}_1)'\,\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}_1) \\ l_2(\boldsymbol{\theta}_2) &= -\frac{Nn}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_2| - \frac{1}{2}tr\,(\boldsymbol{y}-\boldsymbol{W\beta}_2)'\,\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{y}-\boldsymbol{W\beta}_2) \end{aligned}$$

In the equations above, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are of dimension $(s \times 1)$,where $s = p+q+r$, vectors that collects unknown fixed and random effects parameters in the null and alternative hypotheses, respectively. The trace (tr) indicates the sum of the diagonal elements of the contained matrix.

If we assume that both fixed and random effects are unknown, then we have:

$$\begin{aligned} \hat{l}_{12} &= l_1\left(\hat{\boldsymbol{\theta}}_1\right) - l_2\left(\hat{\boldsymbol{\theta}}_2\right) \\ &= -\frac{1}{2}\log\left|\hat{\boldsymbol{\Sigma}}_1\right| + \frac{1}{2}\log\left|\hat{\boldsymbol{\Sigma}}_2\right| \\ &\quad -\frac{1}{2}tr\left[\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)'\hat{\boldsymbol{\Sigma}}_1^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)\right] + \frac{1}{2}tr\left[\left(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)'\hat{\boldsymbol{\Sigma}}_2^{-1}\left(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)\right], \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{y}$ under H_1, and under H_2,

$$\hat{\boldsymbol{\beta}}_2 = \left(\boldsymbol{W}'\hat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\hat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{y};$$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_1 &= \frac{1}{Nn}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)';\\ \hat{\boldsymbol{\Sigma}}_2 &= \frac{1}{Nn}\left(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)\left(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)'.\end{aligned}$$

Notably, $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ do not have closed form expressions (Pinheiro et al., 1994). Then, using properties of the trace of a matrix **(Appendix A)**, we have

$$\begin{aligned}\hat{l}_{12} &= \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|} - \frac{1}{2}tr\left[\hat{\boldsymbol{\Sigma}}_1^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)'\right]\\ &\quad + \frac{1}{2}tr\left[\hat{\boldsymbol{\Sigma}}_2^{-1}\left(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)\left(\boldsymbol{y}-\boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)'\right]\\ &= \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|} - \frac{Nn}{2}tr\left(\hat{\boldsymbol{\Sigma}}_1^{-1}\hat{\boldsymbol{\Sigma}}_1\right) + \frac{Nn}{2}tr\left(\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\Sigma}}_2\right)\\ &= \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|} - \frac{Nn}{2}tr\boldsymbol{I} + \frac{Nn}{2}tr\boldsymbol{I}\end{aligned}$$

Thus, $\hat{l}_{12}$ reduces to the following equation.

$$\hat{l}_{12} = \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|}.$$

Now, to compute the second term of T_1 according to equation 7, and applying properties of probability limits **(Appendix A)**, we have:

$$Nn\left(\text{plim}_{Nn\to\infty}\frac{\hat{l}_{12}}{Nn}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_1} = Nn\,\text{plim}_{Nn\to\infty}\left[\frac{1}{2Nn}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_1} = \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_{21}\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|},$$

where $\boldsymbol{\Sigma}_{21}$, the probability limit of $\boldsymbol{\Sigma}_2$ under H_1, is equivalent to $\frac{1}{Nn}\left[\boldsymbol{\Sigma}_1 + \boldsymbol{\beta}_1'\bar{\boldsymbol{\Sigma}}\boldsymbol{\beta}_1\right]$.

Similar to Araujo et al. (2005), we assume the following matrices exist, and that the first two are nonsingular and the third a non-zero matrix:

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{X}} &\equiv lim_{Nn\to\infty}\frac{1}{Nn}\boldsymbol{X}'\boldsymbol{X},\\ \boldsymbol{\Sigma}_{\boldsymbol{W}'\boldsymbol{W}} &\equiv lim_{Nn\to\infty}\frac{1}{Nn}\boldsymbol{W}'\boldsymbol{W},\\ \boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{W}} &\equiv lim_{Nn\to\infty}\frac{1}{Nn}\boldsymbol{X}'\boldsymbol{W},\end{aligned}$$

so that $\bar{\boldsymbol{\Sigma}}_1 = \boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{X}} - \boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{W}}\boldsymbol{\Sigma}_{\boldsymbol{W}'\boldsymbol{W}}\boldsymbol{\Sigma}_{\boldsymbol{W}'\boldsymbol{X}}$.

Combining the two terms of T_1, we have

$$T_1 = \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_{21}\right|}, \tag{15}$$

In order to use this proposed testing strategy, the distribution of the test statistic T_1 must be determined. Cox (1962) showed that asymptotically, under H_1, T_1 has an expected value of 0 and variance given by $Var\left(T_1\right) = V_1\left[l_{12}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{21}\right)\right] - \mathbf{G}_1{}'I_1{}^{-1}\mathbf{G_1}$, where $l_{12}\left(\cdot\right)$ is defined as before, and in the first term the variance is computed under the null hypothesis H_1. Also, $\boldsymbol{\theta}_{21}$ represents the probability limit ($plim$) of $\hat{\boldsymbol{\theta}}_2$ under H_1, $\mathbf{G}_1 = Nn\frac{\partial}{\partial\boldsymbol{\theta}_1}\left[plim_{Nn\to\infty}\frac{l_{12}\left(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right)}{Nn}\right]$, and I_1 is the information matrix of $\boldsymbol{\theta}_1$.

Beginning with the first term of $Var(T_1)$, we have

$$\begin{aligned}V_1\left[l_{12}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{21}\right)\right] = V_1\Bigg\{&\frac{1}{2}\log\frac{\left|\boldsymbol{\Sigma}_{21}\right|}{\left|\boldsymbol{\Sigma}_1\right|} + \frac{1}{2}tr\left[\left(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}_{21}\right)'\boldsymbol{\Sigma}_{21}^{-1}\left(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}_{21}\right)\right]\\ &-\frac{1}{2}tr\left[\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_1\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_1\right)\right]\Bigg\}.\end{aligned}$$

To evaluate the variance of the above expression, we note that the first term drops off since the expression is a constant that does not depend on a random quantity. So we have

$$\begin{aligned} V_1\left(l_{12}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{21}\right)\right) &= \frac{1}{4} V_1\left(tr\left[\left(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}_{21}\right)' \boldsymbol{\Sigma}_{21}^{-1}\left(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}_{21}\right)\right]\right) \\ &+ \frac{1}{4} V_1\left(tr\left[\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_1\right)' \boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_1\right)\right]\right) \\ &= \frac{1}{4}\left[2tr\left(\boldsymbol{\Sigma}_{21}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{21}^{-1}\boldsymbol{\Sigma}_1\right) + 4\left(\boldsymbol{X}-\boldsymbol{W}\boldsymbol{\beta}_{21}\right)' \boldsymbol{\Sigma}_{21}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{21}^{-1}\left(\boldsymbol{X}-\boldsymbol{W}\boldsymbol{\beta}_{21}\right)\right] \\ &+ \frac{Nn}{2} \end{aligned}$$

Applying lemmata and properties used in Araujo et al. (2005) and described in **Appendix A**, we arrive at an expression for the first term of $Var(T_1)$ as follows:

$$\begin{aligned} V_1\left[l_{12}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{21}\right)\right] &= tr\left[\left(\boldsymbol{X}\boldsymbol{\beta}_1-\boldsymbol{W}\boldsymbol{\beta}_{21}\right)' \boldsymbol{\Sigma}_{21}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{21}^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}_1-\boldsymbol{W}\boldsymbol{\beta}_{21}\right)\right] \\ &+ \frac{Nn}{2} tr\left[\left(\boldsymbol{\Sigma}_{21}^{-1}-\boldsymbol{\Sigma}_1^{-1}\right)\boldsymbol{\Sigma}_1\right]^2 \end{aligned}$$

Now, to compute the second term of $Var(T_1)$, we start with

$$\begin{aligned} \mathbf{G_1} &= Nn\frac{\partial}{\partial\theta_1}\left(\operatorname{plim}_{Nn\to\infty}\frac{\hat{l}_{12}}{Nn}\right) \\ &= \frac{Nn}{2}\frac{\partial}{\partial\theta_1}\left[\log\frac{|\boldsymbol{\Sigma}_{21}|}{|\boldsymbol{\Sigma}_1|}\right] \\ &= \frac{Nn}{2}\left[\frac{\partial}{\partial\theta_1}\log|\boldsymbol{\Sigma}_{21}| - \frac{\partial}{\partial\theta_1}\log|\boldsymbol{\Sigma}_1|\right]. \end{aligned}$$

Again, applying results from Araujo et al. (2005), we continue with

$$\begin{aligned} \mathbf{G_1} &= \begin{pmatrix} Nn\frac{\partial}{\partial\boldsymbol{\beta}_1}\operatorname{plim}_{Nn\to\infty}\frac{\hat{l}_{12}}{Nn} \\ Nn\frac{\partial}{\partial\boldsymbol{\Sigma}_1}\operatorname{plim}_{Nn\to\infty}\frac{\hat{l}_{12}}{Nn} \end{pmatrix} \\ &= \begin{pmatrix} Nn\, vec\left(\bar{\boldsymbol{\Sigma}}_1\boldsymbol{\beta}_1\boldsymbol{\beta}_{21}^{-1}\right) \\ \frac{Nn}{2}\left[2vec\left(\boldsymbol{\Sigma}_{21}^{-1}-\boldsymbol{\Sigma}_1^{-1}\right) - vech\, diag\left(\boldsymbol{\Sigma}_{21}^{-1}-\boldsymbol{\Sigma}_1^{-1}\right)\right] \end{pmatrix}, \end{aligned}$$

where $\bar{\boldsymbol{\Sigma}}_1$ is defined as before.

Next, we find an expression for the information matrix of $\boldsymbol{\theta}_1$ represented by

$$\begin{aligned} I_1^{-1} &= \begin{pmatrix} \frac{\partial^2}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1} & \frac{\partial^2}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\Sigma}_1} \\ \frac{\partial^2}{\partial \boldsymbol{\Sigma}_1 \partial \boldsymbol{\beta}_1} & \frac{\partial^2}{\partial \boldsymbol{\Sigma}_1 \partial \boldsymbol{\Sigma}_1} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} Nn\boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{X}} \otimes \boldsymbol{\Sigma}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{Nn}{2}\tilde{\boldsymbol{\Sigma}}_1 \end{pmatrix}, \end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}}_1$ is a product of matrices dependent on the columns and rows of matrices $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_1$.

Now, it's possible to put the terms of $Var(T_1)$ together, replacing (without loss of generality) all unknown quantities with their corresponding maximum likelihood estimates; this expression, however, is dependent upon matrices whose sizes are directly related to the sample size of observed data. This complicates the practical application of computing an asymptotic distribution for T_1. As an alternative to using the derived expression estimating the variance of the test statistic, here we consider the use of bootstrapping to estimate the distribution of T_1 (Godfrey, 2007).

Dameus et al. (2002) outline a strategy for estimating a p-value for the Cox test using parametric bootstrapping in the following steps: (1) estimate the two models under consideration using the actual observed data set; (2) compute the observed log-likelihood ratio of the two models; (3) estimate a distribution function for the original data under the null hypothesis, and, based on this estimate, generate a large number (B) of datasets of the same size; (4) re-estimate the two models for each of the generated datasets; (5) compute the simulated log-likelihood ratio for each simulated dataset compared to the observed data set; (6) compare the true and simulated log-likelihood ratios to compute the p-value according to the formula below:

$$p - value = \frac{\left[l_1 \left(\hat{\boldsymbol{\theta}}_{1b}, \boldsymbol{y}_b \right) - l_2 \left(\hat{\boldsymbol{\theta}}_{2b}, y_b \right) \leq l_{12} \quad \forall b = 1, \ldots, B \right] + 1}{B + 1}$$

Dameus et al. (2002) encountered difficulties generating simulated data with inherent correlations among observations. In our application, we will adapt this methodology for determining bootstrapped p-values for T_1 and T_2. The analogous derivation of T_2 and its asymptotic distribution for comparing linear mixed models with nonnested fixed effects can be found in the next section.

2.3.2 Deriving T_2, computing $Var(T_2)$, and deriving the distribution of T_2

Cox proposed a bi-directional test where the hypotheses may be interchanged such that the model from H_2 in **Section** 2.3.1 is now considered the null hypothesis, and the model from H_1 is the alternative hypothesis. If we consider the same models from (11), then we now consider the hypotheses as follows:

$$H_1 : \boldsymbol{y} \sim N \left(\boldsymbol{W\beta}_2, \boldsymbol{\Sigma}_2 \right)$$

$$H_2 : \boldsymbol{y} \sim N \left(\boldsymbol{X\beta}_1, \boldsymbol{\Sigma}_1 \right)$$

To derive the statistic for this test, T_2, we start with:

$$T_2 = \hat{l}_{21} - N \left(\text{plim}_{N \to \infty} \frac{\hat{l}_{21}}{N} \right)_{\boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2}$$

Computing the first term of T_2, we begin by defining the difference l_{21}.

$$\begin{aligned} l_{21}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) &= l_2\left(\hat{\boldsymbol{\theta}}_2\right) - l_1\left(\hat{\boldsymbol{\theta}}_1\right) \\ &= \frac{1}{2}\log\left|\boldsymbol{\Sigma}_1\right| - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_2\right| \\ &\quad + \frac{1}{2}tr\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right) - \frac{1}{2}tr\left(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)'\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right), \end{aligned}$$

where, under H_1, we have

$$\hat{\boldsymbol{\beta}}_2 = \left(\boldsymbol{W}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{y}, \hat{\boldsymbol{\Sigma}}_2 = \frac{1}{Nn}\left(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)\left(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_2\right)';$$

and, under H_2 we assume

$$\hat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{y}, \hat{\boldsymbol{\Sigma}}_1 = \frac{1}{Nn}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)'.$$

Continuing, we have $l_{21}\left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) = \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}$, noting that the trace terms drop off similarly as in the derivation of T_1.

The second term of T_2 is defined by

$$Nn\left(plim_{Nn\to\infty}\frac{\hat{l}_{21}}{Nn}\right)_{\boldsymbol{\theta}_2=\hat{\boldsymbol{\theta}}_2} = Nn\left[\frac{1}{2Nn}log\frac{\left|\hat{\boldsymbol{\Sigma}}_1\right|}{\left|\hat{\boldsymbol{\Sigma}}_2\right|}\right]_{\boldsymbol{\theta}_2=\hat{\boldsymbol{\theta}}_2} = \frac{1}{2}log\frac{\left|\hat{\boldsymbol{\Sigma}}_{12}\right|}{\left|\hat{\boldsymbol{\Sigma}}_2\right|},$$

where $\boldsymbol{\Sigma}_{12}$, the probability limit of $\boldsymbol{\Sigma}_1$ under the null hypothesis, is equivalent to

$$\frac{1}{Nn}\left(\boldsymbol{\Sigma}_2 + \boldsymbol{\beta}_2'\tilde{\boldsymbol{\Sigma}}\boldsymbol{\beta}_2\right);$$

and $\bar{\boldsymbol{\Sigma}}_2 = \boldsymbol{\Sigma}_{\boldsymbol{W}'\boldsymbol{W}} - \boldsymbol{\Sigma}_{\boldsymbol{W}'\boldsymbol{X}}\boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{X}}\boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{W}}$, where all terms are as defined before.

Putting both terms together, we find an expression for T_2 as follows

$$T_2 = \frac{1}{2} log \frac{\left|\hat{\mathbf{\Sigma}}_1\right|}{\left|\hat{\mathbf{\Sigma}}_{12}\right|} \tag{16}$$

As with T_1, we must determine the asymptotic distribution of T_2 in order to apply it. The variance of T_2, denoted by $Var\left(T_2\right)$, is given by

$$Var\left(T_2\right) = V_2\left[l_{21}\left(\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_2\right)\right] - \mathbf{G}_2{}' I_2{}^{-1} \mathbf{G_2}$$

where $l_{21}\left(\cdot\right)$ is defined as before, and in the first term the variance is computed under the null hypothesis (Model 2). Also, $\boldsymbol{\theta}_{12}$ represents the *plim* of $\hat{\boldsymbol{\theta}}_1$ under the null hypothesis. So the first term is given by

$$\begin{aligned} V_2\left[l_{21}\left(\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_2\right)\right] &= V_2\left[\frac{1}{2} tr\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{12}\right)' \mathbf{\Sigma}_{12}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{12}\right)\right] \\ &\quad - V_2\left[\frac{1}{2} tr\left(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}_2\right)' \mathbf{\Sigma}_2^{-1}\left(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}_2\right)\right] \\ &= tr\left[\left(\boldsymbol{W}\boldsymbol{\beta}_2 - \boldsymbol{X}\boldsymbol{\beta}_{12}\right)' \mathbf{\Sigma}_{12}^{-1} \mathbf{\Sigma}_2 \mathbf{\Sigma}_{12}^{-1}\left(\boldsymbol{W}\boldsymbol{\beta}_2 - \boldsymbol{X}\boldsymbol{\beta}_{12}\right)\right] \\ &\quad + \frac{Nn}{2} tr\left[\left(\mathbf{\Sigma}_{12}^{-1} - \mathbf{\Sigma}_2^{-1}\right)\mathbf{\Sigma}_2\right]^2 \end{aligned}$$

To compute the second term of $Var(T_2)$, note that

$$\mathbf{G}_2 = \left(N \frac{\partial}{\partial \boldsymbol{\theta}_2}\left[plim_{N\to\infty} \frac{l_{21}\left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right)}{N}\right]\right)$$

and I_2 is the information matrix of $\boldsymbol{\theta}_2$; all calculations are taken under the null hypothesis.

$$
\begin{aligned}
G_2 &= Nn\frac{\partial}{\partial\theta_2}\left(\text{plim}_{Nn\to\infty}\frac{\hat{l}_{21}}{Nn}\right)\\
&= \frac{Nn}{2}\frac{\partial}{\partial\theta_2}\left(log\frac{|\boldsymbol{\Sigma}_{12}|}{|\boldsymbol{\Sigma}_2|}\right)\\
&= \frac{Nn}{2}\left[\frac{\partial}{\partial\theta_2}log\,|\boldsymbol{\Sigma}_{12}| - \frac{\partial}{\partial\theta_2}log\,|\boldsymbol{\Sigma}_2|\right]\\
&= \begin{pmatrix} Nn\frac{\partial}{\partial\boldsymbol{\beta}_2}\,\text{plim}_{Nn\to\infty}\frac{\hat{l}_{21}}{Nn}\\ Nn\frac{\partial}{\partial\boldsymbol{\Sigma}_2}\,\text{plim}_{Nn\to\infty}\frac{\hat{l}_{21}}{Nn}\end{pmatrix}\\
&= \begin{pmatrix} Nn\,vec\left(\bar{\boldsymbol{\Sigma}}_2\boldsymbol{\beta}_2\boldsymbol{\beta}_{12}^{-1}\right)\\ \frac{Nn}{2}\left[2vec\left(\boldsymbol{\Sigma}_{12}^{-1}-\boldsymbol{\Sigma}_2^{-1}\right) - vech\,diag\left(\boldsymbol{\Sigma}_{12}^{-1}-\boldsymbol{\Sigma}_2^{-1}\right)\right]\end{pmatrix}
\end{aligned}
$$

An expression for the information matrix of $\boldsymbol{\theta}_2$ is represented by

$$
\begin{aligned}
I_2^{-1} &= \begin{pmatrix} \frac{\partial^2}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_2} & \frac{\partial^2}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\Sigma}_2}\\ \frac{\partial^2}{\partial\boldsymbol{\Sigma}_2\partial\boldsymbol{\beta}_2} & \frac{\partial^2}{\partial\boldsymbol{\Sigma}_2\partial\boldsymbol{\Sigma}_2}\end{pmatrix}^{-1}\\
&= \begin{pmatrix} Nn\boldsymbol{\Sigma}_{\boldsymbol{W}'\boldsymbol{W}}\otimes\boldsymbol{\Sigma}_2^{-1} & \mathbf{0}\\ \mathbf{0} & \frac{Nn}{2}\tilde{\boldsymbol{\Sigma}}_2\end{pmatrix},
\end{aligned}
$$

where, like $\tilde{\boldsymbol{\Sigma}}_2$ from the previous section, $\tilde{\boldsymbol{\Sigma}}_2$ is a product of matrices dependent on the columns and rows of matrices $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_2$. As with T_1 and $Var(T_1)$, while it is possible to obtain a closed form expression for $V_2\left(T_2\right)$ by putting together all the terms outlined above, we opt to apply bootstrapping techniques to determine the distribution of T_2 to avoid complex computations of large matrices.

2.4 Application to Data

An example was introduced that related to identifying a most appropriate anthropometric measure of body fat. Body mass index (BMI) has been the overwhelming standard for the past few decades amidst much criticism due to its inability to distinguish lean fat from detrimental fatty mass and for its gross misclassification of individuals into the overweight and obese categories (Bozeman et al., 2012; Pedersen et al. 2011; Kennedy et al., 2009; Freedman et al., 2009). Researchers have increasingly considered other measures such as waist circumference and waist-to-hip ratio, but have not had statistically rigorous techniques to formally assess the 'best' measure. Here we assess the Cox test's performance to ascertain which body fat measure - BMI or waist circumference - more accurately explains cardiometabolic risk among an aging elderly population.

2.4.1 Comparing nonnested fixed effects models in the NC EPESE data

The Established Populations for Epidemiologic Studies of the Elderly (hereafter, EPESE) was a National Institute on Aging study that took place from 1986 until 1998 (Blazer et al., 1991). Subjects were followed for up to four in person visits, each about four years apart, with some telephone interviews completed between visits. There were five study sites, including one in Central NC; for this investigation, we only use data from this site, which observed more than $4,000$ participants at baseline. Hereafter, we refer to this subset of data as *NC EPESE*. Data were publicly available and additional information was obtained with approval through the Inter-University Consortium for Political and Social Research (ICPSR). This resource contains demographic, clinical, physical functioning, and quality of life data for thousands of elderly subjects, allowing for rich assessments of the trajectory of health among this population. For the case of assessing the Cox tests of separate hypotheses for the linear mixed model, we rely on complete and balanced data through the third in person visit as attrition at the fourth wave of data collection and changes to variable coding do not allow for complete data across all four waves.

2.4.2 Specification of LMMs with nonnested fixed effects

Consider the following two models that result from the data set described above:

$$SBP = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 BMI + \boldsymbol{\alpha}_2 Sex + \boldsymbol{\alpha}_3 Time + a_0 + a_1 Time + e_1$$

$$SBP = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 WC + \boldsymbol{\beta}_2 Sex + \boldsymbol{\beta}_3 Time + b_0 + b_1 Time + e_2$$

In both models, the outcome variable is systolic blood pressure (SBP), a continuous measure used as a proxy for cardiometabolic risk. Typically, an SBP measurement above 120 mmHg is considered unhealthy and indicative of risk for hypertension among other cardiovascular risks. In the EPESE study, two consecutive sitting blood pressure measurements were taken, and the variable used here represents the arithmetic mean of each subject's systolic blood pressure measurements. The first model (Model I) includes baseline body mass index (BMI) as a covariate, while the second model (Model II) uses baseline waist circumference (WC). Since waist circumference measurements were not taken in the EPESE study, we simulated these values based on a piecewise linear function of subjects' baseline weight, height, and gender; thus, BMI and WC are positively correlated but neither can be obtained via a direct linear function of the other. Common to both models is the inclusion of sex (male or female) and time of in person observation as covariates (first through third waves of in person observation); thus, the design matrices for the two models (previously denoted as $\boldsymbol{X}$ and $\boldsymbol{W}$, respectively, have some columns in common and neither can be fully obtained as a subset of the other. Additionally, both models contain a random intercept and a random slope for time of in person visit $(1, 2, 3)$. As noted previously, we assume complete and balanced data; that is, each subject has data for the first three in person visits and there are no missing observations. In total, the models above have nonnested fixed effects since restrictions must be applied to both models in order to achieve a nested mean model structure. By comparing these models, we can assess whether BMI or WC is more predictive of longitudinal cardiometabolic risk.

2.4.3 Results

Descriptive information at baseline for the NC EPESE data is outlined in table 2. All NC EPESE participants in this subset of data were of age 65 years or older at the first wave, about two-thirds were male, and on average participants were overweight as categorized by their BMI. Baseline systolic blood pressure indicates that this subset of the population, on average, were hypertensive.

Table 2: Descriptive statistics at baseline for the NC EPESE data

Variable	Mean (SD)	(Min, Max)
SBP (mm Hg)	142.4(19.49)	$(72.5, 230.0)$
Age (years)	72.2(5.65)	$(65.0, 94.7)$
BMI (kg/m^2)	26.3(4.59)	$(13.6, 44.5)$
WC (inches)	34.2(4.84)	$(20.6, 51.3)$

Variable	N (%)	
Male Sex	610(33.4%)	

For each of the hypothesized models above, a linear mixed model with corresponding fixed and random effects was fit to the data. Each model assumes a compound symmetric covariance model; that is, it is assumed that correlations between subjects' blood pressure measurements vary according to a common pattern, and the correlation does not change with distance between observations. Table 3 lists fit statistics for each model, and table 4 displays fixed effects and covariance estimates for each model. The log-likelihood functions (transformed to $-2l$) for Models I and II are very close in value, and both the AIC and BIC favor Model I since their values are smaller than corresponding values for Model II. Additionally, assuming a significance level of 0.05, the fixed effects estimate for BMI is statistically significant in Model I, while the estimate for WC in Model 2 is not significant. These results suggest that, among this dataset, BMI is more explanatory of SBP than waist circumference.

Table 3: Fit statistics for Models 1 and 2

	$-2l$	AIC	BIC
Model I: BMI	48115.9	48121.9	48132.9
Model II: WC	48118.6	48124.6	48135.6

Table 4: Mixed model estimates, standard errors (SE) and p-values for Models 1 and 2; both models specify a random intercept and slope with compound symmetric covariance structure

					Cov. estimates	
Model	Fixed effect	Estimate	SE	p-value	Random effects	Error
I	Intercept	137.49	2.17	< 0.0001	$\hat{\sigma}^2_{b1} = 18.79$	$\hat{\sigma}^2_e = 257.9$
	BMI	0.16	0.08	0.038		
	Sex	-1.57	0.76	0.039		
	Time	0.75	0.28	0.007		
II	Intercept	138.51	2.71	< 0.0001	$\hat{\sigma}^2_b = 18.80$	$\hat{\sigma}^2_e = 258.01$
	WC	0.10	0.08	0.213		
	Sex	-2.10	0.82	0.011		
	Time	0.77	0.28	0.007		

To assess the Cox test of separate hypotheses, we computed values for test statistics T_1, which assumes that Model I is the null hypothesis and Model II is the alternative; and T_2, which assumes that Model II is the null hypothesis with Model I as the alternative hypothesis. From 1, we can use each statistic to determine which model we prefer for predicting cardiometabolic risk. In the previous section, we determined that we could bootstrap the distributions of T_1 and T_2, as use these estimates to complete the tests and make decisions about a preferred model. As in Monfardini (2003), we resampled $B = 100$ datasets from our original NC EPESE data, preserving the correlation among individual observations and maintaining complete and balanced data. For each resampled dataset, we computed values for T_1 and T_2; the variance of each test statistic was estimated by the sample variance of the test statistic values for each resampled dataset. Table 5 displays the results of our computation.

For the hypothesis test that uses T_1, we find that the observed value of the statistic is 0.975; assuming that T_1 asymptotically follows a standard normal distribution, then the

Table 5: Values of T_1, T_2, their bootstrap variance estimates, and corresponding bootstrap p-values

Model I (BMI)	Model II (WC)
$T_1 = 0.975$	$T_2 =$ -3.612
$\hat{Var}_{boot}(T_1) = 1.877$	$\hat{Var}_{boot}(T_2) = 3.322$
p-value$_{boot} = 0.2441$	p-value$_{boot} < 0.0001$

bootstrap p-value indicates that we should not reject Model I - the null hypothesis in this case. Similarly, the results for T_2 lead to a rejection of its null hypothesis, Model II, in favor of the alternative. From Table 1, we have the case where T_1 is not significant and T_2 is significant with a negative sign; so, we do not reject Model I. That is, these results support what we observe from using traditional model and variable selection techniques leading to the selection of the model containing BMI as its measure of body fat. Overall, values of T_1 and T_2 did not vary greatly across the datasets generated from parametric bootstrapping. Similarly, model I and II fit statistics and parameter estimates were also comparable across datasets.

2.5 Discussion and Conclusions

The case of comparing nonnested models has received minimal attention in statistical literature since the pioneering work of Sir David Cox on tests of separate families of hypotheses. Much of the subsequent work has focused on comparisons of models applied to problems in econometrics and quantitative psychology. Two important extensions of Cox's work created formulations of hypothesis tests to compare nonnested linear regression models and nonnested multivariate regression models, building the case for deriving test statistics for a new extension of this methodology to compare linear mixed models with nonnested fixed effects. Having to account for correlated observations within subjects, some conforming assumptions make determining a closed expression for the variance of test statistics a tedious task.

In this investigation, the proposed Cox test of separate hypotheses was applied to linear mixed models with nonnested fixed effects. Particularly, we sought to compare models with competing measures of body fat (BMI vs. waist circumference) using observed epidemiological data to determine which model including one of the body fat measures was most explanatory of systolic blood pressure, a proxy for cardiometabolic risk. Our investigation has demonstrated that the Cox test is viable for comparing linear mixed models with nonnested fixed effects. Since the application of Cox's work to the linear mixed model is new, there is not much understanding of the performance of the tests under unsuitable conditions. Our requirement of the existence of matrices $\boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{X}}$, $\boldsymbol{\Sigma}_{\boldsymbol{X}'\boldsymbol{W}}$, $\boldsymbol{\Sigma}_{\boldsymbol{W}'\boldsymbol{X}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{W}'\boldsymbol{W}}$ makes assumptions that may not always hold. More work is needed to understand the limitations of this approach.

Some limitations to this investigation include, but are not limited to: intensive computation required to perform the tests and lack of missing data to assess robustness of the methodology. Future investigations should consider other types of nonnested linear mixed models, including models with nonnested random effects, models with fixed and random effects that are nonnested, as well as models whose outcome (or dependent) variables represent similar phenomena but take on nonnested forms. Here, we have demonstrated that it is possible to construct and implement a test of separate hypothesis for linear mixed models with nonnested fixed effects.

CHAPTER 3: SELECTING BETWEEN LINEAR MIXED MODELS WITH NONNESTED RANDOM EFFECTS

3.1 Introduction

Analyses of repeated measures data assume that the collection of observations for an individual are correlated; the linear mixed model allows one to model this correlation structure to determine individual trajectories as well as population averages. With the advent of availability of patient data via electronic health records and other emerging technologies that open up individual data for analysis, many healthcare providers are interested in monitoring the health of individual patients as well as cohorts of patients over time. As changes in the health of individuals can be just as informative as tracking patterns among a population of patients over time, the linear mixed model facilitates separation of individual trajectories from population averages in the separate modeling of the mean model and covariance model. In this section, we draw attention to the structure of the covariance model of the random effects which capture unobserved heterogeneity among individuals, not captured in fixed effects.

Similarly for linear mixed models with nonnested fixed effects, literature describing approaches for comparing models with nonnested random effects is greatly limited. In many cases, the commonly proposed approach of comparing nonnested random effects models by using information criteria AIC and BIC seems to trivialize the importance of the covariance structure in the validity and interpretation of the linear mixed model. Several papers suggest that an appropriate covariance structure must be identified prior to selecting fixed effects (Keselman et al., 1998; Littell et al., 1996; Wolfinger, 1993), but do not offer alternative approaches to selecting among structures other than the previously highlighted information criteria. Econometrics and quantitative sociology literature examine the theoretical basis for comparing nonnested random effects (Mebane and Sekhon, 1996; Levy and Hancock, 2007; Timm et al., 2002), but the ties to statistical notation and application to public health research questions have not been made clear.

Of interest here, Morrell et al. (2009) explore how the structure of a model can impact the interpretation. Specifically, among a cohort of patients from the Baltimore Longitudinal Study of Aging (BLSA), the authors sought to determine if it was necessary to model patient age at baseline or whether to keep age as a dynamic variable. There are justifications for choosing either parameterization of age, depending on the interest in observing cross-sectional differences in a specificied factor by age or modeling longitduinal changes in a factor across subjects who entered the study at various ages. The cadre of models considered in the authors' investigation give rise to several examples of models with nonnested random effects, which we will apply to similar data on a longitudinal cohort of elderly subjects from the NC EPESE study (Blazer et al., 1991).

In **Chapter 2**, we established the viability of using methodology first proposed by Cox (1961; 1962) to formulate statistical tests of linear mixed models with nonnested fixed effects. Here, we wish to formulate a viable test for comparing models with nonnested random effects. Below, we describe some of the more common covariance structures applied to the linear mixed model to characterize random effects and review recommended techniques for selecting the appropriate structure for observed data. Next, we derive statistical tests of separate hypotheses to compare models with covariance structures that are nonnested, and apply the tests to models of real data. Finally, we will highlight the benefits and limitations of this approach.

Table 6: Common covariance structures applied to the linear mixed model, assuming k repeated measurements

Name	Abbrev.	# of Unknown Parameters
Variance Components	VC	1
Compound Symmetry	CS	2
Autoregressive, First Order	AR(1)	2
Toeplitz	TOEP	$k - 1$
Unstructured	UN	$\frac{k(k+1)}{2}$

3.1.1 Common covariance structures in the LMM

Table 6 describes some of the most common covariance structures used for the linear mixed model. Here, we assume that there are k repeated measurements for each subject. The table above is not an exhaustive list of covariance structures for linear mixed models; many other structures exist, can be modeled, or are under development (Kincaid, 2005; Fitzmaurice et al., 2004; McCullagh and Nelder, 1989; Laird and Ware, 1982). The MIXED procedure in SAS uses variance components (VC) as its default. The unstructured covariance is the least restrictive and can be considered to indirectly ”contain” all other possible covariance structures. However, it requires estimation of the largest number of unknown parameters of all structures. It can be useful when there are few repeated measurements among individual subjects, a large sample size relative to the number of repeated measurements, no apparent correlation structure that fits with other structures listed above, and when there is no interest in assuming a specific variance-covariance pattern among observed data. On the other extreme is the variance components structure structure, which doesn't assume any change in covariance across repeated measurements. The VC structure can be considered nested within all other possible covariance structures. The first-order auto-regressive structure is a special case of, and thus nested within, the Toeplitz structure; it is most useful when data are balanced and equally spaced. From table above, it can be noted that the variance components structure requires the estimation of only one unknown parameter. One drawback to using this structure is that it may be too simple to accurately describe patterns of within-subject correlation. Choosing a structure that is too simple results in increased Type I errors. For the other structures listed, the number of unknown parameters increases as the number of observed repeated measurements increases and as the structure of the matrix wanes. If choosing the unstructured model, one must sacrifice statistical power as this structure uses up resources to estimate a larger number of unknown parameters. For instance, if considering the unstructured covariance to model data with 7 repeated measurements, one would be required to estimate 28 unknown parameters.

Strategies for selecting covariance structures: Ignoring or misspecifying the variability within persons or sampling units within the linear mixed model has serious consequences, mostly introducing bias in estimated the mean model's fixed effects. Thus, the determination of an adequate representation of random effects in the linear mixed model is important; when selecting among covariance structures, one must take a strategic approach. While the literature offers some strategies for selecting the best covariance structure for a given set of observed data (Cheng et al., 2010), Kincaid (2005) outlines four basic approaches towards selecting a covariance structure for hierarchical data or data with repeated measurements of a continuous variable. These four approaches are listed below, as well as a fifth approach not considered by Kincaid and not often addressed in general, the use of statistical tests. One may opt to use a single approach, or several in combination, to determine the structure that is most representative of patterns of variation among observed data.

Parsimony: First, one may aim for parsimony, choosing the simplest possible structure. One benefit of this approach is that the covariance structure is easy to interpret and explain, and fewer unknown parameters must be specified. However, choosing a structure that is too simple for the data being considered may increase Type I error rates in selecting fixed effects. A structure that is highly complex and requires estimation of many unknown parameters can cloud interpretation of the covariance pattern and require inefficient use of computing resources.

Contextual Knowledge: In selecting a covariance structure, one may also draw upon contextual knowledge to determine which structure most adequately represents patterns of variation among observed data. This approach is particularly useful if the researcher has an in-depth understanding of the study design, general data structure, and/or the phenomenon being modeled. A potential drawback is that the most appropriate covariance structure may be difficult to model, or require the estimation of too many unknown parameters.

Graphical Tools: The ability to visualize patterns in variance and covariance within and between subjects offers great insight into the inherent correlation structure of repeated measures data. In particular, Kincaid (2005) discusses the semi-variogram and lag observation graphs as useful graphical tools for ascertaining correlation and potential outliers among subjects' measurements. These graphical tools are best used in tandem with another strategy to determine the most adequate covariance structure for repeated measures data.

Information Criteria: This approach is most commonly used, since most statistical software provides three default IC which are generally straightforward to interpret. There is also the caveat that IC are subjective in nature, and do not offer a formal comparison of structures (including a p-value) so differences in magnitude give no indication of statistical significance. Studies using IC to select covariance structures (Dimova et al., 2011; Timm et al, 2002) have led to inconclusive decisions as to which IC is best for a variety of scenarios.

Statistical Tests: As with fixed effects models, one may opt to use a statistical test to compare covariance structures. Most tests are based on Neyman-Pearson theory which assumes that one model (or structure) is nested within the other. Likelihood ratio tests are applied to test a model containing one or more random effects vs a model that excludes a subset of those effects (while not adding additional effects). By our previous definition of nested models, one model can be obtained as a simple limit of the other. For example, one may compare a mixed model with specified fixed effects design and a random subject-specific intercept to a model with the same set of fixed effects and a random intercept and random slope to allow completely subject-specific trajectories and not just different intercepts. These comparisons are made under restricted maximum likelihood estimation (REML), and no decisions are made concerning the choice of fixed effects since they are common in both models. For covariance structures, one must assess whether the two structures are nested. For some pairs of structures, it is simple to recognize that they are nested. Consider the Toeplitz ($TOEP$) structure, which is setup as a diagonal-constant matrix, meaning that variances are constant along the diagonal and along each descending diagonal; as a special case, the first-order autoregressive $AR(1)$ structure adds additional restrictions to the $TOEP$ by relating adjacent observations by a multiplicative factor.

3.1.2 Selecting among nonnested covariance structures

In the previous section, we discussed cases where covariance structures for two models are nested. For cases where covariance structures are not nested and traditional approaches for selecting a most appropriate covariance structure do not hold. This case where structures are nonnested remain elusive to many researchers, and the blanket recommendation of using information criteria is frequently offered as the primary choice for comparing models of this type. Below, we address nonnested structures and propose methodology to formulate a statistical test to be used for these cases. Additionally, one may encounter nonnested random effects as in Morrell et al. (2009). We define *nonnested* models as we did in previous chapters; random effects models are considered nonnested if one cannot obtain a candidate model by applying a linear restriction to the other candidate model. As an example, one might be interested in comparing a model with a random intercept for each subject and a random slope for visit (or repeated observation) with a model with only the random intercept; these models are clearly nested, and a likelihood ratio test comparing the two models, following mixture chi-square distribution could be applied. However, if one wanted to compare a random effects model with only the intercept to a model with a random slope, it would not be possible to use the likelihood ratio test to compare them. The literature does not adequately address cases of nonnested covariance structures; again, most investigations appeal to the use of information criteria for all comparisons of covariance structures to evade distinguishing between nested and nonnested structures.

3.2 Cox Test of Separate Hypotheses for LMMs with Nonnested Random Effects

We have previously covered the formulation of Cox's test of separate hypotheses **(Section 1.4.1)**. Here, we wish to derive similar test statistics for comparing linear mixed models with nonnested random effects according to this methodology.

Consider the following set of hypotheses:

$$\begin{aligned} H_1: &\quad \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}\mathbf{b_1} + \mathbf{e_1} \\ H_2: &\quad \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{W}\mathbf{b_2} + \mathbf{e_2} \end{aligned}$$

First, the notation for the models listed above is not specified as in Chapter 2, though similar. Both hypotheses assume a linear mixed model with fixed effects parameters captured by the $(p \times 1)$ vector $\boldsymbol{\beta}$, along with fixed effects design matrix $\boldsymbol{X}$ of dimension $(Nn \times p)$. Keeping the fixed effects common facilitates the selection of random effects. The null model assumes random effects captured by $(q \times 1)$ vector $\mathbf{b_1}$, while the alternative model assumes a different - nonnested - set of random effects captured by vector $\mathbf{b_2}$ of the same dimension. That is, $\boldsymbol{W}$ and $\boldsymbol{Z}$ are not subsets of each other, but both are subsets of $\boldsymbol{X}$. In this initial scenario, we assume that both models have the same covariance structure, though the estimates of unknown parameters associated with each structure will likely differ. Since our focus is on comparing random effects, we use restricted maximum likelihood (REML) estimation (Fitzmaurice et al., 2004). The REML log likelihood functions of each model are listed below:

$$l_{1,R} = -\frac{Nn-p}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_1| - \frac{1}{2}tr\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right) - \frac{1}{2}\log\left|\boldsymbol{X}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{X}\right|$$
$$l_{2,R} = -\frac{Nn-p}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_2| - \frac{1}{2}tr\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)'\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right) - \frac{1}{2}\log\left|\boldsymbol{X}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{X}\right|$$

Deriving R_1, $Var(R_1)$, and its distribution

Recall the statistic formulated previously, T_1, defined as:

$$T_1 = \hat{l}_{12} - Nn\left[\text{plim}_{Nn\to\infty}\frac{\hat{l}_{12}}{Nn}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_1}.$$

To distinguish this new formulation where we compare models with nonnested random effects, we refer to the corresponding statistic as R_1. So,

$$R_1 = \hat{l}_{12} - Nn\left[\text{plim}_{Nn\to\infty}\frac{\hat{l}_{12}}{Nn}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_1}.$$

Below, we will derive an expression for R_1 for the hypotheses specified above. We will apply the same properties as used in the previous section and described in **Appendix A** to arrive at simplified expressions for all quantities derived below. First, taking the difference of the specified log-likehood functions evaluated at their respective maximum likelihood estimates, $\hat{l}_{12}$, we have

$$\begin{aligned}\hat{l}_{12} &= \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|} + \frac{1}{2}tr\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\Sigma}}_2^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) - \frac{1}{2}tr\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\Sigma}}_1^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\\ &\quad + \frac{1}{2}\log\frac{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{X}\right|}{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}\right|}\\ &= \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|} + \frac{1}{2}\log\frac{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{X}\right|}{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}\right|}\\ &\quad + \frac{1}{2}tr\hat{\boldsymbol{\Sigma}}_2^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' - \frac{1}{2}tr\hat{\boldsymbol{\Sigma}}_1^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\\ &= \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|} + \frac{1}{2}\log\frac{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{X}\right|}{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}\right|}\end{aligned}$$

The equation reduces down since the terms including the trace cancel out. Now, to find the second term of R_1, we evaluate the probability limit of $\hat{l}_{12}$ under the null hypothesis, H_1, and replace unknown quantities with estimates under H_1. Since we are evaluating under REML functions, the previous ML estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ no longer hold.

$$Nn\left[\text{plim}_{Nn\to\infty}\frac{\hat{l}_{12}}{Nn}\right]_{\boldsymbol{\theta_1}=\hat{\boldsymbol{\theta}}_1} = \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_{21}\right|}{\left|\hat{\boldsymbol{\Sigma}}_1\right|} + \frac{1}{2}\log\frac{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{21}^{-1}\boldsymbol{X}\right|}{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}\right|},$$

where $\boldsymbol{\Sigma}_{21}$ is the probability limit of $\boldsymbol{\Sigma}_2$ under H_1 evaluated under $\hat{\boldsymbol{\theta}}_1$, and its expression follows similarly from the quantity $\boldsymbol{\Sigma}_{1|0}$ from Araujo et al. (2005). Specifically,

$$\hat{\boldsymbol{\Sigma}}_{21} = \hat{\boldsymbol{\Sigma}}_1 + \frac{1}{Nn}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_1' - \hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_1' - \hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_2' + \hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_2'\right)\boldsymbol{X}',$$

where $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ represent model-specific estimates of fixed effects. Since we specify the same fixed effects in each model, the expectation is that both quantities will not differ much.

In the analogous derivation of R_2, the quantity $\hat{\boldsymbol{\Sigma}}_{12}$ is estimated by the expression:

$$\hat{\boldsymbol{\Sigma}}_{12} = \hat{\boldsymbol{\Sigma}}_2 + \frac{1}{Nn}\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_2' - \hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_2' - \hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_1' + \hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_1'\right)\boldsymbol{X}',$$

where $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are defined as before.

A final expression for R_1 is found by putting its two terms together,

$$R_1 = \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_2\right|}{\left|\hat{\boldsymbol{\Sigma}}_{21}\right|} + \frac{1}{2}\log\frac{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{X}\right|}{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{21}^{-1}\boldsymbol{X}\right|}. \tag{17}$$

It is easy to see that, given the expression for $\hat{\boldsymbol{\Sigma}}_{21}$, the expected value of R_1 under the null hypothesis is 0.

Note that the formula for R_2, which tests the interchanged hypotheses with the second model specified previously as the null hypothesis and the first model as the alternative hypothesis, can be similarly derived as R_1 as follows.

$$R_2 = \frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Sigma}}_1\right|}{\left|\hat{\boldsymbol{\Sigma}}_{12}\right|} + \frac{1}{2}\log\frac{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}\right|}{\left|\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{12}^{-1}\boldsymbol{X}\right|}, \tag{18}$$

where $\boldsymbol{\Sigma}_{12}$ is the probability limit of $\boldsymbol{\Sigma}_1$ under the new null evaluated under $\hat{\boldsymbol{\theta}}_2$.

The variance of R_1 is given by the formula below.

$$Var\left(R_1\right) = V_1\left[l_{12}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{21}\right)\right] - \mathbf{G}_1{}' I_1{}^{-1}\mathbf{G_1},$$

where G_1 and I are defined as before in **Section 2.3.1**. Beginning with the first term, we want to evaluate the variance of $l_{12}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{21}\right)$ under H_1.

$$l_{12}\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{21}\right) = \frac{1}{2}\log\frac{\left|\boldsymbol{\Sigma}_{21}\right|}{\left|\boldsymbol{\Sigma}_1\right|} + \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) + \frac{1}{2}\log\frac{\left|\boldsymbol{X}'\boldsymbol{\Sigma}_{21}^{-1}\boldsymbol{X}\right|}{\left|\boldsymbol{X}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{X}\right|}$$

So, using the above quantity, we begin with the first term of $Var(R_1)$.

$$\begin{aligned} V_1 \left[l_{12} \left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{21} \right) \right] &= V_1 \left[\frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{X\beta} \right)' \left(\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1} \right) \left(\boldsymbol{y} - \boldsymbol{X\beta} \right) \right] \\ &= \frac{1}{4} V_1 \left[\frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{X\beta} \right)' \left(\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1} \right) \left(\boldsymbol{y} - \boldsymbol{X\beta} \right) \right] \\ &= \frac{1}{2} tr \left[\left(\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1} \right) \boldsymbol{\Sigma}_1 \left(\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1} \right) \boldsymbol{\Sigma}_1 \right] \\ &= \frac{Nn}{2} + \frac{1}{2} tr \left(\boldsymbol{\Sigma}_{21}^{-1} \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_{21}^{-1} \boldsymbol{\Sigma}_1 \right) - tr \left(\boldsymbol{\Sigma}_{21}^{-1} \boldsymbol{\Sigma}_1 \right) \end{aligned}$$

Continuing with the second term of $Var(R_1)$,

$$\begin{aligned} \mathbf{G_1} &= Nn \frac{\partial}{\partial \theta_1} \left(\text{plim}_{Nn \to \infty} \frac{\hat{l}_{12}}{Nn} \right) \\ &= \frac{Nn}{2} \frac{\partial}{\partial \theta_1} \left[\log \frac{|\boldsymbol{\Sigma}_{21}|}{|\boldsymbol{\Sigma}_1|} + \log \frac{\left| \boldsymbol{X}' \boldsymbol{\Sigma}_{21}^{-1} \boldsymbol{X} \right|}{\left| \boldsymbol{X}' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X} \right|} \right] \\ &= \frac{Nn}{2} \frac{\partial}{\partial \theta_1} \left[\log |\boldsymbol{\Sigma}_{21}| - \log |\boldsymbol{\Sigma}_1| + \log \left| \boldsymbol{X}' \boldsymbol{\Sigma}_{21}^{-1} \boldsymbol{X} \right| - \log \left| \boldsymbol{X}' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X} \right| \right] \\ &= \frac{Nn}{2} \left[2\boldsymbol{\Sigma}_{21}^{-1} - diag \left(\boldsymbol{\Sigma}_{21}^{-1} \right) \right] - \frac{Nn}{2} \left[2\boldsymbol{\Sigma}_1^{-1} - diag \left(\boldsymbol{\Sigma}_1^{-1} \right) \right] \\ &\quad + \frac{Nn}{2} \left[2\boldsymbol{X}' \boldsymbol{\Sigma}_{21}^{-1} \boldsymbol{X} - diag \left(\boldsymbol{X}' \boldsymbol{\Sigma}_{21}^{-1} \boldsymbol{X} \right)^{-1} \right] - \frac{Nn}{2} \left[2\boldsymbol{X}' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X} - diag \left(\boldsymbol{X}' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X} \right)^{-1} \right] \\ &= Nn \left(\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{X}' \left(\boldsymbol{\Sigma}_{21}^{-1} - \boldsymbol{\Sigma}_1^{-1} \right) \boldsymbol{X} + diag \left(\boldsymbol{X}' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X} \right)^{-1} - diag \left(\boldsymbol{X}' \boldsymbol{\Sigma}_{21}^{-1} \boldsymbol{X} \right)^{-1} \right) \end{aligned}$$

The expression for $\mathbf{G_1}$ is a constant since none of the containing terms depend directly on $\boldsymbol{\beta}$. Furthermore, to find I_1^{-1}, it suffices to obtain $\frac{\partial^2}{\partial \boldsymbol{\Sigma}_1 \partial \boldsymbol{\Sigma}_1}$.

$$\begin{aligned} I_1^{-1} &= \frac{\partial^2 l_1 \left(\theta_1 \right)}{\partial \boldsymbol{\Sigma}_1 \partial \boldsymbol{\Sigma}_1} \\ &= \frac{\partial}{\partial \boldsymbol{\Sigma}_1} \frac{1}{2} diag \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_1^{-1} - \frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{X\beta} \right) \left(\boldsymbol{y} - \boldsymbol{X\beta} \right)' + \frac{1}{2} diag \left(\boldsymbol{X}' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X} \right)^{-1} \\ &\quad + \left(\boldsymbol{X}' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X} \right)^{-1} \end{aligned}$$

From here, one can obtain the full expression for $Var(R_1)$ by combining the above terms;

the derivation of $Var(R_2)$ follows similarly. The computation of the resulting expression, however, is tedious, yielding complex expressions of large sized matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_{21}$. In lieu of computing the closed form expressions for $Var(R_1)$ and $Var(R_2)$, we explore the approach of bootstrapping the variances and determine the asymptotic distributions of R_1 and R_2 (Shao, 1988).

3.2.1 Application to Data

In this section, we demonstrate the viability of the tests of separate hypotheses for models with nonnested random effects proposed in the previous sections using practical data. We again employ a subset of the NC EPESE data as used in **Section 2.4.1**, which contains data on 1829 subjects, observed at three equally spaced time points, each about four years apart, with information collected on their systolic blood pressure (SBP), body mass index (BMI), gender, and Center for Epidemiologic Studies Depression ($CESD$) scale score.

Application to Data - NC EPESE

In SAS, we use PROC MIXED under the default estimation method, REML. For each model, the RANDOM statement is included to specify random intercepts and slopes. From **Chapter 2**, we determined that BMI was preferred over WC for predicting cardiometabolic risk. Here, we introduce another variable, $CESD$ score that measures clinical depression among the study population. A score higher than 16 is indicative of clinical depression. So, each candidate model contains fixed effects for BMI, gender, $CESD$ score, and time of observation. Both models contain a random intercept, and a random slope in either $CESD$ score (Model III) or time of observation (Model IV). The models are specified below:

$$SBP = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 BMI + \boldsymbol{\beta}_2 Sex + \boldsymbol{\beta}_3 CESD + \boldsymbol{\beta}_4 Time + b_0 + b_1 CESD + e_1$$

$$SBP = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 BMI + \boldsymbol{\beta}_2 Sex + \boldsymbol{\beta}_3 CESD + \boldsymbol{\beta}_4 Time + b_0 + b_1 Time + e_1$$

Table 7: Fit statistics from candidate models fit to observed data

	$-2l$	AIC	BIC
Model III: Random slope in $CESD$ score	48105.3	48113.3	48135.3
Model IV: Random slope in $Time$	48054.0	48062.0	48084.1

Generally, it is problematic to include two time-varying covariates in a linear mixed model; in this scenario, $CESD$ score is time varying, and thus can be considered as a complex function of time. However, unlike the time covariate, a subject's $CESD$ score can fluctuate up and down, making its interpretation difficult as a 'rate of change'. This example is still important for demonstrating two models with nonnested random effects, but we acknowledge the need for additional examples. Summary statistics from the generated data are found in Table 2.1. For these models, table 7 gives a summary of fit statistics computed under $REML$ estimation.

The log-likelihood functions for models III and IV are relatively close in value, with Model IV having a lower value among the two. Both the AIC and BIC both prefer the model with a random slope in time of observation (Model IV), since these values are smaller in Model IV than those for Model III. Table 8 displays fixed effects and covariance parameter estimates; recall that both models assumed an unstructured covariance pattern across repeated measurements, making no expectations about the level of correlation between a subject's observations across time.

We assumed the same set of fixed effects in both models; that is, design matrix $\boldsymbol{X}$ is common in both models. As expected, fixed effects parameter estimates are very similar in both models, and each covariate is statistically significant at the $\alpha = 0.05$ level.

To estimate the asymptotic distributions of R_1 and R_2, we generated 100 bootstrapped data sets from the original NC EPESE data set (Monfardini, 2003; Shao, 1998). For each bootstrap data set, the two candidate models were fit; information criteria, covariance model estimates, standard errors, and p-values were captured; and each test statistic (R_1 and R_2)

Table 8: Mixed model estimates, standard errors (SE) and p-values for Models III and IV; Model III specifies a random intercept and slope in $CESD$ with unstructured covariance; Model IV specifies a random intercept and slope in time of observation with unstructured covariance

					Cov. estimates	
Model	Fixed effect	Estimate	SE	p-value	Random effects	Error
III	Intercept	132.63	2.82	< 0.0001	$\hat{\sigma}^2_{b1} = 380.76$	$\hat{\sigma}^2_e = 251.49$
	BMI	0.17	0.08	0.048	$\hat{\sigma}^2_{b2} = -13.22$	
	CESD	0.25	0.09	0.006	$\hat{\sigma}^2_{b3} = 0.86$	
	Time	0.75	0.26	0.005		
IV	Intercept	132.32	2.75	< 0.0001	$\hat{\sigma}^2_{b1} = 187.93$	$\hat{\sigma}^2_e = 229.60$
	BMI	0.18	0.08	0.0322	$\hat{\sigma}^2_{b2} = -25.09$	
	CESD	0.25	0.09	0.004	$\hat{\sigma}^2_{b3} = 26.55$	
	Time	0.75	0.28	0.007		

was computed. Average run time for computing both test statistics for each data set was 11 minutes and 59 seconds, which is similar to the corresponding run times for T_1 and T_2 (test statistics for comparing models with nonnested fixed effects). Table 9 summarizes aspects of the bootstrapped data sets and resulting model statistics. Now, for the Cox test statistics, R_1 has a large positive value which could lead to a decision to reject the null hypothesis, the model with a random slope in time, in favor of the model including $CESD$ score as a random effect, given the values of the bootstrapped standard error and p-value of R_1. Correspondingly, R_2 takes on a relatively large negative value for our observed data and candidate models, but its bootstrapped standard error and p-value make it marginally not significant. Again, referring to Table 1, the results for R_1 and R_2 are classified as a scenario that is possibly inadmissible. A slightly smaller p-value for R_2 would reclassify our decision to reject both hypotheses. Overall, these results support decisions that would be made using the common information criteria AIC and BIC, favoring the model with a random slope in time.

Unlike the simulations for T_1 and T_2 in **Chapter 2**, it is not clear from this simulation whether the distributions of R_1 and R_2 are distributed normally. The expressions of $\hat{\mathbf{\Sigma}}_{21}$

Table 9: Values of R_1 and R_2, for $N = 100$ bootstrapped data sets

Metric	Observed Value	Mean (SE_{boot})	$p - value_{boot}$
R_1	11.67	10.29 (4.27)	< 0.0001
R_2	-6.457	-5.30 (4.22)	0.063

and $\hat{\boldsymbol{\Sigma}}_{12}$ include a term consisting of combinations of products of fixed effects parameter estimate vectors $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. Thus, assuming that the fixed effects estimates are similar between models, we expect this second term to cancel out, expecting the value of $\hat{\boldsymbol{\Sigma}}_{21}$ under the null hypothesis that specifies Model III to be $\boldsymbol{\Sigma}_1$. Similarly for

3.3 Discussion

The case of comparing nonnested models has received minimal attention in statistical literature, particularly a rigorous treatment of comparing nonnested random effects in the linear mixed model. Since the pioneering work of Cox on tests of separate families of hypotheses. Formulations of hypothesis to compare nonnested linear regression models and nonnested multivariate regression. This investigation helps to give a practical example of techniques available for rigorously approach the selection of random effects in the linear mixed model.

The choice of covariance structure to model random effects in the linear mixed model is very important. Moreover, failing to strategically address this aspect of the linear mixed model often leads to biased fixed effects estimates and faulty inference. We have demonstrated the viability of a test of separate hypotheses for linear mixed models with nonnested random effects. Particularly, cases where covariance structures are not nested between models ; other examples of models with nonnested random effects exist, with one particular example addressed in the next chapter (Morrell et al., 2009). Future work will use real data to assess the performance of the test statistics to select the correct covariance structure in situations of unbalanced, missing, and otherwise imperfect data.

CHAPTER 4: COMPARING THE COX TESTS WITH THE *EIC* FOR SELECTING NONNESTED LMMS

4.1 Introduction

Previous chapters have demonstrated the need for more rigorous approaches to comparing nonnested models, particularly nonnested linear mixed models which are often applied to longitudinal data. In statistical, econometrics, and other literature, the most commonly recommended approach for comparing nonnested models of various types (including, but not limited to linear mixed models) is to use information criteria to select the favored model. Only recently have researchers begun to critically examine the adequacy of these criteria for comparing complex models, and to broaden the scope of nonnested model selection to consider approaches that are less subjective and more statistically grounded. The investigations in **Chapters 2** and **3** demonstrated the plausibility of the Cox test of separate hypotheses for linear mixed models with nonnested fixed or random effects (but not the case where both sets of effects are nonnested). This investigation seeks to compare the performance of the Cox Test against information criteria - particularly, a variant of the Akaike Information Criterion (AIC), the Extended Information Criterion (EIC) - in selecting among nonnested linear mixed models. The vast majority of previous work on model selection techniques for linear mixed models has considered comparisons among information criteria, and most have not considered EIC (Wang and Schaalje, 2009; Dimova et al., 2011) nor have they often considered comparisons of nonnested models. Those that do include this criteria in their comparisons among nested models have shown it to outperform the AIC and several AIC variants among small samples (Dimova et al., 2011; Pan, 1999; Yafune et al., 2005). Li and Wong (2010) compared semiparametric models from longitudinal data using both likelihood ratio tests (to compare nested covariance structures only) and information criteria.

Here we wish to add to the scarcity of research comparing the performance of both statistical tests and information criteria to compare models arising from repeated measures and longitudinal data. In the following sections, we specify the formula for the EIC, provide examples of nonnested linear mixed models, and present results for applying both the EIC and bidirectional Cox tests to compare models. Application to observed data will show which (if any) of the two techniques performs above the other, as well as how they fare against other widely used model selection techniques.

4.2 Applying the *EIC* to nonnested models arising from repeated measures data

When fitting a linear mixed model using the *MIXED* procedure in SAS (version 9.3, Cary, NC), there exists the option to print model fit statistics such as -2 times the observed log likelihood (or $-2\,l$), the AIC, and the BIC. Naturally, it is of interest to persuade developers to include the EIC in this list of fit statistics. Here, we outline the steps to compute the EIC for a single linear mixed model in SAS. The computational ease and efficiency of this method is also evaluated.

Recall the formula for EIC:

$$EIC = -2\,l\,\left(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}\right) + 2\hat{C}^*$$

where $\hat{C}^*$ is given by

$$\hat{C}^* = \frac{1}{B}\sum_{b=1}^{B}\left[l\,\left(\boldsymbol{y}_b{}^* \mid \hat{\boldsymbol{\theta}}_b^*\right) - l\,\left(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_b^*\right)\right]$$

The steps required to compute the EIC for a linear mixed model follow directly from Yafune et al. (2005).

Step 1. From the observed data $\boldsymbol{y} = [\boldsymbol{y}_1{}^*, \ldots, \boldsymbol{y}_N{}^*]$, draw B independent random samples of size N. That is, sample with replacement among each subject's vector of repeated outcome measurements. In SAS, this is facilitated using the *SURVEYSELECT* procedure.

Step 2a. Fit the candidate mixed model to the original data using *PROC MIXED*. Using the *Output Delivery System*, or *ODS*, extract fixed and random effects parameter estimates, as well as the value of the log-likelihood function (under maximum likelihood or some other estimation method, such as REML as Yafune et al. assumed) and fit statistics from the *MIXED* output summary.

Step 2b. Complete step [2a.] for each of the B resampled datasets.

Step 3. Define the first term of the EIC formula by the originally observed data log likelihood, $l\left(\boldsymbol{y}\middle|\hat{\boldsymbol{\theta}}\right)$.

Step 4. Compute $\hat{C}^*$ by replacing log-likelihood values for each corresponding term. To determine the value of $l\left(\boldsymbol{y}\middle|\hat{\boldsymbol{\theta}}_b^*\right)$, that is, the likelihood of the originally observed data conditional upon bootstrap parameter estimates, use the Interactive Matrix Language (*IML*) to facilitate these computations.

Step 5. Repeat steps [1] through [4] for all candidate models.

Step 6. The model with the lowest EIC value should be selected as most suitable, according to the 'smaller is better' principle.

EIC computation in R follows similar steps. Data resampling is facilitated using the *sample* function. Before fitting a mixed model, one should first load the *nlme* package, and use the contained function *lmer* to fit the model. The *summary* function allows viewing of the results and extraction of relevant elements. Results from R computations nor the limitations of fitting linear mixed models in this software are not discussed here.

A motivating example arises from the recent analysis by Morrell et al. (2009), who used data from the Baltimore Longitudinal Study of Aging (BLSA), an observational study conducted by the National Institute on Aging (NIA) beginning in 1958 (and in 1978 for female participants). Participants were healthy volunteers at baseline and were observed approximately every two years and were followed for an average of about three repeated measurements, (up to six observations per subject). The age at baseline of BLSA participants varied greatly, making it important to discern cross-sectional vs. longitudinal trends in cholesterol among study participants.

One comparison of note explored by the authors is the following exploration of the association between (male only) participants' age and cholesterol levels:

$$y_i = \beta_0 + \beta_1 \left(FAge + Time\right)_i + b_{0i} + b_{1i} \left(FAge + Time\right) + \epsilon$$

$$y_i = \beta_0 + \beta_1 FAge_i + \beta_2 Time_i + b_{0i} + b_{1i} FAge_i + b_{2i} Time_i + \epsilon$$

where $FAge$ represents a participant's age (in years) at baseline and $Time$ is the time of observation. Both models allow for a random intercept (b_{0i}; that is, each subject's baseline cholesterol is allowed to vary around an average estimated by the fixed effects intercept, β_0. The first model contains random slopes in both baseline age and time, while the second model only contains a random slope in time. So, the first model essentially only considers a subject's age at time of observation, ignoring any differences in cholesterol attributable to the subject's age at baseline. That is, the model similarly considers a subject who was 65 years of age at the time they entered the study, and a 59 year-old subject who has been observed by the study for six years. The second model, however, separates subjects' baseline age from the time of observation, allowing one to better distinguish the subjects purported in the example above. This difference is evident in the mathematical parameterization of the models which specify different fixed effects, making the models nonnested as one model cannot be obtained as a simple limit of the other. While one could not use a likelihood ratio test to compare these models, the authors note that these models could be compared under maximum likelihood estimation using information criteria AIC and BIC. However, like others, they emphasize the uncertainty in using these measures to compare models of this type.

Using the SAS macro described in the previous section, we assess the performance of the *EIC* in selecting between a pair of nonnested linear mixed models, using the NC EPESE data (described in **Section 2.4.1**) which is similar to that from the BLSA.

4.3 EIC for Nonnested Models from Sections 2.4.2 and 3.2.1

We have covered the formulation of the Cox's methodology for comparing separate families of hypotheses (1961, 1962) in previous chapters. Below, we revisit the two scenarios covered in Sections 2.4.2 and 3.2.1, the cases of comparing models with nonnested fixed effects and models with nonnested random effects, respectively. A version of the macro to compute the *EIC* is found in Appendix 0.18. The results of these computations are found below.

Recall the models with textitnonnested fixed effects compared in **Section 2.4.2**:

$$SBP = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 BMI + \boldsymbol{\alpha}_2 Sex + \boldsymbol{\alpha}_3 Time + a_0 + a_1 Time + e_1$$

$$SBP = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 WC + \boldsymbol{\beta}_2 Sex + \boldsymbol{\beta}_3 Time + b_0 + b_1 Time + e_2$$

The above model specify repeated systolic blood pressure measurements as a function of participants' sex, time of observation, and a body fat measure (*BMI* in Model I, and *WC* in Model II). Both models specify a random intercept and a random slope in time.

For each model in each pair above, the *EIC* was computed

In **Section 3.2.1**, we compared the following models with *nonnested random effects*:

$$SBP = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 BMI + \boldsymbol{\beta}_2 Sex + \boldsymbol{\beta}_3 CESD + \boldsymbol{\beta}_4 Time + b_0 + b_1 CESD + e_1$$

$$SBP = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 BMI + \boldsymbol{\beta}_2 Sex + \boldsymbol{\beta}_3 CESD + \boldsymbol{\beta}_4 Time + b_0 + b_1 Time + e_1$$

Based on the results from **Section 2.4.2**, we preferred the use of *BMI* over *WC* for

Table 10: Applying the EIC to LMMs with Nonnested Fixed Effects for assessing the impact of body fat on SBP

	Model I: Fixed Effect for BMI	**Model II: Fixed Effect for WC**
	$-2l = 48115.9$	$-2l = 48118.6$
	$AIC = 48121.9$	$AIC = 48124.6$
	$BIC = 48132.9$	$BIC = 48135.6$
B	EIC	EIC
100	49742.54	49849.05
200	49725.57	49831.78
500	49703.33	49811.46
1000	49677.06	49775.12
	T_1 (p-value)	T_2 (p-value)
	0.975(0.2441)	$-3.612(< 0.0001)$

Table 11: Applying the EIC to LMMs with Nonnested Random Effects: comparing random slopes

	Model III: $CESD$ score	**Model IV: Time**
	$-2l = 48105.3$	$-2l = 48054.0$
	$AIC = 48113.3$	$AIC = 48062.0$
	$BIC = 48135.3$	$BIC = 48084.1$
B	EIC	EIC
100	48133.85	48027.33
200	48033.49	48027.93
500	48006.73	48001.07
1000	47983.14	47985.62
	R_1 (p-value)	R_2 (p-value)
	$11.67(< 0.0001)$	$-6.46(0.063)$

characterizing body fat among NC EPESE participants; so the models above contain comon fixed effects for BMI, sex, $CESD$ score, and time of observation. While each model specifies a random intercept, the first model contains a random slope in CESD score, while the second model has a random slope in time. For both pairs of models, we wish to determine if the EIC will support our choices of models made in previous sections.

In tables 10 and 11, it is clear that the outcome of the EIC simulations correspond as expected with the AIC and BIC as well as the results from applying the Cox tests of separate hypotheses. That is, for each pair of models the smaller EIC value corresponds to the model favored by the other information criteria, regardless of the number of bootstrap resamples used in computation. It is worth noting that, for both models, the absolute value of. Additionally, the scale of units of the EIC in each case correspond with the other criteria.

One benefit of using the EIC is that its computing time is much faster than that required for computing Cox test statistics. The bulk of computing time required for obtaining the EIC relates to the generation of bootstrap data resamples, while model fitting and computation of the EIC takes relatively less time. For $B = 100$, total computing time (data resampling, fitting two mixed models, and extracting important information for EIC computation) is about 4 minutes. The average runtime for the data resampling algorithm for $B = 1000$ is approximately 40 minutes. In all, this motivating example demonstrates that the EIC can be readily applied to comparing linear mixed models with nonnested fixed or random effects.

4.4 Discussion

The benefit of applying a viable statistical test to compare nonnested model clearly shows great promise for expanding available model selection techniques for linear mixed models and other models applied to longitudinal data. One major benefit of the EIC is that it can be applied under ML and non-ML estimation methods (Yafune et al., 2005). This is especially beneficial for the linear mixed model, where much debate has centered around the use of ML vs. REML for estimating fixed and random effects. Since the authors' paper, other investigations (Gurka, 2006; Dimova et al., 2011; Pan, 1999) have more rigorously assessed the performance of various information criteria under both estimation methods, though very few of these investigations have included the EIC in their analyses, nor have they considered the application of a statistical test. In this investigation, both the EIC and the Cox tests require bootstrap resampling of observed data, but the computation of EIC is much faster than producing Cox test statistics.

Other investigations that have compared the performance of a statistical test to the use of information criteria to select among covariance models have only applied the statistical test (usually, the likelihood ratio test) to nested models. Li and Wong (2010) assert than when results of the two procedures are inconclusive, one should rely on results from the information criteria. Additionally, many investigations of existing techniques for selecting among nested or nonnested models have involved small-sample data. In practice, longitudinal studies often have larger samples, as in BLSA and EPESE. This investigation explored how effective the Cox Test and EIC are at selecting among nonnested linear mixed models arising from data with larger sample sizes, and simply demonstrated that both approaches could be applied to comparing models with nonnested random effects. It remains to evaluate the EIC under more practical scenarios with imperfect conditions of observed data (unbalanced/unequally-spaced observations, missing data, etc.). As a start, we have demonstrated that the EIC is a viable model selection technique for nonnested linear mixed models, but offer no conclusive decision regarding the comparative performance in selecting among nonnested linear mixed models between the EIC and the Cox tests.

CHAPTER 5: CONCLUSION

- This work draws more attention to model selection as a critical step in the analysis of longitudinal data. Particularly for the case of selecting among nonnested linear mixed models, the two proposed techniques expand one's options for determining the most adequate model among several candidates that cannot be compared using traditional techniques built off of theory for nested models. We built the case for extending the seminal work of Sir David Cox to comparing linear mixed models with separate hypotheses related to their fixed effects or random effects. We also explored the case of using a variant of the AIC, the Extended Information Criterion (EIC), to compare nonnested linear mixed models.

An important complement to the Cox tests of separate hypotheses statistics proposed are the computational algorithms required to compute the statistics, their corresponding distributions and p-values. Aspects of the computation of both the Cox Tests and the EIC rely on parametric bootstrap resampling. In both cases, it is of interest to rigorously determine the optimal number of bootstrap resamples (B) required to produce consistent estimates. Future refinement of these computational techniques will test the proposed approaches in situations of incomplete, unbalanced, and otherwise imperfect repeated measures data, as well as assess their adequacy in application to additional practical examples.

Future directions of this work include exploring the derivation of test statistics for comparing nonnested generalized linear mixed models (GLMMs) as well as models arising from generalized estimating equations (GEE). These classes of models are often applied to data similar to repeated observations considered here. Particularly for GEE, which enjoys the advantage of overcoming misspecification of the correlation structure, it would be interesting to explore whether the approaches proposed here would enhance its utility. Another approach to consider is the case of comparing a linear mixed model to a nonlinear mixed effects model. These models are nonnested because their functional forms do not allow for one model to be derived as a simple limit of the other.

It is important to note that more than fifty years later, much of the literature around nonnested models stemming from Cox's seminal work lies in the field of econometrics, though its potential for practical application is much more vast. Recently, Cox (2013) revisited his seminal papers to discuss the varied applications of his original work. A symposium included comments from some researchers who have made the most notable extensions to the methodology. This work seeks to build a bridge from these important econometrics theoretical discoveries and applications to the statistical literature. Data observed or collected over time among individual subjects sets the stage for many interesting research questions in fields beyond econometrics, but differences in notation cloud these apparent ties to other disciplines. The practical application of the proposed methodology are applicable to myriad problems in public health and medical research. Important areas of application include research on aging populations, when there is a concern about how to parameterize age in models arising from repeated measures data; cardiovascular diseases, where the comparison of models including similar fat or adiposity measures is of interest; nutrition, in selecting the most appropriate set of dietary fat measures associated with various clinical outcomes; brain imaging, when analyzing longitudinal data on functional magnetic resonance imaging (fMRI) to assess changes in brain function over time; and in genetics, when comparing nonnested segregation analysis models. All of the above mentioned applications tie back to the study of progression of life and physiology over time, an issue that has become increasingly important in the United States and many other countries as life expectancy increases and demands for improved quality of health care for aging populations have skyrocketed.

An important goal of this research is to produce software that can be widely used by researchers who build linear mixed models with longitudinal data, providing them an expanded suite of techniques to compare linear mixed models that are nonnested. The proposed statistical tests and information criterion have been shown to be viable for use among linear mixed models; with some refinement, the resulting code could eventually become pre-programmed into popular software programs for advanced statistical modeling which are not currently well-equipped to address selection among nonnested linear mixed models. The development of well-tested and easy-to-use software will facilitate widespread use of the proposed methodology across a variety of research areas, and inform future areas of improvement related to theoretical methodology and computation.

APPENDIX A: IMPORTANT PROPERTIES OF RANDOM VARIABLES AND MATRICES

Derivations in **Chapters 2** and **3** rely on some important properties of random variables and matrices. Here, define some of these properties and demonstrate their application to equations found throughout this document.

Trace of a matrix

The trace of a square matrix, denoted by tr is defined as the sum of the elements along the diagonal (Hazewinkel, 2001). That is, if we suppose that square matrix M is of dimension $(m \times m)$, then

$$tr\,(M) = \sum_{i=1;j=1}^{m} M_{ij}$$

Properties of the trace

Suppose A and B are square matrices both having dimension $(k \times k)$, where B is invertible, and that x is a scalar. The following trace properties are true.

1. $tr\,(A) = tr\,(A')$

2. $tr\,(AB) = tr\,(BA)$

3. $tr\,(B^{-1}AB) = tr\,(A)$

4. $tr\,(A+B) = tr\,(A) + tr\,(B)$

5. $tr\,(xA) = x\,tr(A)$

6. $\log\,(|A|) = tr\,(\log A)$

Consider the maximum log-likelihood function of a linear mixed model:

$$l(\hat{\boldsymbol{\theta}}) = -\frac{Nn}{2}\log 2\pi - \frac{1}{2}\log\left|\hat{\boldsymbol{\Sigma}}\right| - \frac{1}{2}tr\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

In taking the trace of the third term, we can apply the properties listed above to rewrite the term as

$$tr\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = tr\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'$$

Now, by definition, $\hat{\boldsymbol{\Sigma}} = \frac{1}{Nn}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'$, so we simplify the expression as follows

$$\begin{aligned} tr\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) &= tr\hat{\boldsymbol{\Sigma}}^{-1}Nn\hat{\boldsymbol{\Sigma}} \\ &= Nntr\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}} \\ &= Nntr\mathbf{I}, \end{aligned}$$

where $\mathbf{I}$ is the $(Nn \times Nn)$ identity matrix whose trace is Nn. Thus, the trace term reduces to $(Nn)^2$.

Probability limits of random variables

An essential part of the formula for each Cox test statistic involves the use of *probability limits*. The term is sometimes used (erroneously) interchangeably with *expected value*. In his book, Dougherty (2011) defines and describes the probability limit and its relationship to expected value in great detail. Essentially, the existence of a probability limit is a necessary condition for defining a consistent estimator. It is widely accepted that, all else constant, use of the probability limit is preferred over expectation.

First, we define a probability limit, given by plim. Let Z_n represent a random variable, and α be the probability limit of Z_n. Then plim $Z_n = \alpha$ implies that:

$$\lim_{n \to \infty} P\left(|Z_n - \alpha| > \epsilon\right) \to 0.$$

Suppose that X, Y, and Z are random variables, b is a constant, and $f(\cdot)$ is a continuous function. The following properties of probability limits hold (Dougherty, 2011).

Properties of Probability Limits

1. $\text{plim } (X + Y + Z) = \text{plim } X + \text{plim } Y + \text{plim } Z$
2. $\text{plim } bX = b \text{ plim } X$
3. $\text{plim } b = b$
4. $\text{plim } XY = \text{plim } X \quad \text{plim } Y$
5. $\text{plim } \frac{X}{Y} = \frac{\text{plim } X}{\text{plim } Y}$, if $\text{plim } Y \neq 0$
6. $\text{plim } f(X) = f(\text{plim } X)$

The first three properties also hold when plim is replaced with expected value, $E(\cdot)$. Only when X and Y are independent of each other does the fourth property hold. Property [5.] is a tedious case for expectation. The final property does not always hold for expectation.

We apply these properties in deriving expressions for terms in T_1, T_2, R_1, R_2, as well as their respective variances.

Properties of quadratic forms of random variables

Working with log-likelihood equations of linear mixed models often requires manipulation of quadratic forms of random variables. Here, we focus on properties related to determining expected values and variances of these quantities (Mathai and Provost, 1992).

Let ϵ represent an $(k \times 1)$ vector, with $E[\epsilon] = \mu$ and $Var\,(\epsilon) = \Sigma$, where Σ is a $(k \times k)$ matrix. Note that one does not have to assume that ϵ follows a Normal distribution. Also, let A be a $(k \times k)$ symmetric matrix.

It can be shown that

$$E\left[\epsilon' A \epsilon\right] = tr[A\Sigma] + \mu' A \mu$$

$$Var\left(\epsilon' A \epsilon\right) = 2tr[A\Sigma A\Sigma] + 4\mu' A\Sigma A\mu.$$

More generally,

$$Cov\left(\epsilon' A \epsilon, \epsilon' B \epsilon\right) = 2tr[A\Sigma B\Sigma] + 4\mu' A\Sigma B\mu.$$

Say that $(k \times 1)$ vector $\gamma = X - \mu$, where $E[X] = \mu$, and let $Var\,(\gamma) = \Sigma$. Symmetric matrix A is as defined above. Then

$$E\left[\gamma' A \gamma\right] = tr[A\Sigma]$$

$$Var\left(\gamma' A \gamma\right) = 2tr[A\Sigma A\Sigma]$$

APPENDIX B: SAS CODE FOR COMPUTING THE *EIC*

```
%macro eic(B);

proc iml;

use ncepese;

read all var _num_ into data;

close ncepese;

/* Fit statistics from original data and bootstrapped data */

use fit;

read all var _num_ into fit;

close fit;

/* Fixed effects parameter estimates */

use est;

read all var _num_ into beta;

close est;


  /* Variance-covariance (random effects) parameter estimates */

use cov;

read all var _num_ into cov;

close cov;


  /* Extract relevant variables from observed data and construct
fixed/random effects design matrices*/

  y   = data[, 9];

  x   = J(nrow(y), 1, 1)||data[, 3]||data[, 4];

  z   = J(nrow(y), 1, 1)||data[, 4];
```

```
p   = nrow(beta)-1;


/* Set dummy matrix for l(y|b*) */
  databoot_ll = j(&B,nrow(y)/4,0);



  do i=1 to &B; /* For each dataset ... */
  b = beta[i+1,];


/* Use first row of ’cov’ to populate covariance matrix) */
d11 = cov[1,1];
d12 = cov[1,2];
d22 = cov[1,3];
sigsqd = cov[1,4];


D   = (d11 // d12) || (d12 d22);
sig = z*D*z‘ + sigsqd*I(nrow(y));

/* Iterate by subject vectors of dim (3 x 1) */
do j=1 to nrow(y) by 3;
m = (j+2)/3;  /* For every subject...*/

ysub =  y[j,]//y[j+1,]//y[j+2,];
xsub =  x[j,]//x[j+1,]//x[j+2,];
zsub =  z[j,]//z[j+1,]//z[j+2,];
   diag  = I(3);
r   = sigsqd*diag;
```

```
   sigma = zsub*D*t(zsub) + r;

dsig = det(sigma);

isig = inv(sigma);

/* Correct for 'close to zero' determinants*/

if dsig < 0.0001 then dsig=0.0001;

/* Compute l(y|b*) for each bootstrapped dataset */

   databootll[i,m]  = -0.5*log(2*3.14)-0.5*log(dsig)

-0.5*t(ysub-(xsub*t(b)))*isig*(ysub-(xsub*t(b)));

end;

  end;

lorig = fit[1,1];

avgbootll = (1/&B)*(fit[,+]-lorig);

term1 = (1/&B)*J(1,&B,1)*avgbootll[2:(&B+1),1];

dbootll = databootll[,+];

avgdatabootll = (1/&B)*J(1,&B,1)*dbootll;

Ck = term1 - avgdatabootll;

eic = lorig + 2*Ck;

print eic;

quit;

%mend;
```

BIBLIOGRAPHY

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory.* (pp. 267-281). Akademiai Kiado.

Araújo, M. I., Fernandes, M., and de B. Pereira, B. (2005) Alternative procedures to discriminate non-nested multivariate linear regression models. *Comm. Stat. Theory Meth.*, **34**: 2047-2062.

Bergman, R.N., Stefanovski, D., Buchanan, T.A., Sumner, A.E., Reynolds, J.C., Sebring, N.G., Xiang, A.H., and Watanabe, R.M. (2011). A better index of body adiposity. *Obesity* **19**, 1083-1089.

Blazer, D., Burchett, B., Service, C., and George, L. (1991). The association of age and depression among the elderly: an epidemiologic exploration. *J. Gerontol.: Med. Sci.* **46**, M210-M215.

Bollen, K. A. Harden, J. J., Ray, S., and Zavisca, J. (2014). Bayesian Information Criterion and Alternative Bayesian Information Criteria in the selection of structural equation models. *SEM: A Multidisc. J.* **21**(1): 1-19.

Bollen, K. A. and Stine, R. A. (1992). Bootstrapping Goodness-of-fit measures in structural equation models. *Soc. Meth. Res.* **21**(2): 205-229.

Bozeman, S. R., Hoaglin, D. C., Burton, T. M., Pashos, C. L., Ben-Joseph, R. H., Hollenbeak, C. S. (2012). Predicting waist-circumference from body mass index. *BMC Medical research methodology* **12**, 115-122.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information theoretic approach, 2nd Edition.* New York: Springer-Verlag.

Cavanaugh, J. E. (1999) A large-sample model selection criterion based on Kullback's symmetric divergence. *Stat. Prob. Lett.*, **42**: 333-334.

Chan, D.C., Watts, G.F., Barrett P.H., and Burke, V. (2003) Waist circumference, waist-to-hip ratio and body mass index as predictors of adipose tissue compartments in men. *QJM* **96**, 441-447.

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

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A., and Muller, K. E. (2010) Real longitudinal data analysis for real people: building a good enough mixed model. *Stat. Med.*, **29**(4): 504-520.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. of Math. Stat.* **25**, 573-578.

Clark, V. R., Greenberg, B., Harris, T. S., and Carson, B. I. (2012). Body mass index and waist circumference predictors of cardiovascular risk in African-Americans. *Ethnicity and Disease* **22**, 162-167.

Clarke, K. A. (2001) Testing nonnested models of international relations. *Am. J. Pol. Sci.* **45**, 724-744.

Clarke, K. A. (2007) A simple distribution-free test for nonnested hypotheses. *Pol. Anal.*, **15**(3): 347-363.

Cole, T. J, Faith, M. S., Pietrobelli, A. and Heo, M. (2005). What is the best measure of adiposity change in growing children: BMI, BMI %, BMI z-score or BMI-centile. *Euro. J. of Clin. Nutr.* **59**, 419-425.

Coulibaly, N., and Brorsen, W. (1999) A Monte Carlo sampling approach to testing separate families of hypotheses: Monte Carlo results. *Econometric Rev.* **18**: 195-209.

Cox, D. R. (1961). Tests of separate families of hypotheses. *Proc. of the 4th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 105-123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Royal Stat. Soc., Series B. (Methodological)* **24**, 406-424.

Cox, D. R. (2013). A return to an old paper: 'Tests of separate families of hypotheses'. *J. Royal Stat. Soc., Series B. (Methodological)* **75**, 207-215.

Dameus, A., Richter, G. C., Brorsen, B. W., and Sukhdial, K. P. (2002) AIDS versus the Rotterdam Demand System: A Cox test with parametric bootstrap. *J. Agricult. Resource Econ.*, **27**(2): 335-347.

Davidson, R., and MacKinnon, J.G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica* **49**, 781-793.

Davidson, R., and MacKinnon, J. G. (1993). *Estimation and inference in econometrics.* New York: Oxford University Press.

Davidson, R., and MacKinnon, J.G. (2007). Improving the reliability of bootstrap tests with the fast double bootstrap. *Computational Statistics and Data Analysis* **51**, 3259-3281.

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

Dimova, R. B., Markatou, M., and Talal, A. H. (2011) Information methods for model selection in linear mixed effects models with application to HCV data. *Comp. Stat. Data Anal.*, **55**(9): 2677-2697.

Dougherty, C. (2011). *Introduction to econometrics.* 4th Edition. Oxford University Press.

Dziak, J.J., and Li, R. (2007). An overview on variable selection for longitudinal data. In D. Hong (Ed.), *Quantitative medical data analysis* (pp. 5-24). Singapore: World Sciences Publishers.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., and Schabenberger, O. (2008) An R^2 statistic for fixed effects in the linear mixed model. *Stat. Med.*, **27**(29): 6137-6157.

Efron, B. (1983) Estimating the error rate of prediction rule: some improvements on cross-validation. *J. Am. Stat. Assoc.*, **78**: 316-331.

Ette, E. (1996) Comparing non-hierarchical models: application to non-linear mixed effects modeling. *Comp. Biol. Med.*, **26**(6): 505-512.

Farasat, S.M., Valdes, C., Shetty, V., Muller, D.C., Egan, J.M., Metter, E.J., Ferrucci, L., and Najjar, S.S. (2010). Is longitudinal pulse pressure a better predictor of 24-hour urinary albumin excretion than other indices of blood pressure? *Hypertension* **55**, 415-421.

Fitzmaurice, G., Laird, N., and Ware, J. (2004). *Applied Longitudinal Analysis.* Hoboken: Wiley-Interscience.

Freedman, D. S., Wang, J., Thornton, J. C., Mei, Z., Sopher, A. B., Pierson Jr., R. N., Dietz, W. H., and Horlick, M. (2009). Classification of body fatness by body mass index for age categories among children. *Archives of Pediatrics and Adolescent Medicine* **163**, 805-811.

Genius, M. and Strazzera, E. (2002) A note about model selection and tests for non-nested contingent valuation models. *Econ. Lett.*, **74**: 363-370.

Godfrey, L. G. (2007). On the asymptotic validity of a bootstrap method for testing nonnested hypotheses. *Econ. Lett.* **94**, 408-413.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional *AIC* in linear mixed models. *Biometrika* **97**, 773-789.

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

Hall, P., and Titterington, D.M. (1989) The effect of simulation order on level accuracy and power of Monte Carlo tests. *J Royal Stat Soc, Series B* **51**: 459-467.

Haskard, K. A., Rawlins, B. G., Lark, R. M. (2010) A linear mixed model, with non-stationary mean and covariance, for soil potassium based on gamma radiometry. *Biogeosci. Discuss.*, **7**: 1839-1862.

Hazewinkel, M., ed. (2001) Trace of a square matrix. *Encyclopedia of Mathematics*, Springer.

Hojgaard, B., Gyrd-Hansen, D., Olsen, K.R., Sogaard, J., Sorensen, T.I. (2008) Waist circumference and body mass index as predictors of health care costs. *PLoS One*, Jul 9; **3**(7): e2619. Epub 2008 Jul 9.

Hotelling, H. (1940) The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Ann. Math. Stat.*, **11**: 271-283.

Huber, P. (1967) The behaviour of maximum likelihood estimates under nonstandard conditions. *In Proc. 5th Berekeley Symp. Mathematical Statistics and Probability (eds L. M. LeCam and J. Neyman)*, pp. 221233. Berkeley: University of California Press.

Hurvich, C. M. and Tsai, C. L. (1989) Regression and time series model selection in small samples. *Biometrika* **78**, 499-509.

Ishiguro, M., Sakamoto, Y., and Kitagawa, G. (1997) Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Stat. Math.*, **49**: 411-434.

Kalilani, L., and Atashili, J. (2006) Measuring additive interaction using odds ratios. *Epidemiol. Persp. Innov.*, **3**(5).

Kennedy, A. P., Shea, J. L., and Sun, G. (2009). Comparison of the classification of obesity by BMI vs. dual-energy X-ray absorpiometry in the Newfoundland population. *Obesity* **17**, 2094-2099.

Keselman, H. J., Algina, J., Kowalchuk, R. K., and Wolfinger, R. D. (1998) A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Comm. Stat. Sim. Comp.*, **27**(3): 591-604.

Kincaid, C. (2005) Guidelines for Selecting the Covariance Structure in Mixed Model Analysis. *In proceedings of the Thirtieth Annual SAS Users Group International Conference, No. 198-30.*

Kullback, S., and Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**(1): 79-86.

Laird, N. M., and Ware, J. H. (1982) Random effects models for longitudinal data. *Biometrics*, **38**(4): 963-974.

Levy, R. and Hancock, G. R. (2007). A framework of statistical tests for comparing mean and covariance structure models. *Multivariate Behavioral Research* **42**, 33-66.

Lewis, C.E., McTigue, K.M., Burke, L.E., Poivier, P., Eckel, R.H., Howard, B.V., Allison, D.B., Kumanyika, S., Pi-Sunyer, F.X. (2009). Mortality, health outcomes, and body mass index in the overweight range: a science advisory from the American Heart Association. *Circulation* **119**, 3263-3271.

Li, J. and Wong, W. K. (2010). Selection of covariance patterns for longitudinal data in semi-parametric models. *Stat. Meth. Med. Res.* **19**, 183-196.

Liang, H., Wu, H., and Zou, G. (2008) A note on the conditional AIC for linear mixed-effects models. *Biometrika*, **95**(3): 773-778.

Liang, Q., Li, H., Mendes, P., Roethig, H., and Frost-Pineda, K. (2009) Using bootstrap method to evaluate the estimates of nicotine equivalents from linear mixed model and generalized estimating equation. *J. App. Stat.*, **36**(4): 453-463.

Lindstrom, M. J., and Bates, D. M. (1988) Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *J. Am. Stat. Assoc.*, **83**: 1014-1022.

Liquet, B., Sakarovitch, C., and Commenges, D. (2003). Bootstrap choice of estimators in parametric and semiparametric families: an extension of EIC. *Biometrics* **59**, 172-178.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996) *SAS System for Mixed Models.* Cary, NC: SAS Institute Inc.

Liu, S. and Yang, Y. (2012). Combining models in longitudinal data analysis. *Ann. Inst. Stat. Math.* **64**, 233-254.

Mathai, A.M. and Provost, S.B. (1992). *Quadratic forms in random variables*, CRC Press.

Mebane, W. and Sekhon, J. (1996). Bootstrap methods for non-nested hypothesis tests. *Paper presented at the 1996 Summer Methods Conference, University of Michigan, Ann Arbor.*

Mizon, G. and Richard, J. F. (1986). The encompassing principle and its application to non-nested hypotheses. *Econometrica* **54**, 657-678.

Monfardini, C. (2003) An illustration of Cox's non-nested testing procedure for logit and probit models. *Comp. Stat. Data Anal.*, **42**: 425-444.

Morrell, C. H., Brant, L. J., and Ferrucci, L. (2009). Model choice can obscure results in longitudinal studies. *J. Gerontol.* **64**, 215-222.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Edition.* London: Chapman and Hall.

Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc., Series A.* **231**, 289-337.

Ngo, L., and Brand, R. (2002). Model selection in linear mixed effects models using SAS Proc Mixed. *SAS Global Forum* **22**.

Pan, W. (1999). Bootstrapping likelihood for model selection with small samples. *J. Comp. Graph. Stat.*, **8**: 687-698.

Pedersen, S. D., Astrup, A. V., and Skovgaard, I. M. (2011). Reduction of misclassification rates of obesity by body mass index using dual-energy X-ray absorpiometry scans to improve subsequent prediction of per cent fat mass in a Caucasian population. *Clinical Obesity* **1**, 69-76.

Pesaran, M. H. (1974). On the general problem of model selection. *Rev. Econ. Stud.* **41**, 153-171.

Pesaran, M. H. and Deaton, A. S. (1978). Testing non-nested nonlinear regression models. *Econometrica* **46**, 677-694.

Pesaran, M. H. and Weeks, M. (2000). Non-nested hypothesis testing: an overview. *Companion to Theoretical Econometrics*, ed. B. H. Baltagi. Oxford: Basil Blackwell.

Pinheiro, J.C., Bates, D.M., and Lindstrom, M.J. (1994). Model building for nonlinear mixed effects models. *Proceedings of the Biopharmaceutical Section, 1994 Joint Statistical Meetings.*

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

Pu, P. and Niu, F. (2006). Selecting mixed-effects models based on generalized information criteria. *J. Multivar. Anal.* **97**, 733-758.

Sakamoto, W. (2011). Selecting variance structure in mixed effects models by information criteria based on Monte Carlo approximations. *Joint Meeting of the 2011 Taipei International Statistical Symposium and 7th conference of the Asian Regional Section of the IASC*, p. 160.

Sawyer, K. R. (1983). Testing separate families of hypotheses: an information criterion. *Journal of the Royal Statistical Society, Series B (Methodological)* **45**, 89-99.

Schork, N. and Schork, M. A. (1989) Testing separate families of segregation hypotheses: bootstrap methods. *Am. J. Hum. Genet.*, **45**: 803-813.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Ann. of Stat.* **6**, 461-464.

Shang, J., and Cavanaugh, J. E. (2008) Bootstrap variants of the Akaike Information Criterion for mixed model selection. *Comp. Stat. Data Anal.*, **52**: 2004-2021.

Shao, J. (1988) A note on bootstrap variance estimation. *Technical Report* $\#88-29$, Purdue University.

Szroeter, J. (2007). Testing non-nested econometric models. *The Current State of Economic Science* **1**, 223-253.

Timm, N. H., and Al-Subaihi, A. A. (2001) Testing model specification in seemingly unrelated regression models. *Comm. Stat. Theory Meth.*, **30**: 579-590.

Timm, N. H. (2002) Testing non-nested multivariate effect size models in meta-analysis. *J. Educ. Behav. Stat.*, **27**(4): 321-333.

Vaida, F., and Blanchard, S. (2005) Conditional Akaike information for mixed effects models. *Biometrika*, **92**: 351-370.

Vasquez, G., Duval, S., Jacobs, D.R., Silventoinen, K. (2007) Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. *Epidemiol. Rev.*, **29**(1): 115-128.

Vuong, Q. H. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**(2): 307-333.

Wang, J. and Schaalje, G. B. (2009) Model selection for linear mixed models using predictive criteria. *Comm. Stat. - Sim. and Comput.*, **38**(4): 788-801.

Watnik, M. R. and Johnson, W.(2002) The behaviour of linear model selection tests under globally non-nested hypotheses. *Sankhya, Series A*, **64**(1): 109-138.

West, B.T., Welch, K.B., Galecki, A.T. (2006) *Linear Mixed Models: A Practical Guide Using Statistical Software*. CRC Press.

White, H. (1982) Regularity conditions for Cox's test of non-nested hypotheses. *J. Economet.*, **19**: 301-318.

Williams, E. J. (1959). The comparison of regression variables. *J. Roy. Stat. Soc., Ser. B.*, **21**:396-399.

Wolfinger, R. D. (1993) Covariance structure selection in general mixed models. *Comm. Stat.: Sim. Comp.*, **22**: 1099-1106.

Yafune, A., Funatogawa, T., and Ishiguro, M. (2005) Extended information criteria (EIC) approach for linear mixed effects models under restricted maximum likelihood (REML) estimation. *Stat. Med.*, **24**: 3417-3429.

Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937-950.