Harvard University
Harvard University Biostatistics Working Paper Series

Year 2004 *Paper* 18

A Functional-Based Distribution Diagnostic for a Linear Model with Correlated Outcomes: Technical Report

E. Andres Houseman* Brent Coull†
Louise Ryan‡

*Harvard School of Public Health, ahousema@hsph.harvard.edu
†Harvard School of Public Health, bcoull@hsph.harvard.edu
‡Harvard School of Public Health and Dana-Farber Cancer Institute, lryan@hsph.harvard.edu

A Functional-Based Distribution Diagnostic for a Linear Model with Correlated Outcomes: Technical Report

E. Andres Houseman, Brent Coull, and Louise Ryan

Abstract

Despite the widespread popularity of linear models for correlated outcomes (e.g. linear mixed modesl and time series models), distribution diagnostic methodology remains relatively underdeveloped in this context. In this paper we present an easy-to-implement approach that lends itself to graphical displays of model fit. Our approach involves multiplying the estimated marginal residual vector by the Cholesky decomposition of the inverse of the estimated marginal variance matrix. Linear functions or the resulting "rotated" residuals are used to construct an empirical cumulative distribution function (ECDF), whose stochastic limit is characterized. We describe a resampling technique that serves as a computationally efficient parametric bootstrap for generating representatives of the stochastic limit of the ECDF. Through functionals, such representatives are used to construct global tests for the hypothesis of normal margional errors. In addition, we demonstrate that the ECDF of the predicted random effects, as described by Lange and Ryan (1989), can be formulated as a special case of our approach. Thus, our method supports both omnibus and directed tests. Our method works well in a variety of circumstances, including models having independent units of sampling (clustered data) and models for which all observations are correlated (e.g., a single time series).

A Functional-Based Distribution Diagnostic for a Linear Model with Correlated Outcomes: Technical Report

E. Andrés Houseman Brent A. Coull Louise M. Ryan

September 12, 2003

Abstract

Despite the widespread popularity of linear models for correlated outcomes (e.g. linear mixed models and time series models), distribution diagnostic methodology remains relatively underdeveloped in this context. In this paper we present an easy-to-implement approach that lends itself to graphical displays of model fit. Our approach involves multiplying the estimated marginal residual vector by the Cholesky decomposition of the inverse of the estimated marginal variance matrix. Linear functions of the resulting "rotated" residuals are used to construct an empirical cumulative distribution function (ECDF), whose stochastic limit is characterized. We describe a resampling technique that serves as an computationally efficient parametric bootstrap for generating representatives of the stochastic limit of the ECDF. Through functionals, such representatives are used to construct global tests for the hypothesis of normal marginal errors. In addition, we demonstrate that the ECDF of the predicted random effects, as described by Lange and Ryan (1989), can be formulated as a special case of our approach. Thus, our method supports both omnibus and directed tests. Our method works well in a variety of circumstances, including models having independent units of sampling (clustered data) and models for which all observations are correlated (e.g., a single time series).

1 Introduction

Correlated data typically arise in public health, biomedical and environmental applications, requiring techniques such as linear mixed models and time series regression. Despite the widespread popularity of these methods, diagnostic methodology for addressing goodness-of-fit in this context remains relatively underdeveloped. Verbeke & Molenberghs (2000) noted that the choice of diagnostic method for linear mixed models is not obvious (pages 151–152), and Agresti (2002) noted that research is needed on model checking and diagnostics for mixed models in general (page 526). Recently, Lin et al. (2002) have proposed graphical techniques for assessing the adequacy of the deterministic portion of a generalized

linear mixed model, but their methodology does not address the random component of model fit. Although the error distribution plays a secondary role in some applications, it remains an important component of the model when there is an interest in prediction or in optimizing efficiency.

Limited methodology is available for assessing the distribution of the error term. Existing work includes Lange & Ryan (1989), who developed methods using estimated random effects from growth curve models, Fraccaro et al. (2000), who discussed residual diagnostic plots in time series regression, and Louis (1988), who discussed an approach similar to the one we describe. Unlike our paper, the latter two papers present no asymptotic theory. Several authors have recently proposed goodness-of-fit tests that apply in the mixed model setting (Hodges, 1998; Jiang, 2001), but the approaches are complex and do not lend themselves well to graphical displays.

For models with subject-specific random effects, Pinheiro & Bates (2000) advocated the use of the standardized residual formed using predictions of subject-specific means and an estimate of residual error (page 239). However, comparing fitted and observed values can be misleading, as these comparisons reflect intended shrinkage of estimates towards overall means. Coull et al. (2001) and Longford (2001) addressed model fit by simulating data sets from the maximum likelihood fit of the model and checking to see whether the observed data set was extreme relative to the reference set. This idea has also been used in spatial statistics (Ripley, 2001) and is similar to the *posterior predictive checking* strategy often used in Bayesian analyses (Gelman et al., 1995).

Houseman et al. (2003a) presented an approach that involves multiplying the estimated marginal residual vector by the Cholesky decomposition of the inverse of the estimated marginal variance matrix. This approach has the advantage of enabling graphical depictions of goodness-of-fit along the lines of a quantile-quantile (Q-Q) plot. In this article, we establish the global asymptotic properties of the ECDF of such residuals and present a graphical goodness-of-fit procedure using a computationally efficient approximation to a parametric bootstrap. Since our procedure gives a computationally efficient means of obtaining the stochastic distribution of the ECDF, it is possible to use it to obtain the distribution of any test that can be written as a continuous functional of our residuals.

Our methods apply to a larger class of models than those studied by Lange & Ryan (1989); they apply not only to linear mixed models, but also linear models for which there are no independent units of sampling (for example, time series and spatial models). In addition, the approach described in Lange & Ryan (1989) can be formulated as a special case of our proposed methodology; thus we extend their work to address global asymptotics. Consequently our proposed method supports both omnibus tests and tests directed towards assessing a particular component of the error term.

We have focused on distribution diagnostics, but remark that the cumulative residual approach of Lin et al. (2002) can be used to check the mean structure, and graphical displays such as a *draftsman's display* (Dawson et al., 1997) and *PRISM* (Zimmerman, 2000) are useful for diagnosing the form of the covariance.

Our paper is organized as follows. Section 2 describes the background of our

proposed methodology, introducing the problem and the heuristics of our solution. Section 3 presents the technical results needed to justify our proposed methods. Section 4 reports simulation results. Section 5 provides two examples of our methods, and Section 6 concludes with general discussion.

2 Rotated Residuals of a Linear Model

In this section we formally describe the context of our proposed methodology and define the concept of a *rotated residual*, which is central to the method. Throughout our exposition, we assume that a linear model has been specified as follows:

$$y = X\beta_0 + \varepsilon, \tag{1}$$

where $\varepsilon \sim N_n(0, V(\gamma_0))$. Here $y \in \mathsf{R}^n$ consists of n outcomes, possibly correlated. The marginal mean of y is determined by an $n \times p$ design matrix X of covariates and a p-dimensional parameter vector $\beta_0 \in \mathsf{R}^p$. The marginal covariance of y is determined by the $n \times n$ matrix $V(\gamma_0)$ having known form depending on an unknown q-dimensional parameter $\gamma_0 \in \mathsf{R}^q$.

For example, a special case of (1) is a mixed effects model, where the marginal variance $V(\gamma_0)$ is block-diagonal, each diagonal block having the form $\sigma^2(\gamma_0)I_h + Z_h\Delta(\gamma_0)Z_h^T$, where Z_h is a matrix of covariates for cluster h, I_h is the identity matrix of corresponding dimension, $\sigma^2(\gamma_0)$ is the error variance, and $\Delta(\gamma_0)$ is the variance of the random effects. In the general random effects model, $\Delta(\gamma_0)$ is a matrix having known structure depending upon the parameter vector γ_0. Another example is time series regression, in which the correlation of outcomes observed

at times s and t depends upon a function of s and t, e.g. the first-order autoregressive process $\rho^{|s-t|}$. Model fitting for (1) is easily achieved using least squares, maximum likelihood (ML), or restricted maximum likelihood (REML). Our focus here is on the problem of goodness-of-fit and detection of outliers. Specifically, we are interested in whether the ECDF of our proposed rotated residuals matches the cumulative distribution function (CDF) of a normal distribution. To this end, we define the vector of residuals $z(\theta)$ as a function of possible parameter values $\theta = (\beta^T, \gamma^T)^T$. Although the residuals $y - X\beta$ can be standardized using the fitted diagonal of $V(\gamma)$, the resulting residuals are still potentially highly correlated, even in large samples. Therefore, we propose an n-dimensional linear transformation (which we informally term a rotation) to yield new ones that are asymptotically uncorrelated. Let $L(\gamma)$ be the Cholesky decomposition of $V(\gamma)^{-1}$, so that $V(\gamma)^{-1} = L(\gamma)L(\gamma)^T$, and define

$$z(\theta) = L(\gamma)^T(y - X\beta). \tag{2}$$

Then $z(\theta_0) \sim N_n(0, I_n)$, where $\theta_0 = (\beta_0^T, \gamma_0^T)^T$ and I_n is the $n \times n$ identity matrix. If θ_0 were known, a goodness-of-fit procedure for the appropriateness of model (1) could be established by comparing $\Phi(x)$, the standard normal CDF evaluated at x, to the ECDF evaluated at x, for various values of x. The ECDF can be expressed as $F_n(x, \theta_0)$, where

$$F_n(x, \theta) = \frac{1}{n} \sum_{i=1}^{n} I\left(x - \pi_i z(\theta)\right), \tag{3}$$

$I(u) = 1$ if $u \leq 0$, $I(u) = 0$ if $u > 0$, and π_i is the canonical projection onto the ith coordinate. Under the null hypothesis that $z_i(\theta_0) = \pi_i z(\theta_0) \sim N_n(0, I_n)$ [equivalent to (1)], a pointwise confidence interval for $F_n(x, \theta_0)$ is easily constructed

by noting that the variance of (3), evaluated at θ_0, is simply $\Phi(x)(1-\Phi(x))$. A global confidence band can be constructed by inverting the Kolmogorov-Smirnov test for the equality of two distributions.

In any practical situation, θ_0 must be estimated from the data. This complicates the construction of confidence intervals in that the variance of (3) tends to shrink when the parameters are replaced by estimates based on the data. Lange & Ryan (1989) described this phenomenon extensively in the context of constructing a Q-Q plot of random effects predicted in growth curve models. These authors derived the adjustment factor needed for the variance of the ECDF for random effects in growth curve models, while Houseman et al. (2003a) provided a more general approach involving marginal residuals of a general linear model for correlated data. Houseman et al. (2003a) demonstrated that in many practical circumstances, (3) has an asymptotically normal distribution when x is fixed. The goal of the present paper is to extend this result by proving weak convergence of the stochastic process indexed by x in an arbitrary compact interval, and to construct a graphical method for examining the global behavior of such stochastic processes. Thus we present a global test of distributional assumptions. Additionally, by generalizing the projection matrix π_i in (3), we address the global behavior of the ECDF of random effects, thereby extending Lange & Ryan (1989).

2.1 Omnibus Test

In this section we describe the heuristics of our method, deferring technical details until Section 3. Let $\widehat{\beta}$ and $\widehat{\gamma}$ be consistent estimators for β_0 and γ_0 respectively,

and denote the joint estimator as $\widehat{\theta} = (\widehat{\beta}, \widehat{\gamma})$, consistent for θ_0. As shown by Houseman et al. (2003a), $F_n(x, \widehat{\theta})$ is asymptotically normal for fixed x when model (1) is correct, subject to regularity conditions that are often met in practice. In particular, $n^{-\frac{1}{2}}\left(F_n(x, \widehat{\theta}) - \Phi(x)\right) \rightsquigarrow N(0, \tau^2)$, where $\rightsquigarrow$ denotes weak convergence,$\tau^2 = \Phi(x)(1 - \Phi(x)) - \delta_0^T W \delta_0$, $W = Var[\widehat{\theta}]$, and δ_0 is a vector that can be consistently estimated by $\delta_n(x, \widehat{\theta}, \widehat{\theta})$, where

$$\delta_n(x, \theta, \theta_0) = \frac{\partial \tilde{\mu}_n}{\partial \theta^T}(x, \theta, \theta_0)$$

and

$$\tilde{\mu}_n(x, \theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{x - \pi_i L(\gamma)^T X(\beta_0 - \beta)}{\sqrt{\pi_i L(\gamma)^T V(\gamma_0) L(\gamma) \pi_i^T}}\right). \tag{4}$$

These formulas are sufficient for constructing pointwise confidence bands, but they do not provide enough information to construct global confidence bands. To this end, we demonstrate weak convergence of the stochastic process implied by letting x vary and characterize its asymptotic behavior. In practical settings, direct application of the result is complicated by the difficulty of tabulating tail probabilities for quantities of interest, such as the supremum of the difference between observed and expected values of the ECDF. Therefore, we construct a global hypothesis test using a resampling technique motivated by Lin et al. (2002).

Consider the score function for estimating θ_0, which may be expressed as a function $U_n(\theta, z)$ of the rotated residuals (2) simply by substituting $L(\gamma)^{-T} z + L(\gamma)^T X\beta$ for y in the score functions expressed in terms of y. For example, if $U_n = (U_n^{(\beta)}, U_n^{(\gamma_1)}, ..., U_n^{(\gamma_q)})$ expresses the score function as a vector of components for estimating β and the q variance component parameters, then for ML estimators,

the component for estimating β is

$$U_n^{(\beta)} = X^T V(\gamma)^{-1}(y - X\beta) = X^T L(\gamma) z(\theta),$$

while the component for estimating γ_j is

$$\begin{aligned} U_n^{(\gamma_j)} &= \frac{1}{2}(y - X\beta)^T V(\gamma)^{-1} \frac{\partial V}{\partial \gamma_j} V(\gamma)^{-1}(y - X\beta) - \mathrm{tr}\left(V(\gamma)^{-1} \frac{\partial V}{\partial \gamma_j} \right) \\ &= \frac{1}{2} z(\theta)^T L(\gamma)^T \frac{\partial V}{\partial \gamma_j} L(\gamma) z(\theta) - \mathrm{tr}\left(V(\gamma)^{-1} \frac{\partial V}{\partial \gamma_j} \right). \end{aligned}$$

Similar expressions exist for REML estimators. When the score is expressed in this manner, it may be used to simulate realizations from the stochastic limit of $F_n(x, \widehat{\theta}_n) - \Phi(x)$. To see this, let $z^* \sim N(0, I_n)$. Since the rotated residuals are independent when evaluated at θ_0, $U_n(\theta_0, z^*)$ has the same distribution as $U_n(\theta_0, z(\theta_0))$, and the ECDF $F_n^*(x)$ of z^* has the same stochastic distribution as $F_n(x, \theta_0)$. Through a Taylor's expansion, letting $J_n(\theta)$ denote the expected information matrix corresponding to $U_n(\theta_0, z(\theta_0))$, we demonstrate in the next section that

$$\hat{F}_n^*(x) - \Phi(x) \equiv F_n(x, \widehat{\theta}_n) - \Phi(x) + \delta_n(x, \widehat{\theta}_n, \widehat{\theta}_n)^T J_n(\widehat{\theta}_n)^{-1} U_n(\widehat{\theta}_n, z^*) \qquad (5)$$

has (conditional on y) approximately the same stochastic distribution as the unconditional distribution of $F_n(x, \widehat{\theta}_n) - \Phi(x)$. Thus, the distribution of a functional applied to $F_n(x, \widehat{\theta}_n)$ may be obtained, via simulation, by tabulating the distribution of the same functional applied to a large number of simulations $\widehat{F}_n^*(x)$. For example, by noting that

$$\sup_{x \in \mathcal{X}} |F_n(x, \widehat{\theta}_n) - \Phi(x)| \approx \sup_{x \in \mathcal{X}} |\widehat{F}_n^*(x) - \Phi(x)|, \qquad (6)$$

we obtain a test resembling the Kolmogorov-Smirnov test for the equivalence of two distributions, but correctly adjusted for estimation of θ. P-values can be

estimated as the proportion of simulations for which the statistic represented by the right hand side of (6) is larger than the corresponding observed value. A test similar to the Cramér-Von Mises test can be obtained by considering the functional

$$\mathcal{J}(F) = \int_{\mathcal{X}} |F(x) - \Phi(x)|^2 d\Phi(x). \tag{7}$$

Lin et al. (2002) considered a process similar to (5) for checking the predictive portion of a model, but their formulation uses the residual in a quite different manner. We note that (5) is, in essence, an approximation to a parametric bootstrap. However, its advantage is that it can be computed quickly. The exact parametric bootstrap would require the solution of a multi-dimensional nonlinear equation, which is computationally expensive, especially when the complexity of computing $V(\gamma)$ is of order $O(n^2)$. Expression (5) requires only the computation of δ_n and J_n, which can be obtained once and reused over multiple simulations. When $\partial V/\partial\gamma_k$ has a closed form expression for each element γ_k of γ, it is easy to show that $\delta_n(x, \theta, \theta_0)$ also has a closed form. The form of δ_n can be obtained by applying the chain rule to the derivative of (4) and noting that

$$\mathrm{tr}\left(\frac{\partial}{\partial\gamma_k} L(\gamma)^T V(\gamma_0) L(\gamma)\Big|_{\gamma=\gamma_0}\right) = -\mathrm{tr}\left(L(\gamma_0)^T \left.\frac{\partial V}{\partial\gamma_k}\right|_{\gamma=\gamma_0} L(\gamma_0)\right). \tag{8}$$

A typical example of computation times involving 1000 resampled values might be 10 minutes for our procedure and several hours for the corresponding parametric bootstrap.

2.2 Directed Tests

The preceding development corresponds to an omnibus goodness-of-fit test for (1). In some situations it may be of more interest (and more powerful) to test specific distributional assumptions. For instance, when one is interested in prediction of random effects in a mixed model, it is of interest to evaluate the validity of the normality assumption for the random effects. We can extend our theory to incorporate such directed tests by generalizing the projection matrix π_i used in (3). In particular, let $\mathcal{P}_n = (P_{n1}, ..., P_{nN_n})$ be an array of $1 \times n$ matrices, orthogonal in the sense that $P_{ni}P_{nj}^T = 0$ when $i \neq j$ and $P_{ni}P_{ni}^T = 1$. For such arrays, $\{P_{ni}z(\theta_0)\}_{i=1,...,N_n}$ are independent and identically distributed as standard normal variables. We study the ECDF of residuals of the form

$$F_n(x, \theta; \mathcal{P}_n) = \frac{1}{N_n} \sum_{i=1}^{N_n} I\left(x - P_{ni}z(\theta)\right), \tag{9}$$

for which $\mathcal{P}_n = (\pi_1, ..., \pi_n)$ is clearly a special case. In addition, when (1) expresses a random effects model in which V is composed of blocks V_h having the form $V_h = Z_h \Delta Z_h^T + \sigma^2 R_h$ (where $h = 1, ..., N$, Δ and R_h are positive definite symmetric matrices, Δ of size $d \times d$ and R_h of size $k_h \times k_h$, and Z_h is a $k_h \times d$ matrix), it is also possible to express the BLUPs of the random effects as a special case. The standardized BLUP of the jth random effect component for cluster h is obtained as $(C_h^{(j)} C_h^{(j)T})^{-1/2} C_h^{(j)} z(\theta)$, where

$$C_h^{(j)} = \tilde{\pi}_j \Delta Z_h^T V_h^{-1} L_h^{-T} \pi_h^* = \tilde{\pi}_j \Delta Z_h^T L_h \pi_h^*,$$

$V_h = L_h L_h^T$, and $\tilde{\pi}_j : \mathsf{R}^d \to \mathsf{R}$ and $\pi_h^* : \mathsf{R}^n \to \mathsf{R}^{k_h}$ are canonical projections. The array containing the matrices

$$P_h^{(j)} = (C_h^{(j)} C_h^{(j)T})^{-1/2} C_h^{(j)} \tag{10}$$

is orthogonal in the sense just described. Thus, the ECDF of the standardized BLUPs discussed by Lange & Ryan (1989) is a special case of (9). Using this framework, both the pointwise confidence intervals described in Lange & Ryan (1989) for random effects and the confidence intervals described in Houseman et al. (2003a) for rotated marginal residuals can be extended to global confidence bands under the same theory.

For the more complicated choices of $\mathcal{P}_n$, equations (4) and (5) hold if each instance of π_i is replaced by P_{ni}. In particular, equation (4) becomes

$$\tilde{\mu}_n(x, \theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \Phi \left(\frac{x - P_{ni} L(\gamma)^T X(\beta_0 - \beta)}{\sqrt{P_{ni} L(\gamma)^T V(\gamma_0) L(\gamma) P_{ni}^T}} \right). \tag{11}$$

The projections described by (10) depend upon the unknown parameter γ and, in practice, $P_h^{(j)}$ must be estimated from the data. Therefore, the θ-derivative of (11) should consider the dependence of P_{ni} on γ. However, when $P_{ni}(\gamma)$ is constrained to have unit length, the dependence of P_{ni} on θ has no first-order effect. That is,

$$\frac{\partial}{\partial \theta^T} \Phi \left(\frac{x - P_{ni}(\gamma) L(\gamma)^T X(\beta_0 - \beta)}{\sqrt{P_{ni}(\gamma) L(\gamma)^T V(\gamma_0) L(\gamma) P_{ni}(\gamma)^T}} \right) \Bigg|_{\theta=\theta_0} = \frac{\partial}{\partial \theta^T} \Phi \left(\frac{x - P_{ni} L(\gamma)^T X(\beta_0 - \beta)}{\sqrt{P_{ni} L(\gamma)^T V(\gamma_0) L(\gamma) P_{ni}^T}} \right) \Bigg|_{\theta=\theta_0}, \tag{12}$$

where in the right hand side of (12), P_{ni} is the constant $P_{ni}(\gamma_0)$. This is easily shown by parameterizing $P_{ni}(\gamma)$ as a vector divided by its norm, as in (10), and applying the chain rule to (12).

Unfortunately, when non-constant projections such as (10) are used, second-order terms may affect the performance of our proposed approximation (6), even in relatively large samples. We discuss this in Section 4. Consequently, we propose a simple modification of (6) that, in our simulations, performs better in smaller samples. Instead of using $F_n^*(x) = N^{-1}\sum_{i=1}^{N} I(x - P_{ni}^{(j)} z^*)$, which treats $P_{ni}^{(j)}$ as if it were known, we propose replacing it with $F_n^*(x) = N^{-1}\sum_{i=1}^{N} I(x - P_{ni}^{*(j)} z^*)$, where $P_h^{*(j)} = (C_h^{*(j)} C_h^{*(j)T})^{-1/2} C_h^{*(j)}$ and

$$C_h^{*(j)} = C_h^{(j)} + \frac{\partial C_h^{(j)}}{\partial \theta^T} J_n(\widehat{\theta}_n)^{-1} U_n(\widehat{\theta}_n, z^*).$$

This replacement simulates the effect on (9) of estimating $P_{ni}^{(j)}$. Note that, in principle, we could use $P_h^{*(j)} = P_{ni}^{(j)}(\gamma^*)$, where γ^* is obtained from $\theta^* = \widehat{\theta} + J_n(\widehat{\theta}_n)^{-1} U_n(\widehat{\theta}_n, z^*)$. However, this requires additional computation time; even worse, with bounded parameterizations of γ, first-order approximation may occasionally produce a value of γ^* that lies outside the boundary of the parameter space, causing the construction of $P_h^{*(j)}$ to fail. Note also that by applying the linear approximation to $C_h^{(j)}$ instead of $P_{ni}^{(j)}$, we guarantee that $P_h^{*(j)}$ always has unit length. Our simulations demonstrate that our proposed modification produces good results.

3 Technical Results

In this section we present the technical results needed to justify the proposed methods, together with their formal proofs. We establish general notation as follows, restating some of the definitions given in the previous section for the sake of completeness. We denote the absolute value and vector norm identically

as $|\cdot|$, and reserve $\|\cdot\|$ for the $\mathcal{L}^2$ matrix norm; that is, if A is a $k \times d$ matrix, then $\|A\| = \min\{a \in \mathsf{R} : |Ax| \leq a|x| \ \ \forall x \in \mathsf{R}^d\}$. The norm of a row vector is sometimes equated with the $\mathcal{L}^2$ norm of its associated linear transformation. We denote the column-vector gradient of a possibly vector valued function $g(\zeta)$ as $\frac{\partial g}{\partial \zeta}$. When its transpose is needed, we denote it as $\frac{\partial^T g}{\partial \zeta^T}$. The derivative of a matrix U with respect to a scalar ζ is denoted $d_\zeta U$. Often, we require boundedness of $\|d_{\zeta_j} U\|$ for all elements of some multidimensional parameter vector $\zeta = (\zeta_1, ...\zeta_k)^T$. In this case, we state this condition informally as a boundedness condition on $\|d_\zeta U\|$. If D is a metric space and $f : \mathsf{D} \to \mathsf{R}$, then $|f|_\infty$ denotes the uniform norm, equal to $\sup_{x \in \mathsf{D}} |f(x)|$, and $\mathcal{L}^\infty(\mathsf{D})$ is the space of all $f : \mathsf{D} \to \mathsf{R}$ with bounded uniform norm. By *stochastic process* we mean a tight stochastic process as defined in van der Vaart (1998), and by an *asymptotically tight* sequence of processes we mean a sequence that satisfies condition (ii) of Theorem 18.14 in van der Vaart (1998) [e.g., see (16) below]. We denote weak convergence, whether for a distribution, random variable, or stochastic process, with the symbol $\rightsquigarrow$. We donote the minimum of s and t as $s \wedge t$ and the maximum as $s \vee t$. Finally, we denote the normal CDF as $\Phi(u)$ and the normal density as $\phi(u)$.

Let $\theta_0 = (\beta_0^T, \gamma_0^T)^T \in \Theta \subset \mathsf{R}^p \times \mathsf{R}^q$, where Θ is an open subset. Assume $V_n(\gamma)$ is a (nonsingular) positive definite symmetric matrix of order n, continuously differentiable in a neighborhood of γ_0, and that $L(\gamma)L(\gamma)^T = V(\gamma)^{-1}$. Note that $L(\gamma)$, a Cholesky decomposition, is also positive definite and continuously differentiable in γ. Let

$$y_n \sim N_n(X_n\beta_0, V_n(\gamma_0)),$$

where X_n is bounded. Fix an array of matrices $\mathcal{P}_n = \{P_1, ..., P_{N_n}\}$, where each P_i is a $1 \times n$ matrix and each row in the array is orthogonal in the sense that $P_{ni}P_{nj}^T = 0$ if $i \neq j$ and $P_{ni}P_{ni}^T = 1$. When the context is clear, the subscript n will be omitted from y_n, X_n, V_n, P_{ni} and $\mathcal{P}_n$.

Let $z(\theta) = L(\gamma)^T(y - X\beta)$ and $z_i(\theta) = P_i z(\theta) = P_i L(\gamma)^T(y - X\beta)$. Let $\mathcal{X}$ be a compact subinterval of R and define the empirical process $F_n : \mathcal{X} \times \Theta \to [0, 1]$ as follows:

$$F_n(x, \theta) = N_n^{-1} \sum_{i=1}^{N_n} I\left(x - z_i(\theta)\right). \tag{13}$$

For any finite subset $\{x_1, ..., x_k\} \subset \mathcal{X}$, denote as $F_n(x_1, ..., x_k; \theta)$ the k-dimensional vector $(F_n(x_1, \theta), ..., F_n(x_k, \theta))^T$, denote as $\varphi(x_1, ..., x_k)$ the expectation of $F_n(x_1, ..., x_k; \theta_0)$, equal to $(\Phi(x_1), ...\Phi(x_k))^T$, and denote as $W_F(x_1, ..., x_k)$ the variance of $F_n(x_1, ..., x_k; \theta_0)$. In addition, define

$$\mu_n(x, \theta) = N_n^{-1} \sum_{i=1}^{N_n} \Phi\left(x_i(x, \theta)\right),$$

where $I(w) = 1$ if $w \geq 0$ and $I(w) = 0$ if $w < 0$,

$$x_i(x, \theta) = \frac{x - m_i(\theta)}{s_i(\theta)},$$

$$m_i(\theta) = P_i L(\gamma)^T X(\beta_0 - \beta),$$

and

$$s_i(\theta)^2 = P_i L(\gamma)^T V(\gamma_0) L(\gamma) P_i^T.$$

Note that $z_i(\theta) \sim N(m_i(\theta), s_i(\theta)^2)$, so

$$z_i^*(\theta) = \frac{z_i(\theta) - m_i(\theta)}{s_i(\theta)} \sim N(0, 1)$$

and

$$E\left[I\left(x - z_i(\theta)\right)\right] = E\left[I\left(x_i(x, \theta) - z_i^*(\theta)\right)\right] = \Phi\left(x_i(x, \theta)\right).$$

Let $r_i(\gamma)$ be the correlation between $z_i(\theta)$ and $z_i(\theta_0)$. This is easily shown to be

$$r_i(\gamma) = Corr\left[z_i(\theta), z_i(\theta_0)\right] = \frac{P_i L(\theta)^T V(\theta_0) L(\theta_0) P_i^T}{s_i(\theta)}. \tag{14}$$

Finally, define

$$a_i(x;\theta;\theta') = I\left(x - z_i(\theta)\right) - I\left(x - z_i(\theta')\right). \tag{15}$$

Note that $E[a_i(x;\theta;\theta')] = \Phi\left(x_i(x,\theta)\right) - \Phi\left(x_i(x,\theta')\right)$, that $a_i(x;\theta;\theta') = 1$ when $z_i(\theta) \leq x < z_i(\theta')$, that $a_i(x;\theta;\theta') = -1$ when $z_i(\theta') \leq x < z_i(\theta)$, and that $a_i(x;\theta;\theta') = 0$ otherwise.

In order to establish desired results, it is necessary to consider the combined limiting behavior both of the parameter estimator and of the residuals. As a consequence, it is necessary to distinguish a fixed θ close to θ_0 apart from an estimator $\widehat{\theta}_n$ which may yield θ as a value. Therefore, quantities that converge in $O(|\theta - \theta_0|)$ or $O(|\theta - \theta_0|^2)$ must be understood to converge uniformly with respect to n, so that convergence is preserved when we replace θ by $\widehat{\theta}_n$. In general, when we write $f_n(\theta) = O(g(\theta))$, we mean that there exists an $M > 0$, independent of n, such that $|f_n(\theta)| \leq Mg(\theta)$ for all n. This notation extends standard O-notation [e.g., described in Serfling (1980)] to address our uniform convergence requirements.

Theorem 1 establishes our main result, namely that the ECDF (13) converges weakly to a Gaussian process. To prove it, we require the following Lemma:

Lemma 1 (Asymptotic Tightness). Let $\{Y_n\}_{n=1,2,\ldots}$ be an asymptotically tight sequence of stochastic processes, each having sample paths in $\mathcal{L}^\infty[\mathcal{X}]$. Additionally, let $\{U_n\}_{n=1,2,\ldots}$ be a sequence of random variables on R^d converging in law

to a random variable U, and let $b : \mathcal{X} \to \mathsf{R}^d$ be a uniformly continuous function. Then the sequence of processes $\{Z_n\}_{n=1,2,\ldots}$ defined by $Z_n(x) = Y_n(x) + b(x)^T U_n$ is asymptotically tight.

Proof. Fix $\epsilon > 0$ and $\eta > 0$. By the continuous mapping theorem, $|U_n| \rightsquigarrow |U|$; consequently, if q is a continuity point of the distribution of $|U|$ and $P\{|U| > q\} \leq \eta/2$, then there is an N such that for all $n > N$, $P\{|U_n| > q\} \leq \eta/2$. By the uniform continuity of b, there is also a $\delta > 0$ such that if $|s - t| < \delta$ then $|b(s) - b(t)| < \epsilon/2q$. Thus, for any $|s - t| < \delta$,

$$\begin{aligned} P\{|b(s)^T U_n - b(t)^T U_n| > \epsilon/2\} &\leq P\{|b(s) - b(t)||U_n| > \epsilon/2\} \\ &= P\{|U_n| > |b(s) - b(t)|^{-1}\epsilon/2\} \\ &\leq P\{|U_n| > q\} \leq \eta/2. \end{aligned}$$

By Theorem 18.14 in van der Vaart (1998) which characterizes weak convergence in $\mathcal{L}^\infty[\mathcal{X}]$ in terms of finite-dimensional distributions and tightness, it is possible to choose $N^* > N$ large enough and $\delta^* < \delta$ small enough so that for $n > N^*$,

$$P\left\{\sup_{|s-t|<\delta^*} |Y_n(s) - Y_n(t)| > \epsilon/2\right\} < \eta/2.$$

Since

$$\sup_{|s-t|<\delta^*} |Z_n(s) - Z_n(t)| \leq \sup_{|s-t|<\delta^*} |Y_n(s) - Y_n(t)| + \sup_{|s-t|<\delta^*} |b(s)^T U_n - b(t)^T U_n|,$$

we have

$$\begin{aligned} P\left\{\sup_{|s-t|<\delta^*} |Z_n(s) - Z_n(t)| > \epsilon\right\} &\leq P\left\{\sup_{|s-t|<\delta^*} |Y_n(s) - Y_n(t)| > \epsilon/2\right\} \\ &\quad + P\left\{\sup_{|s-t|<\delta^*} |b(s)^T U_n - b(t)^T U_n| > \epsilon/2\right\} \\ &< \eta/2 + \eta/2 = \eta \end{aligned}$$

for any $n > N^*$. This shows that

$$\limsup_{n\to\infty} P\left\{\sup_{|s-t|<\delta^*} |Z_n(s) - Z_n(t)| > \epsilon\right\} < \eta. \tag{16}$$

□

Theorem 1 (Stochastic Convergence). Suppose $\widehat{\theta}_n$ is a consistent and asymptotically efficient estimator of θ_0 and that for every finite subset $\{x_1, ..., x_k\} \subset \mathcal{X}$,

$$n^{\frac{1}{2}} \begin{bmatrix} \widehat{\theta}_n - \theta_0 \\ F_n(x_1, ..., x_k, \theta_0) - \varphi(x_1, ..., x_k) \end{bmatrix} \rightsquigarrow N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} W_\theta & W_{\theta F} \\ W_{\theta F}^T & W_F(x_1, ..., x_k) \end{bmatrix}\right). \tag{17}$$

Suppose also that

$$R_n \equiv n^{\frac{1}{2}} \sup_{x\in\mathcal{X}} \left|F_n(x, \widehat{\theta}_n) - \mu_n(x, \widehat{\theta}_n) - F_n(x, \theta_0) + \Phi(x)\right| \xrightarrow{\mathcal{P}} 0. \tag{18}$$

In addition, assume there exists a continuously differentiable function $\mu^* : \mathcal{X} \times \Theta \to \mathsf{R}$, also continuously twice-differentiable in the second argument, such that for all n,

$$\mu_n(x, \theta) = \mu^*(x, \theta) + \epsilon_n(x)^T(\theta - \theta_0) + O(|\theta - \theta_0|^2), \tag{19}$$

where $\epsilon_n(x)$ is constant with respect to θ, $\sup_{x\in\mathcal{X}} \epsilon_n(x) \to 0$ as $n \to \infty$, and the boundedness of $O(|\theta - \theta_0|^2)$ is uniform in x and n. Then

$$n^{\frac{1}{2}}[F_n(\widehat{\theta}_n) - \Phi(x)] \rightsquigarrow \mathcal{G}(\tau^2(s, t)), \tag{20}$$

where $\mathcal{G}(\tau^2)$ is a Gaussian process having autocovariance function

$$\tau^2(s, t) = \Phi(s \wedge t)[1 - \Phi(s \vee t)] - \delta(s, \theta_0)^T W_\theta \delta(t, \theta_0), \tag{21}$$

and

$$\delta(x, \theta) = \frac{\partial^T \mu^*}{\partial\theta}(x, \theta). \tag{22}$$

Proof. Since

$$\mu^*(x,\theta_0) = \mu_n(x,\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\Phi(x) = \Phi(x),$$

we have

$$\begin{aligned}
\mu_n(x,\widehat{\theta}_n) - \Phi(x) &= \mu_n(x,\widehat{\theta}_n) - \mu^*(x,\widehat{\theta}_n) + \mu^*(x,\widehat{\theta}_n) - \mu^*(x,\theta_0) + \mu^*(x,\theta_0) - \Phi(x) \\
&= \varepsilon_n(x)^T(\widehat{\theta}_n - \theta_0) + O(|\widehat{\theta}_n - \theta_0|^2) + \mu^*(x,\widehat{\theta}_n) - \mu^*(x,\theta_0) + 0 \\
&= \mu^*(x,\widehat{\theta}_n) - \mu^*(x,\theta_0) + \varepsilon_n(x)^T(\widehat{\theta}_n - \theta_0) + O(|\widehat{\theta}_n - \theta_0|^2) \\
&= \delta(x,\theta_0)^T(\widehat{\theta}_n - \theta_0) + \varepsilon_n(x)^T(\widehat{\theta}_n - \theta_0) + O(|\widehat{\theta}_n - \theta_0|^2),
\end{aligned}$$

where the last equality follows by Taylor expansion. By Slutzky's Theorem, both $n^{\frac{1}{2}}|\widehat{\theta}_n - \theta_0|^2$ and $n^{\frac{1}{2}}\varepsilon_n(x)^T(\widehat{\theta}_n - \theta_0)$ converge in probability to zero, the latter as a stochastic process. Consequently,

$$\begin{aligned}
n^{\frac{1}{2}}[\mu_n(x,\widehat{\theta}_n) - \Phi(x)] &= n^{\frac{1}{2}}\delta(x,\theta_0)^T(\widehat{\theta}_n - \theta_0) + n^{\frac{1}{2}}\varepsilon_n(x)^T(\widehat{\theta}_n - \theta_0) + O(n^{\frac{1}{2}}|\widehat{\theta}_n - \theta_0|^2) \\
&= n^{\frac{1}{2}}\delta(x,\theta_0)^T(\widehat{\theta}_n - \theta_0) + o_p(1).
\end{aligned}$$

Now,

$$\begin{aligned}
n^{\frac{1}{2}}[F_n(x,\widehat{\theta}_n) - \Phi(x)] &= n^{\frac{1}{2}}[F_n(x,\widehat{\theta}_n) - \mu_n(x,\widehat{\theta}_n) - F_n(x,\theta_0) + \Phi(x) + \\
&\qquad \mu_n(x,\widehat{\theta}_n) + F_n(x,\theta_0) - \Phi(x) - \Phi(x)] \\
&= R_n + n^{\frac{1}{2}}[\mu_n(x,\widehat{\theta}_n) - \Phi(x)] + n^{\frac{1}{2}}[F_n(x,\theta_0) - \Phi(x)] \\
&= n^{\frac{1}{2}}[F_n(x,\theta_0) - \Phi(x)] + n^{\frac{1}{2}}\delta(x,\theta_0)^T(\widehat{\theta}_n - \theta_0) + o_p(1).
\end{aligned}$$

This result shows that for arbitrary $\{x_1,...,x_k\} \subset \mathcal{X}$,

$$n^{\frac{1}{2}}[F_n(x_1,...,x_k,\widehat{\theta}_n) - \varphi(x_1,...,x_k)]^T = n^{\frac{1}{2}}[F_n(x_1,...,x_k,\theta_0) - \Phi(x)] + n^{\frac{1}{2}}D(x_1,...,x_k,\theta_0)^T(\widehat{\theta}_n - \theta_0) + o_p(1) \tag{23}$$

where $D(x_1, ..., x_k, \theta_0)$ denotes the matrix $[\delta(x_1, \theta_0), ..., \delta(x_k, \theta_0)]$. From (17) it is clear that (23) converges in law to a normal distribution.

Now $F_n(x, \theta_0)$, an ECDF constructed from independent standard normal variables, weakly converges to a Gaussian process. [See, e.g., Chapter 19 of van der Vaart (1998).] Therefore, it is an asymptotically tight sequence of stochastic processes. By Lemma 1, the uniform continuity of $\delta(x, \theta_0)^T$ (which follows from continuity of $\delta(x, \theta_0)^T$ and compactness of $\mathcal{X}$) insures that the non-remainder portion of

$$n^{\frac{1}{2}}[F_n(x, \widehat{\theta}_n) - \Phi(x)] = n^{\frac{1}{2}}[F_n(x, \theta_0) - \Phi(x)] + n^{\frac{1}{2}}\delta(x, \theta_0)^T(\widehat{\theta}_n - \theta_0) + o_p(1) \quad (24)$$

is also an asymptotically tight sequence of stochastic processes. Therefore, (24) weakly converges to a Gaussian process. The form of its autocovariance function is established in a manner similar to the latter portion of the proof of Theorem 1 in Houseman et al. (2003a). □

Note that (17), (18) and (20) are standardized by $n^{1/2}$. If $\mathcal{P}_n$ contains $N_n < n$ elements, then it may be more convenient to standardize these expressions by $N_n^{1/2}$. In practical settings, it may be assumed that $n^{-1}N_n \to \alpha$ for some $\alpha \leq 1$, in which case standardizing by $N_n^{1/2}$ is equivalent to standardizing by $n^{1/2}$. For example, if $\mathcal{P}_n$ obtains the standardized BLUPs for random effects in a mixed effects model, $n^{-1}N_n$ converges to the inverse of the limiting "average cluster size". We remark that condition (18) precludes cases where the number of observations per cluster grows too quickly. In such a case, condition (32) of Theorem 4 below would not be met; indeed, Theorem 5 requires bounded cluster sizes.

Theorem 1 establishes the theoretical autocovariance τ^2 of the estimator F_n. However, τ^2 must be estimated if the theorem is to be of practical use. If consistent estimators $\widehat{W_\theta}$ and $\widehat{\delta}$ of W_θ and $\delta(x, \theta_0)$ are available, then τ^2 is consistently estimated by substituting $\widehat{W_\theta}$ for W_θ and $\widehat{\delta}$ for $\delta(x, \theta_0)$ in (21). Standard methods for obtaining $\widehat{W_{11}}$ are available using ML and REML [see, for example, McCulloch & Searle (2001) or Diggle et al. (2002)]. To justify a consistent estimator for $\delta(x, \theta_0)$, we present Theorem 2, which uses $\mu_n(\beta, \gamma)$ as an approximation to $\mu^*(\beta, \gamma)$. Since $\mu_n(\beta, \gamma)$ implicitly depends on θ_0, we must justify the substitution of $\widehat{\theta}_n$ for θ_0, as well as the differentiation of $\mu_n(\beta, \gamma)$.

Theorem 2 (Estimation of $\delta(x, \theta_0)$). Let $\delta(x, \theta_0)$ be defined as in (22) and suppose condition (19) in Theorem 1 holds. In addition, let $\tilde{\mu}_n(x, \theta, \theta_0)$ denote $\mu_n(x, \theta)$ considered as a function of both the estimate θ and the true parameter θ_0. Assume $\tilde{\mu}_n(x, \theta, \theta_0)$ is continuously twice-differentiable in the second argument, and define

$$\delta_n(x, \theta, \theta_0) = \frac{\partial^T \tilde{\mu}_n}{\partial \theta}, \tag{25}$$

where the derivative is taken with respect to the second argument. Finally, suppose that for all n and for all $x \in \mathcal{X}$ and all θ sufficiently close to θ_0, there is an $M > 0$ such that

$$|\delta_n(x, \theta, \theta') - \delta_n(x, \theta, \theta_0)| \leq M|\theta' - \theta_0|. \tag{26}$$

Then, if $\widehat{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$, then

$$\sup_{x \in \mathcal{X}} |\delta_n(x, \widehat{\theta}_n, \widehat{\theta}_n) - \delta(x, \theta_0)| \xrightarrow{\mathcal{P}} 0.$$

Proof. Let

$$\zeta_n(x,\theta) = \tilde{\mu}_n(x,\theta,\theta_0) - \mu^*(x,\theta) = \varepsilon_n(x)^T(\theta-\theta_0) + O(|\theta-\theta_0|^2), \tag{27}$$

where the second equality is true by condition (19). Then ζ_n is continuously twice-differentiable in the second parameter and, by Taylor's Theorem,

$$\zeta_n(x,\theta) = \zeta_n(x,\theta_0) + \left.\frac{\partial\zeta_n}{\partial\theta^T}\right|_{\theta=\theta_0}(\theta-\theta_0) + O(|\theta-\theta_0|^2).$$

Since $\tilde{\mu}_n(x,\theta_0,\theta_0) = \mu^*(x,\theta_0)$, this expression, combined with (27), shows that

$$\left.\frac{\partial\zeta_n}{\partial\theta^T}\right|_{\theta=\theta_0} = \varepsilon_n(x)^T,$$

and therefore that

$$\frac{\partial\zeta_n}{\partial\theta^T} = \varepsilon_n(x)^T + O(|\theta-\theta_0|).$$

In other words, $\delta_n(x,\theta,\theta_0) - \delta(x,\theta_0) = \varepsilon_n(x) + O(|\theta-\theta_0|)$, where convergence of the latter term is uniform in n and x. Condition (26) implies that $\delta_n(x,\theta,\theta') - \delta_n(x,\theta,\theta_0) = O(|\theta'-\theta_0|)$, uniformly in n and x. Therefore,

$$\begin{aligned}
\delta_n(x,\widehat{\theta}_n,\widehat{\theta}_n) - \delta(x,\theta_0) &= \delta_n(x,\widehat{\theta}_n,\widehat{\theta}_n) - \delta_n(x,\widehat{\theta}_n,\theta_0) + \delta_n(x,\widehat{\theta}_n,\theta_0) - \delta(\theta_0) \\
&= O(|\widehat{\theta}_n-\theta_0|) + \varepsilon_n(x) + O(|\widehat{\theta}_n-\theta_0|) \\
&= \varepsilon_n(x) + O(|\widehat{\theta}_n-\theta_0|).
\end{aligned}$$

By consistency of $\widehat{\theta}_n$ and uniform convergence in x and n, the supremum of the last expression converges to 0. □

We focus now on the justification of our proposed resampling technique. Let $z^* \sim N(0, I_n)$ and let $U_n(\theta, z)$ be the estimating function for θ, as described in Section 2. Since the rotated residuals are independent when evaluated at θ_0, $U_n(\theta, z^*)$ has the same distribution as $U_n(\theta_0, z(\theta_0))$, and the ECDF $F_n^*(x)$ of z^*

has the same stochastic distribution as $F_n(x, \theta_0)$. We present Theorem 3 to justify the combination of $F_n^*(x)$ and U_n to approximate the behavior of (20).

Theorem 3 (Score Functions). Assume the conditions of Theorems 1 and 2. Let $U_n : \Theta \to \mathsf{R}^{p+q}$ be the score function corresponding to θ and J_n its corresponding derivative matrix. Assume that $n^{-1/2}U_n(\theta_0)$ converges to a normal distribution and

$$n^{1/2}(\widehat{\theta}_n - \theta') = n^{1/2} J_n(\theta')^{-1} U_n(\theta') + O_p(n^{1/2}|\widehat{\theta}_n - \theta'|^2) \tag{28}$$

for all θ' in a neighborhood of θ_0. Additionally, assume $J_n(\theta')$ is continuous. Then $n^{1/2}[F_n(x, \widehat{\theta}_n) - \Phi(x)]$ has the same stochastic limit as

$$n^{1/2}[F_n(x, \theta_0) - \Phi(x) + \delta_n(x, \widehat{\theta}_n, \widehat{\theta}_n)^T J_n(\widehat{\theta}_n)^{-1} U_n(\theta_0)]. \tag{29}$$

Proof. Note that $n^{1/2}|\widehat{\theta}_n - \theta_0|^2 = o_p(1)$ by the consistency of $\widehat{\theta}_n$. Consequently, it is clear from (24) that $n^{1/2}[F_n(x, \widehat{\theta}_n) - \Phi(x)]$ has the same stochastic limit as $n^{1/2}[F_n(x, \theta_0) - \Phi(x) + \delta_n(x, \theta_0)^T J(\theta_0)^{-1} U_n(\theta_0)]$, which differs from the right hand side of (29) only by $n^{1/2}[\delta(x, \theta_0)^T J_n(\theta_0)^{-1} - \delta_n(x, \widehat{\theta}_n, \widehat{\theta}_n)^T J_n(\widehat{\theta}_n)^{-1}] U_n(\theta_0) + o_p(1)$. By Theorem 2, the consistency of $\widehat{\theta}_n$, the continuity of J_n, and the asymptotic normality of $U_n(\theta_0)$, this difference is of stochastic order $o_p(1)$. □

Equating $\tilde{U}_n(\theta_0)$ and $U_n(\theta_0, z(\theta_0))$, expression (29) can be used to approximate $F_n(x, \widehat{\theta}_n) - \Phi(x)$. By Theorem 2 and the Continuous Mapping Theorem,

$$n^{1/2}[F_n^*(x) - \Phi(x) + n^{1/2}\delta_n(x, \widehat{\theta}_n, \widehat{\theta}_n)^T J_n(\widehat{\theta}_n)^{-1} U_n(\theta_0, z^*)] \tag{30}$$

is asymptotically equivalent to (29). However, since θ_0 is unknown, $\widehat{\theta}_n$ must be used as an estimate of θ_0 in (30). It is possible to justify such a substitution using, for example, the methods described in Chapter 23 ["Bootstrap"] of van der Vaart

(1998). The details are somewhat too lengthy to include in the present exposition. Letting

$$\widehat{F}_n^*(x) \equiv F_n(x, \widehat{\theta}_n) + \delta_n(x, \widehat{\theta}_n, \widehat{\theta}_n)^T J_n(\widehat{\theta}_n)^{-1} U_n(\widehat{\theta}_n, z^*), \tag{31}$$

it follows that realizations of (31), conditional on y, serve as approximations to the distribution of (20). Although (31) is an approximation to a parametric bootstrap, simulation results presented in Section 4 demonstrate that it is adequate for modest sample sizes.

The asymptotic normality condition (17), along with the corresponding efficiency requirement, can be established for many types of models. For clustered data, standard methods [e.g. van der Vaart (1998), Chapter 5] may be used to establish the result. In this context, Searle (1970) describes an explicit form for W_{11}. For autoregressive/moving-average (ARMA) time series models, Durbin (1960) is a useful reference. With regularity conditions on the form of $V(\gamma)$, asymptotic normality (with the corresponding convergence rate) can be proven using a Hájek projection [van der Vaart (1998), Chapter 11], although this is beyond the scope of the present work. The joint asymptotic normality can be derived using the Cramér-Wold device.

Conditions (18), (19), and (26) are difficult to verify directly. We defer a discussion of (19) and (26) until later in this section, but now present arguments to establish (18), which justifies the substitution of μ_n for F_n in asymptotic expresssions. [This leads to an asymptotic differentiability condition critical for the proof of Theorem 1; for a definition see van der Vaart (1998).] Theorem 4 is a general statement, but condition (32), which applies to the indicator residuals a_i

(not to z_i), is also difficult to verify. The condition turns out to hold for a large class of interesting models, and we discuss some special cases later in this section. The proof of Theorem 4 follows by combining Chebychev's inequality with the definition of convergence in probability.

Theorem 4 (Asymptotic Differentiability). Let $\widehat{\theta}_n$ be a consistent estimator of θ_0. If

$$n^{-1} \sum_{i,i'=1}^{n} \left| Cov \left[a_i(x, \theta, \theta_0), a_{i'}(x, \theta, \theta_0) \right] \right| = O\left(|\theta - \theta_0| \right), \tag{32}$$

where the boundedness of $O\left(|\theta - \theta_0| \right)$ is uniform in n and for $x \in \mathcal{X}$, then

$$R_n = n^{\frac{1}{2}} \sup_{x \in \mathcal{X}} \left| F_n(x, \widehat{\theta}_n) - \mu_n(x, \widehat{\theta}_n) - F_n(x, \theta_0) + \Phi(x) \right| \xrightarrow{\mathcal{P}} 0.$$

Proof. Let

$$R_n^*(x, \theta) = n^{\frac{1}{2}} \left(F_n(x, \theta) - \mu_n(x, \theta) - F_n(x, \theta_0) + \Phi(x) \right),$$

so that $R_n = \sup_{x \in \mathcal{X}} |R_n^*(x, \widehat{\theta}_n)|$. For the moment, fix $\theta \in \Theta$. Note that

$$F_n(x, \theta) - F_n(x, \theta_0) = n^{-1} \sum_{i=1}^{n} a_i(x, \theta, \theta_0)$$

and

$$E\left[F_n(x, \theta) - F_n(x, \theta_0) \right] = n^{-1} \sum_{i=1}^{n} E\left[a_i(x, \theta, \theta_0) \right] = \mu_n(x, \theta) - \Phi(x).$$

Therefore $E[R_n^*(x, \theta)] = 0$. Also, by (32),

$$Var[R_n^*(x, \theta)] = n^{-1} Var\left[\sum_{i=1}^{n} a_i(x, \theta, \theta_0) \right] \leq n^{-1} \sum_{i=1}^{n} \sum_{i'=1}^{n} \left| Cov\left[a_i(x, \theta, \theta_0), a_{i'}(x, \theta, \theta_0) \right] \right|.$$

Thus $Var[R_n^*(x, \theta)] = O\left(|\theta - \theta_0| \right)$. Together with the Chebychev inequality, this implies that there exists an $M > 0$ (independent of x and n) such that

$$P(R_n^*(x, \theta) > \varepsilon) \leq \frac{Var[R_n^*(x, \theta)]}{\varepsilon^2} \leq \frac{M}{\varepsilon^2} |\theta - \theta_0|.$$

Now let Π_n be the probability measure for the random variable $\widehat{\theta}_n$. For $\delta > 0$, it is possible to select n_0 large enough so that $\Pi_n \{|\theta - \theta_0| > \varepsilon^2\delta/M\} = P\left(|\widehat{\theta}_n - \theta_0| > \varepsilon^2\delta/M\right) \leq \delta$ whenever $n \geq n_0$. For such n,

$$\begin{aligned} P\left(R_n > \varepsilon, |\widehat{\theta}_n - \theta_0| \leq \frac{\varepsilon^2\delta}{M}\right) &= \int_\Theta 1\left\{|\theta - \theta_0| \leq \frac{\varepsilon^2\delta}{M}\right\} P\left(\sup_{x\in\mathcal{X}} R_n^*(x,\theta) > \varepsilon|\theta\right) d\Pi_n(\theta) \\ &\leq \int_\Theta 1\left\{|\theta - \theta_0| \leq \frac{\varepsilon^2\delta}{M}\right\} \frac{M}{\varepsilon^2}|\theta - \theta_0| d\Pi_n(\theta) \\ &\leq \frac{M}{\varepsilon^2}\frac{\varepsilon^2\delta}{M} \int_\Theta 1\left\{|\theta - \theta_0| \leq \frac{\varepsilon^2\delta}{M}\right\} d\Pi_n(\theta) \\ &\leq \delta \end{aligned}$$

and

$$P\left(R_n > \varepsilon, |\widehat{\theta}_n - \theta_0| > \frac{\varepsilon^2\delta}{M}\right) \leq P\left(|\widehat{\theta}_n - \theta_0| > \frac{\varepsilon^2\delta}{M}\right) \leq \delta.$$

Therefore, $P(R_n > \varepsilon)$ can be made arbitrarily small by choosing n large enough.

□

In practice, condition (32) may be difficult to verify directly, but Theorem 4 is still useful for proving that (18) holds in special cases. For example, it is used to prove Theorem 5, which addresses condition (18) for a large class of block-diagonal (clustered) designs, and Theorems 6 and 7, which address some specific models whose marginal covariance is not block-diagonal.

A critical consideration in the proofs of these theorems is a bound for the variance and covariance of the a_i quantities defined in (15); specifically, we require that

$$Var[a_i(\beta, \gamma; x, \theta_0, x)] = O(|\theta - \theta_0|). \tag{33}$$

The proof of (33) is complicated by the fact that the variance of $a_i(x, \theta, \theta_0)$ is not differentiable at θ_0 because its behavior depends upon whether or not $\gamma = \gamma_0$;

consequently, standard Taylor's Theorem arguments do not apply. We outline the heuristics of an alternative argument that applies generally to models that may not have block diagonal variance matrices. The variance and covariance of the a_i quantities can be decomposed into probabilities of events involving both $z_i(\theta)$ and $z_i(\theta_0)$. For example, $a_i(x, \theta, \theta_0, x) = -1$ precisely when $z_i(\theta_0) \leq x < z_i(\theta)$.

As $\theta \to \theta_0$, the correlation $r_i(\theta)$ of $z_i(\theta)$ and $z_i(\theta_0)$ approaches 1. This leads to singularity in the joint distribution of $z_i(\theta)$ and $z_i(\theta_0)$, and differentiation of $Var[a_i]$ becomes impossible. However, though the variance and covariance of the a_i quantities lead to expressions that have no derivative at θ_0, the derivatives near θ_0 are bounded, and the maximum of the derivatives as $\theta \to \theta_0$ from any direction suffices for the bound implied in (33). To describe conditions under which the derivatives near θ_0 are bounded, we exploit a geometric interpretation of $r_i(\theta)$ as an angle between two vectors, from which a connection between the conditions on $V(\gamma)$ and the boundedness of the derivatives is established.

The arguments are organized in Lemmas 2 through 7. Lemma 2 establishes a bound for the derivative of a vector function that returns the angle between the argument and a fixed vector; this result has no one-dimensional analog, for in one dimension the arccosine function is unbounded near 1. Lemma 3 establishes a geometric interpretation for $r_i(\theta)$ and uses the bound determined by Lemma 2 to bound $r_i(\theta)$. The geometry and corresponding bound are used in Lemma 4 and Lemma 5 to bound the probabilities of certain events involving $z_i(\theta)$ and $z_i(\theta_0)$. These probabilities are used in Lemmas 6 and 7 to bound the variance of a_i the covariance of a_i and a_j.

Lemma 2. Let $e_i \in \mathsf{R}^d$ denote the unit vector in the ith direction, and define $g : \mathsf{R}^d - \{e_i\} \to \mathsf{R}$ by

$$g(u) = \cos^{-1}\left(\frac{e_i^T u}{|u|}\right).$$

Then $|\partial g/\partial u| \to 1$ as $u \to e_i$.

Proof. See Houseman et al. (2003b), Lemma 1. □

Lemma 3. Let $r_i(\theta)$ be defined as in (14). If there is a neighborhood of γ_0 over which $\|V(\gamma)\|$ is bounded and bounded away from zero, and over which $\|d_\gamma L\|$ is bounded, then $cos^{-1} r_i(\theta)$ has bounded gradient near γ_0. Additionally, if the bounds for V and $d_\gamma L$ are uniform in n, then so is the bound for $cos^{-1} r_i(\theta)$.

Proof. See Houseman et al. (2003b), Lemma 2. □

Lemma 4. Let

$$A_i(x, \theta) = P\left(z_i(\theta_0) \leq x < z_i(\theta)\right)$$

and

$$B_i(x, \theta) = P\left(z_i(\theta) \leq x < z_i(\theta_0)\right),$$

and suppose the conditions of Lemma 3 are met. Then there is a neighborhood U of θ_0 and an $M_n > 0$ such that for all $\theta \in U$ and for all $x \in \mathcal{X}$,

$$A_i(x, \theta) \leq M_n |\theta - \theta_0|$$

and

$$B_i(x, \theta) \leq M_n |\theta - \theta_0|.$$

Moreover, if the uniformity conditions in Lemma 3 are met, and if the gradient of $m_i(\theta)$ is uniformly bounded (in n), then there is an $M > 0$ such that $M_n \leq M$ for all n.

Proof. Let $t_i(\theta) = \cos^{-1} r_i(\theta) \geq 0$. A straightforward probability calculation, based on well-known facts about the bivariate normal distribution, produces $A_i(x,\theta) = A_i^{(x)}(m,s,t)$, where

$$A_i^{(x)}(m,s,t) = \int_x^\infty \Phi\left(\frac{x-m-rsw}{s\sqrt{1-r^2}}\right)\phi(w)dw = \int_x^\infty \Phi\left(\frac{x-m-\cos(t)sw}{s\sin(t)}\right)\phi(w)dw, \tag{34}$$

$m = m_i(\theta)$, $s = s_i(\theta)$, $r = r_i(\theta)$, and $t = t_i(\theta)$. Note that $A_i^{(x)} \to 0$ as $m \to 0$, $s \to 1$, and $t \downarrow 0$. For each value of $t > 0$, there is a neighborhood containing t over which the integrand in (34) is bounded. Therefore, the derivatives of $A_i^{(x)}$ with respect to m, s, and t can be obtained by differentiating the integrand and integrating the result. $A_i^{(x)}$ does not possess a derivative at the point $(m,s,t) = (0,1,0)$, but it is differentiable for $t > 0$ and the corresponding derivatives remain bounded as $(m,s,t) \to (0,1,0)$. Furthermore, they are uniformly bounded with respect to x, as long as x ranges over a compact set. Consequently, for all $x \in \mathcal{X}$, the derivatives are bounded in a convex neighborhood U' of $(0,1,0)$ by a constant $M' > 0$. Thus, for any fixed $t' > 0$ with $(0,1,t') \in U'$, we have

$$A_i^{(x)}(m,s,t) = A_i^{(x)}(0,1,t') + \left.\frac{\partial A_i^{(x)}}{\partial m}\right|_{m=m^*} m + \left.\frac{\partial A_i^{(x)}}{\partial s}\right|_{s=s^*} (s-1) + \left.\frac{\partial A_i^{(x)}}{\partial t}\right|_{t=t^*} (t-t'),$$

where m^* is between 0 and m, s^* is between 1 and s, and t^* is between t' and t. [See Serfling (1980), Section 1.12, Theorem B.] From the properties of matrix norms, it follows that

$$|A_i^{(x)}(m,s,t) - A_i^{(x)}(0,1,t')| \leq \left\| \left(\left.\frac{\partial A_i^{(x)}}{\partial m}\right|_{m=m^*}, \left.\frac{\partial A_i^{(x)}}{\partial s}\right|_{s=s^*}, \left.\frac{\partial A_i^{(x)}}{\partial t}\right|_{t=t^*} \right) \right\| |(m, s-1, t-t')^T|.$$

Letting $t' \downarrow 0$, and using the bound for the derivatives and the convexity of U', we have

$$|A_i^{(x)}(m,s,t)| \leq M' |(m, s-1, t)^T|.$$

Along the hyperplane $t = 0$, a similar probability calculation produces a bound that, without loss of generality, can be assumed to be $\leq M'$. (Otherwise, take M' to be the minimum of the two bounds.)

By Lemma 3, $t_i(\theta)$ has bounded gradient near γ_0. Since

$$\frac{\partial s_i}{\partial_{\gamma_k}} = \frac{1}{2} s_i(\theta)^{-1} \left(P_i (d_{\gamma_k} L^T) V(\gamma_0) L(\gamma) P_i^T + P_i L(\gamma)^T V(\gamma_0) (d_{\gamma_k} L) P_i^T \right),$$

the gradient of $s_i(\theta)$ is bounded by $|s_i(\theta)|^{-1}$, $\|d_{\gamma_k} L^T\|$, $\|V(\gamma_0)\|$, and $\|d_{\gamma_k} L\|$. The latter two quantities are bounded by assumption; being upper-triangular, $d_{\gamma_k} L^T$ has the same norm as its transpose; and $s_i(\theta)$ can be made arbitrarily close to 1. Finally,

$$\frac{\partial m_i}{\partial_{\gamma_k}} = P_i (d_{\gamma_k} L^T) X (\beta - \beta_0)$$

and

$$\frac{\partial m_i}{\partial_\beta} = P_i L(\gamma)^T X,$$

and X has bounded elements, so the gradient of $m_i(\theta)$ is bounded (although not necessarily uniformly in n). It follows that there is a neighborhood U of θ_0 and an $M'' > 0$ such that $|m_i(\theta)| < M''|\theta - \theta_0|$, $|s_i(\theta) - 1| < M''|\theta - \theta_0|$, and $|t_i(\theta)| < M''|\theta - \theta_0|$. With $M_n = 3M'M''$, we have

$$|A_i(x, \theta)| \leq M' |(m_i(\theta), s_i(\theta) - 1, t_i(\theta)^T| \leq 3M'M''|\theta - \theta_0| \leq M_n |\theta - \theta_0|.$$

If $\|V(\gamma)\|$ is uniformly bounded and bounded away from zero, and if $\|d_\gamma L\|$ is uniformly bounded, then the gradients of $s_i(\theta)$, and $t_i(\theta)$ are uniformly bounded. For $t_i(\theta)$, this follows from Lemma 3, and for $s_i(\theta)^2$ it follows from the previous paragraph. If, in addition, $m_i(\theta)$ is uniformly bounded, then $3MM''$ is a bound for all n. A similar proof establishes the same result for $B_i(x, \theta)$. □

Lemma 5. Let

$$AA_{ij}(x,\theta) = P\left(z_i(\theta_0) \le x < z_i(\theta), z_j(\theta_0) \le x < z_j(\theta)\right),$$

$$AB_{ij}(x,\theta) = P\left(z_i(\theta_0) \le x < z_i(\theta), z_j(\theta) \le x < z_j(\theta_0)\right),$$

$$BA_{ij}(x,\theta) = P\left(z_i(\theta) \le x < z_i(\theta_0), z_j(\theta_0) \le x < z_j(\theta)\right),$$

and

$$BB_{ij}(x,\theta) = P\left(z_i(\theta) \le x < z_i(\theta_0), z_j(\theta) \le x < z_j(\theta_0)\right).$$

Also, let $A_i(x,\theta)$ and $B_i(x,\theta)$ be defined as in Lemma 4, and let

$$\sigma_{ij}(\theta) = \begin{bmatrix} \sigma_{ij0} \\ \sigma_{ij1} \\ \sigma_{ij2} \end{bmatrix} = \begin{bmatrix} Cov[z_i(\theta), z_j(\theta)] \\ Cov[z_i(\theta), z_j(\theta_0)] \\ Cov[z_i(\theta_0), z_j(\theta)] \end{bmatrix}.$$

If the conditions of Lemma 3 are met, then

$$AA_{ij}(x,\theta) - A_i(x,\theta)A_j(x,\theta) = O(|\sigma_{ij}|),$$

and similar relationships hold for $AB_{ij}(\theta)$, $BA_{ij}(\theta)$, and $BB_{ij}(\theta)$. The bound for these quantities is uniform in i and in $x \in \mathcal{X}$. They are uniform in n if the bounds for $m_i(\theta)$ and the bounds in Lemma 3 are uniform in n.

Proof. As in the proof of Lemma 4, we use a limiting Taylor's Theorem argument. We first express $AA_{ij}^{(x)} = AA_{ij}(x,\theta)$ as the integral of the joint density of $z_i(\theta)$, $z_i(\theta_0)$, $z_j(\theta)$, and $z_j(\theta_0)$. Defining the covariance and mean of the corresponding

multivariate normal, let

$$W_{ij} = \begin{bmatrix} s_i^2 & s_i^2\cos(t_i) & \sigma_{ij0} & \sigma_{ij1} \\ s_i^2\cos(t_i) & 1 & \sigma_{ij2} & 0 \\ \sigma_{ij0} & \sigma_{ij2} & s_j^2 & s_j^2\cos(t_j) \\ \sigma_{ij1} & 0 & s_j^2\cos(t_j) & 1 \end{bmatrix}, \tag{35}$$

$$W_{ij}^* = \begin{bmatrix} s_i^2 & s_i^2\cos(t_i) & 0 & 0 \\ s_i^2\cos(t_i) & 1 & 0 & 0 \\ 0 & 0 & s_j^2 & s_j^2\cos(t_j) \\ 0 & 0 & s_j^2\cos(t_j) & 1 \end{bmatrix},$$

and $\xi_{ij} = (m_i, 0, m_j, 0)^T$, where t_i and t_j are defined as in the proof of Lemma 4 and the functional dependence on (θ) has been omitted from the notation. Then

$$AA_{ij}^{(x)} = \frac{|W_{ij}|^{-1/2}}{(2\pi)^2}\int_x^\infty\int_{-\infty}^x\int_x^\infty\int_{-\infty}^x \exp\left(-\frac{1}{2}(w-\xi_{ij})^T W_{ij}^{-1}(w-\xi_{ij})\right)dw \tag{36}$$

and

$$A_i^{(x)}A_j^{(x)} = \frac{|W_{ij}^*|^{-1/2}}{(2\pi)^2}\int_x^\infty\int_{-\infty}^x\int_x^\infty\int_{-\infty}^x \exp\left(-\frac{1}{2}w^T W_{ij}^{*-1}w\right)dw. \tag{37}$$

Note that $\sigma_{ij1} \to 0$, $\sigma_{ij2} \to 0$, $\sigma_{ij0} \to 0$, $s_i \to 1$, $s_j \to 1$, $t_i \to 0$, $t_j \to 0$, $m_i \to 0$, and $m_j \to 0$ as $\theta \to \theta_0$. We first show that the derivatives of $AA_{ij}^{(x)}$ with respect to σ_{ij1}, σ_{ij2}, and σ_{ij0} are bounded near $\theta = \theta_0$. Note that W_{ij} and W_{ij}^* are singular at this limit. We compute the derivative using formal matrix differentiation; simplification produces

$$d_k AA_{ij}^{(x)} = \frac{|W_{ij}|^{-1/2}}{(2\pi)^2}\int_x^\infty\int_{-\infty}^x\int_x^\infty\int_{-\infty}^x \exp\left(-\frac{1}{2}w^T W_{ij}^{-1}w\right)h_k(w)dw, \tag{38}$$

where

$$h_k(w) = \frac{1}{2}\left[w^T W_{ij}^{-1}d_k W_{ij} W_{ij}^{-1}w - \mathrm{tr}(W_{ij}^{-1}d_k W_{ij})\right]$$

and d_k denotes the matrix derivative with respect to one of the σ_{ij} parameters appearing in (35). It is easily shown that

$$\mathrm{tr}(W_{ij}^{-1} d_k W_{ij}) = 0$$

when $\sigma_{ij1} = \sigma_{ij2} = \sigma_{ij0} = 0$. After a change of variables to $u = U_{ij} w$, where $U_{ij}^T U_{ij} = W_{ij}^{-1}$, (38) produces

$$d_k A A_{ij}^{(x)} = \frac{1}{2(2\pi)^2} \int_\Omega \exp\left(-\frac{1}{2} u^T u\right) u^T (U_{ij} d_k W_{ij} U_{ij}^T) u du, \tag{39}$$

where $\Omega \subset \mathsf{R}^4$ represents the appropriately transformed region. It can be shown that when $\sigma_{ij1} = \sigma_{ij2} = \sigma_{ij0} = 0$, the integral in (39) can be expressed as the sum of integrals of the form

$$\frac{\zeta(t_i, t_j)}{2} \int_x^\infty \int_{-\infty}^{\alpha_i x - \beta_i u_2} u_1^{d_i} u_2^{1-d_i} \phi(u_1)\phi(u_2) du_1 du_2 \int_x^\infty \int_{-\infty}^{\alpha_j x - \beta_j u_4} u_3^{d_j} u_4^{1-d_i} \phi(u_3)\phi(u_4) du_3 du_4,$$

where $\alpha_i = \sin(t_i)^{-1}$, $\beta_i = \cos(t_i)\sin(t_i)^{-1}$, $\alpha_j = \sin(t_j)^{-1}$, $\beta_j = \cos(t_j)\sin(t_j)^{-1}$, $d_i \in \{0, 1\}$, $d_j \in \{0, 1\}$, and $\zeta(t_i, t_j) = O(\sin(t_i)^{-d_i} \sin(t_j)^{-d_j})$. Direct computation of the double integrals shows that the entire integral is bounded. Thus, (38) remains bounded as $\theta \to \theta_0$. Since (38) is a continuous function of x, it is uniformly bounded for $x \in \mathcal{X}$.

Now, $AA_{ij}^{(x)} - A_i^{(x)} A_j^{(x)} = 0$ when $\sigma_{ij1} = \sigma_{ij2} = \sigma_{ij0} = 0$, whatever the values of s_i, s_j, t_i, t_j, m_i, and m_j may be. With the restriction $t_i > 0$ and $t_j > 0$, $AA_{ij}^{(x)} - A_i^{(x)} A_j^{(x)}$ is a continuously differentiable function, so derivatives with respect to other parameters in (35) are zero at $\sigma_{ij1} = \sigma_{ij2} = \sigma_{ij0} = 0$. Consequently,

$$AA_{ij}^{(x)} - A_i^{(x)} A_j^{(x)} = \left.\frac{\partial A A_{ij}^{(x)}}{\partial \sigma_{ij}}\right|_{\theta=\theta^*} \sigma_{ij} = D(x, \theta^*)^T \sigma_{ij},$$

where $D(x, \theta)^T$, the derivative with respect to σ_{ij}, is evaluated near $s_i = s_j = 1$, $m_i = m_j = 0$, and $t_i > 0$, $t_j > 0$. Since $D(x, \theta)$ remains bounded as $t_i \to 0$ and

$t_j \to 0$, an argument similar to that used in the proof of Lemma 4 shows that there is an $M_{ij} > 0$ such that

$$|AA_{ij}^{(x)} - A_i^{(x)} A_j^{(x)}| < M_{ij}|\sigma_{ij}|.$$

As in the proof of Lemma 4 the bound M_{ij} is uniform if the bounds for $m_i(\theta)$ and the bounds in Lemma 3 are uniform. The proofs for AB_{ij}, BA_{ij} and BB_{ij} are identical, except that the ranges of integration in (36) and (37) differ. □

Lemma 6. Let $a_i(x, \theta, \theta')$ be defined as in (15). If the conditions of Lemma 3 are met, then there is a neighborhood of θ_0 and an $M_n > 0$ such that $Var[a_i(x, \theta, \theta_0)] \leq M_n|\theta - \theta_0|$ for all $x \in \mathcal{X}$. If the gradients of $m_i(\theta)$, $s_i(\theta)$, and $\cos^{-1} r_i(\theta)$ are uniformly bounded in n, then there is a uniform bound $M = M_n > 0$ for all n.

Proof. Let $A_i(x, \theta)$ and $B_i(x, \theta)$ be as in Lemma 4, and assume for the moment that x and θ are fixed. Write $a_i = a_i(x, \theta, \theta_0)$, $A_i = A_i(x, \theta)$, and $B_i = B_i(x, \theta)$. Note that A_i and B_i are probabilities of mutually exclusive events, and that $a_i^2 = 1$ when either event holds. Computing the variance directly,

$$\begin{aligned} Var[a_i] &= E[a_i^2] - E[a_i]^2 \\ &= (A_i + B_i) - (A_i - B_i)^2 \\ &= A_i(1 - A_i) + B_i(1 - B_i) + 2A_iB_i \\ &\leq A_i + B_i + 2\max(A_i, B_i) \\ &\leq 4M_n|\theta - \theta_0|, \end{aligned}$$

where M_n is the bound determined in Lemma 4. If the gradients are uniformly bounded, as in Lemma 4, the bound for $Var[a_i]$ is uniform for all n. □

Lemma 7. Let $a_i(\theta, \theta')$ be defined as in (15), and let σ_{ij} be defined as in Lemma 5. There is a neighborhood of θ_0 and an $M_n > 0$ such that $Cov[a_i(x, \theta, \theta_0), a_j(x, \theta, \theta_0)] \leq M_n|\sigma_{ij}|$ for all $x \in \mathcal{X}$. If the gradients of $m_i(\theta)$, $s_i(\theta)$, and $\cos^{-1} r_i(\theta)$ are uniformly bounded in n, then there is a uniform bound $M = M_n > 0$ for all n.

Proof. Let $A_i(x, \theta)$ and $B_i(x, \theta)$ be as in Lemma 4, and let $AA_{ij}(x, \theta)$, $AB_{ij}(x, \theta)$, $BA_{ij}(x, \theta)$, and $BB_{ij}(x, \theta)$ be as in Lemma 5. Assume for the moment that x and θ are fixed. Suppressing the functional notation, we compute the covariance directly:

$$\begin{aligned} Cov[a_i, a_j] &= E[a_i a_j] - E[a_i]E[a_j] \\ &= (AA_{ij} - AB_{ij} - BA_{ij} + BB_{ij}) - (A_i - B_i)(A_j - B_j) \\ &= AA_{ij} - A_i A_j - AB_{ij} + A_i B_j - BA_{ij} + B_i A_j + BB_{ij} - B_i B_j. \end{aligned}$$

The result now follows from Lemma 5. □

Theorem 5 (Clustered Designs). Let $V(\gamma) = (V_1(\gamma), ..., V_N(\gamma))$ have block diagonal structure, with N blocks having dimension at most $k \times k$, and for each h let $L_h(\gamma)L_h(\gamma)^T = V_h(\gamma)^{-1}$. Assume that, near γ_0, the $\mathcal{L}^2$ norm $\|V_h(\gamma)\|$ is uniformly bounded and bounded away from zero, both in block number h and in sample size n, and that for each $j = 1, ..., q$, $\left\|d_{\gamma_j} L_h\right\|$ is uniformly bounded in both h and n. If $\widehat{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$, then the conditions of Theorem 4 hold and the conclusion (18) follows.

Proof. For this proof, we employ double subscripting notation. In particular, let $y = (y_1^T, ..., y_N^T)^T$, where $Var[y_h] = V_h(\gamma)$ and y_h and $y_{h'}$ are independent if $h \neq h'$, and let d_h be the dimension of $V_h(\gamma)$. Let $P_{hi}^* : \mathsf{R}^{d_h} \to \mathsf{R}$ be projection

onto the ith component in block h. Define

$$z_{hi}(\theta) = P^*_{hi} L_h(\gamma)^T (y_h - X_h \beta),$$

which is consistent with the corresponding definition above; define $a_{hi}(\theta; \beta', \gamma')$, and all other quantities analogously (with double subscript notation).

The block-diagonal structure of $V(\gamma)$ insures that, locally, $V(\gamma)$ is uniformly bounded in n, for the norm of $V(\gamma)$ can be no more than $\max_h \|V_h(\gamma)\|$. Similarly, near γ_0, $V(\gamma)$ is uniformly bounded away from zero. It follows that $L(\gamma)$ is also uniformly bounded and bounded away from zero. Similar considerations apply to $d_\gamma L$. Since

$$|\partial m_{hi}/\partial \beta| \leq \|L_h(\gamma)^T\| \|X_h\|$$

and

$$|\partial m_{hi}/\partial \gamma_k| \leq \|d_{\gamma_k} L_h^T\| \|X_h\| |\beta_0 - \beta|,$$

a uniform bound is obtained for $m_{hi}(\theta)$ using the bounds for $\|L_h(\gamma)^T\|$ and $\|d_{\gamma_k} L_h\|$ and the fact that X_h has bounded elements and dimensions. Thus the uniformity conditions of Lemma 6 are met.

Writing $a_{hi} = a_{hi}(x, \theta, \theta_0)$, it follows from Lemma 6 that there is an $M > 0$ such $Var[a_{hi}] \leq M|\theta - \theta_0|$. Consequently, $Cov[a_{hi}, a_{h'i'}] \leq Var[a_{hi}] \leq M|\theta - \theta_0|$ if $h = h'$ and $Cov[a_{hi}, a_{h'i'}] = 0$ if $h \neq h'$. Thus,

$$\begin{aligned}
n^{-1} \sum_{h,i,h',i'} |Cov[a_{hi}, a_{h'i'}]| &= n^{-1} \sum_{h=h', i=i'} Var[a_{hi}] + n^{-1} \sum_{h \neq h' \cup i \neq i'} |Cov[a_{hi}, a_{h'i'}]| \\
&= n^{-1} \sum_{h=h', i=i'} Var[a_{hi}] + n^{-1} \sum_{h=h' \cap i \neq i'} |Cov[a_{hi}, a_{h'i'}]| \\
&\leq n^{-1} \cdot n \cdot M|\theta - \theta_0| + n^{-1} \cdot Nk^2 \cdot M|\theta - \theta_0| \\
&\leq M|\theta - \theta_0| + k^2 M|\theta - \theta_0|
\end{aligned}$$

$$= \quad (1+k^2)M|\theta - \theta_0|.$$

The bound M is uniform for all $x \in \mathcal{X}$ and for all n, so this establishes condition (32) in Theorem 4. □

Classical mixed effects models (Laird & Ware, 1982) fall into the class of models described by Theorem 5, as long as the design matrix Z_n corresponding to the random effects is uniformly bounded and bounded away from zero, and the covariance matrix Δ of the random effects satisfies requirements similar to those required by $V(\gamma)$ in the statement of the theorem. Corollary 1 describes the formal result.

Corollary 1 (Longitudinal Mixed Effects Models). Let $V(\gamma) = (V_1(\gamma), ..., V_N(\gamma))$ have block diagonal structure, where for each h, $V_h(\gamma)$ has dimension $d_h \leq k$,

$$V_h(\gamma) = \gamma_1 I_{d_h} + Z_h \Delta(\gamma_2, ..., \gamma_q) Z_h^T,$$

and Z_h is a $d_h \times (q-1)$ matrix that is bounded and bounded away from zero. Furthermore, assume that $\gamma_1 > 0$, that $\Delta(\gamma_2, ..., \gamma_q)$ has bounded norm near γ_0, and that $d_\gamma \Delta$ is bounded in a neighborhood of γ_0. If $\theta_n \xrightarrow{\mathcal{P}} \theta_0$, then the conditions of Theorem 4 hold and the conclusion (18) follows.

Proof. Follows directly from Theorem 5. □

Condition (18) also applies in models that do not have block diagonal variance. However, (33) is not a sufficient condition in the general case. The covariance $Cov[a_i, a_j]$ for $i \neq j$ must also be bounded. The following theorem, which establishes (18) for linear model with ARMA errors, is a typical example of the utility of our method for unclustered designs:

Theorem 6 (ARMA Errors). Let $\varepsilon(\beta) = y - X\beta$ be a time-series vector from an autoregressive moving-average process of order $(\tilde{p}, \tilde{q})$. Thus, for $i > \max(\tilde{p}, \tilde{q})$,

$$B(\gamma)\tilde{\varepsilon}_i(\beta) = A(\gamma)\tilde{z}_i(\beta, \gamma), \tag{40}$$

where $\tilde{\varepsilon}_i(\beta) = (\varepsilon_{i-\tilde{p}}(\beta), ..., \varepsilon_i(\beta))^T$, $\tilde{z}_i(\beta, \gamma) = (z_{i-\tilde{q}}(\beta, \gamma), ..., z_i(\beta, \gamma))^T$, and $A(\gamma)$ and $B(\gamma)$ are conformable matrices, depending on γ, such that $A_1(\gamma) \neq 0$, $A_{\tilde{q}+1}(\gamma) = 1$, $B_1(\gamma) \neq 0$, and $B_{\tilde{p}+1}(\gamma) = -1$. If $\widehat{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$, then the conditions of Theorem 4 hold and the conclusion (18) follows.

Proof. We first demonstrate that the conditions of Lemma 6 and Lemma 7 are met. Because the covariance terms of $V(\gamma)$ die off exponentially for widely separated observations [see Davidson (1994), page 215], the $\mathcal{L}^2$ norm of $V(\gamma)$ is uniformly bounded in n. In addition, the variance of $\varepsilon_i(\beta)$ is a positive value (and constant after finitely many observations), so $||V(\gamma)||$ is bounded away from zero and $||L(\gamma)||$ is bounded. The matrix $L(\gamma)$ must produce residuals that satisfy (40), so that $A(\gamma)E[\tilde{z}_i(\beta, \gamma)] = B(\gamma)\tilde{X}_i(\beta_0 - \beta)$, where $\tilde{X}_i$ is an appropriate submatrix of X. It follows that $m_i(\beta, \gamma)$ is uniformly bounded in n. Therefore the lemmas apply. Equation (40) implies that for $i, i' > \max(\tilde{p}, \tilde{q})$,

$$Cov[z_i(\beta, \gamma), z_{i'}(\beta', \gamma')] = B(\gamma)Cov[\tilde{\varepsilon}_i(\beta), \tilde{\varepsilon}_{i'}(\beta')]B(\gamma')^T.$$

Since $B(\gamma)$ and $B(\gamma')$ are bounded near γ_0 and $Cov[\varepsilon_i(\beta), \varepsilon_{i'}(\beta')] = Cov[\varepsilon_i(\beta_0), \varepsilon_{i'}(\beta_0)]$ approaches zero exponentially as $i - i' \to \infty$ (after finitely many terms), it follows that $Cov[z_i(\beta, \gamma), z_{i'}(\beta', \gamma')] = O(\rho^{|i-i'|})$ for some nonnegative $\rho < 1$ that depends only on γ_0. Lemmas 6 and 7 imply that $|Cov[a_i(\beta, \gamma, \beta_0, \gamma_0), a_{i'}(\beta, \gamma, \beta_0, \gamma_0)]| =$

$O(\rho^{|i-i'|}|\theta-\theta_0|)$, thus

$$\begin{aligned} n^{-1}\sum_{i,i'=1}^{n}|Cov[a_i(\beta,\gamma,\beta_0,\gamma_0),a_{i'}(\beta,\gamma,\beta_0,\gamma_0)]| &= n^{-1}\sum_{i,i'=1}^{n}O(\rho^{|i-i'|}|\theta-\theta_0|) \\ &= O(|\theta-\theta_0|)\left(1+n^{-1}\sum_{i\neq i'}O(\rho^{|i-i'|})\right) \\ &= O(|\theta-\theta_0|). \end{aligned}$$

The last equality follows from that fact that

$$\sum_{i\neq i'}\rho^{|i-i'|}=2\sum_{i<i'}\rho^{|i-i'|}\leq 4(n-1)\sum_{i=1}^{\infty}\rho^i=4(n-1)\rho(1-\rho)^{-1}.$$

Therefore condition (32) of Theorem 4 is satisfied. □

In practical situations involving more complicated models, it may be difficult to verify (32) algebraically. However, (33) is true under mild conditions, so that if a plausible qualitative argument that $Corr[a_i,a_j]=O(r^{|i-j|})$ for some r can be made, Theorem 4 can be used to justify (18). For example, it may be desirable to apply our methods to higher order ARMA models. The following theorem justifies the method for models whose correlation is based on the Kronecker product of correlations from different components, although (41) can still be challenging to verify.

Theorem 7 (Products). Let $V_1(\gamma)$ be a $k\times k$ covariance matrix (fixed with respect to n) and let $V_2(\gamma)$ be a covariance matrix whose norm is bounded and bounded away from zero. Let $L_1(\gamma)$ and $L_2(\gamma)$ be the corresponding Cholesky matrices, and assume that the norm of $L_2(\gamma)$ has bounded gradient. Finally, assume that $\|L_2(\gamma)^TX\|$ is bounded and that there is an $M>0$ such that

$$\sum_{i\neq j}Cov[z_i(\theta),z_j(\theta')]\leq M|\theta-\theta_0|, \tag{41}$$

where $\theta' = \theta$ or $\theta' = \theta_0$. If $V(\gamma) = V_1(\gamma) \otimes V_2(\gamma)$, then the conditions of Theorem 4 apply to $V(\gamma)$ and the conclusion (18) follows.

Proof. Note that the Cholesky matrix of $V_1(\gamma) \otimes V_2(\gamma)$ is $L_1(\gamma) \otimes L_2(\gamma)$. Since $V_1(\gamma)$ has fixed dimension, it is easy to show that $V_1(\gamma) \otimes V_2(\gamma)$ has a norm that is bounded and bounded away from zero. A proof similar to that used for Theorem 5 can be used to establish condition (32) of Theorem 4. □

Theorems 4, 5, 6, and 7 address condition (18) in Theorem 1. However, condition (19) is also difficult to verify in practice. Theorem 8, proven by Taylor expansion, gives plausible conditions that insure (19).

Theorem 8. In a neighborhood of θ_0, let $\|V_n(\gamma)\|$ be uniformly bounded and uniformly bounded away from zero. Assume, in addition, that $L_n(\gamma)$ is continuously twice-differentiable, and that for each $j, j' = 1, ..., q$, $\|d_{\gamma_j} L_n(\gamma)\|$ and $\|d_{\gamma_j} d_{\gamma_{j'}} L_n(\gamma)\|$ are uniformly bounded near θ_0. Finally, suppose that there is a constant vector c_0 such that

$$n^{-1}\phi(x) \sum_{i=1}^{n} \left(\frac{x}{2} \left. \frac{\partial s_i^2}{\partial \theta^T} \right|_{\theta=\theta_0} + \left. \frac{\partial m_i}{\partial \theta^T} \right|_{\theta=\theta_0} \right) \to c_0(x)^T \tag{42}$$

uniformly for $x \in \mathcal{X}$ as $n \to \infty$. Then condition (19) of Theorem 1 is satisfied.

Proof. The proof is identical to the proof of Theorem 7 in Houseman et al. (2003b). Uniform convergence of $\varepsilon_n(x)$ follows directly from the uniform convergence of (42). □

Condition (42) is difficult to verify for arbitrary sequences, but is plausible when there is some uniformity to the model sequence. For example, if there are

exchangeable units of sampling, then (42) will be true. Note that $\partial m_i/\partial\beta = P_i L(\gamma)^T X$, $\partial s_i^2/\partial\beta = 0$,

$$\frac{\partial s_i^2}{\partial\gamma} = P_i[(d_\gamma L^T)V(\gamma_0)L(\gamma) + L(\gamma)^T V(\gamma_0)(d_\gamma L)]P_i^T,$$

and

$$\frac{\partial m_i}{\partial\gamma} = P_i[d_\gamma L^T]X(\beta_0 - \beta).$$

Evaluated at θ_0, $\partial m_i/\partial\beta = P_i L(\gamma_0)^T X$ and $\partial m_i/\partial\gamma = 0$. Thus, (42) reduces to plausible regularity in the covariate matrix X and in the correlation structure implied by $V(\gamma)$.

Similar considerations apply to condition (26) of Theorem 2. We refrain from stating another theorem, but remark that conditions identical to those described for $V(\gamma)$ and $L(\gamma)$ in Theorem 8, together with a condition similar to (42), lead to the desired conclusion (26). The result is proven by differentiating $\delta_n(\theta, \theta')$ with respect to the second argument and showing that the resulting expression is bounded.

It is easy to lose sight of the basic methodology among the many technical conditions such as (18), (19), (26), and (42). We have proven results for longitudinal mixed effects models, for AR-1 time series regression, and for Kronecker products. In practice, our proposed methods work well for mixed effects models and time-series regression, as we demonstrate in the following section.

4 Simulations

Houseman et al. (2003b) reported simulations demonstrating that for many types

of correlation models, pointwise standard errors and Wald confidence intervals based on $\tau^2(x, x)$, as defined in (21), behave adequately. Simulations were presented for clustered AR-1 models, equicorrelated clusters (equivalent to random-intercept models), AR-1 time series regression, and more complicated correlation structures for which (18) could not be verified. In each of these cases, analytical standard errors matched the corresponding simulation standard errors and coverage probabilities of Wald confidence intervals matched their nominal levels. Additionally, when non-normal errors were supplied, the ECDF estimates were biased and coverage probabilities less than their nominal level, demonstrating that the method is sensitive to non-normal errors for the marginal distribution of (1). In this section we present simulations that demonstrate the behavior of the global tests formed using the resampling technique suggested by (5), (6) and (7).

Our first set of simulations investigated the performance of our proposed rotated marginal residuals, formed with $P_i = \pi_i$. For each of three time series cases, 1000 simulated analyses were conducted. The cases included an AR-1 time series model with normal errors, skewed (standardized χ^2_3) errors, and heavy-tailed (standardized t_3) errors. In addition, six random slope/intercept models were examined: random effects having normal, skewed, heavy-tailed, and binary (standardized binomial with success probability equal to 0.5) distributions, each with normal conditional errors; and random effects having normal distribution with conditional errors having skewed and heavy-tailed distributions. For each of these cases 500 simulated analyses were conducted. For each AR-1 case, the covariate matrix X consisted of rows of the form $(1, u)$, where u was a uniform

variable, and the corresponding regression parameter was $\beta = (10, .5)^T$. For each random effects case, the covariate matrix X consisted of rows of the form $(1, u, obs)$, where u was a uniform variable, obs was the observation number (1 through 5), and the corresponding fixed effect parameter was $\beta = (10, .5, 0)^T$. In the AR-1 time series case, $n = 250$ and $\gamma = (\sigma^2, \rho) = (1, .5)$. In the random slope/intercept model, $N = 50$ clusters, each having 5 observations, for a total of $n = 250$; additionally, $\Delta = \text{diag}(4, .25)$ and Z consisted of rows having the form $(1, obs)$. For each simulated analysis, we computed the sup functional described in (6) and the Cramer-Von-Mises functional described (7), with $\mathcal{X} = (-2.5, 2.5)$, for the observed data and for 1000 resampled processes as described by (5). For convenience we denote this functional as $\mathcal{J}$. Quantiles and rejection probabilities from the omnibus test for normality are presented in the Tables 1 through 3.

When the error (and random effect) was normal, rejection probabilities matched their nominal level and the resampled functionals had quantiles that matched the quantiles of the observed values. For non-normal errors, rejection probabilities were larger than nominal and the observed quantiles were larger than their resampled counterparts. Using the Kolmogorov-Smirnov type functional (6), the power was reasonable for AR-1 time-series models at $n = 250$ and at $\alpha = 0.05$, with about 82% probability of rejecting normality when the errors were skewed and 68% probability of rejecting normality when the errors were heavy-tailed. The corresponding probabilities using the Cramer-Von-Mises type functional (7) were better, about 90% and 79%. However, the power to detect non-normal random effects in the random slope/intercept models was disappointingly small, generally

below 20% (and sometimes close to 5%) for $n = 50 \times 5$ and $\alpha = 0.05$. In random effects models having normal effects but non-normal errors, power was reasonable (40% to 70% at $\alpha = 0.05$) but not as large as for the AR-1 models.

To address the lack of power with the marginal residuals in detecting non-normal random effects, we propose using the resampling technique applied to the standardized BLUPs recommended by Lange & Ryan (1989). Tables 4 through 7 present simulation results for the same six random slope/intercept models described in the previous paragraph using the projections described by (10). In general, tests were slightly conservative when the random effects were normal. When the random effects were not normal, the power to detect non-normality was somewhat better than for the omnibus test presented in Tables 2 and 3. At $\alpha = 0.05$, the Kolmogorov-Smirnov test had about 47% power to detect a skewed intercept, 41% to detect a skewed slope, 24% to detect a heavy-tailed intercept, and 19% to detect a heavy-tailed slope. At $\alpha = 0.05$, the Cramer-Von-Mises test had about 64% power to detect a skewed intercept, 53% power to detect a skewed slope, 36% to detect a heavy-tailed intercept, and 29% to detect a heavy-tailed slope. It is interesting to note that the standardized BLUPs had much greater power than the marginal residuals to detect a binary random effect, over 60% compared with 2% for the Cramer-Von-Mises test at $\alpha = 0.05$! When the random effects were normal, but the errors were not normal, the rejection probabilities were close to their nominal values; thus, standardized BLUPs appear to be effective in distinguishing between non-normal random effects and non-normal errors. Other authors have raised concerns that the methods of Lange & Ryan

(1989) or Houseman et al. (2003a) may fail to distinguish between non-normal random effects and non-normal errors (Verbeke & Molenberghs, 2000; Agresti et al., 2003), but we have found that their concerns may be unwarranted. However, we remark that sample sizes much larger than $n = 50 \times 5$ may be required for adequate power in testing the hypothesis of normal random effects. The lack of power at modest sample sizes may result from the fact that the random effects are never observed directly, but rather must be inferred from the appropriate posterior distribution given y.

As mentioned in Section 2, expression (10) requires knowledge of Δ and V_h, which in practice must be estimated using $\widehat{\theta}_n$. We examined the effect of assuming that expression (10) is a known matrix when the null hyptothesis is true. We used the six random effects models described previously to simulate the rejection probabilities for a resampling technique that did not use the correction proposed in Section 2. As Tables 8 and 9 show, the Kolmogorov-Smirnov test may be equally conservative whether or not the correction is used. However, the Cramer-Von-Mises test produces rejection probabilities that are closer to their nominal values when the correction is used. We also examined whether their was any improvement in the rejection probabilities when the sample size was doubled. We conducted a set of simulations using $N = 100$ instead of $N = 50$; the results also appear in Tables 8 and 9. The Kolmogorov-Smirnov tests showed some very modest improvement, but it is difficult to see this from the tables. The improvement is easer to see in Figures 1 through 3, which illustrate Q-Q plots of the simulated P-values against their hypothesized uniform distribution. Each

plot also depicts a band obtained by simulating 100 Q-Q plots, each obtained from a sample of 500 uniform variables (since 500 simulated data sets were used in each case studied). The distribution of simulated P-values is closer to the band of uniform Q-Q plots for $N = 100$ than for $N = 50$. Note that the figures also show that the marginal residuals and the Cramer-Von-Mises tests (using the correction) produce uniform P-values when the null hypothesis is true.

Finally, we remark that the Cramer-Von-Mises test performs better than the Kolmogorov-Smirnov test in every respect. Under the null hypothesis, the Cramer-Von-Mises P-values are closer to their nominal values, and the Cramer-Von-Mises test appears to be more powerful under every alternative considered. This is to be expected, as expression (7) averages differences over the entire range of $\mathcal{X}$, unlike expression (6) which simply selects the largest difference. For this reason, we recommend the Cramer-Von-Mises test over the Kolmogorov-Smirnov test.

5 Applications

In this section we present two applications of our proposed resampling technique.

5.1 Pig Weights

Diggle et al. (2002) present data on the weights (in kilograms) of 48 pigs measured in nine successive weeks. The authors motivated a random slope-intercept model for the data. That is, the weight y_{ht} of pig h at time t can be modeled as

$$y_{ht} = \beta_{01} + a_{h1} + (\beta_{02} + a_{h2})\, t + \varepsilon_{ht},$$

where $\beta_0^T = (\beta_{01}, \beta_{02})$, $\gamma_0^T = (\gamma_{01}, \gamma_{02}, \gamma_{03}, \gamma_{04})$, $a_h^T = (a_{h1}, a_{h2})$, $\varepsilon_{ht} \sim N_1(0, \gamma_{04})$, $a_h \sim N_2(0, \Delta)$, and

$$\Delta = \begin{bmatrix} \gamma_{01} & \gamma_{03} \\ \gamma_{03} & \gamma_{02} \end{bmatrix}.$$

Such a design fits within the class of models described by Theorem 5. Figure 4 displays normal Q-Q plots both for three types of residuals (rotated marginal, and standardized BLUPs for the random slope and intercept), superimposed upon resampled processes using (31). The rotated marginal residuals and the random intercepts appear to be normally distributed, but there is slight evidence that the random slopes may not be normally distributed. P-values were computed as the fraction of 1000 resampled processes whose maximum distance from $\Phi(x)$ over the interval $(-2, 2)$ was greater than the corresponding observed value. For the rotated marginal residual, the Cramer-Von-Mises P-value was 0.20; for the random intercept BLUP, the P-value was 0.19, and for the random slope BLUP, the P-value was 0.13. The Kolmogorov-Smirnov P-values were similar for the marginal residual and random intercept, but was significant ($P = 0.01$) for the random slope. Since the Cramer-Von-Mises test appears to be more reliable in general, we conclude that the errors and random effects are normally distributed for the pig weight data.

5.2 Pollen Counts

Stark et al. (1997) and Brumback et al. (2000) described pollen counts associated with meteorologic data. Between 1991 and 1994, pollen counts were collected seven days a week during the pollen season in Kalamazoo, MI. While these papers

present a predictive model based on Poisson regression, we use the 1991 data to illustrate the behavior of our methods when they are applied to non-normal time series data.

We present two analyses. Both employ the model described by Theorem 6, with a design matrix composed of an intercept and six covariates: *rain* (1 if there were at least 3 hours of steady rain or brief but intense rain, 0 othersie), *day* (day in season), $\ln(day)$, *wind* (wind speed in knots), *temp trend* (a smooth curve fit to daily temperatures in °F) and *temp resid* (residuals of *temp trend*). The choice of covariates is motivated by Stark et al. (1997). In the first analysis we use the raw pollen counts as the outcome. In the second analysis, we use the square root of the pollen counts, since the square root is the variance-stabilizing transformation for the Poisson distribution.

Figure 5 illustrates the Q-Q plots for each analysis, using the rotated marginal residuals. It is clear from the graphical displays that failing to transform the pollen counts results in a residual vector that deviates substantially from normality. On the other hand, the square root transformation may be adequate to transform the residuals to normality. Note that the corresponding Cramer-Von-Mises P-values, computed as described in Section 5.1, were 0.17 for the transformed counts and less than 0.01 for the untransformed counts.

6 Conclusions

Our proposed methodology makes use of *rotated* residuals, formed as the product of the Cholesky decomposition of the estimated marginal variance with the

estimated residual vector, to construct an empirical CDF and pointwise standard errors. It thereby accommodates natural graphical display techniques. Furthermore, we present a resampling technique that provides a global test and renders a graphical depiction of global fit. The theoretical justification for our proposed methods involves technical details that are motivated by Lange & Ryan (1989), and we extend the methodology presented by these authors to address global stochastic behavior of the ECDF of standardized BLUPs.

The simulations we have conducted suggest that our proposed method enjoys good properties in a variety of circumstances. It appears to work well for models having independent units of sampling (clustered data) and for many models in which all observations are correlated. Additionally, our method displays adequate sensitivity to non-normal marginal errors. However, when only the random effect components of a mixed model are non-normal, large sample sizes may be required for adequate power.

An issue that arises in distribution diagnostics is whether slight deviations from normality lead to substantially different conclusions. For a large enough sample size, a test will be able to detect small deviations from normality that have little impact on the scientific question of interest. At the very least, one might check the impact of normality assumptions using methods with less stringent distributional assumptions, such as GEEs or quantile regression, or by considering transformations of the outcome variable. The graphical methods we propose here can be useful in suggesting which assumptions should be relaxed, and therefore which alternative strategies may be preferable. For example, the suggestion of

a heavy-tailed marginal error might lead one to consider median regression. For a case study that uses our methods to direct the analysis, see Houseman et al. (2003c).

Extensions of our theory to the generalized linear mixed model setting would be be of interest. For example, a graphical method for detection overdispersion in Poisson models would have great utility. Efforts towards this goal could be directed in developing generalizations to rotated "working residuals" (McCullagh & Nelder, 1989), or constructing a stochastic process that serves as a summary statistic for detecting non-normal random effects.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: John Wiley and Sons, 2nd ed.

Agresti, A., Caffo, B. & Ohman-Strickland, P. (2003). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies.

Brumback, B. A., Ryan, L. M., Schwartz, J. D., Neas, L. M., Stark, P. C. & Burge, H. A. (2000). Transitional regression models, with application to environmental time series. *Journal of the American Statistical Association* **95**, 16–27.

Coull, B. A., Hobert, J. P., Ryan, L. M. & Holmes, L. B. (2001). Crossed

random effects models for multiple binary outcomes in a study of teratogenesis. *Journal of the American Statistical Association* **96**, 1194–1204.

Davidson, J. (1994). *Stochastic Limit Theory.* New York: Oxford University Press.

Dawson, K. S., Gennings, C. & Carter, W. H. (1997). Two graphical techniques ueful in detecting correlation structure in repeated measures data. *The American Statistician* **51**, 275–283.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data.* Oxford: Oxford University Press, 2nd ed.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **22**, 139–153.

Fraccaro, R., Hyndman, R. J. & Veevers, A. (2000). Residual diagnostic plots for checking for model mis-specification in time series regression. *Australian and New Zealand Journal of Statistics* **42**, 463–477.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **60**, 497–521.

Houseman, E. A., Ryan, L. M. & Coull, B. A. (2003a). Cholesky residuals

for assessing normal errors in a linear model with correlated outcomes. Under consideration by the Journal of American Statistical Association.

Houseman, E. A., Ryan, L. M. & Coull, B. A. (2003b). Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. Tech. rep., Harvard School of Public Health.

Houseman, E. A., Ryan, L. M. & Coull, B. A. (2003c). Distribution diagnostics for linear models with correlated outcomes: Applications to environmental time series. Tech. rep., Harvard School of Public Health.

Jiang, J. M. (2001). Goodness-of-fit tests for mixed model diagnostics. *The Annals of Statistics* **29**, 1137–1164.

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

Lange, N. & Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics* **17**, 624–642.

Lin, D. Y., Wei, L. J. & Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58**, 1–12.

Longford, N. T. (2001). Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society Series A-Statistics in Society* **164**, 259–273.

Louis, T. A. (1988). General methods for analysing repeated measures. *Statistics in Medicine* **7**, 29–45.

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

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley Series in Probability and Statistics.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Ripley, B. D. (2001). Modeling spatial patterns. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **39**, 172–212.

Searle, S. R. (1970). Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrics* **26**, 505–524.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley series in probability and mathematical statistics.

Stark, P. C., Ryan, L. M., McDonald, J. L. & Burge, H. A. (1997). Using meteorologic data to predict daily ragweed pollen levels. *Aerobiologia* **13**, 177–184.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer Verlag.

Zimmerman, D. L. (2000). Viewing the correlation structure of longitudinal data through a prism. *The American Statistician* **54**, 310–318.

Table 1: Simulations for AR-1 Time Series Regression

Kolmogorov-Smirnov Test

Distribution[1]	Nominal α	Fraction Rejected[2] Estimated	95 % CI	Quantile Comparisons[3] $p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Normal	0.50	0.50	(0.473, 0.534)	0.036	0.036
	0.10	0.09	(0.073, 0.108)	0.050	0.050
	0.05	0.05	(0.037, 0.064)	0.054	0.055
	0.025	0.02	(0.013, 0.031)	0.058	0.059
Skewed	0.50	0.99	(0.986, 0.997)	0.067	0.036
	0.10	0.89	(0.872, 0.911)	0.086	0.050
	0.05	0.82	(0.795, 0.842)	0.093	0.055
	0.025	0.73	(0.700, 0.755)	0.098	0.059
Heavy-Tailed	0.50	0.95	(0.936, 0.963)	0.064	0.036
	0.10	0.76	(0.729, 0.782)	0.102	0.050
	0.05	0.68	(0.649, 0.707)	0.125	0.055
	0.025	0.59	(0.563, 0.624)	0.156	0.059

Cramer-Von-Mises Test

Distribution[1]	Nominal α	Fraction Rejected[2] Estimated	95 % CI	Quantile Comparisons[3] $p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Normal	0.50	0.51	(0.476, 0.537)	2.08E-04	2.04E-04
	0.10	0.08	(0.068, 0.103)	3.98E-04	4.15E-04
	0.05	0.04	(0.032, 0.057)	4.98E-04	5.05E-04
	0.025	0.02	(0.014, 0.032)	5.59E-04	5.95E-04
Skewed	0.50	1.00	(0.991, 0.999)	1.07E-03	2.04E-04
	0.10	0.95	(0.932, 0.960)	1.80E-03	4.14E-04
	0.05	0.90	(0.879, 0.916)	2.11E-03	5.04E-04
	0.025	0.86	(0.841, 0.884)	2.37E-03	5.94E-04
Heavy-Tailed	0.50	0.97	(0.959, 0.980)	1.00E-03	2.04E-04
	0.10	0.84	(0.820, 0.865)	3.32E-03	4.15E-04
	0.05	0.79	(0.759, 0.810)	5.53E-03	5.04E-04
	0.025	0.73	(0.703, 0.758)	8.47E-03	5.94E-04

1000 simulated analyses for AR-1 regression models under various error distributions. For each case, $X = (1, u)$ where u is a vector of uniform variables, $n = 250$, $\beta = (10, .5)^T$, and $\gamma = (\sigma^2, \rho) = (1, .5)$.

[1]Skewed variables were standardized χ^2_3 and heavy-tailed variables were standardized t_3.
[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.
[3]$\mathcal{J}$ represents the functional described in (6) or (7), with $\mathcal{X} = (-2.5, 2.5)$ and 1000 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (6) or (7) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (6) or (7) estimated over all simulations.

Table 2: Simulations for Random Slope/Intercept Models: Kolmogorov-Smirnov Test on Marginal Residuals

Distribution[1]	Nominal α	Fraction Rejected[2] Estimated	95 % CI	Quantile Comparisons[3] $p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Normal	0.50	0.51	$(0.463, 0.551)$	0.037	0.037
Random Effects	0.10	0.10	$(0.073, 0.125)$	0.050	0.051
and	0.05	0.05	$(0.036, 0.075)$	0.055	0.055
Normal Errors	0.025	0.03	$(0.015, 0.044)$	0.059	0.060
Skewed	0.50	0.69	$(0.652, 0.732)$	0.042	0.037
Random Effects	0.10	0.26	$(0.228, 0.305)$	0.059	0.051
and	0.05	0.16	$(0.131, 0.195)$	0.065	0.055
Normal Errors	0.025	0.10	$(0.079, 0.132)$	0.069	0.060
Heavy-Tailed	0.50	0.66	$(0.623, 0.705)$	0.041	0.037
Random Effects	0.10	0.21	$(0.181, 0.253)$	0.058	0.051
and	0.05	0.13	$(0.100, 0.158)$	0.064	0.055
Normal Errors	0.025	0.08	$(0.059, 0.107)$	0.071	0.060
Binary	0.50	0.63	$(0.592, 0.676)$	0.040	0.037
Random Effects	0.10	0.11	$(0.088, 0.143)$	0.051	0.051
and	0.05	0.05	$(0.038, 0.078)$	0.055	0.055
Normal Errors	0.025	0.02	$(0.014, 0.042)$	0.060	0.060
Normal	0.50	0.97	$(0.958, 0.986)$	0.058	0.037
Random Effects	0.10	0.72	$(0.683, 0.761)$	0.079	0.051
and	0.05	0.59	$(0.549, 0.635)$	0.085	0.055
Skewed Errors	0.025	0.45	$(0.408, 0.495)$	0.089	0.060
Normal	0.50	0.92	$(0.895, 0.942)$	0.053	0.037
Random Effects	0.10	0.57	$(0.525, 0.612)$	0.078	0.051
and	0.05	0.44	$(0.396, 0.483)$	0.088	0.056
Heavy-Tailed Errors	0.025	0.37	$(0.326, 0.410)$	0.097	0.060

500 simulated analyses for AR-1 regression models under various error distributions. For each case, $X = (1, u, obs)$ where u is a vector of uniform variables and obs is a vector consisting of the observation numbers within each cluster, $n = 250$ (50 clusters of 5 observations), $\beta = (10, .5, 0)^T$, $\sigma^2 = 1$, $\Delta = \text{diag}(4, .25)$, and $Z = (1, obs)$.

[1]Skewed variables were standardized χ^2_3 and heavy-tailed variables were standardized t_3.
[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.
[3]$\mathcal{J}$ represents the functional described in (6), with $\mathcal{X} = (-2.5, 2.5)$ and 500 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (6) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (6) estimated over all simulations.

Table 3: Simulations for Random Slope/Intercept Models: Cramer-Von-Mises Test on Marginal Residuals

	Nominal	Fraction Rejected[2]		Quantile Comparisons[3]	
Distribution[1]	α	Estimated	95 % CI	$p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Normal	0.50	0.50	(0.457, 0.545)	2.10E-04	2.09E-04
Random Effects	0.10	0.10	(0.075, 0.128)	4.24E-04	4.27E-04
and	0.05	0.05	(0.033, 0.071)	4.96E-04	5.20E-04
Normal Errors	0.025	0.03	(0.015, 0.044)	6.13E-04	6.13E-04
Skewed	0.50	0.71	(0.672, 0.751)	3.06E-04	2.10E-04
Random Effects	0.10	0.28	(0.247, 0.326)	6.73E-04	4.28E-04
and	0.05	0.20	(0.164, 0.234)	7.97E-04	5.20E-04
Normal Errors	0.025	0.13	(0.105, 0.165)	9.59E-04	6.14E-04
Heavy-Tailed	0.50	0.68	(0.643, 0.725)	2.80E-04	2.11E-04
Random Effects	0.10	0.25	(0.216, 0.292)	6.46E-04	4.29E-04
and	0.05	0.16	(0.134, 0.199)	8.25E-04	5.23E-04
Normal Errors	0.025	0.12	(0.091, 0.147)	1.03E-03	6.15E-04
Binary	0.50	0.66	(0.621, 0.703)	2.59E-04	2.10E-04
Random Effects	0.10	0.13	(0.105, 0.165)	4.76E-04	4.26E-04
and	0.05	0.07	(0.048, 0.091)	5.39E-04	5.20E-04
Normal Errors	0.025	0.02	(0.014, 0.042)	6.01E-04	6.12E-04
Normal	0.50	0.99	(0.974, 0.994)	6.92E-04	2.10E-04
Random Effects	0.10	0.81	(0.773, 0.842)	1.34E-03	4.27E-04
and	0.05	0.69	(0.649, 0.730)	1.61E-03	5.20E-04
Skewed Errors	0.025	0.58	(0.537, 0.624)	1.85E-03	6.12E-04
Normal	0.50	0.93	(0.902, 0.947)	6.13E-04	2.10E-04
Random Effects	0.10	0.72	(0.678, 0.757)	1.68E-03	4.27E-04
and	0.05	0.59	(0.549, 0.635)	2.20E-03	5.20E-04
Heavy-Tailed Errors	0.025	0.52	(0.473, 0.561)	2.94E-03	6.14E-04

500 simulated analyses for random slope/intercept models under various error distributions. For each case, $X = (1, u, obs)$ where u is a vector of uniform variables and obs is a vector consisting of the observation numbers within each cluster, $n = 250$ (50 clusters of 5 observations), $\beta = (10, .5, 0)^T$, $\sigma^2 = 1$, $\Delta = \text{diag}(4, .25)$, and $Z = (1, obs)$.

[1]Skewed variables were standardized χ^2_3 and heavy-tailed variables were standardized t_3.
[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.
[3]$\mathcal{J}$ represents the functional described in (7), with $\mathcal{X} = (-2.5, 2.5)$ and 500 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (7) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (7) estimated over all simulations.

Table 4: Simulations for Random Slope/Intercept Models: Kolmogorov-Smirnov Test on Predicted Random Intercepts

Distribution[1]	Nominal α	Fraction Rejected[2] Estimated	95 % CI	Quantile Comparisons[3] $p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Normal	0.50	0.38	(0.339, 0.424)	0.077	0.083
Random Effects	0.10	0.06	(0.041, 0.082)	0.105	0.115
and	0.05	0.03	(0.017, 0.047)	0.116	0.126
Normal Errors	0.025	0.01	(0.006, 0.026)	0.128	0.136
Skewed	0.50	0.91	(0.884, 0.934)	0.124	0.083
Random Effects	0.10	0.60	(0.554, 0.639)	0.172	0.116
and	0.05	0.47	(0.424, 0.511)	0.186	0.127
Normal Errors	0.025	0.36	(0.320, 0.404)	0.199	0.137
Heavy-Tailed	0.50	0.68	(0.637, 0.719)	0.099	0.083
Random Effects	0.10	0.34	(0.302, 0.385)	0.162	0.116
and	0.05	0.24	(0.205, 0.280)	0.202	0.127
Normal Errors	0.025	0.18	(0.153, 0.221)	0.235	0.138
Binary	0.50	1.00	(0.994, 1.000)	0.146	0.083
Random Effects	0.10	0.92	(0.897, 0.944)	0.179	0.115
and	0.05	0.81	(0.775, 0.843)	0.191	0.126
Normal Errors	0.025	0.67	(0.627, 0.709)	0.202	0.136
Normal	0.50	0.38	(0.343, 0.428)	0.077	0.083
Random Effects	0.10	0.06	(0.042, 0.085)	0.109	0.115
and	0.05	0.04	(0.026, 0.061)	0.119	0.126
Skewed Errors	0.025	0.02	(0.011, 0.036)	0.130	0.136
Normal	0.50	0.38	(0.337, 0.422)	0.077	0.083
Random Effects	0.10	0.05	(0.031, 0.068)	0.106	0.115
and	0.05	0.03	(0.015, 0.044)	0.113	0.126
Heavy-Tailed Errors	0.025	0.01	(0.006, 0.026)	0.127	0.137

500 simulated analyses for random slope/intercept models under various error distributions. For each case, $X = (1, u, obs)$ where u is a vector of uniform variables and obs is a vector consisting of the observation numbers within each cluster, $n = 250$ (50 clusters of 5 observations), $\beta = (10, .5, 0)^T$, $\sigma^2 = 1$, $\Delta = \text{diag}(4, .25)$, and $Z = (1, obs)$.

[1]Skewed variables were standardized χ^2_3 and heavy-tailed variables were standardized t_3.
[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.
[3]$\mathcal{J}$ represents the functional described in (6), with $\mathcal{X} = (-2.5, 2.5)$ and 500 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (6) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (6) estimated over all simulations.

Table 5: Simulations for Random Slope/Intercept Models: Cramer-Von-Mises Test on Predicted Random Intercepts

Distribution[1]	Nominal α	Fraction Rejected[2] Estimated	95 % CI	Quantile Comparisons[3] $p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Normal	0.50	0.54	(0.497, 0.584)	1.13E-03	1.08E-03
Random Effects	0.10	0.11	(0.086, 0.141)	2.34E-03	2.28E-03
and	0.05	0.05	(0.034, 0.073)	2.71E-03	2.82E-03
Normal Errors	0.025	0.02	(0.011, 0.036)	3.17E-03	3.38E-03
Skewed	0.50	0.96	(0.946, 0.979)	3.64E-03	1.09E-03
Random Effects	0.10	0.76	(0.722, 0.797)	7.21E-03	2.30E-03
and	0.05	0.64	(0.598, 0.682)	8.61E-03	2.84E-03
Normal Errors	0.025	0.55	(0.511, 0.598)	1.05E-02	3.43E-03
Heavy-Tailed	0.50	0.80	(0.762, 0.832)	2.12E-03	1.09E-03
Random Effects	0.10	0.45	(0.404, 0.491)	6.62E-03	2.34E-03
and	0.05	0.36	(0.324, 0.408)	1.05E-02	2.91E-03
Normal Errors	0.025	0.29	(0.252, 0.332)	1.65E-02	3.53E-03
Binary	0.50	1.00	(0.994, 1.000)	5.84E-03	1.08E-03
Random Effects	0.10	0.99	(0.983, 0.998)	8.72E-03	2.28E-03
and	0.05	0.97	(0.958, 0.986)	9.52E-03	2.81E-03
Normal Errors	0.025	0.94	(0.920, 0.961)	1.06E-02	3.38E-03
Normal	0.50	0.55	(0.507, 0.594)	1.15E-03	1.08E-03
Random Effects	0.10	0.10	(0.080, 0.134)	2.27E-03	2.28E-03
and	0.05	0.04	(0.026, 0.061)	2.67E-03	2.82E-03
Skewed Errors	0.025	0.02	(0.010, 0.034)	3.15E-03	3.39E-03
Normal	0.50	0.56	(0.515, 0.602)	1.18E-03	1.08E-03
Random Effects	0.10	0.09	(0.070, 0.121)	2.20E-03	2.28E-03
and	0.05	0.04	(0.026, 0.061)	2.64E-03	2.82E-03
Heavy-Tailed Errors	0.025	0.02	(0.008, 0.031)	3.00E-03	3.40E-03

500 simulated analyses for random slope/intercept models under various error distributions. For each case, $X = (1, u, obs)$ where u is a vector of uniform variables and obs is a vector consisting of the observation numbers within each cluster, $n = 250$ (50 clusters of 5 observations), $\beta = (10, .5, 0)^T$, $\sigma^2 = 1$, $\Delta = \text{diag}(4, .25)$, and $Z = (1, obs)$.

[1]Skewed variables were standardized χ^2_3 and heavy-tailed variables were standardized t_3.
[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.
[3]$\mathcal{J}$ represents the functional described in (7), with $\mathcal{X} = (-2.5, 2.5)$ and 500 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (7) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (7) estimated over all simulations.

Table 6: Simulations for Random Slope/Intercept Models: Kolmogorov-Smirnov Test on Predicted Random Slopes

Distribution[1]	Nominal α	Fraction Rejected[2] Estimated	95 % CI	Quantile Comparisons[3] $p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Normal	0.50	0.40	(0.361, 0.446)	0.077	0.083
Random Effects	0.10	0.06	(0.044, 0.087)	0.109	0.115
and	0.05	0.02	(0.008, 0.031)	0.118	0.126
Normal Errors	0.025	0.01	(0.004, 0.023)	0.122	0.137
Skewed	0.50	0.88	(0.850, 0.907)	0.119	0.083
Random Effects	0.10	0.54	(0.495, 0.582)	0.161	0.116
and	0.05	0.41	(0.370, 0.457)	0.174	0.127
Normal Errors	0.025	0.29	(0.251, 0.330)	0.192	0.138
Heavy-Tailed	0.50	0.60	(0.556, 0.641)	0.091	0.084
Random Effects	0.10	0.26	(0.220, 0.297)	0.146	0.117
and	0.05	0.19	(0.162, 0.231)	0.167	0.129
Normal Errors	0.025	0.14	(0.111, 0.171)	0.186	0.140
Binary	0.50	0.93	(0.902, 0.947)	0.118	0.083
Random Effects	0.10	0.53	(0.489, 0.576)	0.152	0.115
and	0.05	0.37	(0.331, 0.416)	0.163	0.126
Normal Errors	0.025	0.24	(0.207, 0.282)	0.172	0.136
Normal	0.50	0.44	(0.402, 0.489)	0.080	0.083
Random Effects	0.10	0.08	(0.059, 0.107)	0.111	0.116
and	0.05	0.04	(0.028, 0.063)	0.124	0.127
Skewed Errors	0.025	0.02	(0.011, 0.036)	0.135	0.137
Normal	0.50	0.38	(0.343, 0.428)	0.077	0.083
Random Effects	0.10	0.07	(0.049, 0.094)	0.105	0.116
and	0.05	0.03	(0.020, 0.051)	0.120	0.127
Heavy-Tailed Errors	0.025	0.01	(0.007, 0.029)	0.128	0.137

500 simulated analyses for random slope/intercept models under various error distributions. For each case, $X = (1, u, obs)$ where u is a vector of uniform variables and obs is a vector consisting of the observation numbers within each cluster, $n = 250$ (50 clusters of 5 observations), $\beta = (10, .5, 0)^T$, $\sigma^2 = 1$, $\Delta = \text{diag}(4, .25)$, and $Z = (1, obs)$.

[1]Skewed variables were standardized χ^2_3 and heavy-tailed variables were standardized t_3.
[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.
[3]$\mathcal{J}$ represents the functional described in (6), with $\mathcal{X} = (-2.5, 2.5)$ and 500 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (6) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (6) estimated over all simulations.

Table 7: Simulations for Random Slope/Intercept Models: Cramer-Von-Mises Test on Predicted Random Slopes

Distribution[1]	Nominal α	Fraction Rejected[2] Estimated	95 % CI	Quantile Comparisons[3] $p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Normal	0.50	0.56	(0.521, 0.608)	1.16E-03	1.08E-03
Random Effects	0.10	0.11	(0.089, 0.145)	2.31E-03	2.29E-03
and	0.05	0.05	(0.036, 0.075)	2.91E-03	2.84E-03
Normal Errors	0.025	0.02	(0.012, 0.039)	3.28E-03	3.44E-03
Skewed	0.50	0.94	(0.918, 0.959)	3.08E-03	1.09E-03
Random Effects	0.10	0.66	(0.623, 0.705)	6.49E-03	2.32E-03
and	0.05	0.53	(0.485, 0.573)	7.49E-03	2.90E-03
Normal Errors	0.025	0.44	(0.398, 0.485)	8.19E-03	3.54E-03
Heavy-Tailed	0.50	0.74	(0.701, 0.778)	1.84E-03	1.11E-03
Random Effects	0.10	0.36	(0.318, 0.402)	5.31E-03	2.42E-03
and	0.05	0.29	(0.251, 0.330)	7.35E-03	3.06E-03
Normal Errors	0.025	0.23	(0.196, 0.269)	1.13E-02	3.81E-03
Binary	0.50	0.99	(0.977, 0.996)	3.37E-03	1.08E-03
Random Effects	0.10	0.77	(0.731, 0.804)	5.56E-03	2.27E-03
and	0.05	0.64	(0.596, 0.680)	6.35E-03	2.81E-03
Normal Errors	0.025	0.49	(0.451, 0.539)	7.20E-03	3.38E-03
Normal	0.50	0.59	(0.545, 0.632)	1.22E-03	1.08E-03
Random Effects	0.10	0.11	(0.089, 0.145)	2.38E-03	2.29E-03
and	0.05	0.06	(0.041, 0.082)	3.01E-03	2.85E-03
Skewed Errors	0.025	0.04	(0.026, 0.061)	3.77E-03	3.44E-03
Normal	0.50	0.56	(0.513, 0.600)	1.14E-03	1.08E-03
Random Effects	0.10	0.11	(0.084, 0.139)	2.26E-03	2.29E-03
and	0.05	0.05	(0.036, 0.075)	2.73E-03	2.85E-03
Heavy-Tailed Errors	0.025	0.02	(0.011, 0.036)	3.31E-03	3.45E-03

500 simulated analyses for random slope/intercept models under various error distributions. For each case, $X = (1, u, obs)$ where u is a vector of uniform variables and obs is a vector consisting of the observation numbers within each cluster, $n = 250$ (50 clusters of 5 observations), $\beta = (10, .5, 0)^T$, $\sigma^2 = 1$, $\Delta = \text{diag}(4, .25)$, and $Z = (1, obs)$.

[1]Skewed variables were standardized χ^2_3 and heavy-tailed variables were standardized t_3
[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.
[3]$\mathcal{J}$ represents the functional described in (7), with $\mathcal{X} = (-2.5, 2.5)$ and 500 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (7) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (7) estimated over all simulations.

Table 8: Simulations for Random Slope/Intercept Models: Performance of Tests on Random Intercept at Null Hypothesis

Kolmogorov-Smirnov Test

	Nominal	Fraction Rejected[2]		Quantile Comparisons[3]	
Description[1]	α	Estimated	95 % CI	$p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Assumed	0.50	0.42	(0.380, 0.467)	0.078	0.081
P_i Known	0.10	0.06	(0.044, 0.087)	0.105	0.112
	0.05	0.03	(0.018, 0.049)	0.116	0.122
$N = 50$	0.025	0.02	(0.011, 0.036)	0.123	0.131
Estimated	0.50	0.38	(0.339, 0.424)	0.077	0.083
P_i	0.10	0.06	(0.041, 0.082)	0.105	0.115
	0.05	0.03	(0.017, 0.047)	0.116	0.126
$N = 50$	0.025	0.01	(0.006, 0.026)	0.128	0.136
Estimated	0.50	0.43	(0.390, 0.477)	0.056	0.059
P_i	0.10	0.06	(0.042, 0.085)	0.077	0.081
	0.05	0.03	(0.017, 0.047)	0.084	0.089
$N = 100$	0.025	0.02	(0.010, 0.034)	0.090	0.095

Cramer-Von-Mises Test

	Nominal	Fraction Rejected[2]		Quantile Comparisons[3]	
Description[1]	α	Estimated	95 % CI	$p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Assumed	0.50	0.63	(0.586, 0.671)	1.17E-03	1.01E-03
P_i Known	0.10	0.14	(0.113, 0.174)	2.29E-03	2.08E-03
	0.05	0.08	(0.056, 0.103)	2.82E-03	2.54E-03
$N = 50$	0.025	0.04	(0.026, 0.061)	3.20E-03	3.01E-03
Estimated	0.50	0.54	(0.497, 0.584)	1.13E-03	1.08E-03
P_i	0.10	0.11	(0.086, 0.141)	2.34E-03	2.28E-03
	0.05	0.05	(0.034, 0.073)	2.71E-03	2.82E-03
$N = 50$	0.025	0.02	(0.011, 0.036)	3.17E-03	3.38E-03
Estimated	0.50	0.53	(0.491, 0.578)	5.50E-04	5.26E-04
P_i	0.10	0.10	(0.080, 0.134)	1.10E-03	1.08E-03
	0.05	0.04	(0.028, 0.063)	1.27E-03	1.32E-03
$N = 100$	0.025	0.02	(0.010, 0.034)	1.49E-03	1.56E-03

500 simulated analyses for random slope/intercept models under various error distributions. For each case, $X = (1, u, obs)$ where u is a vector of uniform variables and obs is a vector consisting of the observation numbers within each cluster, $n = N \times 5$ (N clusters of 5 observations), $\beta = (10, .5, 0)^T$, $\sigma^2 = 1$, $\Delta = \text{diag}(4, .25)$, and $Z = (1, obs)$.

[1]The first case assumed that the projection matrix P_i mapping the rotated residual to a random effect was known and equal to its predicted value. The second two cases account for the estimation of P_i by the method described at the end of Section 2.

[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.

[3]$\mathcal{J}$ represents the functional described in (7), with $\mathcal{X} = (-2.5, 2.5)$ and 500 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (7) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (7) estimated over all simulations.

Table 9: Simulations for Random Slope/Intercept Models: Performance of Tests on Random Slope at Null Hypothesis

Kolmogorov-Smirnov Test

	Nominal	Fraction Rejected[2]		Quantile Comparisons[3]	
Description[1]	α	Estimated	95 % CI	$p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Assumed	0.50	0.40	(0.361, 0.446)	0.076	0.081
P_i Known	0.10	0.07	(0.051, 0.096)	0.105	0.111
	0.05	0.03	(0.018, 0.049)	0.115	0.122
$N = 50$	0.025	0.01	(0.004, 0.023)	0.123	0.131
Estimated	0.50	0.40	(0.361, 0.446)	0.077	0.083
P_i	0.10	0.06	(0.044, 0.087)	0.109	0.115
	0.05	0.02	(0.008, 0.031)	0.118	0.126
$N = 50$	0.025	0.01	(0.004, 0.023)	0.122	0.137
Estimated	0.50	0.41	(0.365, 0.451)	0.056	0.059
P_i	0.10	0.07	(0.054, 0.101)	0.077	0.081
	0.05	0.04	(0.026, 0.061)	0.084	0.089
$N = 100$	0.025	0.03	(0.018, 0.049)	0.097	0.095

Cramer-Von-Mises Test

	Nominal	Fraction Rejected[2]		Quantile Comparisons[3]	
Description[1]	α	Estimated	95 % CI	$p_{1-\alpha}\mathcal{J}(\widehat{F})$	$E[p_{1-\alpha}\mathcal{J}(F^*)]$
Assumed	0.50	0.59	(0.547, 0.633)	1.15E-03	1.01E-03
P_i Known	0.10	0.12	(0.095, 0.152)	2.22E-03	2.08E-03
	0.05	0.06	(0.044, 0.087)	2.58E-03	2.54E-03
$N = 50$	0.025	0.03	(0.017, 0.047)	3.17E-03	3.00E-03
Estimated	0.50	0.56	(0.521, 0.608)	1.16E-03	1.08E-03
P_i	0.10	0.11	(0.089, 0.145)	2.31E-03	2.29E-03
	0.05	0.05	(0.036, 0.075)	2.91E-03	2.84E-03
$N = 50$	0.025	0.02	(0.012, 0.039)	3.28E-03	3.44E-03
Estimated	0.50	0.53	(0.491, 0.578)	5.41E-04	5.24E-04
P_i	0.10	0.09	(0.072, 0.123)	1.04E-03	1.08E-03
	0.05	0.06	(0.039, 0.080)	1.38E-03	1.32E-03
$N = 100$	0.025	0.03	(0.017, 0.047)	1.63E-03	1.56E-03

500 simulated analyses for random slope/intercept models under various error distributions. For each case, $X = (1, u, obs)$ where u is a vector of uniform variables and obs is a vector consisting of the observation numbers within each cluster, $n = N \times 5$ (N clusters of 5 observations), $\beta = (10, .5, 0)^T$, $\sigma^2 = 1$, $\Delta = \text{diag}(4, .25)$, and $Z = (1, obs)$.

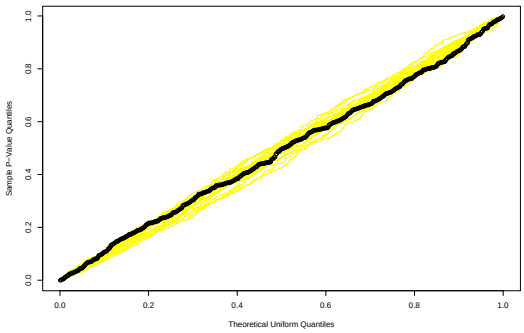
[1]The first case assumed that the projection matrix P_i mapping the rotated residual to a random effect was known and equal to its predicted value. The second two cases account for the estimation of P_i by the method described at the end of Section 2.

[2]Fraction of simulations where null hypothesis of normality was rejected using $p_{1-\alpha}\mathcal{J}(F^*)$ as critical value.

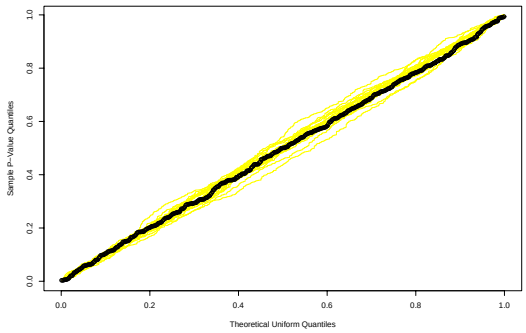
[3]$\mathcal{J}$ represents the functional described in (7), with $\mathcal{X} = (-2.5, 2.5)$ and 500 resampled values. $E[p_{1-\alpha}\mathcal{J}(F^*)]$ represents the average $1-\alpha$ quantile of the right hand side of (7) over all simulations; $p_{1-\alpha}\mathcal{J}(\widehat{F})$ represents the quantile of the left hand side of (7) estimated over all simulations.

Figure 1: Q-Q Plots for P-Values Under Null Hypothesis

Marginal Residuals

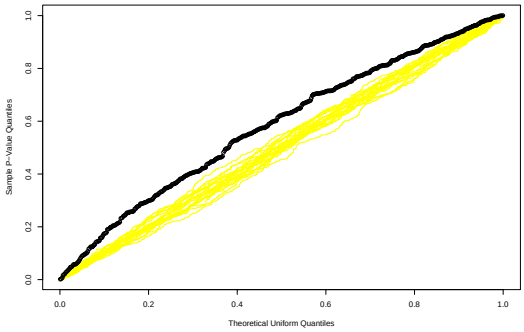


a) Kolmogorov-Smirnov P-Values

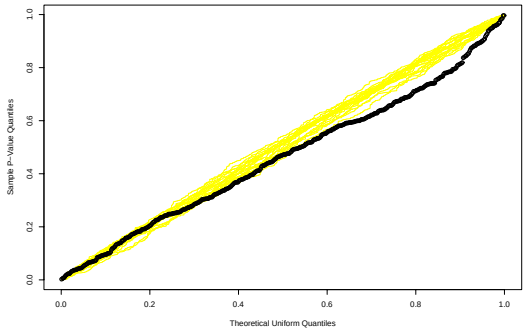


b) Cramer-Von-Mises P-Values

Random Intercept

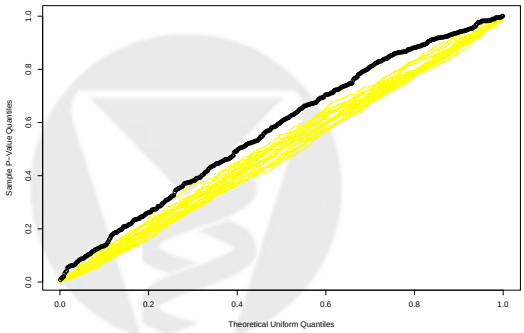


a) Kolmogorov-Smirnov P-Values

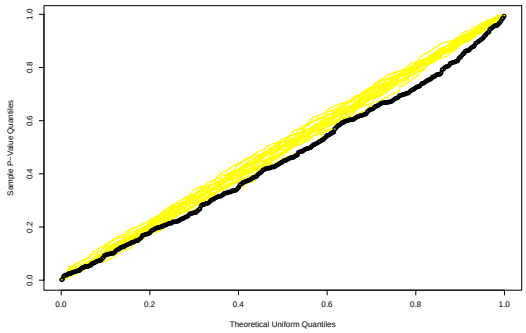


b) Cramer-Von-Mises P-Values

Random Slope



a) Kolmogorov-Smirnov P-Values

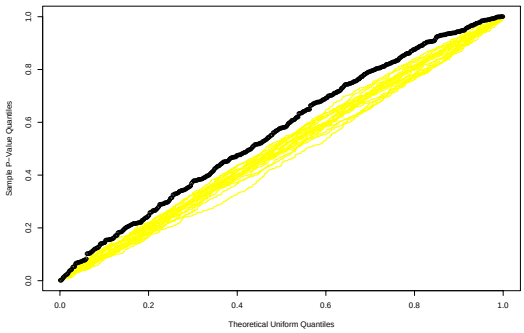


b) Cramer-Von-Mises P-Values

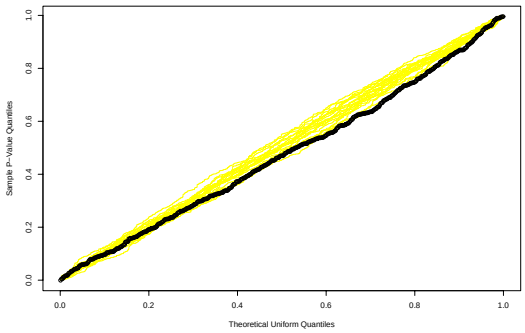
Quantile-quantile plots of P-values under a null hypothesis of normal random effects. Each case was constructed from 500 simulated data sets, and each plot contains a band composed of 100 yellow lines, each representing the Q-Q plot constructed from 500 uniform variables. Thus the band suggests a region through which a Q-Q plot of P-values constructed from 500 simulated data sets could pass if the P-values were truly drawn from a uniform distribution. Each simulation was obtained using the methods described in Section 2 and data sets described in Section 4. Note that $N = 50$.

Figure 2: Q-Q Plots for P-Values Under Null Hypothesis, $N = 100$

Random Intercept

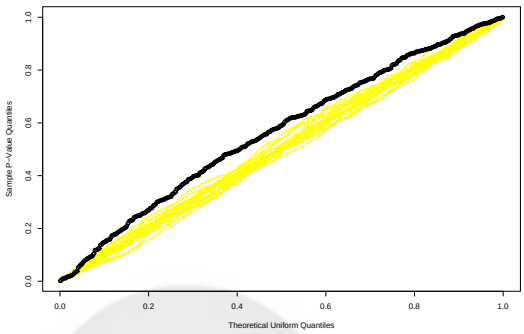


a) Kolmogorov-Smirnov P-Values

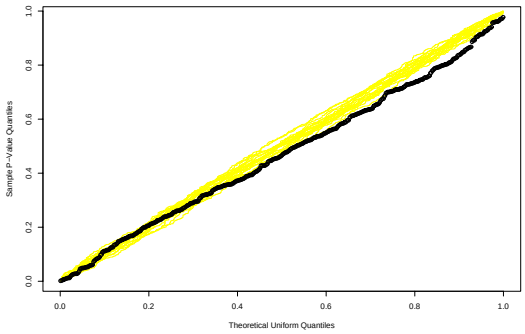


b) Cramer-Von-Mises P-Values

Random Slope



a) Kolmogorov-Smirnov P-Values

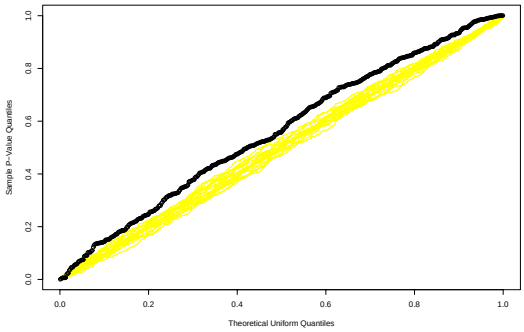


b) Cramer-Von-Mises P-Values

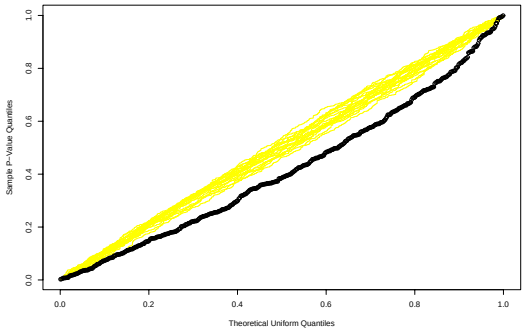
Quantile-quantile plots of P-values under a null hypothesis of normal random effects. Each case was constructed from 500 simulated data sets, and each plot contains a band composed of 100 yellow lines, each representing the Q-Q plot constructed from 500 uniform variables. Thus the band suggests a region through which a Q-Q plot of P-values constructed from 500 simulated data sets could pass if the P-values were truly drawn from a uniform distribution. Each simulation was obtained using the methods described in Section 2 and data sets described in Section 4. Note that $N = 100$.

Figure 3: Q-Q Plots for P-Values Under Null Hypothesis, Without Correcting for Estimation of P_i

Random Intercept

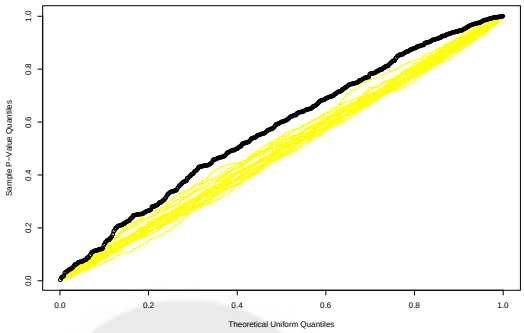


a) Kolmogorov-Smirnov P-Values

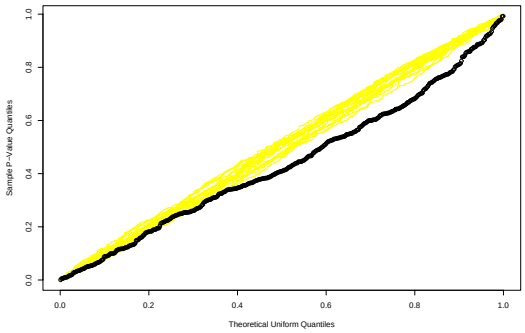


b) Cramer-Von-Mises P-Values

Random Slope



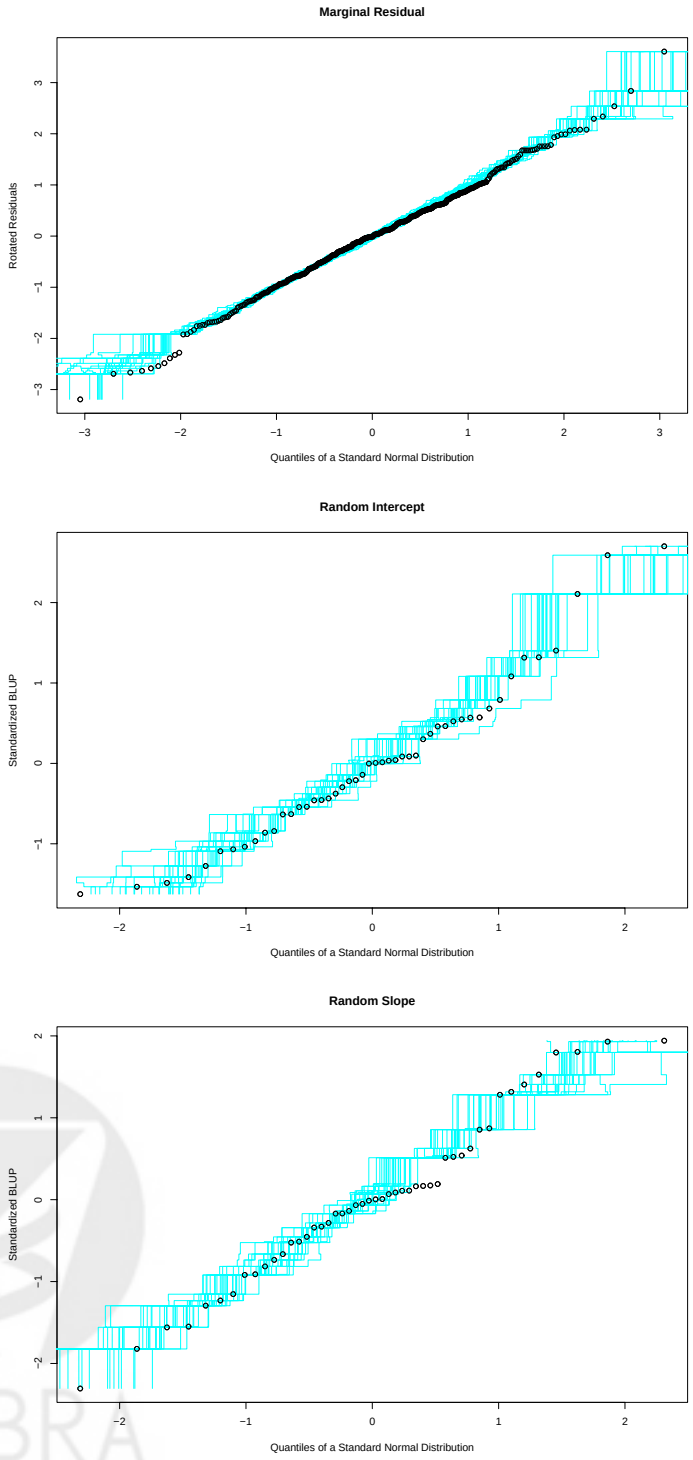
a) Kolmogorov-Smirnov P-Values



b) Cramer-Von-Mises P-Values

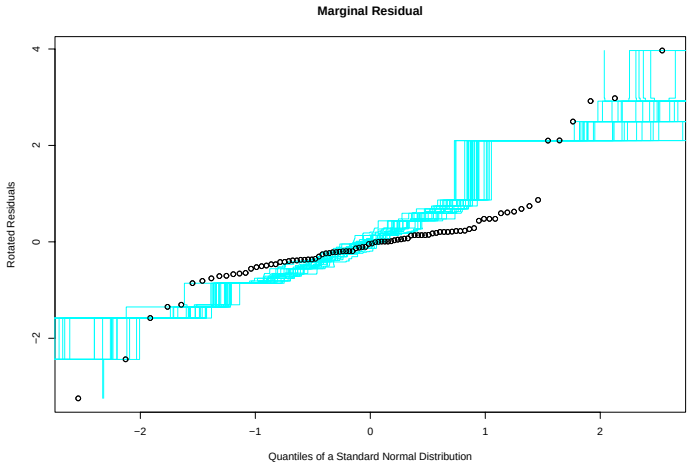
Quantile-quantile plots of P-values under a null hypothesis of normal random effects. Each case was constructed from 500 simulated data sets, and each plot contains a band composed of 100 yellow lines, each representing the Q-Q plot constructed from 500 uniform variables. Thus the band suggests a region through which a Q-Q plot of P-values constructed from 500 simulated data sets could pass if the P-values were truly drawn from a uniform distribution. Each simulation was obtained using the methods described in Section 2, but without correcting for the estimation of the projection matrix P_i, and data sets described in Section 4. Note that $N = 50$.

Figure 4: QQ Plots for Pig Weight Data

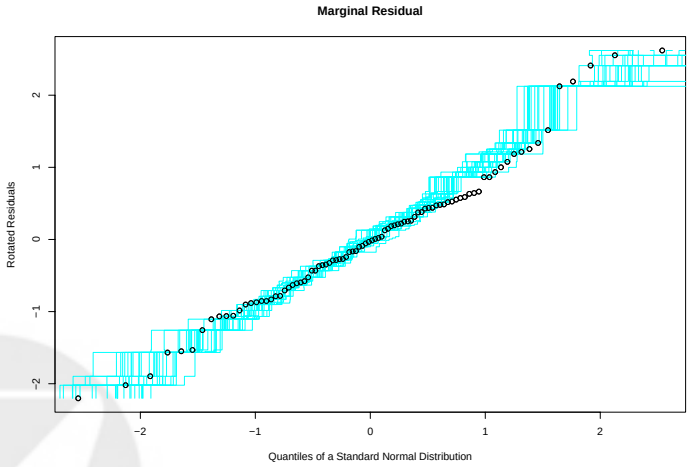


Quantile-quantile plots for a random slope-intercept model applied to the pig weight data from Diggle et al. (2002). Light blue lines represent realizations from the asymptotic distribution under the null hypothesis of normality. Corresponding P-values are (a) 0.13 (Kolmogorov-Smirnov) and 0.20 (Cramer-Von-Mises) for marginal residuals; (b) 0.11 (Kolmogorov-Smirnov) and 0.19 (Cramer-Von-Mises) for random intercepts; and (c) 0.01 (Kolmogorov-Smirnov) and 0.13 (Cramer-Von-Mises) for random slopes.

Figure 5: QQ Plots for Pollen Data



Untransformed Pollen Counts



Square-root Transformed Pollen Counts

Quantile-quantile plots for AR-1 time-series regression model applied to the pollen data from Stark et al. (1997). Light blue lines represent realizations from the asymptotic distribution under the null hypothesis of normality. Corresponding P-values are (a) < 0.01 (Kolmogorov-Smirnov and Cramer-Von-Mises) for transformed counts; (b) 0.08 (Kolmogorov-Smirnov) and 0.17 (Cramer-Von-Mises) for untransformed counts.