Linear Mixed Model Robust Regression

Megan J. Waterman, Jeffrey B. Birch, and Oliver Schabenberger [*]

November 5, 2006

Abstract

Mixed models are powerful tools for the analysis of clustered data and many extensions of the classical linear mixed model with normally distributed response have been established. As with all parametric models, correctness of the assumed model is critical for the validity of the ensuing inference. An incorrectly specified parametric means model may be improved by using a local, or nonparametric, model. Two local models are proposed by a pointwise weighting of the marginal and conditional variance-covariance matrices. However, nonparametric models tend to fit to irregularities in the data and provide fits with high variance. Model robust regression techniques estimate mean response as a convex combination of a parametric and a nonparametric model fits to the data. It is a semiparametric method by which incomplete or incorrect specified parametric models can be improved through adding an appropriate amount of the nonparametric fit. We compare the approximate integrated mean square error of the parametric, nonparametric, and mixed model robust methods via a simulation study, and apply these methods to monthly wind speed data from counties in Ireland.

KEYWORDS: semiparametric, nonparametric, mixed effects, robust

1 INTRODUCTION

The general linear mixed regression model (Laird and Ware, 1982) is commonly expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{1}$$

[*] Megan J. Waterman is employed as a mathematician for the Department of Defense. Email: meganwaterman@earthlink.net. Jeffrey B. Birch is Professor, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. Email: jbbirch@vt.edu. Oliver Schabenberger is a Senior Research Statistician at SAS Institute Inc., Cary, NC. Email: Oliver.Schabenberger@sas.com.

where $\mathbf{Y}$ is an (n x 1) vector of responses, $\mathbf{X}$ and $\mathbf{Z}$ are (n x p) and (n x q) model matrices, $\boldsymbol{\beta}$ is a (p x 1) vector of fixed effects, $\mathbf{b}$ is a (q x 1) vector of random effects, and $\boldsymbol{\epsilon}$ is an (n x 1) vector of random disturbances. We assume that the random effects and errors are normal variates with zero expectation and $Var(\mathbf{b}) = \mathbf{B}$, $Var(\boldsymbol{\epsilon}) = \mathbf{R}$, and $Cov(\mathbf{b}, \boldsymbol{\epsilon}) = \mathbf{0}$.

The estimate of the fixed effects and prediction of the random effects

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \tag{2}$$

$$\widehat{\mathbf{b}} = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \tag{3}$$

are the solutions to the mixed model equations, where $Var(\mathbf{Y}) = \mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{B}\mathbf{Z}'$. The variance-covariance matrix $\mathbf{V}$ is typically unknown and parametrized as $\mathbf{V}(\boldsymbol{\theta})$. After estimating $\boldsymbol{\theta}$ by (restricted) maximum likelihood, ANOVA, or some other method, the estimated variance-covariance matrix $\widehat{\mathbf{V}} = \mathbf{V}(\widehat{\boldsymbol{\theta}})$ is substituted in (2), (3), and in other expressions that depend on $\boldsymbol{\theta}$. We suppress this dependency for brevity.

We are concerned with linear mixed modeling for clustered data. Let i=1,.., s index clusters and write (1) as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i. \tag{4}$$

Quantities arising from the marginal distribution of $\mathbf{Y}_i$ will be referred to as population average (PA) quantities; those arising from the distribution of $\mathbf{Y}_i|\mathbf{b}_i$ are referred to as conditional or cluster-specific.

In this paper, the effects of model misspecification on the linear mixed model are analyzed. Two local methods, developed to alleviate the bias problem of misspecified parametric models, will be presented. Although the local methods result in less bias, there is a tendency for overfitting. The model robust mixed model, a hybrid combination of the parametric and local mixed models, is shown to minimize the integrated mean square error when compared to the parametric and local methods, while retaining important features of the data.

This paper is organized as follows. Sections 2 and 3 introduce the nonparametric and semiparametric (model robust) mixed models. The asymptotic results for the model robust mixed model estimate of the mean function using the theoretically optimal and a data-driven estimate of the mixing parameter are discussed in Section 4. Section 5 offers a simulation study to compare approximate integrated mean square errors of the parametric, nonparametric, and model robust methods; these methods are applied in

Section 6 to data resulting from monthly wind speed readings taken from twelve locations in Ireland. Section 7 contains a discussion of our results.

2 THE LOCAL MIXED MODEL

Let Y_{ij} denote the j^{th} observation from the i^{th} cluster for j=1,..,n_i, i=1,..,s, and $\sum_{i=1}^{s} n_i = n$. The vectors $\mathbf{x}'_{ij}$ and $\mathbf{z}'_{ij}$ are vectors from row j of $\mathbf{X}_i$ and $\mathbf{Z}_i$, the model matrices associated with cluster i for the fixed and random effects, respectively. In this paper, we consider polynomials in a single regressor, extensions to multiple regressors are straightforward. Our local models are found by pointwise fitting a weighted version of the Laird-Ware model. The weights depend on the point of estimation, a bandwidth, and a kernel function and are constructed with respect to the conditional or marginal distribution of Y_i.

2.1 The Conditional Local Mixed Model

Consider the d$^{\text{th}}$ order polynomial mixed model for estimation at $\tilde{x}_0$ for the i^{th} cluster

$$Y_{i0} = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0 + \tilde{\mathbf{z}}'_{i0}\mathbf{b}_0 + \epsilon_{i0} \tag{5}$$

where Y_{i0} is the response at $\tilde{x}_0$ for the i^{th} cluster, $\tilde{\mathbf{x}}'_{i0} = \begin{bmatrix}1 & \tilde{x}_{i0} & ...\tilde{x}_{i0}^{d}\end{bmatrix}$ and $\tilde{\mathbf{z}}'_{i0}$ contains the regressors in $\tilde{x}_0$ corresponding to the random effects. The vectors $\boldsymbol{\beta}_0$ and $\mathbf{b}_0$ are the fixed and random effects parameter vectors at $\tilde{x}_0$, respectively. The vector of random errors, $\boldsymbol{\epsilon}_0$, is assumed to be from a multivariate Gaussian distribution with zero mean and variance-covariance matrix $\mathbf{K}_0^{-\frac{1}{2}}\mathbf{R}\mathbf{K}_0^{-\frac{1}{2}}$, where $\mathbf{K}_0^{-\frac{1}{2}}$ is an (n x n) diagonal weight matrix containing the inverse square root of the Nadaraya-Watson (Nadaraya, 1964; Watson, 1964) weights at $\tilde{x}_0$. This weighting scheme is motivated by local polynomial regression for the fixed effect case. We label this approach the conditional local mixed model (CLMM) because the weighting is applied to the variance of $\mathbf{Y}|\mathbf{b}$. We will assume that the distances used in the weights have been standardized appropriately so that the weights will sum to one across a data set, and that the weights have been assigned without regard to cluster.

The estimator $\widehat{\boldsymbol{\beta}}_0^C$ and the predictor $\widehat{\mathbf{b}}_0$ at the point $\tilde{x}_0$ can be found by incorporating the weight matrix in Henderson's joint likelihood expression (1950) and solving the mixed model equations for estimation at $\tilde{x}_0$. With $\tilde{\mathbf{X}}$ of full rank, solutions to the equations yield the estimator

$$\widehat{\boldsymbol{\beta}}_0^C = (\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{Y} \tag{6}$$

and the predictor

$$\widehat{\mathbf{b}}_0 = \mathbf{B}\tilde{\mathbf{Z}}'\mathbf{V}_{\mathbf{0}}^{*\mathbf{-1}}(\mathbf{Y} - \tilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}_{\mathbf{0}}^{\mathbf{C}}), \tag{7}$$

where $\mathbf{V}_0^* = \mathbf{K}_0^{-\frac{1}{2}}\mathbf{R}\mathbf{K}_0^{-\frac{1}{2}} + \tilde{\mathbf{Z}}\mathbf{B}\tilde{\mathbf{Z}}'$. The expressions given above are of similar form to those given by the parametric mixed model except that different estimates and predictions of the parameter vectors are realized at each $\tilde{x}_0$.

The population average fit (PA) at $\tilde{x}_0$ is simply

$$\widehat{Y}_{PA,0} = \tilde{\mathbf{x}}_0'\widehat{\boldsymbol{\beta}}_0^C = \sum_k h_{PA,0,k}^C y_k, \tag{8}$$

where y_k is the k^{th} element of $\mathbf{Y}$ and where $h_{PA,0,k}^C$ is the k^{th} element of $\tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}$. The cluster specific fits (CS) at $\tilde{\mathbf{x}}_0$ for the i^{th} cluster (i=1,...,s) are

$$\widehat{Y}_{CS,i,0} = \tilde{\mathbf{x}}_{i0}'\widehat{\boldsymbol{\beta}}_0^C + \tilde{\mathbf{z}}_{i0}'\widehat{\mathbf{b}}_0. \tag{9}$$

The cluster specific fits can likewise be expressed as the weighted sum of squares

$$\widehat{Y}_{CS,i,0} = \sum_k h_{CS,0,k}^C y_k, \tag{10}$$

with weight $h_{CS,0,k}^C$, the k^{th} element of $\tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1} + \tilde{\mathbf{z}}_{i,0}'\mathbf{B}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1} - \tilde{\mathbf{z}}_{i,0}'\mathbf{B}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}$, a formula resulting directly from (6) and (7).

2.2 The Marginal Local Mixed Model

Localization through weighting can also be accomplished by targeting the marginal variance-covariance matrix. Consider the following model for estimation at $\tilde{x}_0$ for the i^{th} cluster

$$Y_{i0} = \tilde{\mathbf{x}}_{i0}'\boldsymbol{\beta}_0 + k_{i0}^{-\frac{1}{2}}\tilde{\mathbf{z}}_{i0}'\mathbf{b}_0 + \epsilon_{i0} \tag{11}$$

where Y_{i0}, $\tilde{\mathbf{x}}_{i0}'$, $\tilde{\mathbf{z}}_{i0}'$, $\boldsymbol{\beta}_0$, $\mathbf{b}_0$, and ϵ_{i0} are as defined for CLMM and $k_{i0}^{-\frac{1}{2}}$ is the i^{th} element of $\mathbf{K}_0^{-\frac{1}{2}}$. The vector of random errors, $\boldsymbol{\epsilon}_0$, is again assumed to follow a multivariate Gaussian distribution with zero mean and variance-covariance matrix $\mathbf{K}_0^{-\frac{1}{2}}\mathbf{R}\mathbf{K}_0^{-\frac{1}{2}}$. We label this model the marginal local mixed model (MLMM). As in the conditional model, the local influence of an observation is directed by its variance. Observations that contribute more (that is, to have larger weight) to the prediction at $\tilde{x}_0$ are considered to have smaller variance. The variance, which is transformed to represent the relative weights of observations, is Var[$\mathbf{Y}_i|b_i$] in the model of §2.1 and Var[$\mathbf{Y}_i$] here. The multiplicative involvement of the Nadaraya-Watson weights in (11) accomplish that as

$$\begin{aligned} Var(\mathbf{Y}) &= \mathbf{K}_0^{-\frac{1}{2}}(\mathbf{R} + \tilde{\mathbf{Z}}\mathbf{B}\tilde{\mathbf{Z}}')\mathbf{K}_0^{-\frac{1}{2}} \\ &= \mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}\mathbf{K}_0^{-\frac{1}{2}} = \mathbf{V}_0^{**}. \end{aligned}$$

The estimator $\widehat{\boldsymbol{\beta}}_0^M$ at the point $\tilde{x}_0$ for the marginal local mixed model is

$$\widehat{\boldsymbol{\beta}}_0^M = (\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\mathbf{Y}, \tag{12}$$

and the population average fit at $\tilde{x}_0$ is

$$\widehat{Y}_{PA,0} = \tilde{\mathbf{x}}_0'\widehat{\boldsymbol{\beta}}_0^M = \sum_k h_{PA,0,k}^M y_k, \tag{13}$$

with $h_{PA,0,k}^M$ the k$^{\text{th}}$ element of $\tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}$. Cluster specific prediction in the marginal model is not appropriate. Model (11) localized Var[$\mathbf{Y}_i$] at the cost of incorrectly representing E[$\mathbf{Y}_i|b_i$]. The random effects regressors in the marginal local mixed model are transformed to correctly weight the marginal variance of the response.

2.3 Bandwidth Selection

The kernel weights used in the conditional and marginal local mixed models depend upon a bandwidth parameter (h). A natural criterion for selection is to choose h in order to minimize a function of the squared error of estimation of mean response to account for bias and variance.

Härdle and Marron (1985) and Härdle (1990) provide a rule for bandwidth selection that chooses the asymptotically optimal bandwidth with respect to a number of criteria, including the average squared error, integrated squared error, and the conditional mean integrated squared error. A bandwidth is chosen by minimizing an estimate of some appropriate criterion as the true mean function used in the criteria is unknown.

Plug-in methods, where unknown quantities in the squared error function are replaced with estimates, are very popular (Ruppert, Sheather, and Wand, 1995). Rule of thumb selectors offer a simple estimate of the bandwidth that is easy to calculate (Fan and Gijbels, 1995; Härdle and Marron, 1995).

We prefer bandwidth estimators based on cross-validation (Craven and Wahba, 1979), in particular penalized, “leave-one-out” statistics.

2.3.1 PRESS

The prediction error sum of squares, or PRESS statistic (Allen, 1974), is defined in the usual regression setting as

$$PRESS = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i,-i})^2, \tag{14}$$

where Y_i is the i^{th} observation and $\widehat{Y}_{i,-i}$ is the estimate of the regression function at $\tilde{x}_i$ with the i^{th} data point removed. The bandwidth selected is the value h that minimizes the PRESS statistic.

In the mixed effects model with clustered data, the notion of "leave-one-out" extends to removal of entire clusters, as clusters represent uncorrelated units. Cluster deletion formulas for the parametric mixed model are given by Hurtado-Rodriguez (1993) and Hilden-Minton (1995). The estimators for CLMM and MLMM at $\tilde{x}_0$ with the i^{th} cluster deleted, denoted by $\widehat{\boldsymbol{\beta}}^{C}_{0,-i}$ and $\widehat{\boldsymbol{\beta}}^{M}_{0,-i}$, respectively, are

$$\widehat{\boldsymbol{\beta}}^{C}_{0,-i} = \widehat{\boldsymbol{\beta}}^{C}_{0} - (\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{U}\widehat{\boldsymbol{\phi}}^{C}_{0,-i} \tag{15}$$

$$\widehat{\boldsymbol{\beta}}^{M}_{0,-i} = \widehat{\boldsymbol{\beta}}^{M}_{0} - (\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\mathbf{U}\widehat{\boldsymbol{\phi}}^{M}_{0,-i}, \tag{16}$$

where U is the $(n \text{ x } n_i)$ matrix that contains the identity matrix for the i^{th} cluster and zeros elsewhere, and

$$\widehat{\boldsymbol{\phi}}^{C}_{0,-i} = (\mathbf{U}'\mathbf{P}^{*}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P}^{*}\mathbf{Y},$$

$$\widehat{\boldsymbol{\phi}}^{M}_{0,-i} = (\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P}^{**}\mathbf{Y},$$

$$\mathbf{P}^{*} = \mathbf{V}_0^{*-1} - \mathbf{V}_0^{*-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1},$$

$$\mathbf{P}^{**} = \mathbf{V}_0^{**-1} - \mathbf{V}_0^{**-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}.$$

Then the population average fit at $\tilde{x}_0$ with the i^{th} cluster deleted is

$$\widehat{Y}_{PA,0,-i} = \tilde{\mathbf{x}}_0'\widehat{\boldsymbol{\beta}}_{0,-i}, \tag{17}$$

where $\widehat{\boldsymbol{\beta}}_{0,-i}$ equals $\widehat{\boldsymbol{\beta}}_{0,-i}^{C}$ and $\widehat{\boldsymbol{\beta}}_{0,-i}^{M}$ for CLMM and MLMM, respectively. The BLUP $\widehat{\mathbf{b}}_{0,-i}$ of the i^{th} cluster at $\tilde{x}_0$ in the CLMM is equal to zero[1], so that cluster specific predictions reduce to population average estimation for cluster deletion.

2.3.2 PRESS**

We adopt here to the mixed model scenario a penalized version of PRESS, proposed in the context of fixed effects models for uncorrelated data by Mays, Birch, and Starnes (2001). For the linear mixed model, PRESS** is defined as

$$PRESS^{**} = \frac{PRESS}{n - trace(\mathbf{H}) + (n - d')(\frac{SSE_{max} - SSE_h}{SSE_{max} - SSE_{\bar{y}}})}, \tag{18}$$

where d' is the number of fixed effects parameters in the local mixed model and

$$\mathbf{H} = \mathbf{H}^{CLMM} = \begin{bmatrix} \tilde{\mathbf{x}}_1'(\tilde{\mathbf{X}}'\mathbf{V}_1^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_1^{*-1} \\ \tilde{\mathbf{x}}_2'(\tilde{\mathbf{X}}'\mathbf{V}_2^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_2^{*-1} \\ \vdots \\ \tilde{\mathbf{x}}_n'(\tilde{\mathbf{X}}'\mathbf{V}_n^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_n^{*-1} \end{bmatrix}$$

for the conditional local mixed model and

$$\mathbf{H} = \mathbf{H}^{MLMM} = \begin{bmatrix} \tilde{\mathbf{x}}_1'(\tilde{\mathbf{X}}'\mathbf{V}_1^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_1^{**-1} \\ \tilde{\mathbf{x}}_2'(\tilde{\mathbf{X}}'\mathbf{V}_2^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_2^{**-1} \\ \vdots \\ \tilde{\mathbf{x}}_n'(\tilde{\mathbf{X}}'\mathbf{V}_n^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_n^{**-1} \end{bmatrix}$$

for the marginal local mixed model. These matrices are the local smoother matrices for the population average. The term SSE_{max} is the sum of squared deviations of the response and the local fit that assigns a constant weight to each response. This would result as $h \to \infty$ and represents the worst possible fit using the nonparametric method. The sum of squares $SSE_{\bar{y}}$ is the accumulative across the regressor locations of the sum of the squared deviations of the responses around the mean response and represents the sum of squares error as $h \to 0$ for the population average model. The expression SSE_h is the sum of squared deviations of the response and the local fit

[1] Removal of some, but not all, of the observations in a cluster results in a BLUP not equal to zero for that cluster. See Hurtado-Rodriguez (1993) for formulas.

using a specific value of the bandwidth h. As $h \rightarrow 0$, $SSE_h \rightarrow 0$ for CLMM and $SSE_h \rightarrow SSE_{\bar{y}}$ for MLMM. Cluster specific and population average fits are used in the sum of squared calculations for PRESS** for CLMM and MLMM, respectively.

The denominator in the PRESS** statistic functions as the penalty term, simultaneously providing protection against choosing h too large or too small. The term $n - trace(\mathbf{H})$, where $\mathbf{H}$ equals either $\mathbf{H}^{CLMM}$ or $\mathbf{H}^{MLMM}$, penalizes small bandwidths as this term approaches 0 as $h \rightarrow 0$ and $n - d'$ as $h \rightarrow \infty$. The second term in the denominator ($(n - d')$ multiplied by the sum of squares ratio) protects against large bandwidths, as the second term approaches $n - d'$ as $h \rightarrow 0$ and 0 as $h \rightarrow \infty$.

3 THE SEMIPARAMETRIC MIXED MODEL

We assume that the user has some knowledge about the underlying model from which the data have been generated, but the model fails over a portion of the data; it has been misspecified in functional form. Relying on a nonparametric model entirely results in loss of information about the model. The nonparametric model also has a tendency to produce highly variable fits. A combination of fits may be advantageous; a nonparametric portion "corrects" areas of poor parametric estimation while retaining the information about the true model contained in the parametric model.

Semiparametric models have been proposed in the literature for the fixed effects model. Partial linear regression (Speckman (1988)) assumes that the response could be modeled as a linear predictor plus some smooth unknown function f dependent on the same set of regressors. Estimates of the parameters in the linear predictor are obtained by regression on partial residuals, and the estimate of f is a nonparametric fit to the residuals.

Burman and Chaudhuri (1992), Einsporn and Birch (1993), and Mays, Birch, and Einsporn (2000) developed a semiparametric method termed Model Robust Regression 1 (MRR1) by Mays, Birch, and Einsporn (2000). The MRR1 fit is a convex combination

$$\widehat{\mathbf{Y}}^{MRR1} = (1 - \lambda)\widehat{\mathbf{Y}}^{P} + \lambda\widehat{\mathbf{Y}}^{NP}, \tag{19}$$

where $\widehat{\mathbf{Y}}^{P}$ and $\widehat{\mathbf{Y}}^{NP}$ are the parametric and nonparametric fits, respectively. The mixing parameter $\lambda \in [0, 1]$ determines the proportion of the nonparametric fit that contributes to the model robust fit. For a correctly specified parametric model, λ will be zero and the MRR1 model reduces to the parametric model; for parametric models that are grossly misspecified, λ will be one or close to one and the MRR1 model is the same or nearly equal to the nonparametric model. Simulation results in Mays, Birch, and Starnes (2001) show that MRR1 has smaller average mean square error of fit than separate parametric and nonparametric fits under low to moderate model misspecification. Under no model misspecification, MRR1 is equivalent, or nearly so, to the parametric fit. Under a high degree of misspecification MRR1 is equivalent, or nearly so, to the nonparametric fit.

The rationale is that this convex combination should take advantage of the low variance and bias of the parametric and local fits, respectively, to decrease mean square error. A simulation study in section 5 investigates this claim for the mixed model.

Semiparametric modeling can be extended to the mixed model setting. The proposed Mixed Model Robust Regression (MMRR) fit is an adaptation of the MRR1 fit for use with the mixel model. Specifically, the MMRR fit is

$$\hat{\mathbf{Y}}^{MMRR} = (1-\lambda)\hat{\mathbf{Y}}^{P} + \lambda\hat{\mathbf{Y}}^{NP}, \tag{20}$$

where $\hat{\mathbf{Y}}^{P}$ is the fit from the parametric linear mixed model and $\hat{\mathbf{Y}}^{NP}$ is a local mixed model fit. The mixing parameter is an element between 0 and 1. A value of $\lambda = 1$ produces an MMRR fit equal to the nonparametric fit; $\lambda = 0$ results in an MMRR fit equal to the parametric fit. Values of λ between 0 and 1 produce MMRR fits that are convex combinations of the two fits.

As in the parametric and local models, the MMRR fit can be population average or cluster specific. There are two population average fits for MMRR. One combines the population average fit for the parametric and the conditional local mixed model; the second combines the fit for the parametric and the marginal local mixed model. The cluster specific fit for MMRR utilizes the cluster specific parametric fit in combination with the cluster specific conditional local mixed model fit.

The mixing parameter λ measures the degree of parametric model misspecification and must be estimated from the data. Notice that (20) can be written as

$$(\widehat{\mathbf{Y}}^{MMRR} - \widehat{\mathbf{Y}}^{P}) = \lambda(\widehat{\mathbf{Y}}^{NP} - \widehat{\mathbf{Y}}^{P}) \tag{21}$$

yielding the least square estimate

$$\widehat{\lambda} = \frac{(\widehat{\mathbf{Y}}^{NP} - \widehat{\mathbf{Y}}^{P})'(\widehat{\mathbf{Y}}^{MMRR} - \widehat{\mathbf{Y}}^{P})}{(\widehat{\mathbf{Y}}^{NP} - \widehat{\mathbf{Y}}^{P})'(\widehat{\mathbf{Y}}^{NP} - \widehat{\mathbf{Y}}^{P})}. \tag{22}$$

Thus, $\widehat{\lambda}$ is an estimate of the slope parameter in a no-intercept model with response $(\widehat{\mathbf{Y}}^{MMRR} - \widehat{\mathbf{Y}}^{P})$ and explanatory variable $(\widehat{\mathbf{Y}}^{NP} - \widehat{\mathbf{Y}}^{P})$. For the uncorrelated, fixed effects model, Burman and Chaudhuri (1992) and Mays, Birch, and Starnes (2001) found the optimal data driven estimate of the mixing parameter, an estimate of the value which minimizes the distance between the model robust estimate and the true regression function, as

$$\widehat{\lambda} = \frac{(\widehat{\mathbf{Y}}_{i,-i}^{NP} - \widehat{\mathbf{Y}}_{i,-i}^{P})'(\mathbf{Y} - \widehat{\mathbf{Y}}^{P})}{(\widehat{\mathbf{Y}}^{NP} - \widehat{\mathbf{Y}}^{P})'(\widehat{\mathbf{Y}}^{NP} - \widehat{\mathbf{Y}}^{P})}. \tag{23}$$

$\widehat{\mathbf{Y}}_{i,-i}^{P}$ and $\widehat{\mathbf{Y}}_{i,-i}^{NP}$ are the parametric and nonparametric estimates of the mean response at x_i computed without the point (x_i, y_i). In the cluster correlated mixed model, $\widehat{\mathbf{Y}}_{i,-i}^{P}$ and $\widehat{\mathbf{Y}}_{i,-i}^{NP}$ in (23) are replaced with the parametric and nonparametric fits for the i^{th} cluster with the i^{th} cluster removed. Burman and Chaudhuri (1992) had suggested the substitution of $\widehat{\mathbf{Y}}_{i,-i}^{P}$ and $\widehat{\mathbf{Y}}_{i,-i}^{NP}$ as a precaution against favoring the nonparametric fit. Notice that $\hat{\mathbf{Y}}^{MMRR}$ is unknown and depends on λ. But $\hat{\mathbf{Y}}^{MMRR}$ approaches $E(\mathbf{Y})$ as the sample size increases, so $\mathbf{Y}$ is used in place of $\hat{\mathbf{Y}}^{MMRR}$ in the estimate. The appendix provides the asymptotic theory for the optimal estimators of λ.

The conclusion to the debate over the use of population average versus cluster specific fits for the mixing parameter will be the same as the conclusion in bandwidth selection. For the conditional local mixed model, emphasis is placed upon cluster specific estimation. Thus, the fits used in choosing $\hat{\lambda}$ for MMRR estimation using the conditional local mixed model will be cluster specific. The conditional local mixed model also yields a population average fit. The $\hat{\lambda}$ used in computing the population average MMRR fit using the CLMM will be the same $\hat{\lambda}$ used in computing the cluster specific MMRR fits using the CLMM. Thus, the mean square error for the population average MMRR estimate will not be optimal when using the CLMM population average fit. However, because we are primarily interested in cluster specific fits for the CLMM population average fit, $\hat{\lambda}$ based on cluster specific fits will be used for all mixed model robust regression estimates that use the conditional local mixed model. Population average fits are used in the estimate of λ for MMRR using the marginal local mixed model, as the marginal local mixed model is only appropriate for the population average.

Model robust regression has been extended to a number of scenarios, including quantal data (Nottingham and Birch, 2000), dual modeling (Robinson and Birch, 2000), outlier resistance modeling (Assaid and Birch, 2000) and generalized estimating equations (Clark, 2002).

4 THEORETICAL BIAS, VARIANCE, AND MSE FORMULAS

The mean square prediction error when using random variables $\hat{y}_0$ to predict the constant function $g(x_0)$ is

$$E(\hat{y}_0 - g(x_0))^2 = Var(\hat{y}_0) + [Bias(\hat{y}_0)]^2. \tag{24}$$

In our context $\hat{y}_0$ is the fit at the point x_0 and $g(x_0)$ is the true mean function evaluated at x_0. Thus, the mean square error depends upon the bias and the variance of the fit. Formulas for the bias and variance can be obtained for the parametric, local, and hence the model robust procedures.

In what follows, it is assumed that $\mathbf{V}$, $\mathbf{B}$, and $\mathbf{R}$ are known, and that the bandwidth and mixing parameter are fixed. Complete derivations of the following results may be found in Waterman (2002). For population average estimation evaluated at the design points, the bias formula can be written as

$$Bias(\hat{\mathbf{Y}}_{PA}) = -(\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f}) \tag{25}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{X}$ is the true fixed effects model matrix stacked by cluster, $\boldsymbol{\beta}$ is the true fixed effects parameter vector, $\mathbf{f} = E(\mathbf{Y}) - \mathbf{X}\boldsymbol{\beta}$ is the misspecified portion, and the smoother matrix $\mathbf{H}$ equals

$$\mathbf{H}_{PA}^{P} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \tag{26}$$

for the parametric mixed model,

$$\mathbf{H}_{PA}^{C} = \begin{bmatrix} \tilde{\mathbf{x}}_1'(\tilde{\mathbf{X}}'\mathbf{V}_1^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_1^{*-1} \\ \tilde{\mathbf{x}}_2'(\tilde{\mathbf{X}}'\mathbf{V}_2^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_2^{*-1} \\ \vdots \\ \tilde{\mathbf{x}}_n'(\tilde{\mathbf{X}}'\mathbf{V}_n^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_n^{*-1} \end{bmatrix} \tag{27}$$

for the conditional local mixed model, and

$$\mathbf{H}_{PA}^{M} = \begin{bmatrix} \tilde{\mathbf{x}}_1'(\tilde{\mathbf{X}}'\mathbf{V}_1^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_1^{**-1} \\ \tilde{\mathbf{x}}_2'(\tilde{\mathbf{X}}'\mathbf{V}_2^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_2^{**-1} \\ \vdots \\ \tilde{\mathbf{x}}_n'(\tilde{\mathbf{X}}'\mathbf{V}_n^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_n^{**-1} \end{bmatrix} \tag{28}$$

for the marginal local mixed model. The row vector $\tilde{\mathbf{x}}_k'$ is the k$^{\text{th}}$ row of $\tilde{\mathbf{X}}$. It can be shown that the bias formula for the parametric linear mixed model simplifies to $-(\mathbf{I} - \mathbf{H}_{PA}^{P})\mathbf{f}$.

The variance-covariance matrices for population average estimation in the parametric, CLMM, and MLMM models are

$$Var(\hat{\mathbf{Y}}_{PA}^{P}) = \mathbf{H}_{PA}^{P}\mathbf{V}{\mathbf{H}_{PA}^{P}}', \tag{29}$$

$$Var(\hat{\mathbf{Y}}_{PA}^{C}) = \mathbf{H}_{PA}^{C}\mathbf{V}{\mathbf{H}_{PA}^{C}}', \tag{30}$$

$$Var(\hat{\mathbf{Y}}_{PA}^{M}) = \mathbf{H}_{PA}^{M}\mathbf{V}{\mathbf{H}_{PA}^{M}}'. \tag{31}$$

The parametric variance-covariance matrix for population average estimation can be simplified to $\mathbf{H}_{PA}^{P}\mathbf{V}$. The population average mean square error at the design points is then found by squaring the bias terms, adding the sum of the squared bias terms to the sum of the variances of the fits (i.e., the trace of the variance-covariance matrix), and then dividing by the number of design points.

For cluster specific prediction, the bias and variance formulas are slightly different now that the random effects are included. We first consider cluster specific mean square prediction error formulas by conditioning on the random effects for a fixed true mean function. The cluster specific bias formula for estimation at the design points is

$$Bias(\hat{\mathbf{Y}}_{CS}|\mathbf{b}) = -(\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{f}) \tag{32}$$

where $\mathbf{Z}$ is the true random effects model matrix, $\mathbf{b}$ is the true vector of random effects, and the cluster specific parametric and CLMM smoother matrices are

$$\mathbf{H}_{CS}^{P} = (\mathbf{I} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1})\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1} \tag{33}$$

$$\mathbf{H}_{CS}^{C} = \begin{bmatrix} (\mathbf{i}_1' - \tilde{\mathbf{z}}_1'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_1^{*-1})\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_1^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_1^{*-1} + \tilde{\mathbf{z}}_1'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_1^{*-1} \\ (\mathbf{i}_2' - \tilde{\mathbf{z}}_2'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_2^{*-1})\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_2^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_2^{*-1} + \tilde{\mathbf{z}}_2'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_2^{*-1} \\ \vdots \\ (\mathbf{i}_n' - \tilde{\mathbf{z}}_n'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_n^{*-1})\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_n^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_n^{*-1} + \tilde{\mathbf{z}}_n'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_n^{*-1} \end{bmatrix}, \tag{34}$$

respectively. The row vectors $\mathbf{i}_k'$ and $\tilde{\mathbf{z}}_k'$ are the k$^{\text{th}}$ rows of the identity matrix and $\tilde{\mathbf{Z}}$. The variance-covariance matrices for cluster specific prediction with fixed true mean function using the parametric and CLMM models are

$$Var(\hat{\mathbf{Y}}_{CS}^{P}|\mathbf{b}) = \mathbf{H}_{CS}^{P}\mathbf{R}{\mathbf{H}_{CS}^{P}}' \tag{35}$$

$$Var(\hat{\mathbf{Y}}_{CS}^{C}|\mathbf{b}) = \mathbf{H}_{CS}^{C}\mathbf{R}\mathbf{H}_{CS}^{C}{}'. \tag{36}$$

The cluster specific mean square error calculations at the design points is analogous to the population average mean square error calculations at the design points.

The bias and variance formulas given above can be used to find the MSE formulas for mixed model robust regression. Three bias and variance formulas must be developed - for the MMRR population average based on the conditional local mixed model, the MMRR population average based on the marginal local mixed model, and the MMRR cluster specific fit. The bias formulas for mixed model robust regression are

$$Bias(\hat{\mathbf{Y}}_{PA}^{MMRR,C}) = -\lambda(\mathbf{I} - \mathbf{H}_{PA}^{C})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{PA}^{MMRR,C})\mathbf{f} \tag{37}$$

$$Bias(\hat{\mathbf{Y}}_{PA}^{MMRR,M}) = -\lambda(\mathbf{I} - \mathbf{H}_{PA}^{M})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{PA}^{MMRR,M})\mathbf{f} \tag{38}$$

$$Bias(\hat{\mathbf{Y}}_{CS}^{MMRR}|\mathbf{b}) = -\lambda(\mathbf{I} - \mathbf{H}_{CS}^{C})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{CS}^{MMRR})(\mathbf{Z}\mathbf{b} + \mathbf{f}), \tag{39}$$

where

$$\mathbf{H}_{PA}^{MMRR,C} = (1-\lambda)\mathbf{H}_{PA}^{P} + \lambda\mathbf{H}_{PA}^{C}, \tag{40}$$

$$\mathbf{H}_{PA}^{MMRR,M} = (1-\lambda)\mathbf{H}_{PA}^{P} + \lambda\mathbf{H}_{PA}^{M}, \tag{41}$$

$$\mathbf{H}_{CS}^{MMRR} = (1-\lambda)\mathbf{H}_{CS}^{P} + \lambda\mathbf{H}_{CS}^{C} \tag{42}$$

are the mixed model robust regression smoother matrices for population average (using CLMM and MLMM) estimation and cluster specific prediction. The mixing parameter appears in the bias expression in two places - as the multiplier of the $\mathbf{X}\boldsymbol{\beta}$ term and in the MMRR smoother matrices.

The variance expressions for the three model robust methods are

$$Var(\hat{\mathbf{Y}}_{PA}^{MMRR,C}) = \lambda\mathbf{H}_{PA}^{MMRR,C}\mathbf{V}\mathbf{H}_{PA}^{C}{}' + (1-\lambda)\mathbf{H}_{PA}^{MMRR,C}\mathbf{V}\mathbf{H}_{PA}^{P}{}', \tag{43}$$

$$Var(\hat{\mathbf{Y}}_{PA}^{MMRR,M}) = \lambda\mathbf{H}_{PA}^{MMRR,M}\mathbf{V}\mathbf{H}_{PA}^{M}{}' + (1-\lambda)\mathbf{H}_{PA}^{MMRR,M}\mathbf{V}\mathbf{H}_{PA}^{P}{}', \tag{44}$$

$$Var(\hat{\mathbf{Y}}_{CS}^{MMRR}) = \lambda\mathbf{H}_{CS}^{MMRR}\mathbf{R}\mathbf{H}_{CS}^{C}{}' + (1-\lambda)\mathbf{H}_{CS}^{MMRR}\mathbf{R}\mathbf{H}_{CS}^{P}{}'. \tag{45}$$

The mean square errors for the model robust fits can be found in the same fashion as the parametric and nonparametric models.

5 SIMULATION STUDY

A Monte Carlo simulation study was conducted with data generated from the cluster specific model

$$Y_{ij} = (2+b_{i1})(X_j - 5.5)^2 + (5+b_{i2})X_j + 10\gamma \left[sin \left(\frac{\pi(X_j-1)}{2.25} \right) \right] + \epsilon_{ij}, \quad (46)$$

where Y_{ij} is the simulated response for the i^{th} cluster at X_j. The regressor takes on integer values from one to ten, inclusive.[2] The random effects, b_{i1} and b_{i2}, are generated independently from normal distributions with mean zero and variance 0.50.

Three variance-covariance structures for the random errors ϵ_{ij} were considered. The first variance-covariance structure was independence with the variance of the errors equal to 16. First-order autoregressive models (AR(1)) with ρ=0.20 and ρ=0.80 were also used with σ^2=16.

The user's model is

$$Y_{ij} = (2+b_{i1})(X_j - 5.5)^2 + (5+b_{i2})X_j + \epsilon_{ij}. \quad (47)$$

The true model is given in (56); the trigonometric component then serves as the misspecification. Values of γ equal to 0 (no misspecification), 0.25, 0.50, 0.75, and 1.0 will be used in the study. A plot of the population average models versus γ is given in Figure 1. The smooth parabola, indicated by the dashed line, occurs with no model misspecification and the solid curve represents the most misspecification at γ=1. The large disparity between the γ=0 and γ=1 models should be reflected in the MSE results from the simulation study.

It is assumed that there is no parametric misspecification in the variance-covariance structure. That is, if the random errors are generated from an AR(1) variance-covariance structure, the parametric model is the quadratic model in (47) with an AR(1) structure for **R**.

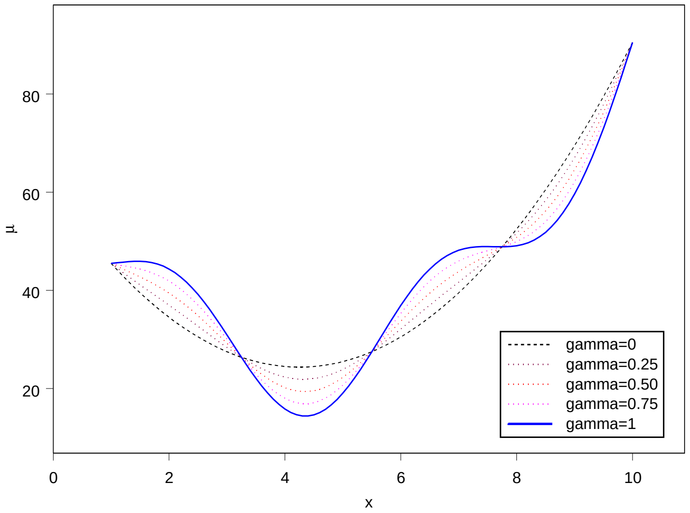
The local model used in the analysis is the local linear mixed model (CLMM or MLMM) with a random intercept:

$$Y_{i0} = \beta_0 0 + \tilde{x}_0 \beta_1 0 + b_{i0} + \epsilon_{i0}. \quad (48)$$

Our extensive Monte Carlo studies indicate that PRESS provides the best mean square error of fit results when using MLMM, while PRESS**

[2] The model studied here is similar to the model of Mays, Birch, and Starnes (2001), except that in this work the cluster correlated, random coefficient case is considered.

Figure 1: Plot of Population Average Underlying Models (where γ is the misspecification parameter



provides best results when using CLMM. Consequently, the bandwidth selectors used in the study were PRESS for MLMM population average estimation (PA MLMM and PA MMRR MLMM) and PRESS** for CLMM population average and cluster specific estimation (PA CLMM, PA MMRR CLMM, CS CLMM, and CS MMRR CLMM). A golden section search was performed over the bandwidth range [0.05,0.30]. These values were selected to minimize the distance covered by the search method as the bandwidths chosen in every scenario fell within these bounds. The estimate of the mixing parameter was found using formula (23). Because no bounds are imposed by this formula, the estimate of the mixing parameter was set to zero or one if the solution to (23) was negative or greater than one, respectively.

Both the bandwidth and the mixing parameter were found by summing over the design points. Using the bandwidth and mixing parameter for a given data set, the integrated mean square error of fit was approximated by calculating the mean square error at 46 points (1 to 10 by 0.20). The mean square errors were calculated for the parametric (population average and cluster specific), CLMM (population average and cluster specific), the MLMM (population average), and the mixed model robust regression models (population average using CLMM, population average using MLMM, and cluster specific using CLMM).

We are interested in the approximate integrated MSE (INTMSE) as a function of cluster size, correlation, and γ. To keep the number of scenarios manageable, s=5 and s=20 clusters per data set are examined for different variance-covariance structures and degrees of misspecification.

Since the data are correlated and parameter estimation is an iterative process, fitting a large number of models requires substantial computing resources. To examine the practically feasible number of needed simulation runs that also provided sufficient precision of Monte-Carlo averages, we examined the standard errors of Monte-Carlo mean square errors. As the number of runs increased, the standard error decreased and leveled off around 250 runs. We decided on 250 simulation runs, attempting to balance between computing time and Monte-Carlo variability.

5.1 Varying Cluster Size

This section presents results for the model in (46) under the assumption of within-cluster independence. It is expected that as the number of clusters increases, the INTMSE for the local population average should decrease since more observations at a given $\tilde{x}_0$ will result in estimates that will be more precise. For small bandwidths, this would mean that non-negligible weight would be given to those observations at $\tilde{x}_0$ only, resulting in a local population average fit that connects the mean response at each value of the regressor. The INTMSE values for the local cluster specific fits should be unaffected by the addition of clusters.

In this work, five clusters will be considered our "small" number of clusters, and twenty clusters our "large" number of clusters. Tables 1 and 2 contain the simulated INTMSE values summed over clusters using an independence within-cluster variance structure over regressor location/cluster size combination.[3] Tables 1 and 2 suggest that MMRR-MLMM should be used for population average estimation, while MMRR-CLMM be utilized for cluster specific prediction. These model robust procedures clearly minimize the INTMSE over the entire range of γ. For example, for the PA fit, we note that MMRR-MLMM is very close to the optimal parametric method (when γ=0) in INTMSE for very small values of γ, is very close to MLMM

[3] Columns 2 through 6 contain population average results, while columns 7 through 9 contain cluster specific results. A population average and a cluster specific INTMSE are highlighted in each row; this bolded value is the minimum population average and cluster specific INTMSE for the given γ value. MLMM and CLMM (and the corresponding model robust procedures) use PRESS and PRESS**, respectively, as bandwidth selectors.

for very large values of γ, and is superior to both for intermediate vales of γ. We note similar results for the CS fit and the MMRR-CLMM technique. MMRR-MLMM and MMRR-CLMM are robust to model misspecification for fitting the PA and CS curves, respectively.

A finer grid between γ=0 and γ=0.30 provides a range of values where parametric, local, and model robust procedures are optimal, information useful to determine the degree of misspecification where model robust procedures would be the most beneficial.

Table 1. Mixed Model Robust Regression using Independence
(10 design points and 5 clusters)

	PA	PA	PA	PA	PA	CS	CS	CS
γ	Parm.	CLMM	MLMM	MMRR	MMRR	Parm.	CLMM	MMRR
				CLMM	MLMM			CLMM
0.00	**13.83**	16.09	15.76	13.87	14.00	**2.65**	7.54	2.73
0.05	**13.95**	16.10	15.81	13.96	14.08	**2.77**	7.55	2.84
0.10	14.30	16.15	15.82	**14.27**	14.36	**3.12**	7.62	3.16
0.15	14.88	16.19	15.87	**14.74**	14.78	3.71	7.73	**3.66**
0.20	15.70	16.25	15.96	15.31	**15.23**	4.55	7.88	**4.29**
0.25	16.75	16.36	16.03	15.94	**15.60**	5.62	8.06	**4.99**
0.30	18.03	16.45	16.03	16.54	**15.87**	6.93	8.28	**5.73**
0.50	25.51	16.77	**16.19**	18.28	16.33	14.63	9.20	**8.35**
0.75	40.11	17.26	**16.57**	19.35	16.69	29.87	**10.48**	10.67
1.00	60.56	17.69	**17.14**	20.33	17.23	51.35	**11.86**	12.68

Table 2. Mixed Model Robust Regression using Independence
(10 design points and 20 clusters)

	PA	PA	PA	PA	PA	CS	CS	CS
γ	Parm.	CLMM	MLMM	MMRR	MMRR	Parm.	CLMM	MMRR
				CLMM	MLMM			CLMM
0.00	**3.37**	4.16	4.08	3.38	3.43	**2.49**	9.56	**2.49**
0.05	**3.49**	4.16	4.08	**3.49**	3.56	**2.60**	9.62	2.61
0.10	3.84	4.17	4.09	3.81	**3.79**	2.95	9.79	**2.94**
0.15	4.43	4.17	4.11	4.28	**3.97**	3.54	10.17	**3.44**
0.20	5.24	4.16	4.12	4.84	**4.08**	4.36	10.51	**4.08**
0.25	6.30	4.19	**4.16**	5.42	**4.16**	5.42	10.73	**4.81**
0.30	7.58	4.22	**4.19**	5.96	4.22	6.72	10.80	**5.56**
0.50	15.06	4.48	**4.42**	7.50	4.45	14.35	11.04	**8.40**
0.75	29.67	4.92	**4.84**	8.96	4.93	29.57	12.84	**11.69**
1.00	50.12	5.47	**5.45**	10.52	5.51	51.42	**14.24**	14.64

A cross-over point is defined as the value of the misspecification parameter at which the minimum INTMSE value switches from parametric to model robust estimation or from model robust to local estimation. For both population average and cluster specific prediction, the minimum INTMSE value changes from the parametric to the model robust method at a γ value between 0.05 and 0.10. The second cross-over point (from model robust to local estimation) occurs much earlier for the population average for both cluster sizes. The second population average cross-over point occurs between γ=0.20 and γ=0.30, whereas the second cluster specific cross-over point occurs for large misspecification– a γ value between 0.75 and 1.0. In addition, the cross-over occurs earlier for larger cluster sizes, a result consistent with Clark (2002).

5.2 AR(1) Correlation Structure

The INTMSE values for the correlated data cases appear in Tables 3-6 for s=5 and s=20. The pattern of the cross-over points in γ across the covariance structures appears to be similar. This pattern suggests that population average mixed model robust regression outperforms MLMM for small amounts of model misspecification, whereas cluster specific mixed model robust regression generally works well for all levels of misspecification, and is outperformed by CLMM only at the extreme cases, as expected.

Table 3. Simulated INTMSE values using AR(1) with ρ=0.20
(10 design points and 5 clusters)

γ	PA Parm.	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parm.	CS CLMM	CS MMRR CLMM
0.00	**13.77**	16.22	15.71	13.79	13.91	**3.74**	9.04	3.77
0.25	16.73	16.38	15.94	15.99	**15.57**	6.74	9.55	**6.05**
0.50	25.53	16.77	**16.09**	18.69	16.28	16.00	10.65	**9.56**
0.75	40.18	17.18	**16.43**	20.15	16.59	31.81	**11.82**	12.12
1.00	60.99	17.57	**16.99**	21.30	17.14	54.01	**13.05**	14.34

Table 4. Simulated INTMSE values using AR(1) with ρ=0.20
(10 design points and 20 clusters)

γ	PA Parm.	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parm.	CS CLMM	CS MMRR CLMM
0.00	**3.42**	4.06	3.97	**3.42**	3.46	**3.34**	10.72	**3.34**
0.25	6.34	4.09	**4.06**	5.42	4.08	6.34	11.80	**5.66**
0.50	15.12	4.36	**4.31**	7.66	4.36	15.74	12.20	**9.36**
0.75	29.77	4.77	**4.72**	9.72	4.81	32.31	13.98	**13.04**
1.00	50.28	5.33	**5.31**	11.88	5.37	55.81	**15.03**	16.30

Table 5. Simulated INTMSE values using AR(1) with ρ=0.80
(10 design points and 5 clusters)

γ	PA Parm.	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parm.	CS CLMM	CS MMRR CLMM
0.00	**14.26**	17.07	16.31	14.44	14.80	**6.48**	14.54	6.60
0.25	17.30	17.21	16.33	16.97	**16.29**	9.41	14.80	**9.14**
0.50	26.24	17.51	**16.41**	19.62	16.83	18.60	15.35	**13.11**
0.75	40.96	17.71	**17.01**	20.44	17.26	34.04	15.90	**15.38**
1.00	61.50	18.06	**17.61**	21.45	17.87	55.93	**16.52**	17.13

Table 6. Simulated INTMSE values using AR(1) with ρ=0.80
(10 design points and 20 clusters)

γ	PA Parm.	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parm.	CS CLMM	CS MMRR CLMM
0.00	**3.68**	4.36	4.25	**3.68**	3.75	**5.78**	14.67	**5.78**
0.25	6.63	4.38	**4.32**	5.62	4.33	8.85	15.04	**8.40**
0.50	15.46	4.64	**4.57**	7.29	4.62	18.13	15.39	**12.28**
0.75	30.13	5.01	**4.99**	8.78	5.03	33.89	16.37	**15.09**
1.00	50.66	5.62	**5.58**	10.72	5.60	56.55	**17.00**	17.56

Notice that the population average MLMM and cluster specific CLMM methods robust models are extremely competitive. For γ=0, the parametric method should have the smallest INTMSE. The model robust procedures obtain INTMSE values very close to the parametric INTMSE values. For γ=1, the local methods should have the smallest INTMSE, and the model robust procedures obtain INTMSE values very close to the local values. As γ increases from zero to one, the INTMSE values for the mixed model robust procedures are either the minimum value or are close in value to the "winning" INTMSE values.

There are some key differences between the independence and correlated cases. On average, the INTMSE values increase as the correlation increases. Consider, for example, the case γ=0, where INTMSE reduces to the approximate integrated variance of fit when conditioned on the values of the random effects. If the same size n remains fixed while ρ increases, the effective sample size decreases. Thus, as the correlation increases, the variance of the fits, and hence the mean square error, must increase.

In virtually all of the cases where s=5, MLMM outperformed CLMM when the local mixed model was the method that minimized mean square error using PRESS as the bandwidth selector. For s=20, MLMM appears to have the smallest INTMSE value on average over differing amounts of γ. As the use of PRESS for choosing bandwidth resulted in smaller INTMSE values for the population average, this suggests that the marginal local mixed model is superior to the conditional local mixed model for population average estimation.

It is clear from the above tables that MMRR-CLMM is superior in cluster specific prediction when measured by INTMSE. For population average estimation, MMRR-MLMM does not achieve the minimum INTMSE value, but its INTMSE value is very close to the minimum value across all values of γ, s, and ρ. Moreover, when the model is correctly specified, MMRR-MLMM always beats MLMM, the preferred population average local method.

5.2.1 Estimation of ρ

One concern in the correlated data case was whether the misspecification term influenced the estimate of ρ. As the estimate of ρ is determined by REML or ML, it is very difficult to determine the expected value of the correlation estimate under model misspecification; such estimates, however, can be examined by varying model misspecification in the previous Monte-Carlo study. Five hundred data sets were generated for different values of γ, and the average estimates of ρ from the parametric analysis over the five hundred data sets were calculated. The data were generated from an AR(1) process using $\rho = 0$, $\rho = 0.10$, $\rho = 0.20$, $\rho = 0.33$, $\rho = 0.80$, and $\rho = 0.90$. The simulation used 10 design points and 20 clusters using REML estimation (Table 7).

Table 7. Average Estimate of ρ from Parametric Estimation
(10 design points, 20 clusters, and 500 iterations)

γ	$\rho = 0$	$\rho = 0.10$	$\rho = 0.20$	$\rho = 0.33$	$\rho = 0.80$	$\rho = 0.90$
0.00	-0.00	0.09	0.19	0.32	0.79	0.89
0.10	0.01	0.10	0.20	0.32	0.76	0.86
0.20	0.05	0.13	0.21	0.32	0.69	0.76
0.25	0.07	0.14	0.22	0.32	0.65	0.71
0.30	0.09	0.16	0.23	0.32	0.61	0.66
0.40	0.13	0.19	0.25	0.32	0.55	0.57
0.50	0.16	0.21	0.26	0.32	0.49	0.50
0.60	0.19	0.23	0.27	0.32	0.45	0.46
0.70	0.21	0.24	0.28	0.32	0.42	0.43
0.75	0.22	0.25	0.28	0.31	0.41	0.41
0.80	0.23	0.25	0.28	0.31	0.40	0.40
0.90	0.24	0.26	0.28	0.31	0.38	0.38
1.00	0.24	0.26	0.28	0.30	0.36	0.36

When the model is correctly specified (γ=0) the estimate of ρ is nearly unbiased. Model misspecification unduly influences the estimate of ρ, as the results indicate that highly (weakly) correlated data appear less (more) correlated for large model misspecification. The value ρ=1/3 provides estimates close to the true value, regardless of γ. This may be an indication that the estimates are converging to a value close to 1/3. We see that model misspecification affects the parametric model's ability to fit as well as its ability to estimate ρ in an AR(1) model.

5.3 Simulated Average and Optimal Bandwidths and Mixing Parameters

Tables 8 and 9 contain the simulated optimal bandwidth and mixing parameter results, respectively, over varying amounts of misspecification, cluster size, correlation structure, and local model. The simulated optimal bandwidth for a given level of misspecification and correlation structure is found by calculating the average mean square error (AVEMSE), a discrete approximation to the integrated mean square error, for each simulated data set over a variety of bandwidths. The AVEMSE values are then averaged over the number of simulated data sets for each bandwidth value, with the optimal bandwidth (h_{opt}) being the value with the smallest average AVEMSE value for the specified levels of γ and ρ. In a similar fashion, the simulated optimal mixing parameter (λ_{opt}) can be found by using a fine grid of λ values and calculating the average AVEMSE values as above. The simulated optimal mixing parameter is the value of λ corresponding to the smallest AVEMSE value for the given γ and ρ.

Table 8. Simulated Optimal Bandwidth h_{opt}

correlation	model	s	γ=0	γ=0.25	γ=0.5	γ=0.75	γ=1
Independence	CLMM	5	0.12	0.11	0.10	0.09	0.08
AR(1), ρ=0.20	CLMM	5	0.12	0.11	0.10	0.09	0.08
AR(1), ρ=0.80	CLMM	5	0.10	0.09	0.08	0.07	0.07
Independence	MLMM	5	0.19	0.09	0.07	0.07	0.06
AR(1), ρ=0.20	MLMM	5	0.19	0.08	0.07	0.06	0.06
AR(1), ρ=0.80	MLMM	5	0.06	0.06	0.05	0.05	0.05
Independence	CLMM	20	0.12	0.11	0.10	0.09	0.08
AR(1), ρ=0.20	CLMM	20	0.12	0.11	0.10	0.09	0.08
AR(1), ρ=0.80	CLMM	20	0.10	0.09	0.08	0.07	0.07
Independence	MLMM	20	0.06	0.06	0.05	0.05	0.05
AR(1), ρ=0.20	MLMM	20	0.07	0.06	0.06	0.05	0.05
AR(1), ρ=0.80	MLMM	20	0.05	0.05	0.05	0.05	0.05

Table 9. Simulated Optimal Mixing Parameter λ_{opt}

correlation	model	s	γ=0	γ=0.25	γ=0.5	γ=0.75	γ=1
Independence	CLMM	5	0.04	0.31	0.70	0.90	0.96
AR(1), ρ=0.20	CLMM	5	0.08	0.31	0.68	0.88	0.95
AR(1), ρ=0.80	CLMM	5	0.14	0.26	0.60	0.79	0.90
Independence	MLMM	5	0.00	0.58	0.88	0.99	0.98
AR(1), ρ=0.20	MLMM	5	0.05	0.61	0.90	0.95	0.98
AR(1), ρ=0.80	MLMM	5	0.00	0.63	0.89	0.95	0.98
Independence	CLMM	20	0.01	0.30	0.71	0.91	0.97
AR(1), ρ=0.20	CLMM	20	0.02	0.28	0.68	0.89	0.95
AR(1), ρ=0.80	CLMM	20	0.05	0.21	0.58	0.78	0.89
Independence	MLMM	20	0.00	0.83	0.95	0.98	0.99
AR(1), ρ=0.20	MLMM	20	0.00	0.84	0.97	0.98	0.99
AR(1), ρ=0.80	MLMM	20	0.05	0.87	0.97	0.99	1.00

Table 10. Average Bandwidth from Simulations $\bar{h}$

selector	model	ρ	s	γ=0	γ=0.25	γ=0.5	γ=0.75	γ=1
PRESS	CLMM	Indep.	5	0.09	0.09	0.07	0.07	0.06
PRESS	MLMM	Indep.	5	0.13	0.11	0.08	0.06	0.06
PRESS	CLMM	Indep.	20	0.07	0.07	0.05	0.05	0.05
PRESS	MLMM	Indep.	20	0.08	0.06	0.06	0.05	0.05
PRESS**	CLMM	Indep.	5	0.11	0.11	0.09	0.08	0.07
PRESS**	MLMM	Indep.	5	0.19	0.17	0.13	0.10	0.08
PRESS**	CLMM	Indep.	20	0.07	0.07	0.07	0.06	0.05
PRESS**	MLMM	Indep.	20	0.13	0.10	0.07	0.07	0.06
PRESS	CLMM	0.20	5	0.09	0.08	0.07	0.06	0.06
PRESS	MLMM	0.20	5	0.12	0.10	0.08	0.07	0.06
PRESS	CLMM	0.20	20	0.07	0.06	0.05	0.05	0.05
PRESS	MLMM	0.20	20	0.07	0.07	0.05	0.05	0.05
PRESS**	CLMM	0.20	5	0.12	0.11	0.09	0.08	0.07
PRESS**	MLMM	0.20	5	0.19	0.18	0.13	0.10	0.08
PRESS**	CLMM	0.20	20	0.08	0.07	0.07	0.05	0.05
PRESS**	MLMM	0.20	20	0.13	0.09	0.07	0.07	0.05
PRESS	CLMM	0.80	5	0.06	0.06	0.05	0.05	0.05
PRESS	MLMM	0.80	5	0.07	0.06	0.05	0.05	0.05
PRESS	CLMM	0.80	20	0.05	0.05	0.05	0.05	0.05
PRESS	MLMM	0.80	20	0.05	0.05	0.05	0.05	0.05
PRESS**	CLMM	0.80	5	0.12	0.11	0.09	0.08	0.07
PRESS**	MLMM	0.80	5	0.18	0.17	0.13	0.09	0.08
PRESS**	CLMM	0.80	20	0.08	0.07	0.06	0.05	0.05
PRESS**	MLMM	0.80	20	0.14	0.09	0.07	0.06	0.05

Table 11. Average λ from Simulations $\bar{\lambda}$

selector	model	ρ	s	γ=0	γ=0.25	γ=0.5	γ=0.75	γ=1
PRESS	CLMM	Indep.	5	0.07	0.25	0.53	0.69	0.77
PRESS	MLMM	Indep.	5	0.18	0.55	0.91	0.97	0.99
PRESS	CLMM	Indep.	20	0.01	0.20	0.46	0.63	0.73
PRESS	MLMM	Indep.	20	0.11	0.84	0.98	0.98	0.99
PRESS**	CLMM	Indep.	5	0.07	0.27	0.61	0.78	0.84
PRESS**	MLMM	Indep.	5	0.15	0.49	0.95	1.00	1.00
PRESS**	CLMM	Indep.	20	0.01	0.20	0.50	0.66	0.73
PRESS**	MLMM	Indep.	20	0.09	0.88	1.00	1.00	1.00
PRESS	CLMM	0.20	5	0.04	0.22	0.49	0.65	0.73
PRESS	MLMM	0.20	5	0.14	0.52	0.90	0.97	0.98
PRESS	CLMM	0.20	20	0.01	0.20	0.45	0.60	0.68
PRESS	MLMM	0.20	20	0.12	0.85	0.97	0.98	0.99
PRESS**	CLMM	0.20	5	0.05	0.23	0.56	0.73	0.79
PRESS**	MLMM	0.20	5	0.12	0.46	0.94	1.00	1.00
PRESS**	CLMM	0.20	20	0.01	0.20	0.48	0.61	0.68
PRESS**	MLMM	0.20	20	0.09	0.88	1.00	1.00	1.00
PRESS	CLMM	0.80	5	0.05	0.24	0.54	0.69	0.76
PRESS	MLMM	0.80	5	0.14	0.58	0.88	0.95	0.97
PRESS	CLMM	0.80	20	0.01	0.24	0.54	0.68	0.74
PRESS	MLMM	0.80	20	0.16	0.87	0.97	0.99	1.00
PRESS**	CLMM	0.80	5	0.06	0.21	0.58	0.76	0.81
PRESS**	MLMM	0.80	5	0.20	0.44	0.95	1.00	1.00
PRESS**	CLMM	0.80	20	0.01	0.24	0.56	0.68	0.74
PRESS**	MLMM	0.80	20	0.09	0.92	1.00	1.00	1.00

The simulated optimal bandwidths can be compared to the average bandwidths from the simulations presented in Table 10. The results suggest that the bandwidth selectors are performing adequately (for example, the average and optimal bandwidths for CLMM, PRESS**, s=5, and ρ are quite close) and that PRESS is appropriate for MLMM and PRESS** for CLMM. Additionally, notice that the simulated optimal bandwidths decrease as ρ increases.

Comparing the results from Tables 9 and 11, it appears that the results for λ_{opt} and $\bar{\lambda}$ for MLMM agree in most cases, with a few exceptions at γ=0. The opposite is true for CLMM. The estimate λ_{opt} performs well for small amounts of misspecification. In fact, $\bar{\lambda}$, on averge, is too small for large misspecification. For misspecified models, CLMM model robust fits does not use enough of the nonparametric fit; it relies too heavily on the parametric part. Alterations to the mixing parameter formula, such as incorporating variance-covariance matrices into the formulas for $\widehat{\lambda}$ or changing the delete cluster portion of the formula for CLMM, are possibilities left for further study.

6 APPLICATION OF METHODS TO A REAL DATA SET

The parametric, nonparametric, and model robust methods discussed above can be applied to the wind speed data set from Haslett and Raftery (1989). Twelve meteorological stations in Ireland were selected and the average wind speed in knots were measured daily during the years 1961 through 1978. This analysis looked at the average weekly wind speeds averaged over the eighteen years. The stations, or clusters, were randomly selected from all such stations in Ireland; consequently, the station is the random effect. Measurements were taken at the same fifty three time points for each station, making a total of 636 observations in the data set.

Because of the parabolic trend, a quadratic model was selected as the parametric model. Two models were considered; a fixed effects model with a quadratic trend in week and a mixed effects model with only a random intercept term. The random intercept term was considered because each cluster had a similar shape. This results in cluster specific fits that are parallel shifts of the population average curve. The within-cluster variation structure for the parametric model in the wind speed example is assumed to be AR(1) and the between-cluster variation is assumed to be of independent structure. Results in Waterman (2002) show that the model containing the random intercept is an improvement over the fixed effects model. Thus, our population average curve is estimated as

$$\hat{\mathrm{E}}(\mathbf{Y}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\begin{bmatrix} 12.6576 \\ -0.2520 \\ 0.004549 \end{bmatrix},$$

and the estimated variance-covariance matrix is $\hat{\mathbf{R}}+(\hat{\sigma}^2_{\mathrm{b_0}})\mathbf{ZZ}' = \hat{\mathbf{R}}+(7.2166)\mathbf{ZZ}'$ where $\hat{\mathbf{R}}_{\mathrm{i}}$ has the estimate of the variance ($\hat{\sigma}^2 = 1.1325$) down the diagonal and the estimated covariances $\hat{c}_{\mathrm{jk}} = \hat{\sigma}^2\hat{\rho}^{|\mathrm{j\text{-}k}|} = (1.1325)(0.5169)^{|\mathrm{j\text{-}k}|}$ in the (j,k) and (k,j) off-diagonal cells.

The conditional distribution of $\mathbf{Y}|\mathbf{b}$ is normal with mean $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zb}$ and variance-covariance matrix $\mathbf{R}$. The cluster specific curves can be estimated as

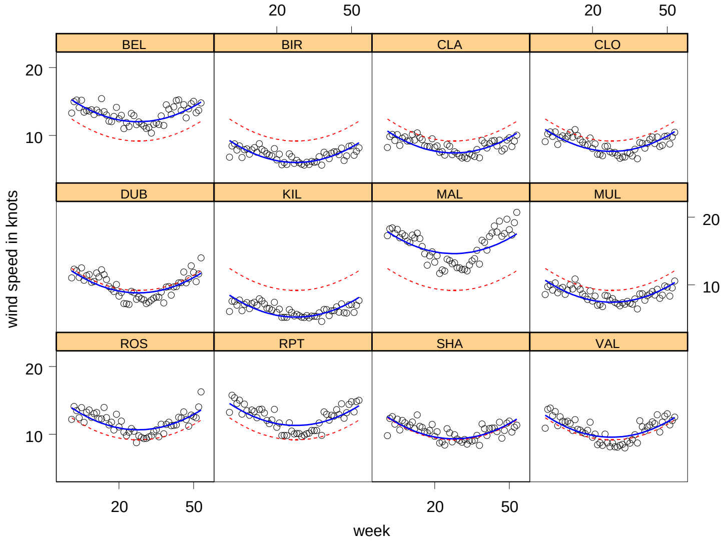
$$\hat{E}(\mathbf{Y}) = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}} = \mathbf{X}\begin{bmatrix} 12.6576 \\ -0.2520 \\ 0.004549 \end{bmatrix} + \mathbf{Z}\begin{bmatrix} 2.8481 \\ -3.1644 \\ -1.7556 \\ -1.5244 \\ -0.3572 \\ -3.9325 \\ 5.4406 \\ -1.7344 \\ 1.4900 \\ 2.1325 \\ 0.1602 \\ 0.3972 \end{bmatrix}$$

and the estimated variance-covariance matrix is given by $\hat{\mathbf{R}}$ as above.

A trellis plot of the population average curve and cluster specific curves by cluster (station) appears in Figure 2. The observations in the cluster are represented by the scatterplot. The dotted curve is the population average curve, and the solid curves are the cluster specific curves. The population average curve is the same for every cluster in this example. As shown in the equations and in plots, the intercepts for the cluster specific curves differ. Thus, the cluster specific fit at each station is a parabola shifted up or down for a particular cluster. Notice that the population average curve fits poorly to some of the clusters, in particular to clusters MAL, KIL, BIR, and MUL. The cluster specific curves are an improvement over the population average curve, as to be expected.

At virtually every station, the wind speeds remain relatively constant through January and February, and then diminish during the spring months. This is followed by a drop in wind speed during the middle of the year. This drop remains during the summer months (with a slight increase in wind speed for some stations during July). For some clusters, such as station BIR, the drop in wind speeds during the summer months is minimal. Other clusters, like station MAL, exhibit a steep drop in wind speed. During the fall and winter months, the wind speeds then increase and then level off. The proposed parametric model is unable to model this type of trend – the level speeds in the winter months, combined with the decreased speeds in the summer months. The parametric model has been misspecified. The bias of the parametric linear model can be rectified through the use of nonparametric, or local, mixed models.

Figure 2: Parametric Linear Mixed Model (Plot of Population Average and Cluster Specific Curves by Station)



The local mixed model used for this data set was the local linear mixed model with a random intercept. Population average and cluster specific curves were found for the conditional local mixed model and the population average curve was found for the marginal local mixed model. The between-cluster and within-cluster variation ($\tilde{\mathbf{B}}$ and $\tilde{\mathbf{R}}$) are assumed to be of independent structure. This differs from the within-cluster structure used in the parametric model. Research by Lin and Carroll (2001) give asymptotic results that suggest the use of the independence structure for local GEE estimation. Both variance-covariance structures were studied for this example and the conclusion was to use independence for the local model due to fewer difficulties with variance component estimation. The independence structure allowed a wider range of bandwidths to be used.

For both the conditional and marginal local mixed models, PRESS and PRESS** both chose a bandwidth of 0.05. It was expected that the bandwidth chosen would be small. The dataset is quite large, so a small bandwidth gives weight to many observations. A small bandwidth is also needed in the conditional local cluster specific analysis to be flexible enough to catch the sudden drop at station MAL.

As in the parametric linear mixed model, a population average curve and cluster specific curves can be found for the conditional local mixed model. The marginal local mixed model will yield a population average curve. Recall that local linear mixed models were calculated at each value of the regressor x̃. Thus, for each $\tilde{x}_0$ in the conditional local mixed model, there is an estimate of parameter vector $\hat{\boldsymbol{\beta}}_0$ and a predictor of random effects vector $\hat{\mathbf{b}}_0$. For example, the CLMM population average fit at x̃=1 using a bandwidth of 0.05 can be expressed, for each of the 12 stations, as

$$\begin{aligned}\widehat{Y}_{PA,0} &= \tilde{\mathbf{x}}_0' \hat{\boldsymbol{\beta}}_0^C \\ &= \begin{bmatrix}1 & 1\end{bmatrix} \begin{bmatrix}10.7398 \\ 0.3793\end{bmatrix}\end{aligned}$$

and the cluster specific fits at x̃=1 are

$$\hat{\mathbf{Y}}_{CS,0} = \tilde{\mathbf{x}}_0' \hat{\boldsymbol{\beta}}_0^C + \hat{\mathbf{b}}_0$$

$$= \begin{bmatrix}1 & 1\end{bmatrix} \begin{bmatrix}10.7398 \\ 0.3793\end{bmatrix} + \begin{bmatrix}2.7599 \\ -3.7212 \\ -2.3118 \\ -1.7334 \\ 0.1229 \\ -4.5850 \\ 6.2676 \\ -2.2272 \\ 1.5486 \\ 3.0417 \\ -0.1867 \\ 1.0247\end{bmatrix}.$$

For each $\tilde{x}_0$ in the marginal local mixed model, there is an estimate of the parameter vector $\hat{\boldsymbol{\beta}}_0$. For a bandwidth of 0.05, the marginal local population average fit at x̃=1 is

$$\begin{aligned}\widehat{Y}_{PA,0} &= \tilde{\mathbf{x}}_0' \hat{\boldsymbol{\beta}}_0^M \\ &= \begin{bmatrix}1 & 1\end{bmatrix} \begin{bmatrix}10.7408 \\ 0.3860\end{bmatrix}.\end{aligned}$$

Figure 3: Conditional Local Mixed Model with h=0.05 (Plot of Population Average and Cluster Specific Curves by Station

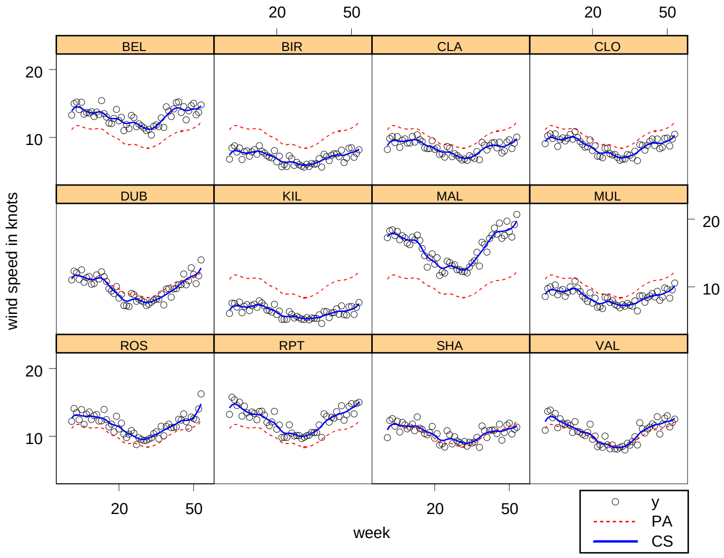


Figure 4: Marginal Local Mixed Model with h=0.05 (Plot of Population Average Curve by Station

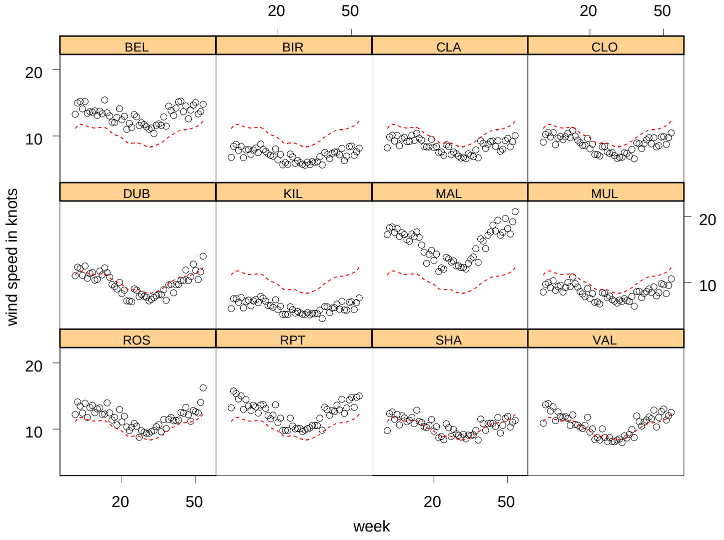
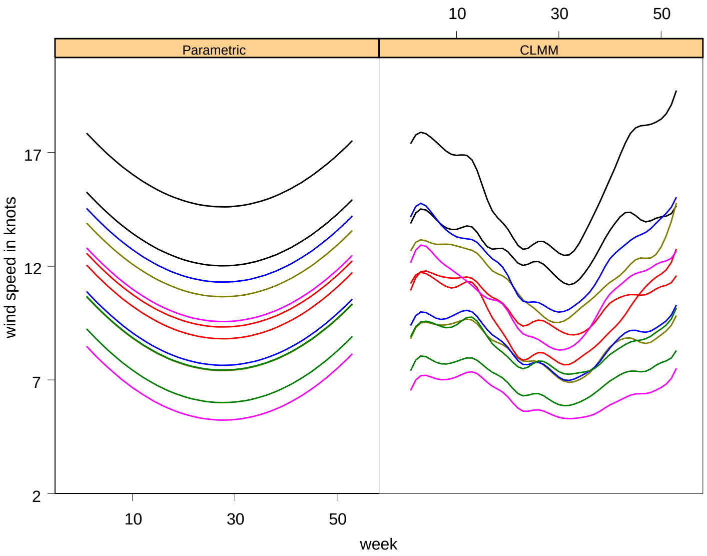


Figure 5: Plot of CLMM and Parametric Cluster Specific Fits (h=0.05)



Notice that the vector $\hat{\boldsymbol{\beta}}_0$ for the two local models are close, indicating that the population average fits at $\tilde{\mathrm{x}}$=1 for the two local models are almost identical.

Trellis plots by cluster appear in Figures 3 and 4. Figure 3 contains the population average and cluster specific curves for the conditional local mixed model using a bandwidth of 0.05. The population average is the dotted line and the cluster specific fit is the solid line. Figure 4 plots the population average curve by station for the marginal local mixed model.

The population average fits are again the same for every cluster. For some of the clusters, the population average is a poor fit. The cluster specific fits are impressive, however. Figure 5 is a comparison of the CLMM cluster specific fits with a bandwidth of 0.05 and the parametric cluster specific fits. The local cluster specific fits are tremendously flexible. Because they are fit pointwise, they no longer follow a particular form. In the parametric model, a random intercept term meant that the cluster specific fits were shifted parabolas; they could never cross. This is not true with the local models. A random intercept term in the local model is also a shift, but it is a shift at a particular point. That shift differs as one moves across the values of the regressor. This allows local fits that potentially could cross.

The local fits are an improvement over the parametric fit. Notice that the drop in wind speed in midyear is captured in both the population average and cluster specific fits, while capturing the level wind speeds in the winter months. The nonparametric mixed model can capture this trend, whereas the specified parametric model was unable to model these trends.

Our mixed model robust regression estimate is

$$\hat{\mathbf{Y}}^{\mathrm{MMRR}} = (1-\lambda)\hat{\mathbf{Y}}^{\mathrm{P}} + \lambda\hat{\mathbf{Y}}^{\mathrm{NP}}$$

where $\hat{\mathbf{Y}}^{\mathrm{P}}$ is the parametric linear mixed model fit and $\hat{\mathbf{Y}}^{\mathrm{NP}}$ is the nonparametric fit. For population average mixed model robust regression using the conditional local mixed model, the parametric and CLMM population average fits are used to find the MMRR population average fits. Mixed model robust regression using the marginal local mixed model uses the parametric and MLMM population average fits. Cluster specific mixed model robust regression uses the parametric and CLMM cluster specific fits in the calculation of the MMRR cluster specific fits.

For mixed model robust regression using CLMM, the estimate of λ was 0.86, and the estimate for MMRR involving MLMM was 1. A λ of 1 corresponds to a mixed model robust regression fit equal to the local fit, so MMRR using the marginal local mixed model is just the marginal local mixed model fit. The MMRR fit using the conditional local mixed model does not strictly use the conditional local fit, as λ does not equal 1. The estimation of λ for MMRR using CLMM involves the cluster specific fits, and the estimate of λ less than 1 suggests that the cluster specific fits may benefit from the smoothness of the parametric regression curve. Because the estimate of λ is so close to one, it would be hard to distinguish the local and mixed model robust fits in a trellis plot like those given above.

Notice that both estimates of lambda are fairly large. This is consistent with our findings given above; the parametric fit can be poor for some clusters, and there is a considerable difference between the parametric and nonparametric fits for some clusters. The nonparametric methods were an improvement over the parametric fits, and the estimates of λ should be close to 1.

7 DISCUSSION

The local mixed model methods offer population average and cluster specific fits with tremendous flexibility. This flexibility is due in part to the fact that they are fit pointwise and therefore able to model trends that the specified parametric model may be incapable of modeling. The local models are typically simple; fitting a local linear, as done in this paper, or a local cubic mixed model with a random intercept at each $\tilde{\mathrm{x}}_0$ value will suffice.

PRESS should be used as the bandwidth selector for population average estimation. Conversely, PRESS** is the bandwidth selector of choice for cluster specific prediction. These conclusions are consistent with the work of Clark (2002) and Mays, Birch, and Starnes (2001). The bandwidth selectors are also performing as expected; evidence of this fact is found by comparing the bandwidths selected from PRESS and PRESS** with the optimal bandwidths from the simulation (results not presented here, but are presented in Waterman (2002)).

The simulation studies indicate that the marginal local mixed model should be used for population average estimation. When using PRESS as the bandwidth selector, the marginal model outperformed the conditional local mixed model in terms of minimizing the integrated mean square error. In addition, the population average model robust mixed model using CLMM has large INTMSE values in comparison to the population average mixed model robust values using MLMM for moderate to large model misspecification. For cluster specific prediction, the conditional local mixed model should be used, as the marginal local mixed model is inappropriate for cluster specific inference.

The mixed model robust methods (using the marginal local mixed model for the population average and the conditional local mixed model for cluster specific inference) are extremely competitive in terms of minimizing the mean square error. With no misspecification, the parametric model should have the smallest INTMSE; the model robust methods are very close to the parametric values for the correctly specified model. For low to moderate misspecification ($0 \leq \gamma \leq 1$) in the simulation study, for example) the mixed model robust methods often have the smallest mean square error when compared to the parametric and local methods. When the model is grossly misspecified (for example, when γ=1 in the simulation study), the local methods have the minimum mean square errors, with the mixed model robust mean squares comparable to the local values.

The mixing parameter estimates for mixed model robust regression are found using the formula from Mays, Birch, and Starnes (2001), but adapted for the cluster correlated random coefficient model. When comparing the average mixing parameter to the simulated optimal mixing parameter for a given combination, $\bar{\lambda}$ for MLMM was close to the optimal value for low to high misspecification.

For the conditional local mixed model, $\bar{\lambda}$ was comparable to the optimal value for little to no model misspecification. However, λ_{opt} and $\bar{\lambda}$ differed for moderate to large misspecification, indicating that when the model has been misspecified, the cluster specific mixed model robust fit favors the parametric fit; it does not use as much of the local fit as it should.

Finally, we can conclude that working with correlated data creates results that may be counterintuitive. Our intuition, often based upon prior work with independent data, was often off the mark due to the lack of consideration of the correlated nature of our data. For example, at first it was counterintutive that the bandwidth in our local models would decrease as the amount of correlation increased. Upon further inspection, we realized that this finding was due to the marginal correlation inherent in the local mixed model. And although we felt that the misspecification term in our simulations would influence the estimate of the correlation, it was unexpected that as γ increased the estimates of ρ in the AR(1) cases either increased or decreased depending upon the magnitude of the correlation; further work indicated that the sinusoidal nature of the misspecifiation term was the reason for this result.

APPENDIX: Aymptotic Theory for the Mixing Parameter

We can write our population average model as

$$y_{ij} = \theta(x_{ij}) + \epsilon_{ij} \tag{49}$$

and the cluster specific model for the i$^{\text{th}}$ cluster as

$$y_{ij} = \theta_i(x_{ij}) + \epsilon_{ij} \tag{50}$$

for i=1,...,s and j=1,...,n$_{\text{i}}$. The functions $\theta(x_{ij})$ and $\theta_i(x_{ij})$ are the true population average and cluster specific mean functions. The asymptotic theory presented here will be for the population average only; we will then assume that "asymptotic"means that the number of observations increases without bound through the number of clusters, as the cluster is the independent unit. In other words, the number of clusters s $\rightarrow \infty$ for fixed values of the regressor. As the data are correlated, asymptotic theory for the cluster specific model is complicated, and has yet to be solved.

We will assume that $E(\boldsymbol{\epsilon}_i) = \mathbf{0}, Var(\boldsymbol{\epsilon}_i) = \mathbf{R}_i$, and $Var(\mathbf{Y}_i) = \mathbf{V}_i$. It is assumed that the values of the regressors x are fixed uniformly on the compact set C in $\aleph$ and that $\boldsymbol{\theta} = \left[\theta(x_{11}), ..., \theta(x_{sn_s})\right]'$ is continuous.

The two estimates used in the MMRR formulation can generically be written as the parametric estimate ($\hat{f}$) and the nonparametric estimate ($\hat{g}$), so that

$$\hat{\theta}(x_{ij}) = (1-\lambda)\hat{f}(x_{ij}, \hat{\beta}) + \lambda\hat{g}(x_{ij}) = (1-\lambda)\hat{f} + \lambda\hat{g}. \tag{51}$$

In this work, the fits $\hat{f}$ and $\hat{g}$ are the population average fits.

We will define the inner product similar to Mays, Birch, and Starnes (2001) and Burman and Chaudhuri (1992), as

$$\langle \mathbf{h}_1, \mathbf{h}_2 \rangle = n^{-1} \sum_{i=1}^{n} h_1(x_i) h_2(z_i), \tag{52}$$

where h_1 and h_2 are two functions of x_i and z_i. The norm is similarly defined as

$$\parallel \mathbf{h}_1 \parallel^2 = \langle \mathbf{h}_1, \mathbf{h}_2 \rangle. \tag{53}$$

Define distances δ_s and γ_s as

$$\delta_s = \{inf \parallel \theta - f(\hat{\beta}) \parallel : \hat{\beta} \in \Re^2\} \tag{54}$$

and

$$\gamma_s^2 = E(\parallel \hat{g}(x_{ij}) - \theta \parallel^2). \tag{55}$$

The first measure is the smallest distance between the parametric fit and the true model. If the infimum is unique, the location of this infimum will be denoted as β^*. The subscript s denotes the fact that this distance measure is dependent upon the number of clusters; note, however, that s→ ∞, so this distance measure approaches an integral. The $lim_{s\to\infty}\delta_s$ is equal to zero if the true model θ is contained in the class of parametric functions under consideration by the user. Otherwise, if the limit of this distance is not zero, the parametric model has been misspecified.

The second measure γ_s is the average squared distance between the nonparametric estimate and the true regression function, with the subscript s indicating the measure's dependency on cluster size. The measure γ_s is the average mean square error (AVEMSE) of Mays (1995).

Now, consider the distance between the model robust estimate and the true regression function

$$\parallel (1-\lambda)\hat{f} + \lambda\hat{g} - \theta \parallel . \tag{56}$$

The value of λ that minimizes this distance is the theoretically optimal mixing parameter. Because this distance is the square root of a sum of squares, the norm is a monotonically increasing function. Thus, the minimum distance is the minimum of the sum of squares. It is easy to show that the minimum is attained at

$$\lambda^* = \frac{\langle \hat{f} - \hat{g}, \theta - \hat{f} \rangle}{\| \hat{f} - \hat{g} \|^2}. \tag{57}$$

The theoretically optimal mixing parameter of course depends on the unknown quantity θ. The estimate of λ^* is the data driven mixing parameter given by

$$\hat{\lambda}^* = \frac{\langle \hat{f}_{-i} - \hat{g}_{-i}, Y - \hat{f} \rangle}{\| \hat{f} - \hat{g} \|^2} \tag{58}$$

where $\hat{f}_{-i}$ and $\hat{g}_{-i}$ are the parametric and nonparametric estimated by deleting the i^{th} cluster. This estimate is an extension of the work by Mays, Birch, and Starnes (2001) and Burman and Chaudhuri (1992) for the fixed effects case, where the data are not clustered and $\hat{f}_{-i}$ and $\hat{g}_{-i}$ refer to the parametric and nonparametric fits with the i^{th} point deleted.

The following three assumptions will be needed for the results that follow:

$$A1.\ \|f(\hat{\beta}, .) - f(\beta^*, .)\| = \mathrm{O_p}(\pi)$$

$$A2.\ \frac{\|\hat{g} - \theta\|^2 - \mathrm{E}(\|\hat{g} - \theta\|^2)}{\mathrm{E}(\|\hat{g} - \theta\|^2)} \overset{\mathrm{P}}{\Rightarrow} 0, \text{ as s } \rightarrow \infty$$

$$A3.\ \lim_{\mathrm{s} \rightarrow \infty} \gamma_{\mathrm{s}}^{-1} \pi = 0.$$

The first assumption provides the parametric convergence rate between the optimal parametric estimate (denoted by $f(\beta^*, .)$) and the user's parametric estimate (given as $f(\hat{\beta}, .)$). The second assumption indicates that the distance $\|\hat{g} - \theta\| = \mathrm{O_p}(\gamma_{\mathrm{s}})$. The third assumption says that the nonparametric estimate has a slower convergence rate than the parametric estimate. (The nonparametric convergence rate is γ_{s}, while the faster parametric rate is π). With these assumptions, we can state the following two lemmas and theorem:

Lemma 1 : Assuming that the assumptions A1- A3 hold,

$$\|\hat{f} - \hat{g}\| = \begin{cases} \mathrm{O_p}(1), & \text{if } \lim_{\mathrm{s} \rightarrow \infty} \delta_{\mathrm{s}} \neq 0 \\ \mathrm{O_p}(\gamma_{\mathrm{s}}), & \text{if } \delta_{\mathrm{s}} = 0. \end{cases}$$

Lemma 2 : Assuming that the assumptions A1- A3 hold,

$$\lambda^* = \begin{cases} \mathrm{O_p}(\gamma_\mathrm{s}), & \text{if } \lim_{\mathrm{s}\to\infty} \delta_\mathrm{s} \neq 0 \\ \mathrm{O_p}(\pi\gamma_\mathrm{s}^{-1}), & \text{if } \delta_\mathrm{s}= 0 \end{cases}$$

Theorem 1 : Assuming that the assumptions A1- A3 hold,

$$\|(1-\lambda^*)\hat{f} + \lambda^*\hat{g} - \theta\| = \begin{cases} \mathrm{O_p}(\gamma_\mathrm{s}), & \text{if } \lim_{\mathrm{s}\to\infty} \delta_\mathrm{s} \neq 0 \\ \mathrm{O_p}(\pi), & \text{if } \delta_\mathrm{s}= 0. \end{cases}$$

Lemma 1 gives the convergence rates of the distance between the parametric and nonparametric estimate. Recall that δ_s is zero if the parametric estimate is correct, and $lim_{s\to\infty}\delta_s$ does not equal zero if the parametric estimate is incorrect. Thus, the distance between $\hat{f}$ and $\hat{g}$ is dependent upon the user's parametric model. Lemma 2 gives the convergence rate of the asymptotically optimal mixing parameter. Notice again the dichotomy - the case where the parametric model is correct, and the case where the parametric model has been misspecified. Theorem 1 states that the distance between the mixed model robust estimate using the asymptotically optimal mixing parameter and the true regression function converges at the faster parametric rate if the parametric model has been correctly specified. Otherwise, the distance between the mixed model robust estimate using λ^* and θ converges at the nonparametric rate.

Asymptotic results are needed for the asymptotically optimal data driven mixing paramter. This is not a straightforward extension of previous work, however. Past asymptotic results for the data driven estimate of the mixing parameter have utilized Whittle's inequality (1960), which assumes independence of the data. This is not the case in our work because the data are marginally correlated. Thus, asymptotic results for the data driven mixing parameter is considered future work.

REFERENCES

Allen, D. (1974), "The relationship between variable selection and data augmentation and a method for prediction", *Technometrics*, **16**, 125-127.

Assaid, C. and Birch, J.B. (2000), “Automatic Bandwidth Selection in Robust Nonparametric Regression”, *Journal of Statistical Computation and Simulation*, **66**, 259-272.

Burman, P., and Chaudhuri, P. (1992), “A hybrid approach to parametric and nonparametric regression”, *Technical Report No. 243*, Division of Statistics, University of California at Davis, Davis, CA.

Clark, S.K. (2002), “Model robust regression based on generalized estimating equations”, *Ph.D. dissertation*, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Craven, P., and Wahba, G. (1979), “Smoothing noisy data with spline functions”, *Numerical Mathematics*, **31**, 377-403.

Einsporn, R.L., and Birch, J.B. (1993), “Model robust regression: using nonparametric regression to improve parametric regression analyses”, *Technical Report 93-5*, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Fan, J. and Gijbels, I. (1995), “Adpative oder polynomial fitting: bandwidth robustification and bias reduction”, *Journal of Computational and Graphical Statistics*, **4**, 213-227.

Härdle, W. (1990), *Applied Nonparametric Regression.* New York : Cambridge University Press.

Härdle, W., and Marron, J. (1985), ”Optimal bandwidth selection in nonparametric regression function estimation”, *Annals of Statistics*, **13**, 1465-1481.

Härdle, W., and Marron, J. (1995), ”Fast and simple scatteplot smoothing”, *Comp. Statist. Data Anal.*, **20**, 1-17.

Haslett, J., and Raftery, A.E. (1989), “Space-time Modelling with Long-memory Dependence: Assessing Ireland’s Wind Power Resources (with Discussion)”, *Applied Statistics*, **38**, 1-50.

Henderson, C.R. (1950), “The estimation of genetic parameters”, *Annals of Mathematical Statistics*, **21**, 309-310.

Hilden-Minton (1995), “Multilevel Diagnostics in Mixed and Hierarchial Linear Models”, *Ph.D. dissertation*, University of California at Los Angeles, Los Angeles, CA.

Hurtado-Rodriguez, G.I. (1993), “Detection of Influential Observations in Linear Mixed Models”, *Ph.D. dissertation*, North Carolina State University, Raleigh, NC.

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

Mays, J.E., Birch, J.B., and Einsporn, R.L. (2000), ”An overview of model robust regression”, *Journal of Statistical Computation and Simulation*, **66**, 79-100.

Mays, J.E., Birch, J.B., and Starnes, B.A. (2001), ”Model robust regression: combining parametric, nonparametric, and semiparametric methods”, *Journal of Nonparametric Statistics*, **13**, 245-277.

Nadaraya, E.A. (1964), “On estimating regression”, *Theory of Probability and its Applications*, **9**, 141-142.

Nottingham, Q. and Birch, J.B. (2000), ”A semiparametric approach to analysing dose-response data”, *Statistics in Medicine*, **19**, 3, 389-404.

Robinson, T.J. and Birch, J.B. (2000) “Dual Model Robust Regression”, *Journal of Statistical Computation and Simulation*, **66**, 113-126.

Speckman, P. (1988), “Kernel smoothing in partial linear models”, *Journal of the Royal Statistical Society B*, **50**, 413-436.

Waterman, M. (2002), “Linear Mixed Model Robust Regression”. Unpublished Ph.D. dissertation, Department of Statistics, Virginia Polytechnic Institute and State University.

Watson, G. (1964), “Smooth regression analysis”, *Sankhya Series A*, **26**, 359-372.

Whittle, P. (1960), “Bounds for the Moments of Linear and Quadratic Forms in Independent Variables", *Theory of Probability and its Applications*, **3**, 302-305.