PARAMETRIC BOOTSTRAP INTERVAL APPROACH TO INFERENCE FOR

FIXED EFFECTS IN THE MIXED LINEAR MODEL

BY

Vincent S. Staggs

Submitted to the graduate degree program in Psychology and the
Graduate Faculty at the University of Kansas in partial fulfillment
of the requirements for the degree of Doctor of Philosophy.

_____
Chairperson

_____

_____

_____

_____

Date defended: _____

The Dissertation Committee for Vincent S. Staggs certifies that this is the approved version of the following dissertation:

PARAMETRIC BOOTSTRAP INTERVAL APPROACH TO INFERENCE FOR FIXED EFFECTS IN THE MIXED LINEAR MODEL

_____
Chairperson

Date approved: _____

Abstract

In mixed models, empirical best linear unbiased estimators of fixed effects generally have mean square errors (MSEs) that cannot be written in closed form. Standard methods of inference depend upon approximation of the estimator MSE, as well as upon approximation of the test statistic distribution by some known distribution, and may not perform well under small samples. The parametric bootstrap interval is presented as an alternative to standard methods of inference. Several parametric bootstrap intervals (Efron percentile, bias-corrected [BC], Hall percentile, and bootstrap-*t*) were compared using simulated data, along with analytic intervals based on the naïve MSE approximation and the Kenward-Roger method. Among the bootstrap methods, the bootstrap-*t* seems especially promising.

Chapter 0. Table of Contents

Chapter 1. Introduction

1.1.Purpose

Standard methods of statistical inference generally involve a statistic, chosen as an estimator of the parameter or effect of interest, and an estimate (often referred to as a *standard error*) of the statistic's standard deviation (SD). The SD is the square root of the statistic's sampling variance—a measure of how dispersed one would expect the values taken by the statistic to be in repeated sampling. The statistic and its standard error are used to form a *test statistic* (such as a *t* or *F*), and statistical inference is carried out based on the known or approximate distribution of the test statistic and its observed value for the given sample. The accuracy of such methods of inference depends in part upon the accuracy with which the statistic's sampling variance can be estimated and, if the distribution of the test statistic is unknown, upon the adequacy of the chosen distributional approximation.

In most mixed models, a straightforward estimate of the variance of the fixed effect estimator is unavailable, and accurate approximation of this variance can be difficult, especially when sample size is small. Moreover, because of this problem, the distribution of the test statistic used for inference must be approximated by some known probability distribution.

These hurdles can be sidestepped entirely by constructing parametric bootstrap confidence limits for the effect of interest; under this approach, accurate inference can be conducted without approximating the estimator variance and without using a known distribution to approximate the distribution of the test statistic. Despite

these advantages and its relative ease of implementation, published applications of this method to the problem of fixed effect estimation are scarce. The primary purpose of this dissertation is to present the parametric bootstrap interval as a viable method of conducting inference for fixed effects in mixed models.

1.2. Overview

The problem of small sample inference in mixed modeling is treated in detail in Chapter 2. The general linear mixed model is presented in Section 2.1, and estimation of both fixed and random effects in the mixed model is discussed in Section 2.2. Analytic methods of MSE approximation are described in Section 2.3, analytic confidence interval construction is covered in Section 2.4, and results from studies of the performance of various analytic approximations are given in Section 2.5.

In Chapter 3, the proposed bootstrap-based solution to the problem of small sample inference is laid out. An introduction to bootstrapping is provided in Section 3.1, and applications of the bootstrap to mixed models are discussed in Section 3.2. An account of bootstrap approaches to the problem of MSE approximation is given in Section 3.3. A general treatment of bootstrap interval construction is presented in Section 3.4, followed by a discussion of parametric bootstrap intervals for mixed models in Section 3.5.

The methods of three simulation studies are detailed in Chapter 4. In the first (Section 4.1), data were generated from a one-way random effects ANOVA model for each of four (2 small sample sizes $\times$ 2 small ICC values) conditions, and five intervals

(two analytic, three parametric bootstrap) were compared based on observed rates of coverage. In the second study (Section 4.2), data were generated under a random coefficient model using estimates from a longitudinal study of personality. As in the first study, five intervals were constructed and compared based on rates of coverage. Data for two sample sizes were generated from a different random coefficient model in the third study (Section 4.3), and three analytic and four bootstrap intervals were compared.

The results of these studies are presented and discussed in Chapter 5. References are listed in Chapter 6, and Chapter 7 is a technical appendix for the reader interested in delving more deeply into the mathematics underlying certain concepts.

Chapter 2. The Problem: Inference under Small Samples

2.1. The Problem in Brief

In linear regression, ordinary least squares (OLS) estimators of the regression coefficients are unbiased—i.e., an OLS estimator takes the value of its target parameter on average in repeated sampling, neither systematically overestimating or underestimating the target. Among linear unbiased estimators, OLS estimators have minimum variance, meaning that no other linear unbiased estimator has a sampling distribution less dispersed about the target. (One might say that a minimum variance estimator "bounces around" its target less from sample to sample, tending more often to take values close to its target than estimators with greater variance.) An unbiased linear estimator with minimum variance is known as a *best linear unbiased estimator*, or *BLUE*.

OLS coefficient estimators have known sampling variance formulas, allowing for straightforward variance estimation and computation of the observed value of the test statistic. Moreover, the test statistic for any regression coefficient has a Student's *t* distribution with a known number of degrees of freedom.

In mixed models, the situation is more complex. Best linear unbiased estimators of fixed effects depend upon the variances of the random terms in the model. These variances are usually unknown and must be approximated. Fixed effect estimators constructed using variance estimates generally have sampling variance that cannot be written in closed form, and a test statistic constructed using a fixed effect estimator and an approximation of its sampling variance has an unknown distribution

for finite sample size. Thus, there are two potential sources of error in any inference based on this standard method.

This problem is laid out in more concrete terms in the following example. After this, the general linear mixed model is presented more formally, and the problem is revisited in theoretical detail.

2.2. An Example

Consider a longitudinal study in which a personality measure is administered to subjects at several time points over a number of years in order to estimate the rate at which personality changes in adulthood (e.g., see Terracciano, McCrae, Brant, & Costa, 2005). The researcher could ignore the clustering of scores within subjects and regress the personality trait score of interest on subject age in a simple linear regression. This would yield an unbiased OLS estimate of the rate of change (slope), but the within-subject errors would be correlated, violating the OLS model assumptions and making the SD estimate for the slope estimator, and thus any inference based upon the SD estimate, unreliable.

A mixed model is more appropriate. Consider the following growth curve model:

$$Y_{ij} = \beta_0 + AGE_{ij}\beta_1 + u_{0i} + AGE_{ij}u_{1i} + e_{ij},$$

where Y_{ij} is the personality trait score for the *i*th subject at the *j*th time point; β_0 is the fixed overall intercept; β_1 is the fixed slope (i.e., the average slope across subjects); AGE_{ij} is the age of the *i*th subject at the *j*th time point, centered about the grand mean; u_{0i} is the random deviation of the *i*th subject's intercept from the mean

intercept, β_0; u_{1i} is the random deviation of *i*th subject's slope from the mean slope, β_1; and e_{ij} is the random error term associated with the *j*th score for the *i*th subject. The e_{ij}s are assumed independent with distribution $N(0, \sigma_e^2)$, and u_{0i} and u_{1i} are assumed independent from the e_{ij}s with joint distribution N(0, G), where G is a 2×2 diagonal variance-covariance matrix.

In this mixed model, the best linear unbiased estimator of the fixed slope depends upon the unknown variances of the random terms in the model. When estimates of these variances are substituted for their targets in the formula for the BLUE, the sampling variance of the resulting estimator cannot be expressed in closed form and must be approximated. Confidence intervals for the slope, and the test of the null hypothesis $\beta_1 = 0$ (i.e., personality score does not change with age), depend upon the accuracy of the estimator variance approximation and upon adequately approximating the distribution of the appropriate test statistic. In the sections that follow, these concepts are developed in theoretical detail, beginning with a general treatment of the mixed linear model in the next section.

2.3. The General Linear Mixed Model

2.3.1. Fixed vs. Random Effects

In a simple linear regression model, say $Y = \beta_0 + \beta_1 x + e$, the intercept and regression coefficient are *fixed effects*—constant, unobservable, "population-averaged" (Demidenko, 2004) quantities. These parameters characterize the entire population; the intercept is the mean of Y at x = 0, and the slope is the mean change in Y per unit increase in x. Other examples of fixed effects are gender effects and

treatment effects (assuming the treatments are selected purposefully rather than randomly).

Unlike a fixed effect, a *random effect* does not characterize the population as a whole, but rather one or more units present in a random sample from the population. The effect is random because its realized value depends upon the particular units sampled from the population. For example, in a repeated measures study, the responses are likely to be determined in part by individual differences among the subjects in the study. This idiosyncratic subject effect can be modeled as a random effect, as the realized values it takes in a given study will depend upon the sample of subjects selected for the study. Similarly, if a sample of classrooms is chosen in a study of student achievement scores, the effect of belonging to a particular classroom can be modeled as a random effect. The error term in linear regression is a random effect.

2.3.2. Mixed Model Notation

Simply stated, a mixed linear model is a linear regression model having one or more random terms in addition to the error term. The general linear mixed model can be expressed as

$$Y = X\beta + Zu + e,$$

where Y is an $n \times 1$ vector of responses, β is a $p \times 1$ vector of unknown fixed effect parameters, u is an $r \times 1$ vector of random effects, and e is an $n \times 1$ vector of random errors. X is a known $n \times p$ design matrix for the fixed effects, comprising a column for each fixed term in the model. Similarly, Z is a known $n \times r$ design matrix for

random the effects vector u. The random terms, u and e, are assumed to be independent with u ~ N(0, G) and e ~ N(0, R), where G and R are variance-covariance matrices known up to a vector θ of unknown variance parameters. According to the model, Y (given X, β, G, and R) is normally distributed with mean Xβ and variance-covariance matrix V = ZGZ' + R.

A *hierarchical model* is a special case of the mixed model involving nested data. In a two-level hierarchical model, sampling units (the lowest level) are nested within clusters (the second level), and the vector Y is constructed by stacking responses by cluster. Let *i* index the *m* clusters ($i = 1, 2, \ldots, m$); let *j* index the observations within a given cluster ($j = 1, 2, \ldots, n_i$); and let Y_{ij} denote the *j*th observation in the *i*th cluster. Then the *ij*th component of Y is Y_{ij}, where *ij* = 11, 12, …, $1n_1$, 21, 22, …, $2n_2$, …, *m*1, *m*2, …, mn_m. The rows of design matrices X and Z are constructed in the same manner, the *ij*th row of each corresponding to the *ij*th response in vector Y. Accommodation of more than two levels of clustering is straightforward.

The two-level hierarchical model can also be written in terms of a representative cluster as

$$Y_i = X_i\beta + Z_i u_i + e_i,$$

where Y_i is the observation vector for the *i*th cluster, u_i is the vector of random effects for the *i*th cluster, and so forth. Let $G_i = \mathrm{Cov}(u_i)$ and $R_i = \mathrm{Cov}(e_i)$. In theory the values of the parameters in the G_is can differ across clusters, but it is generally assumed that they do not (Bryk & Raudenbush, 1992; Wolfinger, 1996). (For an example of a

study in which the use of two different G_i matrices might be appropriate, see Lee and Bryk, 1989).

The variance-covariance matrix of the error terms in the *i*th cluster, R_i, is generally assumed to be the same for clusters of the same size. An *homogeneous* or *homoscedastic* model (or covariance structure) is one in which the errors within each cluster are assumed to have equal variance. This corresponds to each entry on the diagonal of R_i having the same value.

The mixed model can accommodate repeated measures models and other designs in which there is spatial or temporal dependence among the errors within a given cluster. In terms of the model written as $Y_i = X_i\beta + Z_iu_i + e_i$, the off-diagonal entries of R_i (the variance-covariance matrix of e_i) are not assumed to be zero when intra-cluster errors may be correlated. See Wolfinger (1996) for a discussion of both homogeneous and heterogeneous covariance structures for repeated measures.

2.3.3. Multivariate Approach vs. Mixed Model Approach

Muller, Edwards, Simpson, and Taylor (2007) noted that in studies in which the data are balanced (i.e., cluster sizes are equal), none of the data is missing or mistimed, and there is no need to model a particular variance-covariance structure, a multivariate approach to statistical tests is preferable to the univariate mixed model approach. Multivariate tests successfully control error rates, even for small samples; and power methods for the multivariate tests are well established and more convenient. However, the conditions under which the multivariate model is applicable

are seldom encountered in social science research, and the flexibility of the univariate mixed model makes it the obvious choice for many researchers.

2.4. Common Mixed Models

2.4.1. One-way Random Effects ANOVA Model

The one-way random effects ANOVA (RANOVA) model (also known as the *one-way random classification model*) does not technically satisfy the definition of a mixed model given above, as it has no fixed predictors, but it is an ideal model with which to begin a discussion of mixed models. According to Scheffé (1956), this model appeared as early as 1861 in the work of an astronomer, Airy, who used it to model repeated telescopic observations of a phenomenon over several nights.

Let Y_{ij} represent the *j*th observation in the *i*th cluster, u_i represent the *i*th (random) cluster effect (or *cluster intercept*), and e_{ij} represent the random error associated with Y_{ij}. The model is given by

$$Y_{ij} = \beta_0 + u_i + e_{ij},$$

where the u_is are *iid* (independent and identically distributed) $N(0, \sigma_u^2)$, the e_{ij}s are *iid* $N(0, \sigma_e^2)$, and the u_is and e_{ij}s are independent. The fixed effect parameter, β_0, is an overall mean. By assumption, $\sigma_e^2 > 0$ and $\sigma_u^2 \geq 0$. Note that $\text{Var}(Y_{ij}) = \sigma_u^2 + \sigma_e^2$.

Intra-class correlation coefficient. The intra-class correlation coefficient (ICC), ρ, is defined for the one-way RANOVA model as $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$. The ICC is the proportion of total variance attributable to the clustering of the observations, or, equivalently, the correlation between any pair of observations from the same cluster (Fisher, 1925). The ICC can be expressed in terms of the *variance ratio*, $\gamma = \sigma_u^2/\sigma_e^2$,

as $\rho = \gamma/(\gamma + 1)$. Note that the ICC and variance ratio are both equal to zero when the between-cluster variance is zero.

2.4.2. Fay-Herriot Model

The Fay-Herriot model (Fay & Herriot, 1979) is used in survey sampling for *small area estimation*—predicting and drawing inferences about the mean of some variable for geographical areas or subpopulations that are *small* in the sense that they contain few (or no) sampling units (Hulting & Harville, 1991). The model can be expressed as

$$Y_i = X_i\beta + u_i + e_i,$$

where the u_is and e_is are mutually independent random variables with $u_i \sim N(0, \sigma_u^2)$ and $e_i \sim N(0, \sigma_i^2)$. The *i*th small area mean to be predicted is $X_i\beta + u_i$, Y_i is the observed survey estimator of this mean (i.e., the sample mean for the *i*th small area), u_i is the random effect associated with the *i*th area, and e_i is the error in sampling the *i*th area (Datta, Rao, & Smith, 2005). The regression predictors in X_i provide area-level (i.e., cluster-level) information, and responses within each small area are used only in computing the small area means, so a two-level model is unnecessary (Prasad & Rao, 1990).

Following Fay and Herriot (1979), survey researchers commonly assume the σ_i^2s to be known for this model (Ghosh & Rao, 1994), as estimates are usually available from a survey organization (Pfefferman & Glickman, 2004). Methods of small area estimation are discussed in Ghosh and Rao (1994) and Rao (2003, 2005).

2.4.3. Nested Error Regression Model

The Fay-Herriot model and one-way RANOVA model are special cases of the nested error regression model, given by

$$Y_{ij} = X_{ij}\beta + u_i + e_{ij},$$

where the u_js and e_{ij}s are mutually independent random variables with $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. This model can be used to incorporate unit-level (i.e., lower-level) predictors in small area models (in contrast to the Fay-Herriot model, which allows for only area-level predictors). An *unconditional ICC* can be estimated for this model by fitting it with all of the fixed effects other than the intercept omitted and substituting the variance component estimates into the ICC formula. If the model is fit with all of the fixed effects included, the resulting ICC estimate is a *conditional* or *residual ICC*.

2.4.4. Random Coefficient Model

The random coefficient model is common in education research (e.g., Goldstein, 1995) and longitudinal studies (e.g., Laird & Ware, 1982). It allows not only for random cluster intercepts, but also for slopes (corresponding to known regression predictors) that vary randomly among clusters. The simplest such model is given by

$$Y_{ij} = \beta_0 + x_{ij}\beta_1 + u_{0i} + x_{ij}u_{1i} + e_{ij},$$

where x_{ij} is the observed regression predictor value for the *j*th unit in the *i*th cluster, u_{0i} is the *i*th (random) cluster intercept, u_{1i} is the (random) regression slope for the *i*th cluster, and β_0 and β_1 are the traditional, fixed regression intercept and slope,

respectively. The growth curve model described in Section 2.2 is an example of this model. Let u_i be the row vector comprising u_{0i} and u_{1i}, and define (as above) $G_i = \mathrm{Cov}(u_i)$. If G_i is unstructured, it includes three parameters: The intercept variance and slope variance, both on the diagonal; and the covariance between the intercept and slope off the diagonal.

The random coefficient model can include multiple predictors, each with a fixed and/or random regression coefficient, and it need not include an intercept. For example, Dempster, Rubin, and Tsutakawa (1981) used a random coefficient model with no intercept and two centered predictors, undergraduate GPA and LSAT score, each with random slope coefficient, to predict the first-year GPA of law school students.

After specifying an appropriate model for a data set, the variance parameters, fixed effects, and random effects can be estimated. Methods of estimation are described in the next section, beginning with estimation of variance parameters.

2.5. Estimation and Prediction under the Mixed Model

2.5.1. Variance Parameter Estimation

The variance parameters in θ are usually unknown in practice and must be estimated in order to carry out inferences regarding the fixed or random effects. There are several methods of variance parameter estimation, including ANOVA methods, maximum likelihood (ML), and restricted (or residual) maximum likelihood (REML). Searle, Casella, and McCullouch (1992) provided a detailed exposition in their book, much of which is devoted to the topic.

ANOVA estimation. Searle et al. (1992) credit the idea of ANOVA estimation of variance parameters to Fisher (1925), who had introduced the terms *variance* and *analysis of variance* in 1918 (Scheffé, 1956). For balanced data, one can compute ANOVA estimates by setting observed values of ANOVA sums of squares equal to their expected value (mean) and solving the resulting equations for the variance parameters. For example, in the one-way classification model with m clusters and n observations per cluster, the mean of the within-cluster ANOVA sum of squares is $E(SSW) = m(n - 1)\sigma_e^2$, and the mean of the between-cluster sum of squares is $E(SSB) = (m - 1)\sigma_e^2 + n(m - 1)\sigma_u^2$. The notation $E(\cdot)$ indicates the expected value, or mean, of the quantity within the parentheses. The ANOVA estimates are the solutions to the following set of equations:

$$SSW = m(n - 1)\sigma_e^2$$

$$SSB = (m - 1)\sigma_e^2 + n(m - 1)\sigma_u^2.$$

Note that the estimate for σ_u^2 may be negative. In the case of unbalanced data, ANOVA estimation is less straightforward. Henderson (1953) proposed three methods that bear his name; see Searle et al. for details.

Likelihood-based methods of estimation. Likelihood-based methods are a popular alternative to ANOVA estimation. Under the distributional assumptions made about the random terms in the mixed model, the distribution of Y, conditional on β and θ, is known. The *likelihood function* is obtained by expressing this distribution as a function of β and θ, given Y. The *maximum likelihood estimates* of β and θ are those values that maximize the likelihood function (or, equivalently, its natural logarithm,

known as the *log-likelihood*) over all possible values of β and θ. In non-technical terms, ML estimates are those parameter values that make the observed value of Y most likely to have been observed in sampling.

In REML estimation (Patterson & Thompson, 1971), the fixed effects are removed from the likelihood function before ML variance parameter estimation is carried out. One can obtain REML estimates by computing the residuals from OLS regression of Y on X and then fitting the mixed model with the OLS residual vector in place of Y (i.e., treating the OLS residuals as the dependent observations). The resulting ML variance parameter estimates are the REML estimates for the original model.

Both ML and REML estimators are *consistent*, or *asymptotically unbiased*, meaning that the expected value of the estimator, say $\hat{\theta}$, converges to the target parameter value as the total sample size goes to infinity. REML estimators have the additional advantage of being essentially corrected for the degrees of freedom lost in estimating fixed effect parameters, whereas ML estimators are not, and for balanced data, REML estimators are unbiased regardless of sample size (Searle et al., 1992). Lacking this correction, ML estimators tend to be biased downward. Moreover, Datta and Lahiri (2000) showed that bias in REML estimators converges to zero more quickly than the bias of ML estimators (see Section 7.1 in the Technical Appendix for discussion of rate of estimator convergence).

As noted, ANOVA methods may yield a negative solution for a variance parameter. Similarly, under ML or REML estimation, it is possible for the likelihood

function to achieve its maximum at a negative value for some variance parameter. In practice, would-be negative variance estimates are set to some non-negative value, and this truncation introduces bias into any estimator that would otherwise be unbiased.

2.5.2. Point Prediction for Mixed Effects

Given variance parameter estimates, one can proceed with estimation of the fixed and/or random effects in the mixed model. A *mixed effect* is an estimable linear combination of fixed and random effects, say $T = k'\beta + m'u$ for some constant vectors k and m. For example, the *i*th small area mean in the Fay-Herriot model, $T = X_i\beta + u_i$, is a mixed effect, and the goal in small area studies is to predict the realized value of T for each small area using the observed data.

Searle et al. (1992) pointed out that prediction of a mixed effect, say $T = k'\beta + m'u$, involves both *prediction* (of the realized value of the random component, m'u) and *estimation* (of the fixed component, k'β), and opted for the term *prediction* in dealing with a mixed effect. Following Searle et al., *estimation* is reserved in this dissertation for fixed effects, and *prediction* is used for mixed effects.

Best linear unbiased prediction. Goldberger (1962) and Henderson (1963) showed that the best linear unbiased predictor (BLUP) of the mixed effect T is $t(\theta) = k'\beta_{GLS} + m'GZ'V^{-1}(Y - X\beta_{GLS})$, where β_{GLS} is a solution for β in the generalized least squares (GLS) equations (Aitken, 1935),

$$X'V^{-1}Y = X'V^{-1}X\beta,$$

say $\beta_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}Y$, with generalized inverses taken as needed. In addition to being the best linear unbiased estimator of β, β_{GLS} is the maximum likelihood estimator of β for any given V (Birkes & Wulff, 2003). Note that $t(\theta)$ depends on the vector of variance parameters, θ, through G and V.

The BLUP is unbiased in the sense that its expected value (mean) is equal to the expected value of its target. The BLUP is *best* in the sense that its mean squared error (MSE), defined as the mean squared difference between the BLUP and its mean, $MSE[t(\theta)] = E[t(\theta) - T]^2$, is minimum among all other linear unbiased predictors (Searle et al., 1992). A predictor's MSE can be expressed as the sum of its variance and the square of its bias; thus, for an unbiased predictor like the BLUP, MSE equals variance.

Empirical best linear unbiased prediction. Because the variance parameters in θ are usually unknown in practice, estimation of β and prediction of T involve a two-step approach. First, variance parameter estimates, say $\hat{\theta}$, are computed and used to form estimates of the variance-covariance matrices G and V, say $\hat{G}$ and $\hat{V}$. Second, the variance parameter estimates are substituted for their respective targets in the formulas for β_{GLS} and t.

Using $\hat{V}$ in place of V in the formula for β_{GLS} yields the *estimated generalized least squares* (EGLS) estimator of β, $\beta_{EGLS} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y$. When $\hat{G}$, $\hat{V}$, and β_{EGLS} are substituted for their targets in the formula for $t(\theta)$, the result is the *empirical best linear unbiased predictor* (EBLUP) of T,

$$t(\hat{\theta}) = k'\beta_{EGLS} + m'\hat{G}Z'\hat{V}^{-1}(Y - X\beta_{EGLS}).$$

Fortunately, substituting an ANOVA, ML, or REML estimate for its target value in the formula for $t(\theta)$ does not change the expected value (Das, Jiang, & Rao, 2004; Kackar & Harville, 1981). Thus, the EBLUP, like the BLUP, is unbiased.

2.2.3. Point Estimation for Fixed Effects

In many applications of the mixed model, the researcher is primarily interested in inference concerning the fixed effects. Estimation of a linear combination of fixed effects is a special case of the mixed effect prediction problem described above; setting m equal to the zero vector in the mixed effect $T = k'\beta + m'u$ yields the fixed effect $T = k'\beta$. Thus, the theoretical results presented for mixed effect case can simply be restated in adjusted form for the fixed effect case. Note that in the absence of random effects, the problem becomes one of estimation, not prediction.

The *best linear unbiased estimator* (BLUE) of the fixed effect $T = k'\beta$ is $t(\theta) = k'\beta_{GLS}$, where as above, β_{GLS} is the generalized least squares estimator of β. As a special case of the BLUP, the BLUE is also unbiased and has minimum variance among all other linear unbiased estimators.

As with the BLUP, β_{GLS} cannot be computed when the variance parameters are unknown, and $\hat{V}$ is used in place of V in the formula for β_{GLS} to obtain the estimated generalized least squares estimator of β, β_{EGLS}. Substituting β_{EGLS} for β_{GLS} in the formula for $t(\theta)$ yields the *empirical best linear unbiased estimator* (EBLUE) of T, $t(\hat{\theta}) = k'\beta_{EGLS}$.

The problem with the EBLUP (and EBLUE) is that for most mixed models, its MSE cannot be written in closed form (Hulting & Harville, 1991). Approximations of this MSE are described in the next section.

2.6. Analytic MSE Approximations

2.6.1. Naïve MSE Approximation

Kackar and Harville (1984) showed that for the most common variance parameter estimators (including ML and REML estimators), the variance of the EBLUP, MSE[$t(\hat{\theta})$], can be expressed as

$$MSE[t(\hat{\theta})] = MSE[t(\theta)] + E[t(\hat{\theta}) - t(\theta)]^2.$$

The term on the far right is the MSE of $t(\hat{\theta})$ as an estimator of $t(\theta)$. For known variance parameters, MSE[$t(\theta)$], the variance of the BLUP, is known, and a common naïve approximation of MSE[$t(\hat{\theta})$] is obtained simply by substituting $\hat{V}$ for V in the formula for MSE[$t(\theta)$]. Let MSE[$t(\theta)|\hat{\theta}$] denote this approximation.

The naïve approximation involves two potential sources of error. First, $\hat{V}$ is used in place of V to estimate MSE[$t(\theta)$]. Second, $\hat{V}$ is used in place of V to obtain $t(\hat{\theta})$, resulting in a discrepancy between $t(\hat{\theta})$ and $t(\theta)$, and the mean square of this discrepancy, $E[t(\hat{\theta}) - t(\theta)]^2$, is ignored completely in the approximation. In other words, the variance in the EBLUP is inflated by sampling variability in the variance parameter estimates used to compute it, and the naïve approximation fails to account for this "variance propagation," as Littell (2002, p. 486) called it.

Because ML and REML methods yield consistent estimators, under these methods $\hat{V}$ tends to improve as an estimator of V as sample size increases, and the expected error of the naïve approximation tends to zero (Littell, 2002). However, for a small or moderate number of clusters, MSE[t($\hat{\theta}$)] can be seriously underestimated, so it is not surprising that for small samples, hypothesis tests based on the naïve approximation can yield inflated rates of Type I error (Catellier & Muller, 2000; Schaalje, McBride, & Fellingham, 2002; Kenward & Roger, 1997). It is also noteworthy that the error in the naïve approximation goes to zero only as the number of *clusters* goes to infinity, not as the number of observations per cluster goes to infinity for a fixed number of clusters (Demidenko, 2004).

2.6.2. Analytic Alternatives to the Naïve Approximation.

Aware of the deficiency of the naïve approximation, a number of researchers (Das, Jiang, & Rao, 2004; Datta et al., 2005; Datta & Lahiri, 2000; Fuller & Harter, 1987; Harville & Jeske, 1992; Kackar & Harville, 1984; Lahiri & Rao, 1995; Prasad & Rao, 1990; Wang & Fuller, 2003) have studied improved analytic approximations of the MSE of the EBLUP based on Taylor series expansions for various mixed models and for various methods of variance parameter estimation. See Das et al. for a review.

In their seminal paper, Kackar and Harville (1984) proposed an MSE approximation based on a first-order Taylor series expansion. Harville and Jeske (1992) modified the Kackar-Harville approximation, using a second-order Taylor series to adjust for bias in MSE[t(θ)|$\hat{\theta}$] as an estimator of MSE[t(θ)]. They referred to

this MSE approximation as the *Prasad-Rao estimator* because it is a generalization of a similar estimator developed by Prasad and Rao (1990) for three specific mixed models. In this dissertation, the term *Prasad-Rao approximation* is used to refer to the Prasad-Rao estimator as described by Harville and Jeske. (The term *estimator* is also appropriate, as analytic MSE approximations are based on unknown parameter values and must be estimated using data in practice.) Details of both the Kackar-Harville and Prasad-Rao approximations are given in Section 7.2 of the Technical Appendix.

Given an MSE approximation, standard methods of inference can be carried out based on an approximation of the distribution of the appropriate test statistic. Confidence interval construction (which subsumes hypothesis testing) is discussed in the next section.

2.7. Analytic Intervals for Fixed Effects

In the fixed effect case, each estimator in β_{EGLS} is a linear combination of the normally distributed components of Y, so the EBLUE, $t(\hat{\theta}) = k'\beta_{EGLS}$ is also normally distributed. The EBLUE is unbiased, so its variance, say σ_t^2, is equal to its MSE. If σ_t^2 were known, the test statistic $[t(\hat{\theta}) - T]/\sigma_t$ would have a standard normal distribution, and a $1-\alpha$ confidence interval for T could be obtained from the equation $P(-z_{1-\alpha/2} < [t(\hat{\theta}) - T]/\sigma_t < z_{1-\alpha/2}) = 1-\alpha$, where $P(\cdot)$ denotes the probability of the statement in the parentheses, and $z_{1-\alpha/2}$ is a standard normal cutoff. Solving the middle of the inequality for T yields the interval $[t(\hat{\theta}) - z_{1-\alpha/2}\sigma_t, t(\hat{\theta}) + z_{1-\alpha/2}\sigma_t]$.

In the case of known σ_T^2, the test statistic $[t(\hat{\theta}) - T]/\sigma_t$ is a *pivot*—a function of the data and parameter(s) whose distribution does not depend upon the value(s) of the parameter(s). Such functions are said to be *pivotal*. In practice, pivotal quantities generally depend on unknown parameters, which must be estimated. The term *quasi-pivot* is used in this dissertation to describe functions that are approximately, but not exactly, pivotal.

When σ_t^2 is unknown, the pivot can be approximated by $[t(\hat{\theta}) - T]/s_t$, where s_t^2 is an approximation of σ_t^2. The distribution of this quasi-pivot is often approximated by a Student's *t* distribution. Given degrees of freedom for the *t* statistic (discussed below), an approximate $1-\alpha$ confidence interval can be constructed using the formula

$$(1) \qquad [t(\hat{\theta}) - t_{1-\alpha/2}s_t,\ t(\hat{\theta}) + t_{1-\alpha/2}s_t],$$

where $t_{1-\alpha/2}$ is a *t* distribution cutoff (Hulting & Harville, 1991).

One option for computing confidence limits under this approach is to take s_t^2 as the naïve approximation and use $n - p$ degrees of freedom (where p is number of fixed terms in the model) for the approximating *t* distribution (Demidenko, 2004). However, the distribution of the quasi-pivot may have heavier tails than this *t* distribution due to the naïve approximation's downward bias and/or due to the degrees of freedom being incorrect (Harville & Carriquiry, 1992). This can lead to interval coverage rates that fall short of the nominal level, as demonstrated in a simulation study by McLean and Sanders (1988).

Instead of using $n - p$ degrees of freedom, Giesbrecht and Burns (1985) adapted Satterthwaite's (1941, 1946) method of approximating degrees of freedom, details of which are given in Section 7.3 of the Technical Appendix. Fai and Cornelius (1996) and Kenward and Roger (1997) developed extensions of this approach.

For testing a set of linear combinations of fixed effects, Kenward and Roger (1997) proposed an *F* statistic based on the Prasad-Rao MSE approximation, as well as an approximation for its denominator degrees of freedom. In the case of a single linear combination of fixed effects, the degree of freedom approximation is the same as the adapted Satterthwaite (1941, 1946) approximation used by Giesbrecht and Burns (1985). For the linear combination $T = k'\beta$, the square root of the Kenward-Roger *F* statistic is the quasi-pivot $[t(\hat{\theta}) - T]/s_t$, where s_t^2 is the Prasad-Rao approximation. This quasi-pivot has an approximate *t* distribution and can be used to construct confidence limits for T using formula (1) above. The Kenward-Roger statistic, its degrees of freedom, and the Prasad-Rao MSE approximation are calculated when the DDFM=KR option is specified in SAS PROC MIXED.

It is important to note that the accuracy of the methods described in this section for the case of unknown σ_t^2 depends both upon the adequacy of the chosen MSE approximation, *and* upon the similarity between the unknown distribution of the quasi-pivot and the Student's *t* distribution used to approximate it. The performance in simulation studies of the naïve, Kackar-Harville, and Prasad-Rao MSE

approximations, as well as that of the Kenward-Roger method, is discussed in the next section.

2.8. Performance of Analytic Approximations

The naïve, Kackar-Harville, and Prasad-Rao approximations were compared in simulation studies by Harville and Jeske (1992), Hulting and Harville (1991), and Singh, Stukel, and Pfeffermann (1998). The naïve and Kackar-Harville approximations tended to be biased downward, the latter less than the former. The Prasad-Rao approximation generally exhibited the least bias of the three methods except for small values of the variance ratio ($\gamma = \sigma_u^2/\sigma_e^2$).

Hulting and Harville (1991) noted that the rightmost term in the equation given by Kackar and Harville (1984),

$$\mathrm{MSE}[t(\hat{\theta})] = \mathrm{MSE}[t(\theta)] + \mathrm{E}[t(\hat{\theta}) - t(\theta)]^2,$$

might be inadequately approximated in both the Kackar-Harville and Prasad-Rao approximations for values of the variance ratio close to or equal to zero (see Section 7.2). This is evidently much more of a problem in prediction than in estimation.

For example, when the effect of interest in the simulation studies cited above was mixed, the relative bias of the Prasad-Rao approximation under small numbers of clusters ($m \leq 21$) was higher than 12% for $\gamma \leq 0.2$ and exceeded 100% in some cases for $\gamma \leq 0.1$. By contrast, when the effect of interest was fixed (a case considered as part of the study conducted by Hulting and Harville, 1991), the bias of the Prasad-Rao approximation for $m = 12$ was only 5.1% for $\gamma = 0$ and 2.3% for $\gamma = 0.2$. Evidently the

presence of a random term in the effect being estimated greatly exacerbates the problem of bias under small values of the variance ratio.

In their study of MSE approximation for the EBLUE, Hulting and Harville (1991) found that the Prasad-Rao approximation tended to become less biased as γ increased, and was less biased than the other two approximations in every condition save $\gamma = 0$ (the Kackar-Harville having smaller bias of –4.4% in this case). Moreover, for all values of γ in the study, the *t*-based confidence intervals constructed using the Prasad-Rao approximation had coverage rates closer to the nominal level than did intervals constructed using the other approximations.

Except, perhaps, for small values of γ, the Prasad-Rao approximation is a better choice for use in standard methods of inference for fixed effects than the other two analytic approximations considered here. The Kenward-Roger method is based on Prasad-Rao approximation, and on a sophisticated degree of freedom approximation, so it can be expected to outperform other standard methods of inference under most conditions.

Kenward and Roger (1997) conducted a simulation study to evaluate the small sample performance of their method under four mixed models. They found that the Prasad-Rao approximation adequately corrected the consistent, and sometimes severe, downward bias of the naïve approximation, and that hypothesis tests based on their method had reasonably accurate rates of Type I error. The Kenward-Roger method also performed well in repeated measures simulations conducted by Schaalje et al. (2002) and by Gomez, Schaalje, and Fellingham (2005), although it led to inflated

Type I error rates for some complex covariance structures when sample size was small.

Because the Kenward-Roger method is based on the Prasad-Rao approximation, very small variance ratios may hinder its performance. Savin, Wimmer, and Witkoský (2003) simulated data under a one-way RANOVA model for small numbers of clusters (m = 2, 5, 11, and 21), several cluster size conditions, and values of γ ranging from zero to four. They then used the Kenward-Roger method to construct 95% confidence intervals for the fixed intercept in the RANOVA model and compared rates of interval coverage under the various conditions.

Under nearly every sample size condition, coverage rates were least accurate for $\gamma = 0$, in most cases due to severe over-coverage. For larger sample sizes (11 or 21 clusters, each with 10 or 30 observations), coverage rates were mostly excellent for all non-zero values of γ (i.e., for $\gamma \geq 0.25$). However, in several small sample conditions, coverage rates were poor for $\gamma = 0.25$; in some of these cases, rate of coverage improved as γ increased, achieving or nearly achieving the nominal level for $\gamma = 4$.

Conclusion. An alternative to standard methods of inference, including the Kenward-Roger method, is worth considering. While the case of $\gamma = 0$ is somewhat trivial, as a mixed model is usually unnecessary if there are no differences between cluster means, small ICCs (corresponding to small variance ratios) are not uncommon in applied research. For example, Hedges and Hedberg (2007) obtained a national sample of academic achievement scores for grades K-12 and found an overall average

ICC (for students nested within schools) of 0.22 (equivalent to a variance ratio of about 0.28), and an average ICC of only 0.09 (equivalent to a variance ratio of about 0.10) for low-achievement schools. Moreover, even in cases for which the Kenward-Roger method is adequate, greater accuracy may be achieved by an alternative, bootstrap-based approach.

Preview. In the next chapter, a general introduction to parametric and nonparametric bootstrapping is provided, the application of bootstrapping to mixed models is discussed, and an account is given of bootstrap-based methods of MSE approximation. Bootstrap intervals are then described, followed by a discussion of the parametric bootstrap interval approach to estimation in mixed models.

Chapter 3. Proposed Solution: Parametric Bootstrap Intervals

3.1. The Bootstrap—Basic Concepts

Efron introduced bootstrapping to the scientific community in 1979 with his seminal paper in *The Annals of Statistics*. In it, he showed that the jackknife (Quenouille, 1949; Tukey, 1958) can be understood as a Taylor series approximation of the bootstrap and demonstrated its effectiveness on a number of estimation problems. (Jackknifing is described briefly in Section 3.1.2 below.) Chernick (2008) provided a recent overview of bootstrapping techniques and applications, including extensive historical notes and a massive bibliography.

The following descriptions of the parametric and nonparametric bootstraps for a single univariate sample are based on Efron (1979). Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample from probability distribution *F*. Let t(X, *F*) be a statistic chosen to estimate some parameter T, where the notation indicates that t may depend both on the sample, X, and on the distribution from which X is drawn, *F*. Standard methods of inference concerning T are based upon the sampling distribution of t, which may be unknown.

3.1.1. Parametric Bootstrap

If the distribution *F* is known up to a number of unknown parameters, it can be estimated using ML estimates of these parameters computed from the observed data. The estimated distribution of *F* based on these estimates is known as the *empirical distribution*. For example, if *F* is known (or assumed) to be normal with unknown mean and variance, the empirical distribution is the normal distribution

centered at the sample mean with variance equal to the ML sample variance estimate. In parametric bootstrapping, *bootstrap samples* are random draws from the empirical distribution, usually approximated by computer-simulated draws. Let X* denote one bootstrap sample.

For each bootstrap sample, a bootstrap estimate of t, say t*, is computed using the observed bootstrap data. A sampling distribution for the statistic t* is constructed by drawing a large number of bootstrap samples and taking the probability of t* falling in a given range to be the proportion of bootstrap samples for which the observed value of t* falls in that range. This *bootstrap distribution* of t* is used as an estimate of the sampling distribution of t.

3.1.2. Nonparametric Bootstrap

If data are drawn from an unknown distribution, and one is unwilling to make assumptions about its form, it can be approximated using a nonparametric bootstrap by assigning equal probability to each observation in the sample and *resampling*—i.e., drawing random (or pseudo-random) samples of size *n* with replacement from the original data. Note that in any one of these bootstrap samples, a given observation may appear more than once, or not at all. (By contrast, jackknifing is carried out by assembling, without replacement, every possible sample of size $n - 1$ from the data; no randomness is involved.) As in the parametric case, the sampling distribution of the statistic t is estimated by the bootstrap distribution of t* across a large number of bootstrap samples.

Residual bootstrap for regression. Efron (1979) showed how the nonparametric bootstrap could be applied to a (possibly nonlinear) regression model to estimate the sampling distribution of a regression coefficient estimate—an instructive case for the problem addressed in this dissertation. (See Ch. 6 in Davison & Hinkley, 1997, for further discussion of bootstrap methods for linear regression.)

Applied to a linear regression model with ordinary least squares (OLS) regression, this procedure (known as the *residual bootstrap*) is carried out as follows. Express the linear regression model in matrix terms as $Y = X\beta + e$. Y is regressed on X to obtain the OLS estimate of β, say β_{OLS}, and the corresponding vector of residuals, say r. Bootstrap samples are drawn from the empirical distribution of the OLS residuals to approximate the (unknown) distribution of the errors. For each bootstrap residual sample, say r*, a bootstrap response vector, Y*, is constructed by adding the bootstrap residual vector to the estimated mean structure, $Y^* = X\beta_{OLS} + r^*$. Then Y* is regressed on X to obtain the bootstrap estimate of β, β_{OLS}^*, for each sample.

The sampling distribution of β_{OLS} (as well as its mean and variance) can be estimated by the bootstrap distribution of β_{OLS}^*. The mean of β_{OLS} can be estimated by taking the average value of β_{OLS}^* across bootstrap samples. Similarly, an estimate of the variance of β_{OLS} is given by

$$\frac{1}{B}\sum_{b=1}^{B}\left(\beta_{OLS}^* - \beta_{OLS}\right)^2,$$

where bootstrap samples are numbered $b = 1, 2, \ldots, B$ (Chernick, 2008).

Cases bootstrap. The residual bootstrap corresponds to the common, though often dubious, assumption of fixed covariates. Efron and Tibshirani (1993) proposed an alternative for the case of random covariates—i.e., regression predictors assumed to take observed values as a result of a random sampling process.

Let Y_i denote the *i*th response and x_i denote the corresponding row vector of random covariates. A bootstrap data set, (X*, Y*), is formed by drawing a sample of size *n* from the set of pairs (x_1, Y_1), (x_2, Y_2), …, (x_n, Y_n), stacking the bootstrapped covariates to form X*, and stacking the bootstrapped responses to form Y*. For each bootstrap sample, Y* is regressed on X* to obtain β_{OLS}*, and the sampling distribution of β_{OLS} is estimated by the bootstrap distribution of β_{OLS}*.

This method of bootstrapping pairs is an example of what is known as a *cases bootstrap*. Whereas the residual bootstrap is based on the common assumption that the error terms are independent of the covariates, the only assumption underlying the cases bootstrap is that the original sample of x_is and Y_is was drawn randomly from some bivariate distribution (Efron & Tibshirani, 1993).

The parametric and nonparametric bootstrap methods above can be adapted for models giving rise to nested data. Bootstrap methods for mixed models are described in the next section.

3.2. Bootstrap Methods for Mixed Models

3.2.1. Parametric Bootstrap

The most straightforward mixed model bootstrap is the parametric bootstrap, which is designed to imitate the sampling of data from the underlying population. As

described below, the parametric bootstrap is based on the assumption of a fixed design matrix, X, as well as on the standard distributional assumptions concerning u and e (van der Leeden et al., 2008).

The method is implemented as follows. The distributions of u and e, which are assumed known up to the variance parameters, are estimated using variance parameter estimates from the original data. A bootstrap random effects vector u* is obtained by simulating *m* random draws (corresponding to the *m* clusters in the original data) from the distribution of u. For each cluster u_i, n_i random errors are simulated by simulating draws from the distribution of e (where n_i is the size of the *i*th cluster in the original data). These bootstrapped errors are stacked by cluster to form a bootstrap error vector e*. The bootstrap observation vector Y* is given by

$$Y^* = \beta_{EGLS}X + Zu^* + e^*,$$

where β_{EGLS} is estimated from the original data. This sampling process is repeated to form a large number of bootstrap data sets, and parameter estimates are obtained for each bootstrap data set using the same estimation methods employed for the original data. The sampling distribution of the statistic of interest can then be estimated by its bootstrap distribution.

3.2.2. Nonparametric Bootstrap

When the distributional assumptions of the mixed model are violated, nonparametric bootstrap methods can be considered. Their application to mixed models, especially involving clustered data, is not as straightforward as for regression

models. See Field and Welsh (2007) for an overview of bootstrap methods for clustered data.

Residual bootstrap. The mixed model residual bootstrap is nonparametric in the sense that it is not based on the usual distributional assumptions associated with the mixed model. In this context, the term *residual* applies to any estimated random term (in u or in e), and the residual bootstrap involves pseudo-random sampling of these residuals. The following description is based on van der Leeden et al. (2008).

There are various methods of obtaining an estimate of u, say $\hat{u}$, which is equivalent to estimating factor scores in factor analysis (see ten Berge, Krijnen, Wansbeek, & Shapiro, 1999). An OLS estimate of u can be obtained by estimating β (by EGLS, say), expressing the mixed model as $Y - X\beta_{EGLS} = Zu + e$, and regressing $(Y - X\beta_{EGLS})$ on Z. Denote the resulting estimate by $\hat{u}$. By assumption, $E(u) = 0$, so $\hat{u}$ must be mean-centered to remove any bias. After centering $\hat{u}$, e is simply estimated by centering $\hat{e} = Y - (X\beta_{GLS} + Z\hat{u})$.

The random effects in u can also be estimated using empirical best linear unbiased prediction. The EBLUP estimator of u is sometimes referred to as a *shrinkage* estimator because compared to estimators obtained when u is treated as fixed (such as the OLS estimator described above), the EBLUP is *shrunken* toward its mean (Robinson, 1991). After centering this estimate of u, an estimate of e is obtained by subtraction as described above for the OLS case.

Given estimates of u and e, the bootstrap vector u* is formed by taking one pseudo-random draw per cluster from the estimated u_is and stacking by cluster. The

bootstrap error vector e* is obtained by simulating *n* random draws from e. Note that independent sampling from u and e yields bootstrap estimates u* and e* that are independent, consistent with the standard assumption that u and e are independent (Carpenter, Goldstein, & Rasbash, 2000). The bootstrap observation vector Y* is given by $Y^* = \beta_{GLS}X + Zu^* + e^*$.

Carpenter et al. (2000) pointed out that without correction, the bootstrap variance parameter estimates obtained using the residual bootstrap under EBLUP (shrinkage) estimation are biased downward because the estimates of u and e (from which bootstrap samples are drawn) tend to be underdispersed—i.e., the variance-covariance matrices of the estimates of u and e, respectively, will tend to be smaller (in a matrix sense) than the estimates of G and R obtained from the original data. They proposed a method for "reflating" the estimates of u and e to correct this problem. (See also Wang, Carpenter, & Kepler, 2006).

Cases bootstrap. There are various versions of the cases bootstrap for clustered data involving sampling of clusters and/or units within clusters (Davison & Hinkley, 1993; van der Leeden et al., 2008). Note that for clustered data, the *ij*th case comprises the observation Y_{ij} and the *ij*th row of the X and Z matrices.

If sampling with replacement is to be carried out at both levels, the procedure is as follows. One begins by drawing (with replacement) a pseudo-random sample of clusters. For each cluster selected, random draws of cases from within the cluster are simulated. Data for the selected cases are stacked by cluster to form bootstrap vectors

Y*, X*, and Z*. The bootstrap model is $Y^* = \beta_{GLS}X^* + Z^*u + e$, and parameter estimates are obtained as usual for each bootstrap sample.

The cases bootstrap is usually less efficient than the residual bootstrap but is more appropriate in cases in which X contains random (rather than fixed) explanatory variables (van der Leeden et al., 2008). Note that for unbalanced data, the total bootstrap sample size may vary depending on which clusters are selected in a given bootstrap sample.

Regardless of the bootstrap method employed, after bootstrap data sets are generated the procedure for obtaining bootstrap estimates is the same. The value of the statistic of interest is computed for each bootstrap sample, and the resulting bootstrap distribution is used to approximate the sampling distribution of the statistic of interest for purposes of inference and/or interval construction.

3.3. Bootstrap Approaches to MSE Approximation in Mixed Models

Given the promise of bootstrap methods and the problem of MSE estimation in mixed modeling, it is not surprising that bootstrap-based approaches to MSE estimation have been proposed. These include jackknife estimation (Jiang, Lahiri, & Wan, 2002), bootstrap-based improvements of analytic approximations (Butar & Lahiri, 2003; González-Manteiga, Lombardía, Molina, Morales, & Santamaría, 2008; Pfefferman & Glickman, 2004), and parametric bootstrap MSE estimation (Hall & Maiti, 2006).

Bootstrap-based MSE estimates computed using these methods can be used in standard methods of inference—for example, to compute confidence limits for a fixed

effect using formula (1). However, standard methods of inference depend not only upon accurate MSE estimation, but also upon the goodness of the approximation to the test statistic distribution; if bootstrapping is to be carried out, there is no reason to limit its use to MSE estimation when it can also be used to estimate the test statistic distribution, thereby eliminating the need to approximate it by some known distribution.

In fact, bootstrap estimation of the test statistic distribution makes MSE estimation unnecessary as well. Under the bootstrap interval approach, discussed in the following sections, confidence limits are based on cutoffs of a bootstrap distribution rather than on *t* or *z* distribution cutoffs, and an MSE estimate is not required. A general presentation of bootstrap intervals is provided in Section 3.4, followed by a discussion of parametric bootstrap intervals for fixed effects in mixed models in Section 3.5.

3.4. Bootstrap Interval Construction

Let statistic t be an estimator of T, where T is a model parameter or some function of model parameters. Let s_t^2 be an estimator for the variance of t. In this dissertation, the statistic of interest is the EBLUE of the fixed effect T, $t = t(\hat{\theta})$, but bootstrap intervals are presented below in more general terms. First, however, the standard analytic approach to interval construction is reviewed.

Standard analytic intervals. If the sampling distribution of t is normal, or becomes so asymptotically, then standard analytic confidence limits for T are based on the quasi-pivot $(t - T)/s_t$, which is assumed to have an approximate standard

normal distribution (or, for samples that are not large, a *t* distribution). An approximate $1-\alpha$ normal-based confidence interval is given by $(t - z_{1-\alpha/2}s_t, t - z_{\alpha/2}s_t)$, which simplifies to the more familiar formula

$$(2) \qquad (t - z_{1-\alpha/2}s_t, t + z_{1-\alpha/2}s_t)$$

by symmetry of the normal distribution.

Bootstrap normal interval. The bootstrap normal interval is constructed simply by substituting a bootstrap estimate of SD(t), say s_t^*, for s_t in the normal-based interval formula (2). Given *B* bootstrap samples and values of t*, a reasonable substitute for s_t is

$$s_t^* = \left[\frac{1}{B} \sum_{b=1}^{B} (t^* - t)^2 \right]^{1/2}.$$

In the mixed model context, where t is the EBLUE, one of the bootstrap-based methods mentioned in Section 3.3 could be used to estimate SD(t).

Because the quasi-pivot distribution is approximated under this method by the normal distribution, not by a bootstrap distribution, the bootstrap normal is not a bootstrap interval in the strictest sense. Note that in constructing the interval, the bootstrap distribution of t* is used only to compute the sampling variance of t*; the other information contained in the bootstrap distribution—i.e., information about its shape—is ignored. This is not the case with the true bootstrap intervals described below.

Efron's percentile interval. Efron's percentile interval (see Efron & Tibshirani, 1993) is based on the premise that the bootstrap distribution of t*

resembles the sampling distribution of t. Under this premise, it seems reasonable that an interval containing 100(1–α)% of the ordered realized values of t* would also contain T in roughly the same percentage of repeated samples from the population (Chernick, 2008). Let t_α^* denote an α cutoff for the bootstrap distribution of t*—i.e., a value that cuts off approximately 100(α)% of the observed values of t* to its left. Based on the approximations

$$(3) \qquad 1-\alpha \approx P(t_{\alpha/2}^* < t^* < t_{1-\alpha/2}^*) \approx P(t_{\alpha/2}^* < t < t_{1-\alpha/2}^*) \approx P(t_{\alpha/2}^* < T < t_{1-\alpha/2}^*)$$

an approximate 1–α confidence interval for T is given by $(t_{\alpha/2}^*, t_{1-\alpha/2}^*)$.

For theoretical reasons beyond the scope of this dissertation (see Chernick, 2008), t_α^* should ideally be defined as the $\alpha(B+1)^{th}$ ordered value of t* in the bootstrap sample, where *B* is chosen such that $\alpha(B+1)$ is an integer. If $\alpha(B+1)$ is not an integer, interpolation can be used to find an approximate value for t_α^* (see Davison & Hinkley, 1993). References to α cutoffs of bootstrap distributions in the remainder of this dissertation are given with this definition in mind.

Efron's percentile interval does not work well for small samples, especially when drawn from asymmetric distributions (Chernick, 2008). Its performance can be improved by modifications discussed below. Note that the interval may not be symmetric about t, unlike standard analytic intervals.

Bias-corrected (BC) interval. If t is a biased estimator of T, the two probabilities on the far right in (3) above will tend to be unequal. Assuming t is a *plug-in estimator* of T—i.e., t is calculated using the formula for T by substituting an

estimate for each unknown parameter—Efron's percentile interval can be bias-corrected by the following method, described by Efron and Tibshirani (1993).

Let $p_{t^*<t}$ denote the proportion of observed values of t* less than the observed value of t. The bias correction z_0 is the point on the standard normal distribution that cuts off $p_{t^*<t}$ of the area under the curve to the left—i.e., $z_0 = \Phi^{-1}(p_{t^*<t})$, where Φ is the standard normal cumulative distribution function. For example, if $p_{t^*<t} = 0.512$, indicating a tendency of t* to underestimate t, $z_0 = 0.03$ (which cuts off 51.2% of the area under the normal distribution to its left).

In constructing the $1-\alpha$ bias-corrected (BC) interval, the usual cutoff proportions, $\alpha/2$ and $1-\alpha/2$, are replaced by the following bias-adjusted proportions:

$$\alpha_{lo} = \Phi(2z_0 + z_{\alpha/2}) \text{ and } \alpha_{up} = \Phi(2z_0 + z_{1-\alpha/2}),$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ denote standard normal cutoffs. As indicated by the notation, α_{lo} is the area under the normal curve to the left of $2z_0 + z_{\alpha/2}$, and α_{up} is the area under the normal curve to the left of $2z_0 + z_{1-\alpha/2}$. The lower endpoint of the BC interval is the value that cuts off α_{lo} of the observed values of t* to its left, say $t^*(\alpha_{lo})$, and the upper endpoint the value that cuts off α_{lo} of the values of t* to its left, say $t^*(\alpha_{up})$.

Note that if the sample median of the bootstrap distribution of t* is equal to t, $p_{t^*<t} = 0.5$, $z_0 = 0$, and no bias adjustment takes place. On the other hand, if $p_{t^*<t}$ is less than 0.5, the observed median bias of t* is positive, and the bias-correction z_0 will be negative, resulting in both the lower and upper confidence limits being adjusted downward. Similarly, a negative median bias will result in adjusting the confidence interval upward.

Bias-corrected accelerated (BC_a) interval. In some cases the standard deviation of t, SD(t), may depend on the true value of T, and the performance of the BC interval can be improved by adjusting for this dependence. The *acceleration* is the rate at which SD(t) changes with respect to T. This value is approximated on a normalized scale by the constant *a*. Given *a*, the following cutoff proportions are computed:

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})} \right), \quad \alpha_2 = \Phi\left(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})} \right).$$

The BC_a interval has lower endpoint $t^*(\alpha_1)$ and upper endpoint $t^*(\alpha_2)$, where each cuts off the indicated proportion of the observed values of t* to its left. This interval performs well in a wide range of cases (Chernick, 2008).

Efron (1987) proposed various methods of estimating the acceleration, as well as a method for reducing a multi-parameter problem to a single-parameter problem using multivariate calculus so that an acceleration estimate can be obtained. This multi-parameter technique can be used with the parametric bootstrap for mixed models but is rather complicated (van der Leeden et al., 2008).

Hall's percentile interval. Hall (1992) proposed a percentile interval based on the bootstrap estimate of the sampling distribution of (t – T). Let q_α^* denote the α cutoff for the bootstrap distribution of $q^* = (t^* - t)$. Assuming $P[q_{\alpha/2}^* < (t - T) < q_{1-\alpha/2}^*] \approx 1-\alpha$, an approximate $1-\alpha$ confidence interval for T is given by $(t - q_{1-\alpha/2}^*, t - q_{\alpha/2}^*)$. In terms of t* as defined above for Efron's percentile interval, q* equals t* – t,

and Hall's interval is given by $(2t - t_{1-\alpha/2}^*, 2t - t_{\alpha/2}^*)$. Thus, Hall's interval seems backwards compared to Efron's, and vice versa.

Bootstrap-t interval. The bootstrap-*t* (or percentile-*t*) interval (Efron & Tibshirani, 1993) is based on the bootstrap estimate of the sampling distribution of the quasi-pivot $(t - T)/s_t$. Let $q^* = (t^* - t)/s_t^*$ denote the bootstrap pivot approximation, where s_t^* is an estimate of SD(t*). Let q_α^* denote the α cutoff for the bootstrap distribution of q*. Assuming

$$1-\alpha \approx P[q_{\alpha/2}^* < (t^* - t)/s_t^* < q_{1-\alpha/2}^*] \approx P[q_{\alpha/2}^* < (t - T)/s_t < q_{1-\alpha/2}^*],$$

an approximate $1-\alpha$ confidence interval for T is given by

$$(4) \qquad (t - q_{1-\alpha/2}^* s_t^*, t - q_{\alpha/2}^* s_t^*).$$

The bootstrap-*t* interval generally performs better than Hall's interval because $(t^* - t)/s_t^*$ behaves more like a pivot (i.e., is less dependent on the parameter estimates) than $(t^* - t)$, the quantity on which Hall's interval is based (van der Leeden, Meijer, & Busing, 2008). Although it converges at a faster rate (see Section 7.1, Technical Appendix) than standard analytic intervals and bootstrap percentile intervals, the bootstrap-*t* is susceptible to influence from outliers and can yield erratic results (Efron & Tibshirani, 1993). The performance of the interval depends upon having a reasonably accurate estimate of SD(t*), and in the absence of such an estimate, the results can be "disastrous" (Hall, 1992, p. 141). If a good analytic estimate is unavailable, a bootstrap estimate of SD(t*) can be computed using a second layer of bootstrapping, but this adds cost in terms of computing time and the complexity of the calculations required to compute the confidence limits.

Comparison of bootstrap intervals. Among the available bootstrap intervals, the bootstrap-*t* seems to provide the best balance between performance and cost, provided SD(t*) can be conveniently and accurately estimated. Like the bootstrap-*t*, the BC_a interval converges at a faster rate than standard analytic intervals and bootstrap percentile intervals (Efron & Tibshirani, 1993), but the difficulty of estimating the acceleration for the BC_a interval seems prohibitive for most researchers.

In those cases in which the problem of estimating SD(t*) makes the bootstrap-*t* a poor choice, Hall's interval seems the better of the two basic percentile intervals. Hall (1992) argued that Efron's interval is inadequate without adjustment. Efron and Tibshirani (1993) acknowledged that Efron's interval could benefit from adjustment but pointed out that neither percentile interval performs well generally and gave two cases for which Efron's interval is more appropriate. In cases for which SD(t*) is relatively constant with respect to t*, the quantity upon which Hall's interval is based, t* – t, closely approximates a pivotal quantity, and Hall's interval should outperform Efron's.

3.5. Parametric Bootstrap Intervals for Mixed Models

Using a parametric bootstrap to construct basic percentile and bootstrap-*t* intervals for a mixed effect T in the mixed model context is relatively straightforward. With the parametric version of the bootstrap there is no need to choose a resampling scheme, total sample size remains constant across bootstrap samples even with unbalanced data, and no programming is necessary to reflate underdispersed

residuals. Its limitation, of course, is that like standard analytic methods (including likelihood-based methods), it is based upon the usual distributional assumptions regarding the random terms in the mixed model.

Applications of the parametric bootstrap to interval construction in mixed modeling have been relatively few. Carpenter et al. (2003) used both a parametric bootstrap and a nonparametric residual bootstrap (modified to correct for underdispersion) to construct Efron percentile intervals for fixed effects using data simulated under a random coefficient model with non-normal random effects and non-normal errors. The nonparametric intervals had coverage levels closer to the nominal rate than the parametric intervals for all parameters for each of the three sample sizes studied, with coverage accuracy for the nonparametric intervals generally improving with increased sample size. This result is not surprising, as the parametric bootstrap is based on the assumption of normally distributed random terms, whereas the nonparametric bootstrap is not.

Harville and Carriquiry (1992) proposed what is essentially a bootstrap-*t* prediction interval based on a parametric bootstrap. Taking this idea further, Chatterjee, Lahiri, and Li (2008) developed a parametric bootstrap method for constructing intervals for mixed effects of the form $T = c'(X\beta + Zu)$, where c is a constant vector. Define (as above) a parametric bootstrap data set by $Y^* = X\beta_{EGLS} + Zu^* + e^*$. Because u^* is known (after simulation) for a given bootstrap sample, a parametric bootstrap (PB) estimate of T is given by $t_{PB}^* = c'(X\beta_{EGLS} + Zu^*)$. Alternatively, T can be estimated by the bootstrap EBLUP, say t_{EBLUP}^*, which is

computed by treating the bootstrap data set like a real data set (i.e., with unobservable random effects and errors that must be estimated) and using the usual EBLUP formula.

Let $\sigma^2_{T|Y}$ denote the conditional variance of T given Y, and let $s_{T|Y}$ denote the plug-in estimator of $\sigma_{T|Y}$ (obtained by substituting variance component estimates for their targets in the formula for $\sigma_{T|Y}$). As an alternative to the standard interval based on the normal pivot $(T - \mu_T)/\sigma_{T|Y}$, Chatterjee et al. (2008) simulated parametric bootstrap samples and computed the quasi-pivot $q^* = (t_{EBLUP}^* - t_{PB}^*)/s_{T|Y}^*$ for each bootstrap sample, where $s_{T|Y}^*$ denotes the bootstrap value of $s_{T|Y}$. Their prediction interval is given by $(t_{EBLUP} + q_1^* s_{T|Y}, t_{EBLUP} + q_2^* s_{T|Y})$, where q_1^* and q_2^* are appropriate quantiles of the bootstrap distribution of the quasi-pivot q^*.

Chatterjee et al. (2008) showed that this interval has a high rate of convergence that is dependent upon total sample size, meaning accuracy can be improved by increasing the number of clusters or by increasing cluster size. By contrast, coverage rates of intervals based on analytic MSE approximations can be increased only by adding clusters. Chatterjee et al. simulated data under the Fay-Herriot model and found that the coverage rates of their intervals were consistently closer to the nominal rate than those of standard intervals based on the Prasad-Rao approximation.

The theoretical and simulation results presented by Chatterjee et al. (2008) are promising, but their method cannot be applied to fixed effect estimation without modification. The denominator of their quasi-pivot, $s_{T|Y}^*$, is an estimator of $\sigma_{T|Y}$, the

conditional variance of T given Y; but when T is fixed (e.g., $T = k'\beta$), its variance (conditional or otherwise) is zero, like that of any constant.

A fixed effect version of the quasi-pivot used by Chatterjee et al. is

$$q^* = (t_{EBLUE}^* - t_{EBLUE})/s_t^*,$$

where t_{EBLUE} is estimated from the original data and s_t^* approximates the SD of t_{EBLUE}^* (i.e., the square root of its MSE as an estimator of t_{EBLUE}). This is identical to the quantity one would use in applying the bootstrap-*t* approach to interval estimation of T, and given a suitable choice for s_t^*, intervals can be constructed using the bootstrap-*t* formula (4), with $t^* = t_{EBLUE}^*$ and $t = t_{EBLUE}$.

As noted above, the performance of the bootstrap-*t* approach depends upon the quality of the estimator of SD(t*) used in the denominator of the quasi-pivot. Use of a bootstrap-based estimator requires a second bootstrap iteration; given the additional programming and computing time associated with an added layer of bootstrapping, an analytic approximation seems preferable. This is the approach taken by Chatterjee et al. (2008), who used a simple analytic plug-in estimator in their study. In fixed effect estimation, the Kenward-Roger method would seem to be the best choice for approximating SD(t*) under most conditions.

Simple alternatives to the bootstrap-*t* approach are Hall's (1992) percentile interval, based on the bootstrap distribution of $(t_{EBLUP}^* - t_{EBLUP})$, and Efron's interval. Unless the SD of the numerator is roughly constant across bootstrap samples, Hall's interval may not perform as well as the bootstrap-*t* (van der Leeden et al., 2008), but its advantage is that it can be computed without an estimate of this SD.

Chapter 4. Methods

Three simulation studies were carried out to compare the performance of several analytic and parametric bootstrap intervals. All parameter estimates were obtained using SAS PROC MIXED with the REML option.

4.1. Simulation Study 1

In the first simulation study, interval methods were compared using data simulated from the one-way RANOVA model $Y_{ij} = \beta_0 + u_i + e_{ij}$, with $\beta_0 = 0$. Two ICC conditions were considered, with the values of σ_u^2 and σ_e^2 chosen such that $\text{Var}(Y_{ij}) = \sigma_u^2 + \sigma_e^2 = 10$ and ICC = 0.05 or 0.10. Data were generated for two sample size conditions, $m = 10$ and $m = 15$. Each group of five clusters comprised four 3-unit clusters and one 6-unit cluster.

Scores for a fixed predictor, x_{ij}, were simulated with random draws from a N(0, 100) distribution, each rounded to the nearest 0.01, and the following nested error regression model was fit for each simulated data set: $Y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{1i} + e_{ij}$. Note that the data were generated independently of the x_{ij}s—i.e., the value of β_1 in the generating model was zero.

For each data set, formula (1) was used to construct two analytic intervals for β_1 at 90%, 95%, and 99% confidence levels. Naïve intervals were computed using *t* distribution cutoffs with $n - p$ degrees of freedom, and Kenward-Roger intervals were computed using the Kenward-Roger (KR) option in SAS PROC MIXED.

Using the random coefficient model, 999 parametric bootstrap samples were generated for each simulated data set and used to construct Hall, Efron, and

bootstrap-*t* intervals for β_1 at the three nominal confidence levels: 90%, 95%, and 99%. The Prasad-Rao MSE approximation, obtained using the KR option in SAS PROC MIXED, was used in calculating the bootstrap-*t* intervals. The observed coverage rate of the five intervals was calculated as the proportion of data sets for which the interval contained zero.

A two-sided, α-level test of the hypothesis H_0: $\beta_1 = 0$ can be conducted by constructing a $1-\alpha$ confidence interval for β_1 and rejecting H_0 if the interval does not contain zero. Because the data in this study were simulated with $\beta_1 = 0$, the Type I error rate of the interval-based hypothesis test (i.e., the probability of rejecting a true null hypothesis) can be estimated by the proportion of simulated data sets for which the interval does not contain zero, or equivalently, by one minus the observed coverage rate of the confidence interval.

Note that an interval may have poor coverage because it is too wide or too narrow, because it is biased upward or downward, or because of some combination of these problems. In order to assess bias in the five intervals, each case in which an interval failed to contain zero was classified as an overestimate (when the lower endpoint exceeded zero) or an underestimate, and the proportion of misses due to overestimation was computed for each interval under each nominal level.

4.2. Simulation Study 2

In the second simulation study, repeated measures data were simulated using a fitted model obtained by Terracciano et al. (2005) in a study of personality change in adulthood. During this 15-year study, researchers administered the Revised NEO

Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) at least once, and as many as 11 times, to each of 1,944 (967 female, 977 male) participants in the Baltimore Longitudinal Study of Aging. The average number of administrations per participant was 2.6; the average time between administrations was 2.8 years. Data were gathered from participants as young as 20 and as old as 96 during the study, but the bulk of the data came from participants over 60. The average age at first administration was 56.7, and average age across all administrations was 64.5.

For each scale in the NEO-PI-R, Terracciano et al. (2005) fit a random coefficient model with age (in decades, centered about its grand mean) as a linear predictor, as well as a random coefficient model with age as a quadratic predictor. After choosing the better of the two models based on model fit using a likelihood ratio test, they tested for effects of gender and cohort. They arrived at the following fitted model for Impulsiveness (a facet of the Neuroticism factor):

$$Y_{ij} = 48.3 + u_{0i} - 0.975(AGE_{ij} - 6.45) + u_{1i}(AGE_{ij} - 6.45) + e_{ij},$$

where Y_{ij} is the Impulsiveness score for the *i*th subject at the *j*th administration, AGE_{ij} is the age in decades of the *i*th subject at the *j*th administration, u_{0i} is the *i*th subject's random deviation from the mean intercept, u_{1i} is the *i*th subject's random deviation from the mean slope, and e_{ij} is the random error term associated with the *j*th score for the *i*th subject. The variances of u_{0i}, u_{1i}, and e_{ij} were estimated to be 52.91, 2.27, and 25.76, respectively. (Terracciano et al. did not model the covariance between u_{0i} and u_{1i}.)

The rationale of the simulation study is as follows. Suppose the fitted model and variance parameter estimates obtained by Terracciano et al. (2005) provide a reasonably accurate estimate of the population model and that the standard mixed model assumptions are valid. If the variance parameter estimates are treated as population parameters, the distributions of the random terms are known: $u_{0i} \sim N(0, 52.91)$, $u_{1i} \sim N(0, 2.27)$, and $e_{ij} \sim N(0, 25.76)$. Given values for the AGE variable, random sampling of subjects and their scores can be simulated by drawing random terms from these distributions for each subject and using these values to compute simulated Impulsiveness scores. After simulating a large number of such data sets, the performance of confidence intervals can be assessed by comparing each interval's nominal coverage rate to the proportion of simulated data sets for which it contains the value of the target parameter—in this study, the fixed slope associated with AGE, which was equal to –0.975.

Data were simulated for 5,000 samples of 15 subjects. For simplicity, three scores were simulated for each subject, with each score corresponding to one of three equally spaced time points. (This is comparable to a real-life study in which each administration occurs on the same day for all subjects.) Time points were 2.8 years apart. Subject ages at the median time point were spaced at equal intervals from 59.5 to 69.5 years (and thereby centered at 64.5 years). (Because subject age is considered fixed in the model, simulating a random sample of ages is unnecessary. The random coefficient model allows a unique regression intercept and slope for each subject, and in theory the ages at which data are gathered for a subject have no effect on the line;

they simply determine the points on the line that are observed.) Ages were converted to decades, mean-centered, and used (along with the simulated random terms) to compute Impulsiveness scores.

For each simulated data set, SAS was used to estimate the fixed intercept and slope, and the variances of the three random terms. Bootstrap Impulsiveness scores were generated using these fixed intercept and slope estimates from PROC MIXED, the subject ages in the simulated data set, and random terms (intercepts, slopes, and errors) obtained by pseudo-random draws from their respective estimated distributions. In other words, bootstrap samples were generated in the same way as the simulated data sets but based on different parameter estimates: Bootstrap data were based on parameter estimates obtained from a simulated data set, and the simulated data were based on estimates from Terracciano et al. (2005).

For each simulated data set, naïve, Kenward-Roger, Hall, Efron, and bootstrap-*t* intervals for the fixed slope were constructed at the three nominal confidence levels. The bootstrap intervals were based on 1,999 parametric bootstrap samples per data set. The Prasad-Rao approximation was used to compute the bootstrap-*t* intervals. As in the first simulation study, proportions of misses due to overestimation were computed along with interval coverage rates.

4.3. Simulation Study 3

In the third simulation study, 3,250 data sets were simulated using the random coefficient model $Y_{ij} = \beta_0 + u_{0i} + \beta_1 u_{1i} + e_{ij}$, with $\beta_0 = 0$, $\beta_1 = 2$, $\mathrm{Var}(u_{0i}) = 1$, $\mathrm{Var}(u_{1i}) = 1$, and $\sigma_e^2 = 4$. The random components of the intercept and slope were generated

independently—i.e., $\text{Cov}(u_{0i}, u_{1i}) = 0$. Scores for a fixed predictor, x_{ij}, were simulated with random draws from a N(0, 1) distribution, each rounded to the nearest 0.01. Each of the data sets comprised either ten or twenty 10-unit clusters.

For each data set, PROC MIXED was used to obtain estimates of the model parameters, including the covariance between the random intercept and slope. The FA0(2) option was used to ensure that each estimated G matrix was a plausible covariance matrix (i.e., convertible to a correlation matrix), and a lower bound of 0.01 was set for both the random intercept variance and random slope variance.

Analytic and parametric bootstrap intervals for β_1 were constructed at 90%, 95%, and 99% confidence levels. In addition to the naïve and Kenward-Roger intervals considered in the first two simulation studies, a third analytic interval was constructed using the naïve MSE approximation and the Satterthwaite degree of freedom approximation developed by Giesbrecht and Burns (1985) and adopted by Kenward and Roger (1997) for use in their method. The Satterthwaite degrees of freedom were obtained using the KR option in PROC MIXED.

Using 2,500 bootstrap samples for each simulated data set, Hall, Efron, BC, and bootstrap-*t* intervals were constructed. Bootstrap samples for which PROC MIXED did not converge were omitted, but this had a negligible effect on the number of bootstrap samples per data set, the minimum being 2,486. In pilot studies, intervals computed using the naïve approximation and Satterthwaite degrees of freedom had more accurate rates of coverage than Kenward-Roger intervals, so the naïve MSE approximation was used to compute the bootstrap-*t* intervals.

Chapter 5. Results and Discussion

5.1. Results of Simulation Study 1

Interval coverage rates and percentages of overestimation, respectively, are presented in Tables 1 and 2 in Chapter 8. Overall, all five intervals performed well in terms of coverage, regardless of sample size or ICC value. Not surprisingly, coverage rates and levels of bias were generally better under the larger sample size. In most cases, either the Kenward-Roger or bootstrap-*t* interval had rate of coverage closest to the nominal level. There was no discernable effect of ICC value on Kenward-Roger coverage rates; in fact, for both sample sizes, coverage rates were better in some cases for the smaller ICC value.

5.2. Results of Simulation Study 2

Tables 3 and 4 (see Chapter 8) contain coverage rates and percentages of overestimation for the second simulation study. The naïve intervals had the best coverage rates for all three nominal levels, with the Hall percentile intervals coming in a close second. The Efron interval coverage rates fell noticeably short of the nominal levels, while the Kenward-Roger coverage rates exceeded them.

5.3. Results of Simulation Study 3

Satterthwaite degrees of freedom could not be computed with PROC MIXED for 121 of the 3,250 simulated data sets with $m = 10$ and for 8 data sets with $m = 20$. Intervals based on the Satterthwaite degrees of freedom were not computed for these data sets, and their reported coverage rates are based on the data sets for which they

could be computed. Coverage rates for the other intervals were based on all 3,250 data sets.

As shown in Table 5, the naïve intervals computed with Satterthwaite degrees of freedom had the best coverage rates, followed by the bootstrap-*t* intervals. Coverage rates for the other intervals fell short of the nominal rate by over one percent in nearly every case. Better coverage was achieved for the larger sample size by all the intervals save the naïve/Satterthwaite intervals, which performed at the same level for both sample sizes.

The performance of the three percentile intervals (Hall, Efron, and BC) was nearly identical. The bias correction used in calculating the BC intervals did little to correct the bias in the Efron intervals and, in fact, increased it in some cases (see Table 6).

5.4. Discussion

As demonstrated in the simulation studies, the parametric bootstrap interval approach represents a viable alternative to standard methods of inference for fixed effects in the mixed model, even with small sample sizes. The bootstrap-*t* method seems especially promising, in spite of its dependence upon an accurate MSE estimate. The question of which bootstrap intervals are best suited for which models and conditions might be addressed in future studies.

The primary limitation of the parametric bootstrap is its dependence upon the mixed model assumption of normally distributed random terms. If this assumption is not met, or cannot be verified, the nonparametric bootstrap is more appropriate than

either standard methods of inference or parametric bootstrap methods. To the author's knowledge, the question is open as to whether there is any advantage, other than convenience, to choosing the parametric bootstrap over a nonparametric bootstrap when the mixed model assumptions are met.

The issue of statistical power is not addressed in this dissertation but merits consideration. A comparison of the power of various bootstrap and standard methods, coupled with further research regarding interval coverage rates, might help to identify a best method for various sample sizes.

Chapter 6. References

Aitken, A. C. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh, 55,* 42-48.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association, 83*(401), 28-36.

Birkes, D., & Wulff, S. S. (2003). Existence of maximum likelihood estimates in normal variance-components models. *Journal of Statistical Planning and Inference, 113*(1), 35-47.

Booth, J. (1995). Bootstrap methods for generalized linear mixed models with applications to small area estimation. In G. U. H. Seeber, B. J. Francis, R. Hatzinger, & G. Steckel-Berger (Eds.) *Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modelling* (pp. 43-51). New York: Springer.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models.* Newbury Park, CA: Sage.

Butar, F. B., & Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference, 112*(1-2), 63-76.

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal Of The Royal Statistical Society Series C, 52*(4), 431-443.

Catallier, D. J., & Muller, K. E. (2000). Tests for Gaussian repeated measures with missing data in small samples. *Statistics in Medicine, 19,* 1101-1114.

Chatterjee, S., Lahiri, P., & Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics, 36*(3), 1221-1245.

Chernick, M. C. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers* (2nd Ed.). Hoboken, NJ: Wiley.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Das, K., Jiang, J. M., & Rao, J. N. K. (2004). Mean squared error of empirical predictor. *Annals of Statistics, 32*(2), 818-840.

Datta, G. S., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica, 10*(2), 613-627.

Datta, G. S., Rao, J. N. K., & Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika, 92*(1), 183-196.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge, UK: Cambridge University Press.

Demidenko, E. (2004). *Mixed Models: Theory and Applications.* New Jersey: John Wiley & Sons, Inc.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association, 76*(374), 341-353.

Efron, B. (1979). Another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171-200.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* London: Chapman & Hall.

Fai, A. H. T., & Cornelius, P. L. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation, 54*(4), 363-378.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: Application of James-Stein procedures to census-data. *Journal of the American Statistical Association, 74*(366), 269-277.

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society—Statistical Methodology, 69*, 369-390.

Fisher, R. A. (1918). The correlation between relatives on the assumption of Mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh, 52*, 399-433.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh and London: Oliver & Boyd.

Fuller, W. A., & Harter, R. M. (1987). The multivariate components of variance model for small area estimation. In R. Platek, J. N. K. Rao, C. E. Sarndal, & M. P. Singh (Eds.) *Small Area Statistics: An International Symposium* (pp. 103-123). Wiley: New York.

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science, 9*(1), 55-76.

Giesbrecht, F. G., & Burns, J. C. (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. *Biometrics, 41*(2), 477-486.

Goldberger, A. S. (1962). Best linear unbiased prediction in generalized linear-regression model. *Journal of the American Statistical Association, 57*(298), 369-375.

Goldstein, H. (1995). *Multilevel Statistical Models* (2nd Ed.). Edward Arnold: London.

Gomez E. V., Schaalje G. B., & Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics—Simulation and Computation, 34*, 377-392.

Gonzalez-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D., & Santamaria, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation, 78*(5), 443-462.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer.

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 68*, 221-238.

Harville, D. A., & Carriquiry, A. L. (1992). Classical and Bayesian prediction as applied to an unbalanced mixed linear-model. *Biometrics, 48*(4), 987-1003.

Harville, D. A., & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association, 87*(419), 724-731.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized experiments in education. *Educational Evaluation and Policy Analysis, 29*, 60-87.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics, 9*, 226-252.

Henderson, C. R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding* (pp. 141-163). National Academy of Sciences, National Research Council, Pub. 982. Washington, DC: U. S. Government Printing Office.

Hulting, F. L., & Harville, D. A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: Computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association, 86*(415), 557-568.

Jiang, J. M., Lahiri, P., & Wan, S. M. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics, 30*(6), 1782-1810.

Kackar, R. N., & Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear-models. *Communications in Statistics—Theory and Methods, 10*(13), 1249-1261.

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear-models. *Journal of the American Statistical Association, 79*(388), 853-862.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983-997.

Lahiri, P., & Rao, J. N. K. (1995). Robust estimation of mean squared error of small-area estimators. *Journal of the American Statistical Association, 90*(430), 758-766.

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

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high-school achievement. *Sociology of Education, 62*(3), 172-192.

Littell, R. C. (2002). Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics, 7*(4), 472-490.

McLean, R. A., & Sanders, W. L. (1988). Approximating degrees of freedom for standard errors in mixed linear models. In *Proceedings of the Statistical Computing Section, American Statistical Association* (pp. 50-59).

Muller, K. E., Edwards, L. J., Simpson, S. L., & Taylor, D. J. (2007). Statistical tests with accurate size and power for balanced linear mixed models. *Statistics in Medicine, 26*, 3639-3660.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58*, 545-554.

Pfefferman, D. & Glickman, H. (2003). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 4167-4178).

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association, 85*(409), 163-171.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society B, 11*, 14-44.

Rao, J. N. K. (2003). *Small Area Estimation*. New York: Wiley.

Rao, J. N. K. (2005). Inferential issues in small area estimation: Some new developments. *Statistics in Transition, 7*(3), 513-526.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science, 6*(1), 15-32.

SAS Institute Inc. (2008). *SAS OnlineDoc® 9.1.3***.** Cary, NC: SAS Institute Inc. Retrieved November 12, 2008 from http://support.sas.com/onlinedoc/913/getDoc/en/statug.hlp/mixed_sect21.htm.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*(6), 110-114.

Savin, A., Wimmer, G., & Witkoský (2003). On Kenward-Roger confidence intervals for common mean in interlaboratory trials. *Measurement Science Review, 3*(1), 53-56.

Schaalje, G.B., McBride, J.B., & Fellingham, G.W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological & Environmental Statistics, 7*(4), 512-524.

Scheffé, H. (1956). Alternative models for the analysis of variance. *The Annals of Mathematical Statistics, 27*(2), 251-271.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components.* New York: Wiley.

Singh, A. C., Stukel, D. M., & Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 60*, 377-396.

ten Berge, J. M. F., Krijnen, W. P., Wansbeek, T., & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and Its Applications, 289*(1-3), 311-318.

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T., Jr. (2005). Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging, 20*(3), 493-506.

Tukey, J. W. (1958). Bias and confidence in not quite large samples [abstract]. *Annals of Mathematical Statistics, 29*, 614.

Van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In R. van der Leeden & E. Meijer (Eds.) *Handbook of Multilevel Analysis* (pp. 401-433). New York: Springer.

Wang, J., Carpenter, J. R., & Kepler, M. A. (2006). Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. *Computer Methods and Programs in Biomedicine, 82*, 130-143.

Wang, J. Y., & Fuller, W. A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association, 98*(463), 716-723.

Wolfinger, R. D. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1*(2), 205-230.

Chapter 7. Technical Appendix

7.1. Rate of Estimator Convergence

In comparing estimators, it is useful to consider the rate at which the estimator converges to its target (or, equivalently, the rate at which the estimator's bias goes to zero) as sample size goes to infinity. Rates of convergence can be expressed succinctly using the Landau symbol $O(\cdot)$, which can be defined as follows: An estimator t_n of target T has rate of convergence $O(n^{-q})$ if there exists some constant k such that $|t_n - T| \leq kn^{-q}$ for constant q and sufficiently large *n*. Informally, one might say that t_n converges at least as quickly as kn^{-q}. This notation facilitates comparison of an estimator's rate of convergence with that of a familiar function of *n*, and with rates of convergence of other estimators.

For example, consider the ML variance estimator. Let X be a random variable, and let $x_1, x_2, \ldots, x_n$ be a random sample drawn from the distribution of X. The ML estimator of $\sigma^2 = \mathrm{Var}(X)$ is given by $\sigma^2_{ML} = \frac{1}{n}\sum_i (x_i - \bar{x})^2$. This estimator has expected value $\sigma^2(1 - 1/n)$, so its bias is given by $\sigma^2(1 - 1/n) - \sigma^2 = \sigma^2/n$. As *n* goes to infinity, this bias goes to zero at same rate as kn^{-1}, where $k = \sigma^2$, so the ML estimator has rate of convergence $O(n^{-1})$.

7.2. Kackar-Harville and Prasad-Rao Approximations

Kackar and Harville (1984) showed $\mathrm{MSE}[t(\hat{\theta})] = \mathrm{MSE}[t(\theta)] + E[t(\hat{\theta}) - t(\theta)]^2$. The term on the far right (which is neglected by the naïve approximation) can be approximated as follows. The first-order Taylor expansion of $t(\hat{\theta})$ about θ is given by

$t(\hat{\theta}) \approx t(\theta) + t'(\theta)(\hat{\theta} - \theta)$, where $t'(\theta)$ is the row vector of partial derivatives of $t(\theta)$ with respect to the components of θ. Subtracting $t(\theta)$ from both sides, squaring, and taking the expected value yields $E[t(\hat{\theta}) - t(\theta)]^2 \approx E[t'(\theta)(\hat{\theta} - \theta)]^2$.

Kackar and Harville (1984) proposed approximating the term on the right (and thus the term on the left) by $tr[A(\theta)B(\theta)]$, where $A(\theta)$ is the variance-covariance matrix of $t'(\theta)$, and $B(\theta)$ is an estimate of $E(\hat{\theta} - \theta)^2$. Computational formulas for $A(\theta)$ are given in Kackar and Harville and, for the fixed effect case, in Kenward and Roger (1997). If ML or REML estimation is used, the asymptotic variance-covariance matrix of $\hat{\theta}$ can be used for $B(\theta)$. The Kackar-Harville MSE approximation is obtained by adding their approximation of $E[t(\hat{\theta}) - t(\theta)]^2$ to the naïve approximation:

$$MSE[t(\hat{\theta})] \approx MSE[t(\theta)|\hat{\theta}] + tr[A(\theta)B(\theta)].$$

Harville and Jeske (1992) modified the Kackar-Harville approximation to adjust for bias in $MSE[t(\theta)|\hat{\theta}]$ as an estimator of $MSE[t(\theta)]$. Let $MSE'[t(\theta)] = (\partial/\partial\theta)MSE[t(\theta)]$, the column vector of first partial derivatives of $MSE[t(\theta)]$ with respect to the components of θ. The second-order Taylor expansion of $MSE[t(\theta)|\hat{\theta}]$ about θ is given by

(5) $$MSE[t(\theta)|\hat{\theta}] \approx MSE[t(\theta)] + (\hat{\theta} - \theta)'MSE'[t(\theta)] + \tfrac{1}{2}(\hat{\theta} - \theta)'C(\theta)(\hat{\theta} - \theta),$$

where $C(\theta)$ is the matrix of second partial derivatives of $MSE[t(\theta)]$ with respect to θ—i.e., $C(\theta)$ has ijth element $c_{ij} = \dfrac{\partial^2 MSE[t(\theta)]}{\partial\theta_i\partial\theta_j}$, where θ_i is the ith component of θ.

Taking the expected value of each side of (5) and ignoring bias in $\hat{\theta}$ yields the approximation

$$E\{MSE[t(\theta)|\hat{\theta}]\} \approx MSE[t(\theta)] + (1/2)tr[C(\theta)B(\theta)].$$

The second term on the right approximates the bias in $MSE[t(\theta)|\hat{\theta}]$ as an estimate of $MSE[t(\theta)]$.

For linear covariance structures (i.e., models for which V can be expressed as a linear combination of the components of θ), $C(\theta) = -2A(\theta)$, where as above, $A(\theta)$ is the variance-covariance matrix of $t'(\theta)$. Thus, the bias in $MSE[t(\theta)|\hat{\theta}]\}$ is approximately $-tr[A(\theta)B(\theta)]$. Subtracting this (negative) bias from the Kackar-Harville approximation, Harville and Jeske (1992) obtained the approximation

$$MSE[t(\hat{\theta})] \approx MSE[t(\theta)|\hat{\theta}] + 2tr[A(\theta)B(\theta)].$$

Note that the bias adjustment, $tr[A(\theta)B(\theta)]$, is exactly the approximation of $E[t(\hat{\theta}) - t(\theta)]^2$ derived by Kackar and Harville (1984); thus the factor of two in the rightmost term. Prasad and Rao (1990) developed a similar estimator for three specific mixed models, but unlike Harville and Jeske (1992), they modified the Kackar-Harville approximation by defining $A(\theta)$ differently.

In both the Kackar-Harville and Prasad-Rao approximations, the term $tr[A(\theta)B(\theta)]$ is used in approximating $E[t(\hat{\theta}) - t(\theta)]^2$. Hulting and Harville (1991) noted that for values of the variance ratio close to or equal to zero, this approximation may be inadequate.

7.3. Satterthwaite's Degree of Freedom Approximation

Satterthwaite's (1941, 1946) method is based on the assumption that the variance quantity for which degrees of freedom are to be approximated has a chi-square distribution when suitably scaled. Setting the true variance of the scaled variance quantity equal to what the variance of the scaled quantity would be under an assumed chi-square distribution allows one to solve for the degrees of freedom.

The method can be used for inference in the mixed model as follows (Littell, 2002). Consider the quasi-pivot $[t(\hat{\theta}) - T]/s_t$, where $t(\hat{\theta})$ is the EBLUE and s_t^2 is an MSE approximation. The scaled MSE approximation $vs_t^2/E(s_t^2)$ is assumed to have chi-square distribution with v degrees of freedom. The variance of $vs_t^2/E(s_t^2)$ is $\text{Var}[vs_t^2/E(s_t^2)] = v^2\text{Var}(s_t^2)/[E(s_t^2)]^2$. Under the chi-square assumption, this variance equals twice the degrees of freedom: $v^2\text{Var}(s_t^2)/[E(s_t^2)]^2 = 2v$. It follows that $v = 2[E(s_t^2)]^2/\text{Var}(s_t^2)$. In practice, this quantity is approximated by $2(s_t^2)^2/g'Dg$, where D is the asymptotic variance-covariance matrix of $\hat{\theta}$, and g (known as the *gradient*) is the column vector obtained by differentiating s_t^2 with respect to each component of θ and evaluating the result at $\hat{\theta}$:

$$g = \left.\frac{\partial s_t^2}{\partial \theta}\right|_{\theta=\hat{\theta}}$$

(SAS Institute Inc., 2008).

Chapter 8. Tables

Table 1
Coverage Percentages—Simulation Study 1

Nominal Rate	Interval	ICC = 0.05		ICC = 0.10	
		m = 10	*m* = 15	*m* = 10	*m* = 15
90%	Naïve	89.5	89.8	88.7	89.4
	Kenward-Roger	90.1	90.4	88.75	90.0
	Efron	90.6	90.6	89.6	89.6
	Hall	90.6	90.9	89.7	89.9
	Bootstrap-*t*	90.0	90.4	89.8	90.0
95%	Naïve	94.4	94.6	93.9	94.4
	Kenward-Roger	95.1	95.0	94.55	95.1
	Efron	95.4	95.1	94.9	94.9
	Hall	95.3	95.1	94.7	94.8
	Bootstrap-*t*	95.1	95.1	94.7	95.0
99%	Naïve	98.7	98.8	98.6	98.9
	Kenward-Roger	98.9	99.0	98.8	99.0
	Efron	98.9	99.0	98.7	99.0
	Hall	98.8	98.9	98.8	99.1
	Bootstrap-*t*	98.9	98.9	98.9	99.2

Table 2
Percentages of Misses due to Overestimation—Simulation Study 1

Nominal Rate	Interval	ICC = 0.05		ICC = 0.10	
		m = 10	*m* = 15	*m* = 10	*m* = 15
90%	Naïve	51.2	52.0	51.1	49.3
	Kenward-Roger	51.3	51.5	50.9	49.1
	Efron	51.3	52.3	51.3	49.6
	Hall	51.3	52.4	51.0	49.6
	Bootstrap-*t*	51.3	51.0	51.2	49.6
95%	Naïve	54.7	52.6	51.2	50.8
	Kenward-Roger	55.4	53.2	50.6	49.9
	Efron	54.2	53.3	49.9	49.3
	Hall	53.8	53.5	49.4	47.3
	Bootstrap-*t*	53.0	53.3	51.8	50.0
99%	Naïve	50.8	55.1	45.7	50.5
	Kenward-Roger	51.4	52.4	47.4	50.0
	Efron	56.9	53.8	48.4	48.5
	Hall	56.0	54.5	45.8	50.0
	Bootstrap-*t*	48.6	55.0	47.4	57.1

Table 3
Coverage Percentages—Simulation Study 2

	Nominal Coverage Rate		
Interval	90%	95%	99%
Naïve	90.4	94.9	99.0
Kenward-Roger	91.9	96.1	99.4
Efron	88.5	92.8	96.8
Hall	90.8	95.1	99.2
Bootstrap-*t*	91.0	95.7	99.3

Table 4
Percentages of Misses due to Overestimation—Simulation Study 2

	Nominal Coverage Rate		
Interval	90%	95%	99%
Naïve	50.4	47.5	47.1
Kenward-Roger	49.4	45.4	42.9
Efron	39.8	38.0	44.1
Hall	48.9	45.9	38.1
Bootstrap-*t*	49.0	46.1	40.0

Table 5
Coverage Percentages—Simulation Study 3

	Nominal Coverage Rate					
	90%		95%		99%	
Interval	$m = 10$	$m = 20$	$m = 10$	$m = 20$	$m = 10$	$m = 20$
Naïve ($n - p$)	87.2	88.4	92.2	93.7	97.2	97.9
Naïve (Satterthwaite)	90.2	90.3	95.2	94.8	98.9	98.9
Kenward-Roger	85.2	87.7	91.4	93.3	97.3	98.1
Efron	87.3	88.5	91.9	93.6	97.2	97.9
BC	87.2	88.6	92.2	93.6	97.1	97.9
Hall	87.2	88.6	92.1	93.6	97.1	97.8
Bootstrap-*t*	89.3	90.5	93.8	94.9	98.2	98.9

Table 6

Percentages of Misses due to Overestimation—Simulation Study 3

	Nominal Coverage Rate					
	90%		95%		99%	
Interval	$m = 10$	$m = 20$	$m = 10$	$m = 20$	$m = 10$	$m = 20$
Naïve ($n - p$)	50.8	50.3	48.8	51.7	48.4	51.5
Naïve (Satterthwaite)	50.0	50.6	50.7	52.1	57.1	54.3
Kenward-Roger	51.9	49.9	47.4	50.5	47.7	51.7
Efron	50.5	49.6	48.5	51.4	46.7	47.8
BC	50.4	49.7	48.0	51.9	45.7	47.8
Hall	51.0	50.7	49.2	53.1	47.4	49.3
Bootstrap-*t*	49.6	50.8	46.8	52.7	49.2	56.8