Linear Mixed Model Robust Regression

Megan Janet Tuttle Waterman

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Approved:

_____

Jeffrey B. Birch, Co-Chairman

_____

Oliver Schabenberger, Co-Chairman

_____

Christine M. Anderson-Cook

_____

Eric P. Smith

_____

George R. Terrell

May 8, 2002
Blacksburg, Virginia

KEYWORDS: semiparametric, mixed effects, robust

Linear Mixed Model Robust Regression

Megan Janet Tuttle Waterman

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. Model robust regression techniques predict mean response as a convex combination of a parametric and a nonparametric model fit to the data. It is a semiparametric method by which incompletely or incorrectly specified parametric models can be improved through adding an appropriate amount of a nonparametric fit. We apply this idea of model robustness in the framework of the linear mixed model. The mixed model robust regression (MMRR) predictions we propose are convex combinations of predictions obtained from a standard normal-theory linear mixed model, which serves as the parametric model component, and a locally weighted maximum likelihood fit which serves as the nonparametric component. An application of this technique with real data is provided.

Acknowledgments

Writing a dissertation requires tremendous amounts of time, energy, and patience, and would be impossible without the help of others. That said, I would like to thank everyone who helped me complete my dissertation. Specifically, I would like to thank:

My Ph.D. advisors, Dr. Jeffrey B. Birch and Dr. Oliver Schabenberger. Dr. Birch's knowledge of model robust regression and Dr. Schabenberger's expertise in mixed models defined this work. You have been role models for teaching and research, and you have taught me so much. You have been supportive of my ideas and decisions, both scholastic and career-wise, and have always been there with sage advice. I marvel at the patience, time, and effort that you devote to me, and I greatly appreciate it. It is because of you that I have developed an appreciation for research and have had much success. I have enjoyed working with you, and look forward to working with you in the future.

My committee members, Dr. Christine M. Anderson-Cook, Dr. Eric P. Smith, and Dr. George R. Terrell. Your enthusiasm and talent for teaching, research, and consulting continue to inspire me. Thank you for agreeing to serve on my committee and for all of your helpful suggestions.

The Department of Statistics at Virginia Polytechnic Institute and State University. Thank you for giving me the opportunity to learn. Thank you to Mike Box, who went out of his way to help me with my simulation studies, and to Linda Breeding and Michele Marini, who offered a listening ear when it was needed.

My friends, both those here at Virginia Tech and elsewhere. I am grateful for all of your support. In particular, I would like to thank Seth Clark, Ilya Lipkovich, and Bennett Sango Otieno. Seth gave many helpful comments, insights, and encouragement, Ilya always supplied a laugh when I needed it, and Bennett provided companionship in the computer lab.

My family. The encouragement from my parents, Earl and Constance Tuttle, my sister Amy, the Waterman family, and my husband Richard, has been unwavering. You have always been there for me, from celebrating my accomplishments to listening to my concerns. Rick, thank you for doing far more than your share of the housework, tolerating my late nights of studying and writing, and putting up with me. I love you!

Thank you all for your guidance and encouragement!

–Meg

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

Regression models are commonly used to describe a relationship between a response variable and a set of regressors. For example, a linear relationship between a response and an explanatory variable may be modeled as

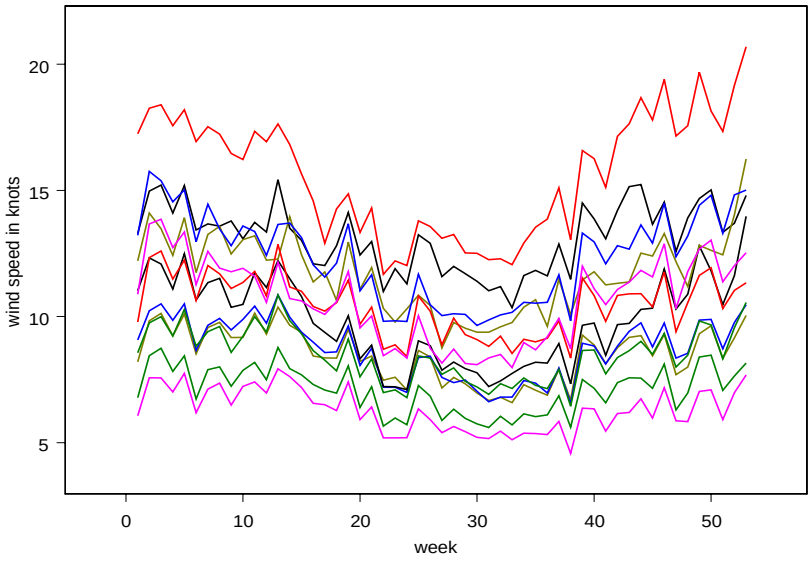
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ for i=1,...,n.} \tag{1.1}$$

In equation (1.1), β_0 and β_1 are unknown regression coefficients. The classical regression model assumes that the coefficients in the model are fixed effects. In other words, the parameters are constants. The data are used to obtain estimates of these constants.

However, there are many situations where the coefficients in a regression model could be thought of as random. In this situation, the coefficients are known as random effects or random coefficients. Random coefficients are randomly selected from a population of coefficients; they are no longer fixed constants but random variables. Inference on random effects relates to the entire population of effects whereas results of fixed effects analysis apply only to those effects chosen. Whether an effect is fixed or random depends upon the interests of the researcher.

Mixed models are models that contain at least one fixed effect and at least two random effects (including the error term). When mixed models arise in practice, the data are often grouped together by a common characteristic. These groups are known as clusters. Clustered data include repeated measures on subjects (where the subject is the cluster) and split-plot agricultural experiments (where the whole-plot is the cluster). An example of clustered data

Figure 1.1: Plot of Wind Speed versus Week by station



is the wind speed data set from Haslett and Raftery (1989). Twelve meteorological stations in Ireland were selected and the average wind speeds in knots were measured daily during the years 1961 through 1978. This analysis looks 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. The wind speeds were measured over time. Measurements were taken at the same fifty three time points for each station. In this example the cluster is the station, hence there are fifty three observations per cluster and a total of 636 observations in the data set. Figure 1.1 is a profile plot of the wind speed data with each line representing wind speed versus week for each cluster. The analysis of the wind speed data will be presented in several later chapters. The weekly wind speed data appear in Appendix A.

Mixed model analysis appeals to the clustered data scenario for many reasons. The mixed model approach permits a population average curve– a curve for the unconditional mean– to be estimated. The population average curve is the same for every cluster. In addition, the cluster specific curves– curves for the mean conditioned on the random effects–

can also be estimated for each cluster. For example, the wind speed data set would have a curve for the entire population of stations based on the fixed effects, and separate curves for each station based on the fixed and random effects.

With clustered data and mixed models the need for estimation of the variance components arises. Variance components are defined here as all unknown parameters of the marginal variance-covariance matrix of the response. It will be shown in Chapter Two that variance components will play a vital role with clustered data.

The classical linear mixed model (conditional and unconditional) assumes a linear parametric form for the mean. Maximum likelihood or restricted maximum likelihood estimates of the fixed effects and solutions of the random effects are then obtained by solving a set of linear equations known as the mixed model equations. Using these equations, estimates of the fixed effects and predictions of the random effects for the model are obtained so that smooth curves for the population average and cluster specific averages can be calculated. However, the true model is most likely unknown. The parametric mixed model may be misspecified in a number of ways, including an incorrect covariance structure, wrong distributional assumptions, wrong effect classification (either fixed or random), and wrong model matrices. Misspecification of the model matrices will be the primary focus in what follows.

The fixed and random effects can also be estimated nonparametrically by using a locally weighted mixed model. The nonparametric fit will often provide a fit superior to its parametric counterpart, capturing structure in the data that a misspecified parametric fit is incapable of modeling. However, nonparametric regression may also follow irregularities in the data and may result in fits that lack smoothness and often are too variable. In addition, if we use a nonparametric mixed model any valid information about the parametric structure known to the modeler is ignored.

The parametric method is superior if the user correctly specifies the parametric form of the linear mixed model. However, if the model is misspecified, the nonparametric method gains appeal. Mays, Birch, and Starnes (2001) have developed methods to combine the

parametric and nonparametric fits in a convex combination. This semiparametric method can yield estimates of the mean function with lower bias than the parametric approach and lower variance than the nonparametric approach. This technique is extended to the mixed model setting in this dissertation.

The dissertation is organized in the following fashion: Ch. 2, Ch. 3, and Ch. 4 give an introduction to the parametric, nonparametric, and semiparametric methods of estimation in fixed effects models, respectively. Ch. 5 through 7 extend these methods to the linear mixed model. Ch. 5 is an overview of the parametric methods for the linear mixed model. Ch. 6 emphasizes the locally weighted or nonparametric mixed model, and Ch. 7 pertains to the estimation, prediction, and inferential procedures of linear mixed model robust regression. Ch. 8 presents the results of a simulation study comparing the parametric, nonparametric, and model robust mixed models. Finally, Ch. 9 contains conclusions and outlook on future work on linear mixed model robust regression.

Chapter 2

Parametric Estimation for the Fixed Effects Model

Regression analysis is used to model relationships between a response and one or more regressors or explanatory variables. One goal of regression analysis is to determine the regression function, or the conditional mean of the response at fixed values of the regressor variable. One way that this can be accomplished is by using a parametric model.

2.1 The Model

One commonly used form of a parametric model is the linear model of conditional means that can be expressed as

$$y_i = E(y_i|x_{1i}, x_{2i}, ..., x_{ki}) + \epsilon_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \tag{2.1}$$

or in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\mathbf{Y}$ is an (n x 1) vector of responses, $\mathbf{X}$ is an (n x (k+1)) model matrix, $\boldsymbol{\beta}$ is a ((k+1) x 1) vector of unknown fixed parameters, and $\boldsymbol{\epsilon}$ is an (n x 1) vector of random errors. The i^{th} row of the model matrix $\mathbf{X}$ will be denoted as $\mathbf{x}_i'$. It is assumed throughout that the mean model is conditioned on the values of the regressor, and such notation will be omitted for the remainder of the dissertation.

2.2 Ordinary Least Squares

Ordinary least squares (OLS) strives to find estimates of the mean response in a fixed effects regression model. The assumptions for (2.2) are that the errors are uncorrelated and identically distributed with zero mean and common finite variance σ^2. One interpretation of the fact that $\mathrm{E}(\boldsymbol{\epsilon})=\mathbf{0}$ is the implication that model (2.2), the so-called "user's model", is the correct model. Under these assumptions, it is easily shown that

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

$$\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}, \tag{2.3}$$

where $\mathbf{I}$ is the (n x n) identity matrix.

The least squares estimator $\hat{\boldsymbol{\beta}}^{\mathsf{OLS}}$ minimizes the sum of squares

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*)'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*) \tag{2.4}$$

over all possible values of vectors in the parameter space, respresented by $\boldsymbol{\beta}^*$. Taking the derivative of (2.4) with respect to $\boldsymbol{\beta}^*$ and setting this result equal to the ((k+1) x 1) zero vector produces the normal equations

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*) = \mathbf{0}. \tag{2.5}$$

Solving (2.5) for $\boldsymbol{\beta}^*$ yields the least squares estimator $\hat{\boldsymbol{\beta}}^{\mathsf{OLS}}$, given by the well known expression[1]

$$\hat{\boldsymbol{\beta}}^{\mathsf{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{2.6}$$

The fitted least squares regression function can be expressed in terms of $\hat{\boldsymbol{\beta}}^{\mathsf{OLS}}$ as

$$\hat{\mathbf{Y}}^{\mathsf{OLS}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{OLS}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}^{\mathsf{OLS}}\mathbf{Y}, \tag{2.7}$$

where $\mathbf{H}^{\mathsf{OLS}}$ is called the ordinary least squares hat matrix. The OLS hat matrix is symmetric and idempotent,

[1] If the inverse of $(\mathsf{X}'\mathsf{X})$ does not exist, a generalized inverse may be used in its place.

$$\mathbf{H}^{\mathsf{OLS}}\mathbf{H}^{\mathsf{OLS}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}^{\mathsf{OLS}}.$$

Because $\mathbf{H}^{\mathsf{OLS}}$ is symmetric and idempotent, the trace of $\mathbf{H}^{\mathsf{OLS}}$ equals the rank of $\mathbf{H}^{\mathsf{OLS}}$. The rank of the OLS hat matrix equals the number of parameters in the model, or k+1. The properties of the hat matrix will be of importance in bandwidth selection and will be discussed in future chapters.

Notice that the ordinary least squares fits at each observed set of regressors, $\mathbf{x}_i'$, can be expressed as

$$\hat{y}_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \sum_j \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j y_j = \sum_j h_{ij}^{\mathsf{OLS}} y_j, \tag{2.8}$$

where $h_{ij}^{\mathsf{OLS}} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j$ is the element in row i and column j of the hat matrix. This means that the ordinary least squares fits are weighted sums of the observations where the weights of $\hat{y}_i$ are determined by the i^{th} row of the hat matrix. It is easy to show that under the given assumptions,

$$\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{\beta}}^{\mathsf{OLS}}) &= \boldsymbol{\beta} \\
\mathrm{Var}(\hat{\boldsymbol{\beta}}^{\mathsf{OLS}}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
\mathrm{E}(\hat{\mathbf{Y}}^{\mathsf{OLS}}) &= \mathrm{E}(\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{OLS}}) = \mathbf{X}\boldsymbol{\beta} \\
\mathrm{Var}(\hat{\mathbf{Y}}^{\mathsf{OLS}}) &= \sigma^2\mathbf{H}^{\mathsf{OLS}}.
\end{aligned}$$

Thus, $\hat{\boldsymbol{\beta}}^{\mathsf{OLS}}$ is an unbiased estimator for $\boldsymbol{\beta}$ if the user's model is correct. By the Gauss-Markov theorem, $\hat{\boldsymbol{\beta}}^{\mathsf{OLS}}$ is furthermore the best linear unbiased estimator (BLUE) under squared error loss.

If, in addition to the assumptions listed above, the normality assumption on the random errors is invoked, the least squares estimator of $\boldsymbol{\beta}$ is also the uniform minimum variance unbiased estimator (UMVU estimator). The UMVU property is a stronger result than the BLUE property because an UMVU estimator is not restricted to the class of linear estimators. In addition to the regression coefficients, the variance of the error terms, σ^2, must be estimated. A common estimator for the variance of the error terms is the error

mean square

$$s^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{OLS}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{OLS}})}{\text{n-(k+1)}}, \tag{2.9}$$

where k+1 is the number of model terms. The error mean square is an unbiased estimator for σ^2, provided that the model is correct. Notice that s^2 is not the maximum likelihood estimator for σ^2, as the denominator of (2.9) is not equal to n.

2.3 Generalized Least Squares

Another parametric model, a slight generalization of (2.2), is the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.10}$$

where again $\mathbf{Y}$ is an (n x 1) response vector, $\mathbf{X}$ is an (n x (k+1)) model matrix, and $\boldsymbol{\beta}$ is a ((k+1) x 1) vector of fixed and unknown parameters. Suppose, however, that $\boldsymbol{\epsilon}$ is an (n x 1) vector of random errors with zero mean and variance-covariance matrix $\mathbf{V}$. For example, it might be the case that the uncorrelated errors have nonconstant variance, or $\text{Var}(\epsilon_i) = \sigma_i^2$. In this case, the variance-covariance matrix $\mathbf{V}$ is a diagonal matrix with the diagonal elements equal to the error variances, or $\mathbf{V}$=diag$(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) = \langle \sigma_i^2 \rangle$. In general, the matrix $\mathbf{V}$ is assumed to be positive definite. For the model given in (2.10) and under the assumption that $\text{Var}(\boldsymbol{\epsilon}) = \mathbf{V}$, it can be shown that

$$\begin{aligned} \text{E}(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\beta} \\ \text{Var}(\mathbf{Y}) &= \mathbf{V}. \end{aligned}$$

The least squares estimators are found by minimizing the generalized sum of squares

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*)'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*) \tag{2.11}$$

over all $\boldsymbol{\beta}^*$. The derivative of the expression in (2.11) with respect to $\boldsymbol{\beta}^*$, when set equal to the ((k+1) x 1) zero vector, results in the "generalized normal equations"

$$\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*) = \mathbf{0}. \tag{2.12}$$

The solution of (2.12) may be expressed as

$$\hat{\boldsymbol{\beta}}^{\mathsf{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \tag{2.13}$$

and is referred to as the generalized least squares (GLS) estimator. The fitted values at the n observed regressor locations $\mathbf{x}_i'$ are

$$\hat{\mathbf{Y}}^{\mathsf{GLS}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{GLS}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \mathbf{H}^{\mathsf{GLS}}\mathbf{Y}. \tag{2.14}$$

The matrix $\mathbf{H}^{\mathsf{GLS}}$ is idempotent as

$$\begin{aligned}\mathbf{H}^{\mathsf{GLS}}\mathbf{H}^{\mathsf{GLS}} &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\\ &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\\ &= \mathbf{H}^{\mathsf{GLS}},\end{aligned}$$

but $\mathbf{H}^{\mathsf{GLS}}$ is not necessarily symmetric. Notice that OLS is a special case of GLS with $\mathbf{V} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the (n x n) identity matrix.

The expectation and variance of both the GLS estimator and GLS fitted values are

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}^{\mathsf{GLS}}) &= \boldsymbol{\beta}\\ \mathrm{Var}(\hat{\boldsymbol{\beta}}^{\mathsf{GLS}}) &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\\ \mathrm{E}(\hat{\mathbf{Y}}^{\mathsf{GLS}}) &= \mathrm{E}(\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{GLS}}) = \mathbf{X}\boldsymbol{\beta}\\ \mathrm{Var}(\hat{\mathbf{Y}}^{\mathsf{GLS}}) &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'.\end{aligned}$$

Using a generalized version of the Gauss-Markov Theorem, it can be shown that $\hat{\boldsymbol{\beta}}^{\mathsf{GLS}}$ is BLUE in terms of variance minimization. Under the additional assumption of Gaussian errors, $\hat{\boldsymbol{\beta}}^{\mathsf{GLS}}$ is the UMVU estimator (Myers 1990, Searle 1971).

Notice that the equations (2.13) and (2.14) rely on knowledge of the variance-covariance matrix. In practice, $\mathbf{V}$ is seldom known. Suppose that $\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{U} = \mathbf{V}$ is known up to a scalar σ^2. This constant can be estimated (Schabenberger and Pierce, 2001) as

$$\hat{\sigma}^2 = \frac{1}{\text{n-rank}(\mathbf{X})}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{GLS}})'\mathbf{U}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{GLS}}). \tag{2.15}$$

Again, notice that $\hat{\sigma}^2$ is not the maximum likelihood estimate, as the denominator of (2.15) does not equal n.

The other possibility is that $\text{Var}(\mathbf{Y}) = \mathbf{V}$ is completely unknown. An unstructured variance-covariance matrix could be assumed with the estimates obtained by the method of moments. See Deaton, Reynolds, and Myers (1983) for a discussion of weighted least squares.

In either case, the estimated variance-covariance matrix $\hat{\mathbf{V}}$ is substituted into the sum of squares expression (2.12). The estimated generalized least squares estimator (EGLS), obtained by minimizing the sum of squares containing $\hat{\mathbf{V}}$, is denoted by $\hat{\boldsymbol{\beta}}^{\mathsf{EGLS}}$. It is imperative, however, that the estimate of $\mathbf{V}$ be consistent in order for $\hat{\boldsymbol{\beta}}^{\mathsf{EGLS}}$ to be consistent.

2.4 Summary

This chapter demonstrated ordinary and generalized least squares, two common parametric techniques for the fixed effects model. When the parametric fixed effects model has been misspecified over a portion of the data, however, other techniques such as nonparametric models must be used. Chapter 3 will discuss nonparametric techniques for the fixed effects model.

Chapter 3

Nonparametric Estimation for the Fixed Effects Model

In Chapter 2, standard parametric regression techniques for the fixed effects model were discussed. If the parametric model is misspecified, however, the estimates and predictions (and hence the fits) are biased. One way to reduce this bias is to estimate the mean function nonparametrically.

3.1 The Model

Suppose that the underlying model is

$$y_i = g(x_i) + \epsilon_i, \tag{3.1}$$

where y_i is the i^{th} observation (i=1,...,n), $g(x_i)$ is an unknown, smooth function of the single regressor x, and the error terms ϵ_i have zero mean. Further assume that the correlation between y_i and y_j for ($i \neq j$) is zero. Nonparametric regression will provide an estimate of $g(x_0)$ at x_0 for an arbitrary $x_0 \in \aleph$ (where $\aleph$ is the x-space). This estimate can be expressed as

$$\hat{g}(x_0) = \sum_{j=1}^{n} u_{0j} y_j, \tag{3.2}$$

where $0 \leq u_{0j} \leq 1$, and u_{0j} is the weight assigned to the j^{th} observation for the estimation at x_0. In other words, the estimate of the mean function at x_0 is a weighted sum of the

responses. Kernel regression and local polynomial regression, two commonly used nonparametric methods, will produce estimates of the mean function at x_0 that are a weighted sum of the responses.

3.2 Kernel Regression

Kernel regression estimates $g(x_0)$, the mean response at x_0 for arbitrary $x_0 \in \aleph$. Specifically, kernel regression intends to find the estimate of $g(x_0)$ by solving

$$\min \sum_{j=1}^{n} k_{0j}(y_j - g(x_0))^2. \tag{3.3}$$

The kernel regression estimate of the regression function at x_0 is a weighted average of the responses

$$\hat{g}(x_0) = \frac{\sum_{j=1}^{n} k_{0j} y_j}{\sum_{j=1}^{n} k_{0j}}. \tag{3.4}$$

The weights k_{0j} for kernel regression must be chosen so that the responses associated with observed x nearest x_0 receive the most weight. A common weighting scheme by Nadaraya (1964) and Watson (1964) propose the weights for estimation of the mean response at x_0 for the j^{th} response as

$$k_{0j} = \frac{K\left(\frac{|x_0 - x_j|}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{|x_0 - x_j|}{h}\right)}, \tag{3.5}$$

where K is the kernel function and h is the bandwidth, often an unknown parameter.[1] A kernel function is a decreasing function of the distance between x_0 and x_j; consequently, observations at locations close to x_0 receive more weight than observations far from x_0. Kernel funtions are often based on symmetric density functions. Some common kernels include the uniform, normal, Epanechnikov, and tricube kernels. (See Härdle (1990) for these and other kernel functions). The kernel function used in the current work is the Gaussian, or normal kernel given by

$$K\left(\frac{|x_0 - x_j|}{h}\right) = e^{-\left(\frac{|\hat{x}_0 - \hat{x}_j|}{h}\right)^2}$$

[1] Other weighting schemes may be found in Priestley and Chao (1972) and Gasser and Müller (1979).

where $\hat{x}_j = \frac{x_j}{x_{max} - x_{min}}$ represents a rescaling of x to be between 0 and 1. Studies have shown that the choice of the kernel function has less impact on the estimate of the regression function compared to the choice of the bandwidth.

The bandwidth controls the smoothness of the estimated regression function, $\hat{g}(x)$. A larger value of the bandwidth results in a smoother function; a smaller value of the bandwidth results in a less smooth function. If the chosen bandwidth is "too large" the resulting $\hat{g}(x)$ may be too smooth, resulting in fits with low variance but high bias. The estimted regression function may miss important features in the data. The opposite is true for a small bandwidth. Thus, the chosen bandwidth must provide a balance between bias and variance. Bandwidth selectors will be the focus of section 3.4.

The estimated regression function at x_0 using kernel regression for a given kernel and bandwidth is

$$\hat{g}(x_0) = \sum_{j=1}^{n} k_{0j} y_j. \tag{3.6}$$

In matrix form, the kernel regression estimate can be expressed as

$$\hat{g}(x_0) = \mathbf{h}_0^{ker\prime} \mathbf{Y}, \tag{3.7}$$

where $\mathbf{h}_0^{ker\prime} = (k_{01}\ k_{02} \dots\ k_{0n})$. The kernel regression estimate at the regressor locations x_1, x_2, ... , x_n can be expressed as

$$\hat{\mathbf{g}} = \hat{\mathbf{Y}} = \mathbf{H}^{ker} \mathbf{Y}, \tag{3.8}$$

where

$$\mathbf{H}^{ker} = \begin{bmatrix} \mathbf{h}_1^{ker\prime} \\ \mathbf{h}_2^{ker\prime} \\ \vdots \\ \vdots \\ \mathbf{h}_n^{ker\prime} \end{bmatrix}$$

is the kernel "hat" matrix or smoother matrix. The rows of the kernel "hat" matrix are $\mathbf{h}_i^{ker\prime} = \begin{pmatrix} k_{i1}\ k_{i2} \dots\ k_{in} \end{pmatrix}$, where k_{ij} is the kernel weight given to the j^{th} response in prediction at x_i.

One disadvantage of kernel regression is that the estimate of the mean at the boundaries, near the first and n^{th} order statistics of the regressor, is often biased. Recall that the kernel function is a decreasing function of the distance from x_0 and x_j. At the first order statistic of the regressor, for example, there are no regressor values to the left of $x_{(1)}$. Only the observations to the right of the response at $x_{(1)}$ receive nonzero weight and thus the mean estimate at $x_{(1)}$ is a weighted average of observations at or to the right of the observation at $x_{(1)}$. If, for example, all of the responses in the data set are greater than the response at $x_{(1)}$, the estimate of the mean at $x_{(1)}$ is a weighted average of values larger than the observed response at $x_{(1)}$. The estimate of the mean at $x_{(1)}$, $\hat{g}(x_{(1)})$ then will most likely be biased. A similar argument holds for $x_{(n)}$, the maximum regressor value. Local polynomial regression improves upon the bias shortcoming of kernel regression.

3.3 Local Polynomial Regression

Local polynomial regression, or LPR (Cleveland, 1979), estimates $\hat{g}(x_0)$, the mean at x_0 for $x_0 \in \aleph$, by fitting a d^{th} order polynomial at x_0

$$\hat{y}_0 = \hat{g}(x_0) = \beta_0 + \beta_1 x_0 + \beta_2 x_0^2 + \cdots + \beta_d x_0^d + \epsilon_0,$$

where β, β_1, β_2, $\cdots$, β_d are fixed, unknown parameters and ϵ_0 is an unknown random error assumed to have zero mean and variance σ^2. The polynomial is commonly of order d=3, local cubic regression, or of order d=1, local linear regression. The polynomial of order d=0 results in kernel regression. However, local linear or local cubic regressions are usually preferred over kernel regression or local polynomials of even order as local polynomials of odd order for mean function estimation have less bias than even order (Fan and Gijbels, 1995).

The unknown parameters are estimated by finding

$$\min_{\beta_0,\beta_1,\beta_2,\ldots,\beta_d} \sum_{j=1}^{n} k_{0j}(y_j - \beta_0 - \beta_1 x_j - \beta_2 x_j^2 - \ldots - \beta_d x_j^d)^2, \tag{3.9}$$

where k_{0j} are the Nadaraya-Watson weights defined in (3.5) for the j^{th} observation for estimation at x_0 and d is the degree of the polynomial. Again, notice for d equal to zero, we have

$$\min_{\beta_0} \sum_{j=1}^{n} k_{0j}(y_j - \beta_0)^2,$$

and the resulting estimate of $\hat{\beta}_0$ is just $\hat{g}(x_0)$ in (3.4), the kernel regression estimate.

The local polynomial fit at x_0 can be expressed in matrix form as

$$\hat{g}(x_0) = \tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{W}_0\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{W}_0\mathbf{Y} = \mathbf{h}_0^{LPR'}\mathbf{Y}, \tag{3.10}$$

where $\tilde{\mathbf{X}}$ is a model matrix for a dth order polynomial

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^d \end{bmatrix}$$

and

$$\tilde{\mathbf{x}}_0' = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^d \end{bmatrix}.$$

Note that the notation $\tilde{\mathbf{X}}$ is used here to distinguish this model matrix from $\mathbf{X}$, the model matrix in the parametric model. The matrices $\mathbf{X}$ and $\tilde{\mathbf{X}}$ may contain different regressors and hence may not be identical. The matrix $\mathbf{W}_0$ is a diagonal matrix containing the kernel weights associated with x_0

$$\mathbf{W}_0 = \begin{bmatrix} k_{01} & 0 & 0 & \dots & 0 \\ 0 & k_{02} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & k_{0n} \end{bmatrix} = \langle k_{0j} \rangle.$$

As in kernel regression, the estimate of the mean function at x_0 is a weighted sum of the responses, but the weights at x_0 for local polynomial regression are different than the weights at x_0 for kernel regression. Specifically, the weight vector associated with x_0 for kernel regression is $\mathbf{h}_0^{ker'} = \begin{pmatrix} k_{01} & k_{02} \dots & k_{0n} \end{pmatrix}$, whereas the weight vector associated with x_0 for local

polynomial regression is $\mathbf{h}_0^{\mathsf{LPR}\prime} = \tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{W}_0\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{W}_0$. Note that local polynomial regression also requires a bandwidth to calculate $\mathbf{W}_0$ as the kernel function depends on the bandwidth.

A weighted least squares regression problem is solved at each value of x_0. In matrix form, the LPR regression estimate at the regressor values $x_1, ..., x_n$, can be expressed as

$$\hat{\mathbf{g}} = \mathbf{H}^{\mathsf{LPR}}\mathbf{Y}, \tag{3.11}$$

where

$$\mathbf{H}^{\mathsf{LPR}} = \begin{bmatrix} \mathbf{h}_1^{\mathsf{LPR}\prime} \\ \mathbf{h}_2^{\mathsf{LPR}\prime} \\ \vdots \\ \mathbf{h}_n^{\mathsf{LPR}\prime} \end{bmatrix}$$

with $\mathbf{h}_i^{\mathsf{LPR}\prime} = \tilde{\mathbf{x}}_i'(\tilde{\mathbf{X}}'\mathbf{W}_i\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{W}_i$. The matrix $\mathbf{H}^{\mathsf{LPR}}$ is the local polynomial "hat" matrix. Further results pertaining to local polynomial regression can be found in Fan and Gijbels (1992, 1995).

3.4 Bandwidth Selection

Both kernel regression and local polynomial regression require the selection of a bandwidth. The bandwidth parameter essentially determines the width of a window where observations that fall outside the window receive negligible weight. A window of width 2ϵ is the set of all points (a,y) such that $a \in (x-\epsilon, x+\epsilon)$, or that a is in the neighborhood of size ϵ of x. A large bandwidth has a large window and so x_j far from x_0 may receive non-negligible weight. A bandwidth that is too large may result in a fit that tends to be smooth and, perhaps, misses important features of the mean function. Such fits are often characterized as having low variance and high bias.

Similarly, a small bandwidth results in a small window so only those x_j close to x_0 receive significant weight. An LPR fit to the data may be overfit for a bandwidth that is too small, resulting in an estimated regression function with more features than in the mean function. Fits of this type characteristically have little bias but high variability.

Bandwidth selection then becomes a tradeoff between bias and variance. Because it is desired to minimize both of these properties, a natural criterion for bandwidth selection

would be to choose h to minimize a function of the squared error of estimation of mean response. 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 criterions, including average squared error (ASE)

$$\text{ASE}=\frac{1}{n}\sum_{i=1}^{n}(\hat{g}(x_i)-g(x_i))^2,$$

integrated squared error (ISE)

$$\text{ISE}=\int_{\aleph}(\hat{g}(x_i)-g(x_i))^2 f(x)dx,$$

and conditional mean integrated squared error (CISE)

$$\text{CISE}=E(\text{ISE}|x_1,...,x_n).$$

Thus the average squared error, which is calculated only at the data points, is a discrete approximation to the integral given in the integrated squared error. The function $g(x_i)$, used in the definition of the distance measures ASE, ISE, and CISE, is unknown. We can instead minimize an estimate of ASE. One estimate of ASE is a cross-validation criterion

$$\text{CV}(h)=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_{i,-i})^2, \tag{3.12}$$

where n is the sample size and $\hat{Y}_{i,-i}$ is the estimate of the regression function at x_i with the i^{th} data point (y_i, x_i) removed. Notice that CV(h) is an estimate of ASE, where Y_i is the estimate of $g(x_i)$ and $\hat{Y}_{i,-i}$ is the estimate of $\hat{g}(x_i)$. The leave-one-out fits for the fixed effects model can easily be found by applying the Sherman-Morrison-Woodbury Theorem (Myers, 1990).

The PRESS statistic (Allen, 1974) is an example of a cross-validation statistic. It is given by

$$\text{PRESS}=\sum_{i=1}^{n}(Y_i-\hat{Y}_{i,-i})^2 \tag{3.13}$$

and is known as the prediction error sum of squares. The bandwidth can be chosen as the value h that minimizes PRESS. Simulation studies for the traditional fixed effects model

for uncorrelated data have shown that bandwidths chosen by PRESS are often too small, resulting in an overfit model.

One solution is to use a penalized cross-validation technique (Craven and Wahba, 1979). Penalized cross-validation incorporates a penalty function into the cross-validation expression. In the case of PRESS, the penalty function must correct for the small bandwidth preference. One penalized cross-validation technique is PRESS*, defined as

$$\text{PRESS}^* = \frac{\text{PRESS}}{\text{n-trace}(\mathbf{H}^{\mathsf{LPR}})}. \tag{3.14}$$

The chosen bandwidth is that which minimizes PRESS*. Notice that the elements along the diagonal of $\mathbf{H}^{\mathsf{LPR}}$ will be large for a small value of the bandwidth. This increases PRESS* and protects against choosing bandwidths that are too small. However, it has been shown in Mays, Birch, and Starnes (2001) that bandwidths chosen by PRESS* are often too large for the uncorrelated fixed effects model.

PRESS** was created to alleviate the problem of extreme bandwidths. PRESS**, also a penalized PRESS statistic, is defined as

$$\text{PRESS}^{**} = \frac{\text{PRESS}}{\text{n-trace}(\mathbf{H}^{\mathsf{LPR}}) + (\text{n-d})\left[\frac{\mathsf{SSE}_{\mathsf{max}} - \mathsf{SSE}_{\mathsf{h}}}{\mathsf{SSE}_{\mathsf{max}}}\right]}, \tag{3.15}$$

where $\mathsf{SSE}_{\mathsf{max}}$ is the largest error sum of squares over all possible values of the bandwidth, $\mathsf{SSE}_{\mathsf{h}}$ is the error sum of squares associated with a particular value of the bandwidth h, and d is the number of parameters estimated at each $\tilde{\mathbf{x}}_0'$ by LPR (i.e., d=1 for kernel regression and 2 for local linear regression). Notice that the term

$$(\text{n-d})\left[\frac{\mathsf{SSE}_{\mathsf{max}} - \mathsf{SSE}_{\mathsf{h}}}{\mathsf{SSE}_{\mathsf{max}}}\right]$$

approaches zero for large bandwidths ($\mathsf{SSE}_{\mathsf{h}}$ approaches the maximum error sum of squares for large bandwidths). Thus, PRESS** guards against overly small and overly large bandwidths. Results in Mays, Birch, and Starnes (2001) show that PRESS** performs well for model robust regression, the topic of Chapter 4.

There are many other bandwidth selectors. Plug-in methods, where unknown quantities in the squared error function are replaced with estimates, are very popular. Ruppert,

Sheather, and Wand (1995) offer a plug-in bandwidth selector for nonparametric regression. Rule of thumb selectors are appealing because of ease of calculation. Rule of thumb bandwidth selectors, which provide simple but crude estimates of the bandwidth, are used when bandwidth choice is not crucial (Fan and Gijbels, 1996). The bandwidth selector used for nonparametric mixed models and mixed model robust regression will be developed in Chapters 6 and 7.

3.5 Summary

Nonparametric regression methods for the fixed effects model were the topic of this chapter. For a misspecified parametric model, nonparametric regression offers the user a method to obtain a regression curve with less bias than the misspecified parametric curve. However, the nonparametric fits may be quite variable. As a result, semiparametric methods have been developed that utilize both parametric and nonparametric fits to either the data or to residuals. Semiparametric regression is discussed in Chapter 4 for the fixed effects model and in Chapter 7 for the mixed model.

Chapter 4

Model Robust Regression for the Fixed Effects Model

The word "semiparametric" is used in the statistics literature in many contexts. Here, a "semiparametric" method combines parametric and nonparametric elements. This chapter summarizes the literature on semiparametric model techniques for the traditional regression setting with an emphasis on model robust regression methods.

4.1 An Introduction to Semiparametric Regression

Semiparametric regression methods utilize both parametric and nonparametric fits to either the data, to the residuals, or to both. Semiparametric regression has recently received considerable attention because of its flexibility and its applicability to disciplines like economics and pharmacology. See, for example, Robinson (1988), Wooldridge (1992), Ullah and Vinod (1993), and Fan and Ullah (1999).

One of the earliest papers on semiparametric regression was Speckman (1988). His method assumed that the response could be modeled as a linear predictor in the regressors $x_1,...,x_k$ plus some smooth unknown function f of regressors $t_1,...,t_p$. The partial linear regression (PLR) method, based on Speckman's model (see Mays, Birch, and Starnes, 2001), assumes that both the linear predictor and f depend on the same 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.

The main disadvantage of PLR is that the entire nonparametric fit to the residuals is added back to the parametric estimate of the linear predictor, even if the parametric fit is adequate. This could lead to a substantial increase in the variance of the fits. Model robust regression, allowing for only a portion of the fits to be added back, will decrease this variance.

4.2 Model Robust Regression 1 (MRR1)

Einsporn (1987) and Einsporn and Birch (1993) developed Model Robust Regression 1, or MRR1. The underlying model is

$$y_i = g(x_i) + \epsilon_i \tag{4.1}$$

for i=1,...n, where y_i is the i^{th} observation, $g(x_i)$ is some smooth, unknown, continuous function of the single regressor x_i, and the errors are assumed to be independent and identically distributed random variables with mean zero and variance σ^2. Model misspecification occurs when the user's specified parametric model is incorrect. For example, the user may specify the model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$. The user's model is correct only if $g(x_i) = \mathbf{x}_i'\boldsymbol{\beta}$, otherwise it is misspecified in functional form, the type of model misspecification considered here.

It is known that the parametric fit yields smooth estimates of the mean response that are low in variance. It may not pick up important structure in the data when the model is misspecified, however. The parametric fit then results in a biased estimate of the mean. Conversely, nonparametric estimates of the mean response, with a reasonable choice of bandwidth, attempt to fit the structure of the data. Thus, the nonparametric estimates result in low bias but may have large variances. The goal of MRR1 is to provide a smooth estimate of the mean response that captures all of the important trends in the data set, thereby reducing bias while simultaneously reducing the variance.

Any appropriate method can be used to obtain the parametric fits. For example, if the data are independently and identically normally distributed, least squares (equivalent to maximum likelihood) is a common parametric method. If the data set contains independent

0-1 responses, quantal regression based on maximum likelihood estimation (for example, logistic regression) would be the parametric method of choice (Myers, 1990). Local polynomial regression with perhaps d=1 or d=3 can be performed for the nonparametric fit. The Model Robust Regression 1 (MRR1) fit is defined as

$$\hat{\mathbf{y}}^{\mathsf{MRR1}} = (1-\lambda)\hat{\mathbf{y}}^{\mathsf{P}} + \lambda\hat{\mathbf{y}}^{\mathsf{NP}}, \tag{4.2}$$

where $\hat{\mathbf{y}}^{\mathsf{P}}$ is the parametric fit, $\hat{\mathbf{y}}^{\mathsf{NP}}$ is the nonparametric fit, and $\lambda \in [0,1]$ is a mixing parameter. Thus, the MRR1 estimate is a convex combination of the parametric and nonparametric fits. MRR1 results in a purely nonparametric fit when λ=1 and a parametric fit when λ=0. Mays, Birch, and Starnes (2001) show through a series of Monte-Carlo simulations that when using a proper mixing parameter selector, a model specified correctly results in an MRR1 estimate equal to the parametric fit. If the model is badly misspecified, the MRR1 estimate is the same or nearly equal to the nonparametric fit. The simulation results also show that MRR1 performs as well as or better than the separate parametric and nonparametric fits in terms of mean square error minimization under low to moderate model misspecification.

The MRR1 method will produce a smooth fit while capturing the structure that the misspecified parametric method is incapable of modeling. Mays, Birch, and Starnes (2001) prove that, under certain conditions given in Starnes (1999), the MRR1 estimate using the data-driven theoretically optimal estimate of the mixing parameter ($\hat{\lambda}^*$) converges to true the mean function at the parametric rate if the model is correctly specified. If the model is incorrectly specified, the MRR1 estimate using $\hat{\lambda}^*$ converges to the true mean at the same rate at which the nonparametric estimate converges to the mean function.

4.3 Model Robust Regression 2 (MRR2)

Model Robust Regression 2 (MRR2) was developed as an alternative to MRR1 (Mays, Birch, and Starnes, 2001). The underlying model and assumptions for MRR2 are the same as those for MRR1. If the parametric model is incorrectly specified, the goal of MRR2 is to again

combine parametric and nonparametric fits to obtain an improved fit to the data.

The parametric fit $\hat{\mathbf{y}}^{\mathsf{P}}$ and the residuals $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}^{\mathsf{P}}$ are obtained using the parametric method appropriate for the assumptions. The residuals represent the structure in the data not captured by the parametric model. This structure can be modeled nonparametrically. Local polynomial regression provides a nonparametric fit to the residuals, $\hat{\mathbf{r}} = \mathbf{H}^{\mathsf{LPR}}\mathbf{r}$, where $\mathbf{H}^{\mathsf{LPR}}$ is the local polynomial "smoother" or "hat" matrix. In MRR2, a portion of this nonparametric residual fit is added back to the parametric fit. The MRR2 fit is then defined as

$$\hat{\mathbf{y}}^{\mathsf{MRR2}} = \hat{\mathbf{y}}^{\mathsf{P}} + \lambda\hat{\mathbf{r}}, \tag{4.3}$$

where $\lambda \in [0, 1]$ is the mixing parameter.

As shown in Mays, Birch, and Starnes (2001), the MRR2 fit is equal, or nearly equal to, the parametric fit when the model has little or no misspecification. When the model is badly misspecified, the MRR2 fit adds a large portion of the nonparametric residual fit to the parametric fit. For low to moderate model misspecification, the MRR2 method outperforms the parametric, nonparametric, and, in general, the MRR1 method in terms of minimizing the mean square error, although MRR1 is very competitive to MRR2.

Model robust regression has been extended to a number of scenarios, including quantal data (Nottingham and Birch, 2000), dual modeling (Robinson, 1997), outlier resistance modeling (Assaid, 1997), and with generalized estimating equations (Clark, 2002).

4.4 Choice of the Mixing Parameter

The mixing parameter for model robust regression is an unknown parameter and must be estimated based upon the sample data. Mays, Birch, and Starnes (2001) propose a data driven, asymptotically optimal estimator of the mixing parameter:

$$\hat{\lambda}^{*}_{\mathsf{MRR1}} = \frac{\sum_{\mathsf{i}}(\hat{\mathrm{y}}^{\mathsf{NP}}_{\mathsf{i},\text{-}\mathsf{i}} - \hat{\mathrm{y}}^{\mathsf{P}}_{\mathsf{i},\text{-}\mathsf{i}})(\mathrm{y}_{\mathsf{i}} - \hat{\mathrm{y}}^{\mathsf{P}}_{\mathsf{i}})}{\sum_{\mathsf{i}}(\hat{\mathrm{y}}^{\mathsf{NP}}_{\mathsf{i}} - \hat{\mathrm{y}}^{\mathsf{P}}_{\mathsf{i}})^2}. \tag{4.4}$$

Here, $\hat{\mathrm{y}}^{\mathsf{NP}}_{\mathsf{i},\text{-}\mathsf{i}}$ and $\hat{\mathrm{y}}^{\mathsf{P}}_{\mathsf{i},\text{-}\mathsf{i}}$ are the nonparametric and parametric fits at x_{i} with the i^{th} data point removed, and $\hat{\mathrm{y}}^{\mathsf{NP}}_{\mathsf{i}}$ and $\hat{\mathrm{y}}^{\mathsf{P}}_{\mathsf{i}}$ are the nonparametric and parametric fits at x_{i} using the entire

data set. Likewise, the asymptotically optimal data driven mixing parameter for MRR2 is

$$\hat{\lambda}^*_{\mathsf{MRR2}} = \frac{\sum_{\mathsf{i}} \hat{\mathrm{r}}_{\mathsf{i}}(\mathrm{y}_{\mathsf{i}} - \hat{\mathrm{y}}^{\mathsf{P}}_{\mathsf{i}})}{\sum_{\mathsf{i}} \hat{\mathrm{r}}_{\mathsf{i}}^{\,2}}, \tag{4.5}$$

where $\hat{\mathrm{r}}_{\mathsf{i}}$ is the local linear fit to the residuals of the parametric fit at x_{i}. That is, $\hat{\mathrm{r}}_{\mathsf{i}} = \mathbf{h}_{\mathsf{i}}^{\mathsf{LLR}'}\mathbf{r}$, where $\mathbf{h}_{\mathsf{i}}^{\mathsf{LLR}'}$ is the vector of local linear weights and $\mathbf{r}$ is the vector of residuals from the parametric model.

Mays, Birch, and Starnes (2001) provide asymptotic results for the MRR2 estimate, specifically convergence rates of the MRR2 estimate. The asymptotic results for MRR2 are similar to those of MRR1.

4.5 Summary

Two methods of model robust regression for the traditional regression setting have been proposed. In addition, an asymptotically optimal data driven mixing parameter has been provided. Both model robust regression estimates may result in lower bias and/or smaller variance than either the parametric or nonparametric estimates for low to moderate misspecification.

Using the parametric and nonparametric mixed model developed in Chapters 5 and 6, a mixed model robust regression (MMRR) estimate will be proposed in Chapter 7. Also under consideration will be mixing parameter estimation for the MMRR estimate.

Chapter 5

Parametric Estimation for the Mixed Model

A mixed model is defined as a model that contains at least one fixed effect and at least two random effects, including the error term. Mixed models are widely used in many disciplines. Henderson (1950), for example, one of the earliest authors on mixed models, applied mixed models to estimation problems in the animal sciences. Laird and Ware (1982) generalized Henderson's work; their formulation is commonly referred to as the Laird-Ware model.

5.1 The Laird-Ware Model

For the fixed effects models of Chapter 2, the response vector $\mathbf{Y}$ was defined as the sum of the linear predictor and error, where the linear predictor was comprised of fixed effects. The Laird-Ware model includes an additional linear combination of the random coefficients. The model consists of a fixed part and a random part.

Clustered data, or data grouped into clusters by a common trait, are often modeled with a mixed model. The wind speed data set from Chapter 1 took repeated measurements on twelve stations, with each station representing a cluster. As will be shown in section 5.5, the wind speed data set can be analyzed as a mixed model.

The matrix formulation for the Laird-Ware model for clustered data is

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{5.1}$$

where subscripts identify clusters. The number of observations for the i^{th} cluster is denoted by n_i. Akin to the fixed effects model, $\mathbf{Y}_i$ is an (n_i x 1) vector of responses, $\mathbf{X}_i$ is an (n_i x p_1) model matrix, and $\boldsymbol{\beta}$ is a (p_1 x 1) vector of unknown fixed parameters. The matrix $\mathbf{Z}_i$ is an (n_i x p_2) model matrix and $\mathbf{b}_i$ is a (p_2 x 1) vector of random variables.

The model for the entire data set can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{5.2}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ : \\ \mathbf{Y}_s \end{bmatrix} \text{ is a stacked vector of cluster response vectors}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ : \\ \mathbf{X}_s \end{bmatrix} \text{ is a matrix of model matrices stacked by cluster}$$

$\boldsymbol{\beta}$ is the (p_1 x 1) vector of fixed parameters given in (5.1)

$$\boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ : \\ \boldsymbol{\epsilon}_s \end{bmatrix} \text{ is a vector of error vectors stacked by cluster}$$

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ : \\ \mathbf{b}_s \end{bmatrix} \text{ is a vector of random effect vectors stacked by cluster}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \mathbf{0} \\ : & : & : \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_s \end{bmatrix}.$$

Note that $\mathbf{Z}$ is a block diagonal matrix with the random effects model matrices for each cluster on the diagonals. The Laird-Ware model includes an additional linear combination of the random coefficients. This means that there is a fixed part ($\mathbf{X}\boldsymbol{\beta}$) and a random part ($\mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$) to the mixed model.

For convenience, stronger distributional assumptions than those given for the fixed effects model of Chapter 2 are often made to permit likelihood-based inference. The random effects vector $\mathbf{b}_i$ is assumed to be normally distributed with zero mean and variance -covariance matrix $\mathbf{D}$. The vector of random errors $\boldsymbol{\epsilon}_i$ contains normal random variates with

mean zero and variance-covariance matrix $\mathbf{R}_i$. The matrix $\mathbf{R}_i$ represents the within-cluster variances and covariances, while the matrix $\mathbf{D}$ contains the variances and covariances of the between-cluster effects. The vectors $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are further assumed to be independent.

In matrix notation, the distributional assumptions can be expressed as

$$\begin{bmatrix} \mathbf{b}_i \\ \boldsymbol{\epsilon}_i \end{bmatrix} \sim \text{MN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_i \end{bmatrix} \right). \tag{5.3}$$

Notice that the expected value of $\mathbf{Y}_i$ conditioned on the random effects for that cluster equals

$$\text{E}(\mathbf{Y}_i|\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \tag{5.4}$$

and is often referred to as the "cluster specific" mean. Also, the marginal expected value of $\mathbf{Y}_i$ may be found by

$$\text{E}(\mathbf{Y}_i) = \text{E}(\text{E}(\mathbf{Y}_i|\mathbf{b}_i)) = \mathbf{X}_i\boldsymbol{\beta}. \tag{5.5}$$

$\text{E}(\mathbf{Y}_i)$ is referred to as the "marginal" mean. The variance of $\mathbf{Y}_i$ can be found using the conditional expectation and conditional variance

$$\text{Var}(\mathbf{Y}_i|\mathbf{b}_i) = \mathbf{R}_i \tag{5.6}$$

as

$$\text{Var}(\mathbf{Y}_i) = \text{E}(\text{Var}(\mathbf{Y}_i|\mathbf{b}_i)) + \text{Var}(\text{E}(\mathbf{Y}_i|\mathbf{b}_i)) = \mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' = \mathbf{V}_i. \tag{5.7}$$

The mixed model, along with the normal assumptions on $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$, implies that the conditional distribution of $\mathbf{Y}_i|\mathbf{b}_i$ is normal with mean $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ and variance-covariance matrix $\mathbf{R}_i$ and that the marginal distribution of $\mathbf{Y}_i$ is normal with mean $\mathbf{X}_i\boldsymbol{\beta}$ and variance-covariance matrix $\mathbf{V}_i$.

5.2 Estimation of Fixed Effects and Prediction of Random Effects

It is often of interest to obtain the maximum likelihood estimate of $\boldsymbol{\beta}$ and the predictor of $\mathbf{b}$. These estimates and predictions can be obtained by maximizing the joint "likelihood" of

$\mathbf{b}$ and $\boldsymbol{\epsilon}$. Since $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}$ and $\mathbf{b}$ are independent and are both normally distributed, the joint density of $\mathbf{b}$ and $\boldsymbol{\epsilon}$ can be expressed as

$$f(\mathbf{b}, \boldsymbol{\epsilon}) = (2\pi)^{-\frac{(n+sp_2)}{2}} \begin{vmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{vmatrix}^{-\frac{1}{2}} \exp -\frac{1}{2} \left\{ \begin{bmatrix} \mathbf{b} \\ \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b} \end{bmatrix}' \begin{vmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{vmatrix}^{-1} \begin{bmatrix} \mathbf{b} \\ \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b} \end{bmatrix} \right\}, \tag{5.8}$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{D} & \mathbf{0} & .. & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & .. & \mathbf{0} \\ : & : & : & : \\ \mathbf{0} & \mathbf{0} & .. & \mathbf{D} \end{bmatrix} \text{ is an } (sp_2 \text{ x } sp_2) \text{ block diagonal matrix and}$$

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} & .. & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & .. & \mathbf{0} \\ : & : & : & : \\ \mathbf{0} & \mathbf{0} & .. & \mathbf{R}_s \end{bmatrix} \text{ is an (n x n) block diagonal matrix.}$$

Setting the derivatives of $\log f(\mathbf{b}, \boldsymbol{\epsilon})$ equal to zero yields the mixed model equations (See Appendix B for derivation)

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & (\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}) \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}. \tag{5.9}$$

Solving the equation in (5.9) yields the estimates[1]

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & (\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}. \tag{5.10}$$

Henderson (1950) and Appendix B show that the estimates of the fixed and random effects can be expressed as

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

$$\hat{\mathbf{b}} = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{5.11}$$

for fixed $\mathbf{R}$ and $\mathbf{B}$ where $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{B}\mathbf{Z}'$. Using the equations in (5.11), the linear parametric cluster specific fits can then be stated as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}. \tag{5.12}$$

[1] If the inverse does not exist, a generalized inverse may be used in its place.

Harville (1976) established optimality of the estimates given in (5.11). Specifically, he showed through an extension of the Gauss-Markov Theorem that $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) and that $\hat{\mathbf{b}}_i$ is the best linear unbiased predictor (BLUP). "Best" in this sense means that the estimator and the predictor minimize mean square error. The vector $\boldsymbol{\beta}$ is a vector of parameters and must be estimated. We denote by $\hat{\boldsymbol{\beta}}$ its estimator. The vector $\mathbf{b}_i$ is a vector of random variabels and must be predicted. We denote by $\hat{\mathbf{b}}_i$ it predictor.

5.3 Variance Component Estimation

Variance components are important in clustered data analysis. With clustered data, a marginal correlation structure is introduced due to the hierarchy of the random effects. This structure occurs even if all of the random effects are independent from one another. For example, in a standard split-plot design, two observations from the same whole-plot are marginally correlated because they share the same whole-plot experimental error. In addition, correlations in the conditional distribution $\mathbf{Y}_i|\mathbf{b}_i$ can be modeled using such common correlation structures as compound symmetry, first-order autoregressive (AR(1)), power, or unstructured, to name but a few.

In practice, the variance-covariance matrices $\mathbf{B}$ and $\mathbf{R}$ are unknown and their unique elements must be estimated. Two commonly used methods are maximum likelihood (ML) and restricted maximum likelihood (REML).

5.3.1 Maximum Likelihood

Consider the vector $\boldsymbol{\theta}$, which contains all unknown variance components and write the variance of $\mathbf{Y}$ as $\mathbf{V}(\boldsymbol{\theta})$. The log likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be written as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{Y}) = -\frac{1}{2}\log\left|\mathbf{V}(\boldsymbol{\theta})\right| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\text{nlog}(2\pi) \tag{5.13}$$

for n equal to the total sample size. The profile likelihood $\ell(\boldsymbol{\theta}|\mathbf{Y})$ can be found by substituting equation (5.11), the expression for $\hat{\boldsymbol{\beta}}$, into $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{Y})$. The profile likelihood is now a function of $\boldsymbol{\theta}$ only and is maximized with respect to $\boldsymbol{\theta}$ to find $\hat{\boldsymbol{\theta}}_{\mathsf{ML}}$. The estimated variance-covariance

matrix, $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_{\mathsf{ML}})$, is substituted into the expression for $\hat{\boldsymbol{\beta}}$ in (5.11) to obtain $\hat{\boldsymbol{\beta}}_{\mathsf{ML}}$.

It is often the case that no closed form solution for $\hat{\boldsymbol{\theta}}_{\mathsf{ML}}$ exists. The variance components are then attained through some iterative process. A commonly used algorithm for variance component estimation is the Newton-Raphson algorithm.

Unfortunately, maximum likelihood estimators of the variance components are typically negatively biased. Restricted maximum likelihood (REML) is an alternative estimation method that results in estimates that have less bias.

5.3.2 Restricted Maximum Likelihood

Consider the linear combination $\mathbf{AY}$, where $\mathbf{A}$ is an (n x n) matrix. The matrix $\mathbf{A}$ is chosen so that $\mathbf{AY}$ does not contain any fixed effects and E($\mathbf{AY}$)=$\mathbf{0}$. The matrix $\mathbf{A}$ is not unique. However, as long as $\mathbf{A}$ is of full rank and $\mathbf{AX}= \mathbf{0}$, the REML variance component estimates will be independent of the choice of $\mathbf{A}$. A proof of this result can be found in Harville (1976). Schabenberger and Pierce (2001) show how to construct a suitable $\mathbf{A}$ matrix.

The REML estimates of the variance components are found by maximizing the likelihood of $\mathbf{AY}$ with respect to the variance components $\boldsymbol{\theta}$. As in maximum likelihood estimation, an iterative procedure such as the Newton-Raphson algorithm is used to obtain the variance component estimates $\hat{\boldsymbol{\theta}}_{\mathsf{REML}}$ if no closed form solution exists.

Restricted maximum likelihood estimates of the variance components usually have less bias than the maximum likelihood variance component estimates. However, restricted maximum likelihood will only estimate the variance components, while maximum likelihood produces estimates of both the variance components and the fixed effects. An EGLS estimate of $\boldsymbol{\beta}$ is obtained by substituting $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_{\mathsf{REML}})$ into (5.11), and is denoted by $\hat{\boldsymbol{\beta}}_{\mathsf{REML}}$.

5.4 Model Selection

A selection criterion to distinguish between competing nested mixed models is necessary. Suppose there is a "full" model that has some number of parameters under consideration and a "reduced" model in which at least one of the parameters is constrained, typically to

zero. Thus, the set of parameters of the reduced model can be thought of as a subset of parameters in the full model. Because of the nested nature of these models, the log likelihood (or -2*log likelihood) can be used to choose between the full and reduced models. The expression $-2\text{lnL}(\hat{\boldsymbol{\beta}}_{\mathsf{ML}}, \hat{\boldsymbol{\theta}}_{\mathsf{ML}})$ denotes twice the negative log of the likelihood evaluated at the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Likewise, the expression $-2\text{lnL}(\hat{\boldsymbol{\beta}}_{\mathsf{REML}}, \hat{\boldsymbol{\theta}}_{\mathsf{REML}})$ is used to denote the negative of two times the natural log of the likelihood evaluted at the restricted maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Then $-2\text{lnL}(\hat{\boldsymbol{\beta}}_{\mathsf{ML}}, \hat{\boldsymbol{\theta}}_{\mathsf{ML}})$ can be used to test hypotheses about fixed effects and variance components. The expression $-2\text{lnL}(\hat{\boldsymbol{\beta}}_{\mathsf{REML}}, \hat{\boldsymbol{\theta}}_{\mathsf{REML}})$, however, is used to test hypotheses about the variance components only. Asymptotically,

$$\left(-2\text{lnL}(\hat{\boldsymbol{\beta}}_{\mathsf{ML}}, \hat{\boldsymbol{\theta}}_{\mathsf{ML}})_{\mathsf{reduced}}\right) - \left(-2\text{lnL}(\hat{\boldsymbol{\beta}}_{\mathsf{ML}}, \hat{\boldsymbol{\theta}}_{\mathsf{ML}})_{\mathsf{full}}\right) \sim \chi^2_{(\mathsf{d})} \tag{5.14}$$

$$\left(-2\text{lnL}(\hat{\boldsymbol{\beta}}_{\mathsf{REML}}, \hat{\boldsymbol{\theta}}_{\mathsf{REML}})_{\mathsf{reduced}}\right) - \left(-2\text{lnL}(\hat{\boldsymbol{\beta}}_{\mathsf{REML}}, \hat{\boldsymbol{\theta}}_{\mathsf{REML}})_{\mathsf{full}}\right) \sim \chi^2_{(\mathsf{d})} \tag{5.15}$$

where d is the number of parameters in the full model minus the number of parameters in the reduced model. A large value of the test statistic would result in the rejection of the null hypothesis that the d parameters excluded from the reduced model are equal to zero.

5.5 An Example of the Parametric Linear Mixed Model

The wind speed data, presented in Chapter One, easily lends itself to linear mixed model analysis. 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 an additional 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 complete data set appears in Appendix A.

Model 1: Fixed Effects Model with fixed X^2, X, and intercept

The fixed effects model can be written in terms of the Laird-Ware model in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

in this case since $\mathbf{Z}$ is the zero matrix. In this model,

$$\mathbf{Y} = \begin{bmatrix} Y_{1,1} \\ : \\ Y_{12,53} \end{bmatrix} = \begin{bmatrix} 13.29 \\ : \\ 12.53 \end{bmatrix}$$

is a (636 x 1) vector of responses,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,1}^2 \\ : & : & : \\ 1 & x_{12,53} & x_{12,53}^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ : & : & : \\ 1 & 53 & 2809 \end{bmatrix}$$

is the (636 x 3) model matrix,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

is a (3 x 1) vector, and $\boldsymbol{\epsilon}$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{1,1} \\ : \\ \epsilon_{12,53} \end{bmatrix}$$

is a (636 x 1) vector of random errors.

We assume that the within-cluster variance-covariance matrix (modeled through $\mathbf{R}_i$) is that of an AR(1) process. The autoregressive lag-one model contains σ^2 on the diagonals of the variance-covariance matrix, and $\text{Cov}(y_{ik},y_{ij}) = \sigma^2\rho^{|k-j|}$ in the (j,k) and (k,j) off-diagonal cells of $\mathbf{R}_i$. The AR(1) model lends itself well to those situations where the observations within a cluster are measured over equally spaced time intervals, as is the case here.

The fixed effects model for the wind speed data set can then be written in scalar form as

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \epsilon_{ij},$$

where x_{ij} is the j^{th} week in which the i^{th} cluster was observed. The marginal distribution of $\mathbf{Y}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix $\mathbf{R}$ where

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} & .. & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & .. & \mathbf{0} & \mathbf{0} \\ : & : & .. & : & : \\ \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{R}_{12} \end{bmatrix},$$

and $\mathbf{R}_i$ has the AR(1) structure described above.

Model 2: Mixed Model with fixed X^2, X, and intercept and a random intercept

The mixed model using model 1 and a random intercept can be written as a Laird-Ware model in matrix form as

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

where $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ are given above, with $\mathbf{Z}$ as the (636 x 12) matrix given by

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & .. & 0 & 0 \\ : & : & .. & : & : \\ 1 & 0 & .. & 0 & 0 \\ 0 & 1 & .. & 0 & 0 \\ : & : & .. & : & : \\ 0 & 1 & .. & 0 & 0 \\ : & : & .. & : & : \\ : & : & .. & : & : \\ 0 & 0 & .. & 0 & 1 \\ : & : & .. & : & : \\ 0 & 0 & .. & 0 & 1 \end{bmatrix},$$

and $\mathbf{b}$ is a (12 x 1) vector of random effects

$$\mathbf{b} = \begin{bmatrix} b_{1,0} \\ b_{2,0} \\ : \\ b_{12,0} \end{bmatrix}.$$

with the "0" subscript denoting the intercept.

This model also assumes that the within-cluster variance is the AR(1) covariance model. The variance-covariance matrix $\mathbf{B}$, measuring the between-cluster variation, is assumed to be diagonal, suggesting that the clusters, or stations, are independent. This may not be the case, as the stations may be spatially correlated; however, due in part to a lack

of information as to the proximity of stations and for the desire for a simple model, the independence structure was selected.

The model for the wind speed data set can then be written as

$$Y_{ij} = (\beta_0 + b_{i,0}) + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \epsilon_{ij},$$

where x_{ij} is the j^{th} week for the i^{th} cluster. The marginal distribution of $\mathbf{Y}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix $\mathbf{R}+\mathbf{ZBZ}'$ where $\mathbf{Z}$ is defined above, $\mathbf{B} = \sigma^2_{b_0}\mathbf{I}_{(12x12)}$ and

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} & .. & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & .. & \mathbf{0} & \mathbf{0} \\ : & : & .. & : & : \\ \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{R}_{12} \end{bmatrix}.$$

As before, $\mathbf{R}_i$ has σ^2 down the diagonal and covariances equal to $\sigma^2\rho^{|j-k|}$ on the (j,k) and (k,j) off-diagonal cells. The distribution of $\mathbf{Y}$ conditioned on the random effects $\mathbf{b}$ is normal with mean $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zb}$ and variance-covariance matrix $\mathbf{R}$, where $\mathbf{Z}, \mathbf{B}$, and $\mathbf{R}$ are defined above.

To decide between the two models, -2lnL($\hat{\boldsymbol{\beta}}_{REML}, \hat{\boldsymbol{\theta}}_{REML}$) for both models was calculated. We consider model 2 to be the "full" model and model 1 to be the "reduced" model, as model 1 is nested within model 2. The hypotheses were $H_0 : \sigma^2_{b_0} = 0$ versus $H_1 : \sigma^2_{b_0} > 0$ with a test statistic equal to -2Log $\text{Likelihood}_{reduced}$ - (-2Log Likelihood_{full}) = 1869.7 - 1765.6 = 104.1. The test statistic is asymptotically distributed as a Chi-square random variable with one degree of freedom, and the approximate p-value is 0.0001. The variance of the intercept is significantly different from zero, and thus the "full" model (the model containing the random intercept) is an improvement over the fixed effects model.

Thus, the parametric model under consideration will be model 2, the model containing a fixed X^2, X, and intercept terms, and a random intercept term. The marginal distribution of $\mathbf{Y}$ is normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix $\mathbf{V}$. The population average curve is estimated as

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

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

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

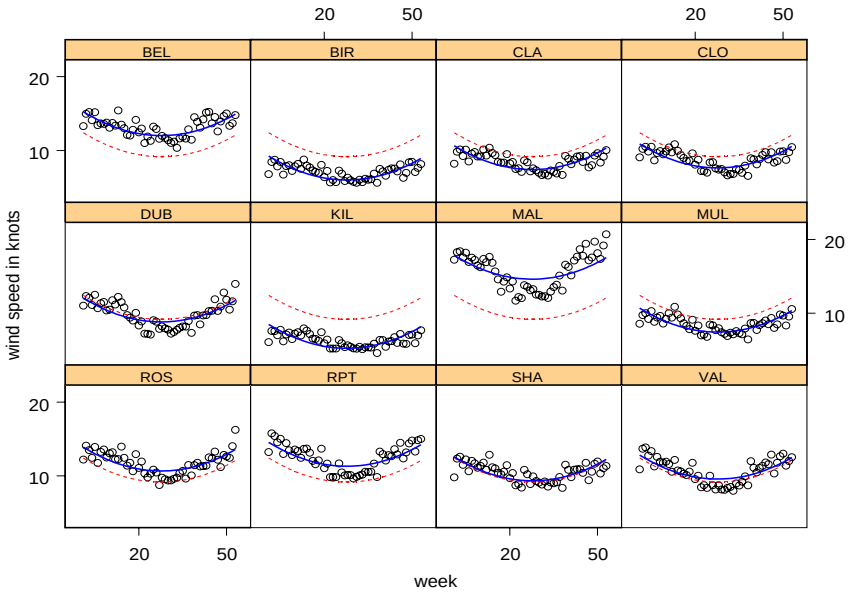
$$\hat{E}(\mathbf{Y}) = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}} = \mathbf{X}\begin{bmatrix} 12.6576 \\ -0.2520 \\ 0.004549 \end{bmatrix} + \mathbf{Z}\begin{bmatrix} 2.8481 \\ -3.1644 \\ -1.7556 \\ -1.5244 \\ -0.3572 \\ -3.9325 \\ 5.4406 \\ -1.7344 \\ 1.4900 \\ 2.1325 \\ 0.1602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

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 5.1. 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

Figure 5.1: Parametric Linear Mixed Model (Plot of Population Average and Cluster Specific Curves by Station)



clusters, like station MAL, exhibit a steep drop in wind speed. During the fall and winter months, the wind speeds appear to increase slightly (left hand portion of the curves) 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.

In order for the user to handle this model misspecification, he or she must consider a more flexible model. The bias of the parametric linear model will be rectified through the use of nonparametric, or local, mixed models.

5.6 Summary

The linear mixed model is very flexible. It allows for both fixed and random coefficients. For clustered data, the linear mixed model permits separate curves for each cluster to be estimated in addition to an average curve for the population. The correlations inherent in many clustered datasets are modeled either directly through $\mathbf{R}_i$ or indirectly through the

choice of random coefficients and random effects.

However, the linear mixed model assumes a parametric form. This model may be misspecified. Since the true model is rarely known, there is often model misspecification. One way to avoid this misspecification is to estimate the mixed model nonparametrically. This concept will be discussed at length in Chapter 6.

Chapter 6

Nonparametric Estimation for the Mixed Model

In Chapter 5, the standard parametric linear mixed model was introduced. The parametric linear mixed model may be misspecified, however. The need for a nonparametric mixed model to reduce the bias incurred by a misspecified model arises. This chapter will begin with an introduction to nonparametric estimation in the mixed model. Two local mixed models and bandwidth selectors for these models will be presented. Finally, an example using the local mixed models will be provided.

6.1 Previous Work on the Nonparametric Estimation of Mixed Models

Much of the literature pertaining to nonparametric estimation in the mixed model has focused on nonparametric estimation of the mixing distribution, or the distribution of the random effects. An early method of obtaining a nonparametric estimate of the mixing distribution is nonparametric maximum likelihood estimation (NPMLE, Laird 1978). The NPMLE method uses the EM algorithm to obtain an estimate of the mixing distribution. It can be shown that the NPMLE will be a step function at discrete points in $\aleph$ (where $\aleph$ is the x-space) with a finite number of steps k.

The smooth nonparametric maximum likelihood estimation method (SNPMLE), proposed by Magder and Zeger (1996), also estimates the distribution of the random effects.

Their estimate of the mixing distribution is a mixture of Gaussian distributions. Magder and Zeger show that the estimated distribution is composed of n or fewer Gaussian distributions, each of which has variance equal to a constant h. The variance h must be specified by the user and thus is akin to a bandwidth in nonparametric regression. It can be shown that the SNPMLE approaches the NPMLE as h approaches zero.

The SNPMLE estimate is preferred to the NPMLE estimate because the SNPMLE does result in a smooth estimate of the mixing distribution. However, SNPMLE also uses the EM algorithm to estimate the mixing distribution. If we have a multivariate mixing distribution, estimation can be computationally expensive because the EM algorithm converges slowly (Magder and Zeger, 1996).

The prediction recursion (PR) method is a relatively new method developed by Newton and Zheng (1999). The PR method also estimates the density of the random effects, using an iterative technique. The updated estimate of the distribution is a convex combination of the previous distribution estimate and the product of the previous distribution estimate and a conditional likelihood. The fixed effects parameters are estimated using the log profile likelihood with the distribution of the random effects replaced by its estimate. Monte-Carlo results in Tao, Palta, Yandell, and Newton (1999) indicated that the PR method performed better than or as well as the SNPMLE and NPMLE estimates. Yet the PR algorithm is also not computationally efficient.

Traditional density estimation can vary from the simple, like the histogram, to the complex, such as mixture methods (for example, using a mixture of Gaussian distributions) or kernel estimation with multiple bandwidths. However, the three density estimation methods for mixed models outlined above differ from traditional density estimation. The three methods for density estimation in the mixed model assume that the errors are normally distributed with zero mean and common variance. The goal then is to estimate the distribution of the random effects. Traditional density estimation makes no distributional assumption on the error terms, as the distribution of the error terms is of concern.

If the linear parametric mixed model is misspecified, a nonparametric method to

estimate the mixed model may be called for. The above methods to estimate the random effects density could be used when the user's specified model has incorrectly specified the mixing distribution. However, our interest lies in the misspecification of either one or both model matrices, not in the misspecification of the random effects density. The sections below outline the model and two locally weighted methods to obtain nonparametric mixed model fits. These nonparametric methods will offer an alternative to the parametric linear mixed model and a solution to model misspecification of the model matrices.

6.2 The Model

In Chapter 3, the estimate of the mean response $g(x_0)$ at x_0 using kernel and local polynomial regression could be expressed as a weighted sum of the responses

$$\hat{g}(x_0) = \sum_{j=1}^{n} u_{0j} y_j,$$

where $0 \leq u_{0j} \leq 1$, and u_{0j} is the weight assigned to the j^{th} observation for the estimation of $g(x_0)$ at x_0. The weights were decreasing in the distance of x_j to x_0. This concept of local weighting will be applied to mixed models by two approaches. One approach, the conditional local mixed model, weights the variance-covariance matrix of the observation vector conditional on the random effects. The second approach, the marginal local mixed model, weights the variance-covariance matrix of the observation vector. The details of each approach are presented below.

6.3 The Conditional Local Mixed Model (CLMM)

Consider the mixed model at $\tilde{x}_0$ for the entire data set[1]

$$\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \tilde{\mathbf{Z}}\mathbf{b}_0 + \boldsymbol{\epsilon}_0,$$

where $\mathbf{Y}$ is an (n x 1) stacked vector of cluster response vectors, $\tilde{\mathbf{X}}$ is an (n x d_1) matrix of model matrices stacked by cluster, and $\tilde{\mathbf{Z}}$ is an (n x d_2) block diagonal matrix with the

[1] This model can be written for a single cluster by including a subscript i on Y, $\tilde{X}$, $\tilde{Z}$, b_0, and ϵ_0 to denote the i^{th} cluster.

random effects model matrices for each cluster on the diagonals. Note that d_1 and d_2 do not have to equal p_1 and p_2, the dimensions for the matrices in Chapter 5. The regressors in the parametric mixed model may differ from the regressors in the nonparametric mixed model, and so a tilde denotes a nonparametric counterpart.

As in the parametric mixed model, the (d_1 x 1) vector $\boldsymbol{\beta}_0$ is a vector of unknown fixed parameters, $\mathbf{b}_0$ is a (d_2 x 1) vector of random effects, and $\boldsymbol{\epsilon}_0$ is a vector of error vectors stacked by cluster. The random effects $\mathbf{b}_0$ are assumed to be normally distributed with mean $\mathbf{0}$ and variance-covariance matrix $\tilde{\mathbf{B}}$, where $\tilde{\mathbf{B}}$ is a block diagonal matrix. The subscript "0" denotes that this model represents the fit at $\tilde{x}_0$. The local mixed model is a pointwise fit, that is, for each point $\tilde{x}_0$, there exists $\boldsymbol{\beta}_0$, $\mathbf{b}_0$, and $\boldsymbol{\epsilon}_0$. Fits can be calculated at each $\tilde{x}_0$ value and then connected to obtain curves. This differs from the parametric linear mixed model, where the curves were obtained from a single $\boldsymbol{\beta}$ and $\mathbf{b}$ for the entire data set.

Recall from Chapter 5 that the variance of $\mathbf{Y}|\mathbf{b}$ was $\mathbf{R}$, a block diagonal matrix of the within-cluster variance-covariance matrices $\mathbf{R}_i$. The conditional local mixed model (CLMM) incorporates weights in the definition of the variance-covariance matrix of $\boldsymbol{\epsilon}_0$, the variance-covariance matrix of $\mathbf{Y}|\mathbf{b}_0$. Thus, the conditional local mixed model is aptly named because it applies weights to the variance of $\mathbf{Y}$ conditioned on the random effects $\mathbf{b}_0$.

In the conditional local mixed model, the errors $\boldsymbol{\epsilon}_0$ are assumed to be normally distributed with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}$, where $\mathbf{K}_0$ is a diagonal weight matrix. The j^{th} element on the diagonal of $\mathbf{K}_0$ equals the Nadaraya-Watson kernel weight

$$k_{i,0,j} = \frac{K\left(\frac{|\frac{\tilde{x}_{i,0}-\tilde{x}_{i,j}}{\tilde{x}_{max}-\tilde{x}_{min}}|}{h}\right)}{\sum_{j=1}^{n_i} K\left(\frac{|\frac{\tilde{x}_{i,0}-\tilde{x}_{i,j}}{\tilde{x}_{max}-\tilde{x}_{min}}|}{h}\right)}, \tag{6.1}$$

where i=1,...,s and j=1,...,n_i. The weights are assigned without regard to cluster, so the weights sum to 1 across the data set. The weights should not sum to 1 across a cluster; it may be the case that at a given $\tilde{x}_0$ value, an extreme cluster may not have observations at regressor locations close to $\tilde{x}_0$. Forcing the weights for estimation at $\tilde{x}_0$ to sum to one for

that cluster would mean that at least one response from that cluster would receive nonzero weight. But if all of the responses for the extreme cluster occur at values of the regressor far from the point $\tilde{\mathrm{x}}_0$, we do not want any of the responses to obtain substantial weight; doing so would result in a biased estimate of both the population average estimate and the cluster specific predictions (since the cluster specific predictions depend upon the population average) at the point $\tilde{\mathrm{x}}_0$.

At $\tilde{\mathrm{x}}_0$ the distributional assumptions can be expressed as

$$\begin{bmatrix} \mathbf{b}_0 \\ \boldsymbol{\epsilon}_0 \end{bmatrix} \sim \mathrm{MN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \tilde{\mathbf{B}} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_0^{-\frac{1}{2}} \tilde{\mathbf{R}} \mathbf{K}_0^{-\frac{1}{2}} \end{bmatrix} \right).$$

As in the parametric linear mixed model, the variance-covariance matrices $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{R}}$ are often unknown. They may be estimated either by maximum likelihood or restricted maximum likelihood, as discussed in Chapter 5.

The estimator $\hat{\boldsymbol{\beta}}_0$ and the predictor $\hat{\mathbf{b}}_0$ at the point $\tilde{\mathrm{x}}_0$ can be found by including the weight matrix in Henderson's joint likelihood expression. The conditional local mixed model equations may be expressed as

$$\begin{bmatrix} \tilde{\mathbf{X}}' \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{R}}^{-1} \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{X}} & \tilde{\mathbf{X}}' \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{R}}^{-1} \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{Z}} \\ \tilde{\mathbf{Z}}' \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{R}}^{-1} \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{X}} & (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}' \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{R}}^{-1} \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{Z}}) \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\mathbf{b}}_0 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}' \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{R}}^{-1} \mathbf{K}_0^{\frac{1}{2}} \mathbf{Y} \\ \tilde{\mathbf{Z}}' \mathbf{K}_0^{\frac{1}{2}} \tilde{\mathbf{R}}^{-1} \mathbf{K}_0^{\frac{1}{2}} \mathbf{Y} \end{bmatrix}. \tag{6.2}$$

The equations in (6.2) are similar to the traditional mixed model equations in (5.9), with the exception of incorporating the weight matrix $\mathbf{K}_0$. The solution to the equations yields the estimator $\hat{\boldsymbol{\beta}}_0^{\mathsf{C}}$

$$\hat{\boldsymbol{\beta}}_0^{\mathsf{C}} = (\tilde{\mathbf{X}}' \mathbf{V}_0^{*-1} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}' \mathbf{V}_0^{*-1} \mathbf{Y} \tag{6.3}$$

where $\mathbf{V}_0^* = \mathbf{K}_0^{-\frac{1}{2}} \tilde{\mathbf{R}} \mathbf{K}_0^{-\frac{1}{2}} + \tilde{\mathbf{Z}} \tilde{\mathbf{B}} \tilde{\mathbf{Z}}'$ and predictor

$$\hat{\mathbf{b}}_0 = \tilde{\mathbf{B}} \tilde{\mathbf{Z}}' \mathbf{V}_0^{*-1} (\mathbf{Y} - \tilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_0). \tag{6.4}$$

Notice that the expression for $\hat{\boldsymbol{\beta}}_0$ is the generalized least squares estimate from Chapter 2, where the variance-covariance matrix now incorporates both the between and within-cluster variation. The derivations of the conditional local mixed model equations, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\mathbf{b}}_0$ from the joint likelihood can be found in Appendix C.

It can be shown that, under the assumption that the linear mixed model is the correct model, that for fixed $\mathbf{K}_j$ and estimation at $\tilde{x}_j$,

$$\begin{aligned} E(Y_{ij}|\mathbf{b}_j) &= \tilde{\mathbf{x}}'_{ij}\boldsymbol{\beta}_j + \tilde{\mathbf{z}}'_{ij}\mathbf{b}_j \\ E(Y_{ij}) &= \tilde{\mathbf{x}}'_{ij}\boldsymbol{\beta}_j. \end{aligned}$$

The expectations given above are for the response from the i^{th} cluster at $\tilde{x}_j$. Of course, interest only lies in the population average and cluster specific fits at $\tilde{x}_j$, since the local models are a pointwise fit. At each value $\tilde{x}_j$, variance-covariance matrices are obtained based upon all of the data. The variance-covariance matrices can be expressed as

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

Local estimation is performed because the parametric mixed model may be incorrect. Thus, the following models can be considered:

$$\begin{aligned} &\text{Nature's Model (True Model):} Y_{ij} = g(x_{ij}, z_{ij}) + \epsilon_{ij} \\ &\text{The user's parametric model:} Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b} + \epsilon_{ij} \\ &\text{The conditional local user's model:} Y_{ij} = \tilde{\mathbf{x}}'_{ij}\boldsymbol{\beta}_j + \mathbf{z}'_{ij}\mathbf{b}_j + \epsilon_{ij}. \end{aligned}$$

Recall that $\mathbf{X}$ and $\tilde{\mathbf{X}}$ may be different model matrices. When the true model equals the parametric and conditional local user's models, then the user's models are correctly specified. Otherwise, the user's models are misspecified and biased. Specifically, the expectation and variance of $\hat{\boldsymbol{\beta}}_0^C$ at a point $\tilde{x}_0$ are

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_0^C) &= (\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}E(\mathbf{Y}) = (\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{g}(\mathbf{X},\mathbf{Z}) \\ Var(\hat{\boldsymbol{\beta}}_0^C) &= (\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}Var(\mathbf{Y})\mathbf{V}_0^{*-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}. \end{aligned}$$

Thus the estimator $\hat{\boldsymbol{\beta}}_0^C$ is an unbiased estimator for $\boldsymbol{\beta}_0^C$ if the local model is correct, or $g(x_{ij}) = \mathbf{x}_j^{*\prime}\boldsymbol{\beta}_j^C$.

As in the parametric linear mixed model, the conditional local mixed model produces cluster specific and population average curves. The population average fit at an arbitrary value of the regressor, $\tilde{x}_0$, for the conditional local mixed model (CLMM) can be expressed as a weighted sum of the observations

$$\hat{\mu}_{PA,0} = \tilde{\mathbf{x}}_0'\hat{\boldsymbol{\beta}}_0^{C} = \tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{Y} = \mathbf{h}_{PA,0}^{C\prime}\mathbf{Y} = \sum_k h_{PA,0,k}^{C} y_k,$$

where $\tilde{\mathbf{x}}_0' = (1\ \tilde{x}_0\ ...\ \tilde{x}_0^d)$, $\mathbf{h}_{PA,0}^{C\prime} = \tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}$, and $h_{PA,0,k}^{C}$ is the k^{th} element of the vector $\mathbf{h}_{PA,0}^{C\prime}$. The population average curve may be graphically displayed by calculating the population average fits for many $\tilde{x}_0$ values covering the range of $\tilde{x}$ and then connecting the fits to obtain a regression curve. As in the parametric linear mixed model, the population average curve is the same for every cluster if each cluster uses the same values of the regressor.

The cluster specific fits for the i^{th} cluster at $\tilde{x}_0$ for the conditional local mixed model (CLMM) are given by

$$\begin{aligned}
\hat{\mu}_{CS,i,0} &= \tilde{\mathbf{x}}_0'\hat{\boldsymbol{\beta}}_0 + \tilde{\mathbf{z}}_{i,0}'\hat{\mathbf{b}}_0 = \tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{Y} + \tilde{\mathbf{z}}_{i,0}'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\
&= \tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{Y} + \mathbf{z}_{i,0}'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{Y}) \\
&= (\tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1} + \tilde{\mathbf{z}}_{i,0}'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1} - \tilde{\mathbf{z}}_{i,0}'\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1})\mathbf{Y} \\
&= \mathbf{h}_{CS,i,0}^{C\prime}\mathbf{Y} = \sum_k h_{CS,i,0,k}^{C} y_k
\end{aligned}$$

where $\tilde{\mathbf{z}}_{i,0}'$ is a row vector that depends on the random terms in the model at $\tilde{x}_0$ for cluster i and $h_{CS,i,0,k}^{C}$ is the k^{th} element of the vector $\mathbf{h}_{CS,i,0}^{C\prime}$. The cluster specific curves are graphically represented by calculating the cluster specific fits for each cluster for many $\tilde{x}_0$ values covering the range of $\tilde{x}$ and then connecting the fits. As in the parametric linear mixed model, the cluster specific curves differ for every cluster unless the predictor $\hat{\mathbf{b}}_0$ is equal to the zero vector. Thus for s clusters, there are s + 1 curves to be estimated: s cluster specific curves and one population average curve. The cluster specific curves can also be expressed as a weighted sum of the observations, where the weight for the k^{th} observation from the i^{th} cluster for the fit at $\tilde{x}_0$ is given by $h_{CS,i,0,k}^{C}$. The population average and cluster specific fits at the j^{th} data point are obtained by replacing $\tilde{\mathbf{x}}_0'$ with $\tilde{\mathbf{x}}_j'$ in the equations above. An example of the conditional local mixed model will be given in section 6.7.

The conditional local mixed model is tremendously flexible. Where the parametric linear mixed model imposes a parametric form (for example, a parametric model with X^2 as highest order term will be a parabola), the nonparametric mixed model does not make such limitations. Since the local regression curves are connected pointwise fits, both the population average and the cluster specific curves do not have to conform to a certain shape. This flexibility is one of the major advantages of the conditional local mixed model. The disadvantage, however, is that the fit may follow the data too closely and may model nonexistent trends in the data set. The smoothness of the fit will be controlled by the bandwidth, the topic of section 6.6.

6.4 The Marginal Local Mixed Model

In the conditional local mixed model, the weight matrix $\mathbf{K}_0$ was used in the variance-covariance matrix of the conditional distribution. The marginal local mixed model incorporates the weights in the variance-covariance matrix of the marginal distribution. From Chapter 5, the marginal distribution of $\mathbf{Y}$ had variance-covariance matrix $\mathbf{V}$. The marginal local mixed model uses the weight matrix $\mathbf{K}_0$ and the variance-covariance matrix $\tilde{\mathbf{V}}$ such that the variance-covariance of $\mathbf{Y}$ in the marginal local mixed model is $\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{V}}\mathbf{K}_0^{-\frac{1}{2}}$ for estimation at $\tilde{x}_0$.

The marginal local mixed model can be expressed as [2]

$$\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 + \boldsymbol{\epsilon}_0,$$

where $\mathbf{Y}$, $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Z}}$, $\boldsymbol{\beta}_0$, $\mathbf{K}_0$, and $\boldsymbol{\epsilon}_0$ are defined as in the conditional local mixed model. The fits are obtained for each $\tilde{x}_0$ and then connected to obtain the regression curve. This model will be referred to as the marginal local mixed model (MLMM) due to the weighting of the marginal variance-covariance matrix of $\mathbf{Y}$. The random effects $\mathbf{b}_0$ are assumed normally distributed with mean $\mathbf{0}$ and variance-covariance matrix $\tilde{\mathbf{B}}$. The random errors are also

[2] This model can be written for a single cluster by including a subscript i on Y, $\tilde{X}$, $\tilde{Z}$, b_0, K_0, and ϵ_0 to denote the i^{th} cluster.

assumed to be from a normal distribution with mean $\mathbf{0}$ but with variance-covariance matrix $\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}$. In summary, the distributional assumptions at $\tilde{\mathrm{x}}_0$ are

$$\begin{bmatrix}\mathbf{b}_0\\ \boldsymbol{\epsilon}_0\end{bmatrix} \sim \mathrm{MN}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\tilde{\mathbf{B}} & \mathbf{0}\\ \mathbf{0} & \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\end{bmatrix}\right),$$

where the variance-covariance matrices $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{R}}$ may be estimated by maximum likelihood or restricted maximum likelihood if the matrices are unknown.

Notice that the marginal local mixed model contains kernel weights in the variance-covariance matrix of $\boldsymbol{\epsilon}_0$ and as a multiplier of the $\tilde{\mathbf{Z}}$ matrix so that

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

The estimator $\hat{\boldsymbol{\beta}}_0$ and the predictor $\hat{\mathbf{b}}_0$ can be found by incorporating the weight matrix into Henderson's joint likelihood of $\mathbf{b}_0$ and $\boldsymbol{\epsilon}_0$. The MLMM equations can be expressed as

$$\begin{bmatrix}\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\\ \tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}})\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_0\\ \hat{\mathbf{b}}_0\end{bmatrix} = \begin{bmatrix}\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}\\ \tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}\end{bmatrix},$$

or simplified to

$$\begin{bmatrix}\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}\\ \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}})\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_0\\ \hat{\mathbf{b}}_0\end{bmatrix} = \begin{bmatrix}\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}\\ \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}\end{bmatrix}. \tag{6.5}$$

The equations in (6.5) are analogous to the standard mixed model equations. The marginal mixed model equations differ from the conditional local mixed model equations in that the weight matrix, in addition to its use in the definition of the variance-covariance matrix of the errors, is a multiplier of the model matrix for the random effects.

Solving the equations in (6.5) for $\hat{\boldsymbol{\beta}}_0$ and $\hat{\mathbf{b}}_0$ yields the estimator

$$\hat{\boldsymbol{\beta}}_0^{\mathsf{M}} = (\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\mathbf{Y} \tag{6.6}$$

and the predictor

$$\hat{\mathbf{b}}_0^{\mathsf{M}} = \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0^{\mathsf{M}}). \tag{6.7}$$

Similar to the conditional local mixed model, $\hat{\boldsymbol{\beta}}_0^{\mathsf{M}}$ is the generalized least squares estimate of $\boldsymbol{\beta}_0$, albeit with a difference variance-covariance matrix. It is interesting to note that the marginal local mixed model incorporates the weight matrix in the between-cluster variation (in $\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}$)and within-cluster variation (in $\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}$) expressions in $\mathbf{V}_0^{**}$. The derivations of the marginal local mixed model equations, $\hat{\boldsymbol{\beta}}_0^{\mathsf{M}}$, and $\hat{\mathbf{b}}_0^{\mathsf{M}}$ can be found in Appendix D.

Some properties of the marginal local mixed, assuming that the mixed linear model is the correct model are (for fixed $\mathbf{K}_0$ and estimation at $\tilde{\mathrm{x}}_\mathrm{j}$),

$$\begin{aligned}
\mathrm{E}(\mathrm{Y}_{\mathrm{ij}}|\mathbf{b}_\mathrm{j}) &= \tilde{\mathbf{x}}_{\mathrm{ij}}'\boldsymbol{\beta}_\mathrm{j} + \mathrm{k}_{\mathrm{ij}}^{-\frac{1}{2}}\tilde{\mathbf{z}}_{\mathrm{ij}}'\mathrm{b}_\mathrm{j} \\
\mathrm{E}(\mathrm{Y}_{\mathrm{ij}}) &= \tilde{\mathbf{x}}_{\mathrm{ij}}'\boldsymbol{\beta}_\mathrm{j}.
\end{aligned}$$

The expectations for responses Y_{ij} from the i^{th} cluster at $\tilde{\mathrm{x}}_\mathrm{j}$. As this is a pointwise fit, only the fits at $\tilde{\mathrm{x}}_\mathrm{j}$ are of interest. As in the conditional local mixed model, variance-covariance matrices at each $\tilde{\mathrm{x}}_\mathrm{j}$ are obtained using the entire dataset and are

$$\begin{aligned}
\mathrm{Var}(\mathbf{Y}|\mathbf{b}_\mathrm{j}) &= \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}} = \mathbf{R}^* \\
\mathrm{Var}(\mathbf{Y}) &= \mathbf{V}_0^{**}.
\end{aligned}$$

Local estimation is performed because the parametric mixed model may be incorrect. The three models are:

$$\begin{aligned}
&\text{Nature's Model (True Model):}\mathrm{Y}_{\mathrm{ij}} = \mathrm{g}(\mathrm{x}_{\mathrm{ij}}, \mathrm{z}_{\mathrm{ij}}) + \epsilon_{\mathrm{ij}} \\
&\text{The user's parametric model:}\mathrm{Y}_{\mathrm{ij}} = \mathbf{x}_{\mathrm{ij}}'\boldsymbol{\beta} + \mathbf{z}_{\mathrm{ij}}'\mathbf{b} + \epsilon_{\mathrm{ij}} \\
&\text{The marginal local user's model:}\mathrm{Y}_{\mathrm{ij}} = \tilde{\mathbf{x}}_{\mathrm{ij}}'\boldsymbol{\beta}_\mathrm{j} + \epsilon_{\mathrm{ij}}
\end{aligned}$$

The models are correctly specified when the parametric and marginal local mixed models equal the true model. The expectation and variance of $\hat{\boldsymbol{\beta}}_0^{\mathsf{M}}$ at $\tilde{\mathrm{x}}_0$ for the marginal local mixed

model can be expressed as

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_0^{\,\mathsf{M}}) = (\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathrm{E}(\mathbf{Y}) = (\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{g}(\mathbf{X},\mathbf{Z})$$
$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_0^{\,\mathsf{M}}) = (\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathrm{Var}(\mathbf{Y})\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}.$$

The estimate of $\boldsymbol{\beta}_0^{\mathsf{M}}$ is unbiased if $\mathrm{g}(\mathrm{x_{ij}}, \mathrm{z_{ij}}) = \tilde{\mathbf{x}}_{\mathsf{ij}}'\boldsymbol{\beta}_{\mathsf{j}}$.

Unlike the parametric linear and conditional local mixed models, the marginal local mixed model will only estimate a population average curve. Nonparametric estimation of the cluster specific curves will be performed exclusively by conditional local mixed model estimation. (See section 6.5). The expression for the population average fit at $\tilde{\mathrm{x}}_0$ for the marginal local mixed model (MLMM) can be expressed as

$$\hat{\mu}_{\mathsf{PA},0} = \tilde{\mathbf{x}}_0'\hat{\boldsymbol{\beta}}_0^{\mathsf{M}} = \tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{Y} = \mathbf{h}_{\mathsf{PA},0}^{\mathsf{M}}{}'\mathbf{Y} = \textstyle\sum_{\mathsf{k}} \mathrm{h}_{\mathsf{PA},0,\mathsf{k}}^{\mathsf{M}}\mathrm{y_k},$$

where $\tilde{\mathbf{x}}_0' = (1\ \mathrm{x}_0\ \ldots\ \mathrm{x}_0^{\mathsf{d}})$, $\mathbf{h}_{\mathsf{PA},0}^{\mathsf{M}}{}' = \tilde{\mathbf{x}}_0'(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}$, and $\mathrm{h}_{\mathsf{PA},0,\mathsf{k}}^{\mathsf{M}}$ is the k^{th} element of the vector $\mathbf{h}_{\mathsf{PA},0}^{\mathsf{M}'}$. The population average curve may be represented graphically for the marginal local mixed model by calculating the population average fits over many values of the regressor and then connecting the fits. The population average fit at the j^{th} data point are obtained by replacing $\tilde{\mathbf{x}}_0'$ with $\tilde{\mathbf{x}}_{\mathsf{j}}'$. The marginal local mixed model technique will be applied to the wind speed data in section 6.7.

One of the advantages of the conditional local mixed model over the parametric linear mixed model is its flexibility. This is true for the marginal local mixed model as well. Because no parametric form has been specified, the marginal local mixed model does not have to fit a specified parametric form. The disadvantage, as in the conditional local mixed model, is that the fit may follow irregularities in the data. In addition, the marginal local mixed model can only be used for population average estimation. Section 6.5 will explain why the marginal local mixed model is not used for cluster specific inference.

6.5 Appropriate Usage of the Local Mixed Models

Two local mixed models, the conditional and marginal local mixed models, were discussed in sections 6.3 and 6.4, respectively. For a given scenario, the question that ultimately arises

is which local model to use. Consider the local model

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

for estimation at $\tilde{x}_0$ at the i^{th} cluster. The conditional and unconditional expectations for the local model can be expressed as

$$E(Y_{i0}|\mathbf{b}_0) = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0 + \tilde{\mathbf{z}}'_{i0}\mathbf{b}_0 \tag{6.8}$$

$$E(Y_{i0}) = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0. \tag{6.9}$$

As stated previously, we will use the conditional local mixed model for obtain both cluster specific and population average fits. Recall from section 6.3 that the conditional local mixed model at $\tilde{x}_0$ for fixed $\mathbf{K}_0$ had expectations

$$E(Y_{i0}|\mathbf{b}_0) = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0 + \tilde{\mathbf{z}}'_{i0}\mathbf{b}_0$$
$$E(Y_{i0}) = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0$$

and variances for the entire data set as

$$\mathrm{Var}(\mathbf{Y}|\mathbf{b}_0) = \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}$$
$$\mathrm{Var}(\mathbf{Y}) = \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}} + \mathbf{ZBZ}' = \mathbf{V}_0^*.$$

For the conditional local mixed model, the variance-covariance matrix of $(\mathbf{Y}|\mathbf{b}_0)$ is localized correctly; that is, the CLMM model by definition weights the between-cluster variances for cluster specific prediction (and is then appropriate for cluster specific prediction). The conditional mean for CLMM, $E(Y_{i0}|\mathbf{b}_0) = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0 + \tilde{\mathbf{z}}'_{i0}\mathbf{b}_0$, equals (6.8), and hence the conditional CLMM mean has been correctly specified. Notice then that the marginal mean is also correct, because $E(E(Y_{i0}|\mathbf{b}_0)) = \tilde{\mathbf{x}}_{i0}\boldsymbol{\beta}_0$, and this equals (6.9). Thus, the CLMM model is appropriate for estimation and prediction of the unconditional and conditional mean.

The marginal local mixed model at $\tilde{x}_0$ from section 6.4 had expectations for the data set

$$E(Y_{i0}|\mathbf{b}_0) = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0 + k_{i0}^{-\frac{1}{2}}\tilde{\mathbf{z}}'_{i0}\mathbf{b}_0$$
$$E(Y_{i0}) = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0$$

and variances for the entire data set as

$$\text{Var}(\mathbf{Y}|\mathbf{b}_0) = \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}$$
$$\text{Var}(\mathbf{Y}) = \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{V}}\mathbf{K}_0^{-\frac{1}{2}} = \mathbf{V}_0^{**}.$$

for the fixed weights and weight matrix k_{i0} and $\mathbf{K}_0$.

For the marginal local mixed model, the variance-covariance matrix of $\mathbf{Y}$ is correctly localized by the MLMM model and the unconditional mean $E(Y_{i0}) = \tilde{x}'_{i0}\boldsymbol{\beta}_0$ is correct. The conditional mean for MLMM does not equal (6.9), however. The random effects term $\mathbf{K}_0^{-\frac{1}{2}}\mathbf{Z}$ is not appropriate to obtain cluster specific predictions in the marginal local mixed model. The $\mathbf{K}_0^{-\frac{1}{2}}\mathbf{Z}$ terms in the MLMM model are not related to $\mathbf{Y}$– the values in $\mathbf{Z}$ are.

The conditional local mixed model offers estimation of cluster specific and population average curves. So what is the advantage of using the marginal local mixed model over the conditional local mixed model? One advantage is that both the conditional and unconditional variances are localized correctly. This is not true of the conditional local mixed model, where only the conditional variance is localized.

We conjecture that the marginal local mixed model will have, on average, a smaller mean square error than the conditional local mixed model when estimating the marginal mean. This concept was investigated through a simulation study. The results of this study appear in Chapter 8.

6.6 Bandwidth Selection for the Local Mixed Model

The kernel weights used in the conditional and marginal local mixed models depend upon a bandwidth, the "smoothing parameter". A data driven method to select a bandwidth is needed. Often, the bandwidth is chosen to minimize an estimated criterion, such as average squared error or integrated squared error. A brief introduction to bandwidth selectors was given in Chapter 3. We consider three bandwidth selectors for the local mixed models– PRESS, PRESS*, and PRESS**.

6.6.1 PRESS

The PRESS statistic is an estimate of the average squared error. For the fixed effects model, the selected bandwidth h is the value which minimizes

$$\text{PRESS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i})^2, \tag{6.10}$$

where $\hat{Y}_{i,-i}$ is the estimate of the regression function at $\tilde{x}_i$ with the i^{th} data point removed. Extending the traditional PRESS statistic to the mixed model case is not straightforward. Many issues arise, including whether to use population average or cluster specific fits, whether to use a delete-point or delete-cluster scheme, how to obtain the deleted fits, and whether to include a variance-covariance matrix in the PRESS expression.

In the marginal local mixed model, we can only calculate the population average. Thus, the population average fits must be used in the calculation of PRESS for the marginal local mixed model. The conditional local mixed model, however, yields both population average and cluster specific fits. One option is to calculate separate PRESS statistics, and find two, possibly different, bandwidths for the conditional local model. One PRESS statistic would involve the population average fits and would be minimized to find the bandwidth for the population average curve. The other PRESS statistic would use the cluster specific fits and would be minimized to find the bandwidth for the cluster specific curves.

However, our interests in the conditional local model lie primarily in the cluster specific fits. The population average curve is an artifact of the weighting scheme. One could use PRESS with the cluster specific fits to find the bandwidth, as above. The population average curve is then found from the cluster specific analysis. In other words, the population average curve would use the same bandwidth as the bandwidth used in the cluster specific analysis. In our proposed research, one PRESS statistic for the conditional local mixed model, using the cluster specific fits, is utilized to find the bandwidth.

Another decision concerns the deletion scheme. In the traditional PRESS statistic (for the fixed model, uncorrelated data case), $\hat{Y}_{i,-i}$ is the estimate of the mean at x_i by removing that data point from the estimation. In the cluster correlated data case, however,

it is possible to remove an entire cluster, not a single data point. There are some advantages to removing clusters rather than points. In the analysis of cluster correlated data, it is really the cluster that is of interest, not a single point. The clusters are commonly assumed to be independent (although this does not have to be the case) and observations from the same cluster are marginally correlated, even if the random effects and random errors are uncorrelated. Removing points that are correlated with the remaining observations is difficult as no generalization of the Sherman-Morrison-Woodbury Theorem is available for that case.

The question then arises about how to obtain the delete cluster fits. Of course, one can calculate the delete cluster fits by deleting a cluster, run the analysis, and repeat the process s times if there are s clusters. However, this is extremely time consuming. For the fixed effects model with uncorrelated errors, formulas to obtain delete point parameter estimates are readily available (Myers, 1990) using the Sherman-Morrison-Woodbury Theorem. Similar formulas for cluster deletion in the mixed model must be developed.

Hurtado-Rodriguez (1993) developed generic formulas for deletion schemes for the mixed model. The formulas can be used to delete a single point or multiple points. The formulas from the multiple point deletion scheme can be used for cluster deletion. Hilden-Minton (1995) also performed research with deleted schemes, developing diagnostics for mixed models. The delete cluster parametric model for the entire data set can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\phi} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \tag{6.11}$$

where $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{Z}$, and $\boldsymbol{\epsilon}$ are as defined previously for the parametric mixed model and $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and $\mathbf{b}$ are the parameter vectors. The vector $\mathbf{Y}$ is sorted by cluster. The matrix $\mathbf{U}$ is an (n x n_i) model matrix where n_i is the number of observations for cluster i. The matrix $\mathbf{U}$ contains the (n_i x n_i) identity matrix for the values in $\mathbf{Y}$ that correspond to the i^{th} cluster and zeros everywhere else. For example, if we want to obtain the deleted fits for the wind speed data set for the first cluster, our model matrix $\mathbf{U}$ would be the (636 x 12) matrix

$$\begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 & 0 \\ : & : & : & \ldots & : & : & : \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ : & : & : & \ldots & : & : & : \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 \end{bmatrix}$$

for $\mathbf{Y}$ sorted by cluster. The matrix $\mathbf{U}$ serves as a selection matrix; it highlights the observations that will be deleted. The vectors $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ are vectors of fixed effects and $\mathbf{b}$ is a vector of random effects. Notice that if $\mathbf{Y}$ is perturbed as $\mathbf{Y}^* = \mathbf{Y} + \mathbf{U}\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is our perturbation, our reparameterized model becomes

$$\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}(\boldsymbol{\phi} + \boldsymbol{\omega}) + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon},$$

as long as $\begin{bmatrix} \mathbf{X} & | & \mathbf{U} \end{bmatrix}$ is of full rank. The estimate of $\boldsymbol{\beta}$ and the predictor of $\mathbf{b}$ do not change with the reparameterization. This is equivalent to saying that the estimator and predictor are independent of those observations selected by $\mathbf{U}$ (Hilden-Minton, 1995). The estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\phi}}$, and $\hat{\mathbf{b}}$ obtained from estimating and predicting quantities in (6.11) are the delete-cluster estimates and predictors, dented by $\hat{\boldsymbol{\beta}}_{\text{-i}}$, $\hat{\boldsymbol{\phi}}_{\text{-i}}$, and $\hat{\mathbf{b}}_{\text{-i}}$ for clarity.

Derivations of $\hat{\boldsymbol{\beta}}_{\text{-i}}$, $\hat{\boldsymbol{\phi}}_{\text{-i}}$, and $\hat{\mathbf{b}}_{\text{-i}}$ for the parametric linear mixed model using Henderson's joint likelihood can be found in Appendix E for fixed $\mathbf{R}$ and $\mathbf{B}$. Specifically, the delete

cluster estimates and predictors for the parametric linear mixed model are

$$\begin{aligned}
\hat{\boldsymbol{\phi}}_{-\mathrm{i}} &= (\mathbf{U}'\mathbf{P}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P}\mathbf{Y} \\
\hat{\boldsymbol{\beta}}_{-\mathrm{i}} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}_{-\mathrm{i}}) \\
&= \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}}_{-\mathrm{i}} \\
\hat{\mathbf{b}}_{-\mathrm{i}} &= \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{-\mathrm{i}} - \mathbf{U}\hat{\boldsymbol{\phi}}_{-\mathrm{i}}) \\
&= \hat{\mathbf{b}} + \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}}_{-\mathrm{i}} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}}_{-\mathrm{i}} \\
&= \hat{\mathbf{b}} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{U}\hat{\boldsymbol{\phi}}_{-\mathrm{i}},
\end{aligned}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. The delete cluster estimate $\hat{\boldsymbol{\beta}}_{-\mathrm{i}}$ and predictor $\hat{\mathbf{b}}_{-\mathrm{i}}$ can be expressed as the estimate or predictor from the full data set minus a correction for deleting the i^{th} cluster.

Similarly, the delete cluster estimate and predictor for the conditional and marginal local mixed models can also be found via the joint likelihood method from Henderson (1950). The delete cluster estimates and predictors for the conditional local mixed model for estimation at x_0 can be expressed as

$$\begin{aligned}
\hat{\boldsymbol{\phi}}_{0,-\mathrm{i}} &= (\mathbf{U}'\mathbf{P}^{*}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P}^{*}\mathbf{Y} \\
\hat{\boldsymbol{\beta}}_{0,-\mathrm{i}} &= (\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}_{0,-\mathrm{i}}) \\
&= \hat{\boldsymbol{\beta}}_0 - (\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{U}\hat{\boldsymbol{\phi}}_{0,-\mathrm{i}} \\
\hat{\mathbf{b}}_{0,-\mathrm{i}} &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{0,-\mathrm{i}} - \mathbf{U}\hat{\boldsymbol{\phi}}_{0,-\mathrm{i}}) \\
&= \hat{\mathbf{b}}_0 + \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{U}\hat{\boldsymbol{\phi}}_{0,-\mathrm{i}} \\
&\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{U}\hat{\boldsymbol{\phi}}_{0,-\mathrm{i}} \\
&= \hat{\mathbf{b}}_0 - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}(\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1})\mathbf{U}\hat{\boldsymbol{\phi}}_{0,-\mathrm{i}}
\end{aligned}$$

and for the marginal local mixed model as

$$\begin{aligned}
\hat{\boldsymbol{\phi}}_{0,\text{-i}} &= (\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^{**}\mathbf{Y} \\
\hat{\boldsymbol{\beta}}_{0,\text{-i}} &= (\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}}) \\
&= \hat{\boldsymbol{\beta}}_0 - (\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}} \\
\hat{\mathbf{b}}_{0,\text{-i}} &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{0,\text{-i}} - \mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}}) \\
&= \hat{\mathbf{b}}_0 + \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}} \\
&\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}} \\
&= \hat{\mathbf{b}}_0 - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1})\mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}}.
\end{aligned}$$

The matrix $\mathbf{P}^*$ equals $\mathbf{V}_0^{*\text{-}1} - \mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}$ for the conditional local mixed model and $\mathbf{P}^{**} = \mathbf{V}_0^{**\text{-}1} - \mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}$ for the marginal local mixed model. The expressions for the delete cluster parametric and nonparametric methods are similar but contain different variance-covariance matrices and may contain different model matrices. The delete cluster parametric mixed model population average and cluster specific fits at x_0 are

$$\begin{aligned}
\hat{\mu}_{\text{PA,-i}} &= \mathbf{x}_0'\hat{\boldsymbol{\beta}}_{\text{-i}} \\
\hat{\mu}_{\text{CS,-i}} &= \mathbf{x}_0'\hat{\boldsymbol{\beta}}_{\text{-i}} + \mathbf{z}_{\text{i},0}'\hat{\mathbf{b}}_{\text{-i}}.
\end{aligned}$$

The delete cluster conditional local mixed model population average and cluster specific fits can then be expressed as

$$\begin{aligned}
\hat{\mu}_{\text{PA,0,-i}} &= \tilde{\mathbf{x}}_0'\hat{\boldsymbol{\beta}}^{\text{C}}_{0,\text{-i}} \\
\hat{\mu}_{\text{CS,0,-i}} &= \tilde{\mathbf{x}}_0'\hat{\boldsymbol{\beta}}^{\text{C}}_{0,\text{-i}} + \tilde{\mathbf{z}}_{\text{i},0}'\hat{\mathbf{b}}_{0,\text{-i}},
\end{aligned}$$

and the delete cluster marginal local mixed model population average fit can be expressed as

$$\hat{\mu}_{\text{PA,0,-i}} = \tilde{\mathbf{x}}_0'\hat{\boldsymbol{\beta}}^{\text{M}}_{0,\text{-i}}.$$

The delete cluster derivations for the conditional and marginal local mixed models appear in Appendices F and G, respectively for fixed $\tilde{\mathbf{V}}$.

While calculating $\hat{\mathbf{b}}_{0,-i}$ for the conditional local mixed model, we noticed that the values for the i^{th} cluster when the i^{th} cluster was deleted were equal to zero. This fact was noted in Hurtado-Rodriguez (1993). Thus, the delete-cluster cluster specific fits are equal to the delete-cluster population average fits. To try to distinguish between the types of mean estimation, an alternate delete cluster BLUP for CLMM, $\hat{\mathbf{b}}_{0,-i,alt} = \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{0,-i})$, was considered. Further research indicated that as the bandwidth approached zero, the delete cluster fit using $\hat{\mathbf{b}}_{0,-i,alt}$ approached $\mathbf{Y}$ for estimation of the i^{th} cluster with the i^{th} cluster deleted. In other words, the PRESS statistic was minimized at the smallest bandwidth attempted. Our conjecture is that the simultaneous estimation of $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{R}}$ for a bandwidth that is approximately zero results in a cluster specific fit that is approximately the observation itself. Consider the simple model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where α_i and ϵ_{ij} have expectation zero and variances σ_α^2 and σ^2, respectively, for i=1,...,a and j=1,...,n. Now using expected mean squares, $\hat{\sigma}^2 = \text{MS(Error)}$ and $\sigma_\alpha^2 = \frac{\text{MS(A) - MS(Error)}}{n}$. Now, let's consider n=1. This is similar to a bandwidth of zero. So,

$$\begin{aligned} \text{SS(A)} &= \sum_{i=1}^{a}\sum_{j=1}^{1}(\bar{y}_{i.} - \bar{y}_{..})^2 \\ \text{SS(Error)} &= \sum_{i=1}^{a}\sum_{j=1}^{1}(y_{ij} - \bar{y}_{i.})^2. \end{aligned}$$

Notice that for n=1,

$$\begin{aligned} \text{SS(A)} &= \sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 \\ \text{SS(Error)} &= 0 \end{aligned}$$

as $y_{ij} = \bar{y}_{i.}$ for n=1. So as the bandwidth goes to zero, the estimate of the within-cluster variation should approach zero. The estimated conditional variance covariance matrix for

estimation at $\tilde{x}_0$ is $\mathbf{K}_0^{-\frac{1}{2}}\hat{\tilde{\mathbf{R}}}\mathbf{K}_0^{-\frac{1}{2}}$. As h$\to$0, if $\lim_{h\to 0}\frac{\hat{\sigma}^2}{k_{ii,0}} \to 0$, then $\hat{\mathbf{V}}_0^* = \mathbf{K}_0^{-\frac{1}{2}}\hat{\tilde{\mathbf{R}}}\mathbf{K}_0^{-\frac{1}{2}} + \tilde{\mathbf{Z}}\hat{\tilde{\mathbf{B}}}\tilde{\mathbf{Z}} \approx \tilde{\mathbf{Z}}\hat{\tilde{\mathbf{B}}}\tilde{\mathbf{Z}}$. So our cluster specific conditional local fit at $\tilde{x}_0$ for the entire data set would be

$$\begin{aligned}\hat{\mathbf{Y}} &= \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \tilde{\mathbf{Z}}\hat{\mathbf{b}}_0 \\ &= \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \tilde{\mathbf{Z}}\hat{\tilde{\mathbf{B}}}\tilde{\mathbf{Z}}'\hat{\mathbf{V}}_0^{*\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\ &= \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \tilde{\mathbf{Z}}\hat{\tilde{\mathbf{B}}}\tilde{\mathbf{Z}}'(\tilde{\mathbf{Z}}\hat{\tilde{\mathbf{B}}}\tilde{\mathbf{Z}}')^{*\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\ &= \mathbf{Y}.\end{aligned}$$

Using $\hat{\mathbf{b}}_{0,\text{-i}}$ has intuitive appeal. If one removes a cluster from the analysis, one can still estimate a population average based upon the remaining clusters. Trying to estimate a BLUP, a cluster specific quantity, for a cluster that has been removed is impossible. Deleting a cluster is equivalent to having no information about that cluster, so the best predictor of $\hat{\mathbf{b}}_{0,\text{-i}}$ is its mean.

Lastly, the remaining issue was whether to use a variance-covariance matrix in the expression of PRESS. Recall that PRESS is an estimate of the mean square error

$$\text{MSE}=(\hat{\mathbf{Y}} - \mathbf{g})'(\hat{\mathbf{Y}} - \mathbf{g}), \tag{6.12}$$

where $\mathbf{g}$ is the true regression function. Of course, we could consider an expression for the mean square error that included $\mathbf{V}$,

$$\text{MSE}= (\hat{\mathbf{Y}} - \mathbf{g})'\mathbf{V}^{\text{-}1}(\hat{\mathbf{Y}} - \mathbf{g}).$$

This of course would make the mean square error depend upon the scale chosen by the researcher. We have decided that we want to see how close our fitted curve is the to true curve, without regard to $\mathbf{V}$. This said, no variance-covariance matrix will be used in the PRESS statistic, as our PRESS statistic will be an estimate of (6.12).

6.6.2 PRESS*

In the traditional regression setting for uncorrelated data, Mays, Birch, and Starnes (2001) found that the PRESS statistic tended to choose values of the bandwidth that were too

small for model robust regression. An alternate bandwidth selector, PRESS* was developed to protect against small bandwidths. The PRESS* statistic can be extended to the linear mixed model case and can be expressed as

$$\text{PRESS*}=\frac{\text{PRESS}}{\text{n-trace}(\mathbf{H}^{\text{CLMM}})} \tag{6.13}$$

for the conditional local mixed and as

$$\text{PRESS*}=\frac{\text{PRESS}}{\text{n-trace}(\mathbf{H}^{\text{MLMM}})} \tag{6.14}$$

for the marginal local mixed model. The matrices $\mathbf{H}^{\text{CLMM}}$ and $\mathbf{H}^{\text{MLMM}}$ are the local smoother matrices defined as

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

for the conditional local mixed model and

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

for the marginal local mixed model. The vector $\tilde{\mathbf{x}}_i'$ is the i^{th} row of $\tilde{\mathbf{X}}$ and the matrices $\mathbf{V}_i^{*\text{-}1}$ and $\mathbf{V}_i^{**\text{-}1}$ are the CLMM and MLMM variance-covariance matrices for estimation at $\tilde{x}_i$. The trace of the ordinary least squares hat matrix in the uncorrelated fixed effects model is p, the number of parameters in the model. Likewise, the trace of the smoother matrix can be thought of as the number of parameters needed for a parametric model to obtain population average fits comparable to the local population average fits for a given value of the bandwidth. Thus, a small value of the bandwidth results in a large trace and a large value of the bandwidth has a small trace of the resulting smoother matrix. So the denominator of the PRESS* statistic should be small for a small bandwidth, thus having a large PRESS* statistic. The bandwidth chosen is the value that minimizes PRESS*, so it is unlikely that extremely small bandwidths would be selected.

6.6.3 PRESS**

Mays, Birch, and Starnes (2001) found that the PRESS* statistic on average chose values of the bandwidth that were too large for model robust regression in the uncorrelated fixed effects model. The selector PRESS** was created to alleviate the extreme bandwidth problem. For the linear mixed model, the PRESS** statistic is

$$\text{PRESS}^{**}=\frac{\text{PRESS}}{\text{n-trace}(\mathbf{H}^{\mathsf{CLMM}})+(\text{n-d})(\frac{\mathsf{SSE}_{\mathsf{max}}-\mathsf{SSE}_{\mathsf{h}}}{\mathsf{SSE}_{\mathsf{max}}})} \tag{6.15}$$

for the conditional local mixed and

$$\text{PRESS}^{**}=\frac{\text{PRESS}}{\text{n-trace}(\mathbf{H}^{\mathsf{MLMM}})+(\text{n-d})(\frac{\mathsf{SSE}_{\mathsf{max}}-\mathsf{SSE}_{\mathsf{h}}}{\mathsf{SSE}_{\mathsf{max}}-\mathsf{SSE}_{\bar{\mathsf{y}}}})} \tag{6.16}$$

for the marginal local mixed model. The value d is the number of parameters in the local mixed model and $\text{SSE}_{\bar{\text{y}}}$ is the sum across the regressor locations of the sum of the squared deviations of the responses around the mean response at a regressor location. The sum of squares $\text{SSE}_{\bar{\text{y}}}$ represents the sum of squares as the bandwidth $\text{h}\rightarrow 0$ for the population average model. This fit connects the average response across the values of the regressor. For the cluster specific model, the sum of squares as the bandwidth $\text{h}\rightarrow 0$ would be zero (a perfect fit to the clusters). This fit connects the average response across the values of the regressor. The term $\text{SSE}_{\mathsf{max}}$ is the sum of squared deviations of the response and the local fit that assigned a constant weight to each response. This would result as the bandwidth $\text{h}\rightarrow \infty$, and would represent the worst possible fit using the nonparametric method. The expression SSE_{h} is the sum of squared deviations of the response and the local fit using a bandwidth h. The sum of squares $\text{SSE}_{\mathsf{max}}$ and SSE_{h} are cluster specific for the conditional local mixed model and are population averages for the marginal local mixed model. The sum of squares expressions, as in the PRESS statistic, are not weighted by a variance-covariance matrix.

The PRESS** statistics for the two local models have different sum of squares ratios. The PRESS** statistic for the conditional local mixed model uses cluster specific fits. The sum of squares associated with cluster specific fits ideally have an upper and lower bound.

The lower bound would be zero, in the case where the fit connects the data points. The upper bound should be SSE_{max}, in the case where each observation receives equal weight.

The PRESS** statistic for the marginal local mixed model uses population average fits. The sum of squares associated with population average fits would be bounded by below by $SSE_{\bar{y}}$, where the fit at $\tilde{x}_0$ would be the average response at $\tilde{x}_0$. The sum of squares should be bounded above by SSE_{max}.

The ratio of these sum of squares should be between zero and one. The ratio should be zero for large bandwidths (when SSE_{max} equals SSE_h)and one for small bandwidths (when SSE_h equals zero for the conditional local model and $SSE_{\bar{y}}$ for the marginal local model). This makes the additional penalty term in the PRESS** denominator, the ratio multiplied by n-d, a value between 0 and n-d. Since our goal is to minimize PRESS**, the additional penalty term avoids bandwidths that are too large. Thus, PRESS** protects bandwidths that are too small (by the penalty term introduced in PRESS*) and those that are too large (by the additional penalty term). Results from Mays, Birch, and Starnes (2001) indicate that PRESS** works well for model robust regression for the uncorrelated fixed effects model.

6.7 An Example of the Conditional Local and Marginal Local Mixed Models

The wind speed data set will be revisited for local mixed model analysis. Recall that in Chapter 5, the parametric mixed model was a quadratic polynomial with a random intercept. The population average curve, the same for every station, was a parabola. The cluster specific curves, which differed for every station, were shifted parabolas. The amount of the shift was determined by the random intercept.

The local mixed model used for this data set was the local linear mixed model with a random intercept. A local quadratic model, although considered, was eliminated. Fan and Gijbels (1995) indicate that local regression of odd order for the fixed effects model tended to have less bias for mean function estimation. As in Chapter 5, 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 conditional local mixed model at $\tilde{x}_0$ for the i^{th} cluster can be expressed as

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

and for the marginal model at $\tilde{x}_0$ as

$$Y_{i0} = \tilde{\mathbf{x}}'_{i0}\boldsymbol{\beta}_0 + \epsilon_{i0}.$$

The vector $\tilde{\mathbf{x}}'_0 = \begin{bmatrix} 1 & \tilde{x}_0 \end{bmatrix}$ reflects the use of the local linear model. The matrix $\boldsymbol{\beta}_0$

$$\boldsymbol{\beta}_0 = \begin{bmatrix} \beta_{00} \\ \beta_{10} \end{bmatrix}$$

contains the fixed effects parameters at $\tilde{x}_0$. The (1 x s) vector $\tilde{\mathbf{z}}_{i0}$ has a one in the i^{th} column for the i^{th} cluster and a zero everywhere else, and the vector of random effects can be expressed as

$$\mathbf{b}_0 = \begin{bmatrix} b_{10} \\ b_{20} \\ : \\ b_{s0} \end{bmatrix}.$$

The value ϵ_{i0} is the random error associated with the response at $\tilde{x}_0$ from the i^{th} cluster.

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. In Chapter 5, the AR(1) covariance structure was used. 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.

All three bandwidth selectors (PRESS, PRESS*, and PRESS**) were used for the conditional and marginal local mixed models. The values of the PRESS, PRESS*, and PRESS** statistics for the bandwidths used in the search appear in Table 6.1. The bold values in the table indicate the minimum value in the column. For both the conditional and

Table 6.1: PRESS, PRESS*, and PRESS** values by bandwidth (h)

h	PRESS CLMM	PRESS MLMM	PRESS* CLMM	PRESS* MLMM	PRESS** CLMM	PRESS** MLMM
0.05	**5423.42**	**5420.21**	**8.54**	**8.53**	**4.69**	**6.95**
0.06	5437.99	5434.58	8.56	8.55	4.74	6.99
0.07	5448.65	5444.83	8.57	5.86	4.78	7.02
0.08	5456.99	5452.30	8.59	8.58	4.82	7.04
0.09	5464.18	5458.02	8.60	8.59	4.85	7.06

marginal local mixed models, PRESS, PRESS*, and PRESS** all 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.

The variance components must be estimated. As in the parametric linear mixed model, restricted maximum likelihood estimation was used to fit the weighted mixed models. Recall that $\tilde{\mathbf{V}} = \tilde{\mathbf{R}} + \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'$, where $\tilde{\mathbf{R}} = \sigma^2\mathbf{I}$ and $\tilde{\mathbf{B}} = \sigma_{\mathsf{b}}^2\mathbf{I}$. It would be interesting to look at the estimates of σ^2 and σ_{b}^2 for the two different local models. So, consider the case at $\tilde{\mathrm{x}}$=1 using a bandwidth of 0.05. For the conditional local mixed model, $\hat{\sigma}^2 = 0.000903$ and $\hat{\sigma}_{\mathsf{b}}^2$=9.8135. For the marginal local mixed model, $\hat{\sigma}^2 = 0.01402$ and $\hat{\sigma}_{\mathsf{b}}^2$=0.001071. Notice that $\hat{\sigma}^2$ at $\tilde{\mathrm{x}}$=1 for the conditional local mixed model is extremely small. In this example and in the simulation study of Chapter 8, it appears that the residual variance in the conditional local mixed model shrinks to zero for small bandwidths. For the marginal local mixed model, on the other hand, $\hat{\sigma}^2$ at $\tilde{\mathrm{x}}$=1 is larger than $\hat{\sigma}_{\mathsf{b}}^2$. An in-depth look at the variance components as a funtion of bandwidth will be provided in Chapter 8.

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 $\tilde{\mathrm{x}}$. Thus, for each $\tilde{\mathrm{x}}_0$ 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 $\tilde{x}$=1 using a bandwidth of 0.05 can be expressed as

$$\begin{aligned}\hat{\mathbf{Y}} &= \tilde{\mathbf{x}}_1'\hat{\boldsymbol{\beta}}_1 \\ &= \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}10.7398 \\ 0.3793\end{bmatrix}\end{aligned}$$

and the cluster specific fits at $\tilde{x}$=1 are

$$\hat{\mathbf{Y}} = \tilde{\mathbf{x}}_1'\hat{\boldsymbol{\beta}}_1\mathbf{1} + \hat{\mathbf{b}}_1$$

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

In this model, $\mathbf{1}$ is an (s x 1) = (12 x 1) column of ones. 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 $\tilde{x}$=1 is

$$\begin{aligned}\hat{\mathbf{Y}} &= \tilde{\mathbf{x}}_1'\hat{\boldsymbol{\beta}}_1 \\ &= \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}10.7408 \\ 0.3860\end{bmatrix}.\end{aligned}$$

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

Trellis plots by cluster appear in Figures 6.1 and 6.2. Figure 6.1 contains the population average and cluster specific curves for the conditional local mixed model using a

Figure 6.1: Conditional Local Mixed Model with h=0.05 (Plot of Population Average and Cluster Specific Curves by Station)

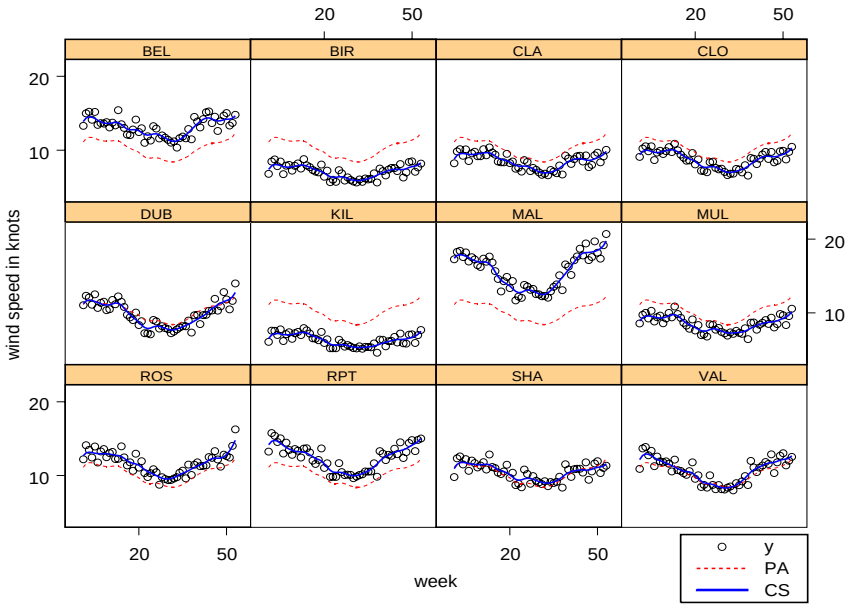


Figure 6.2: Marginal Local Mixed Model with h=0.05 (Plot of Population Average Curve by Station)

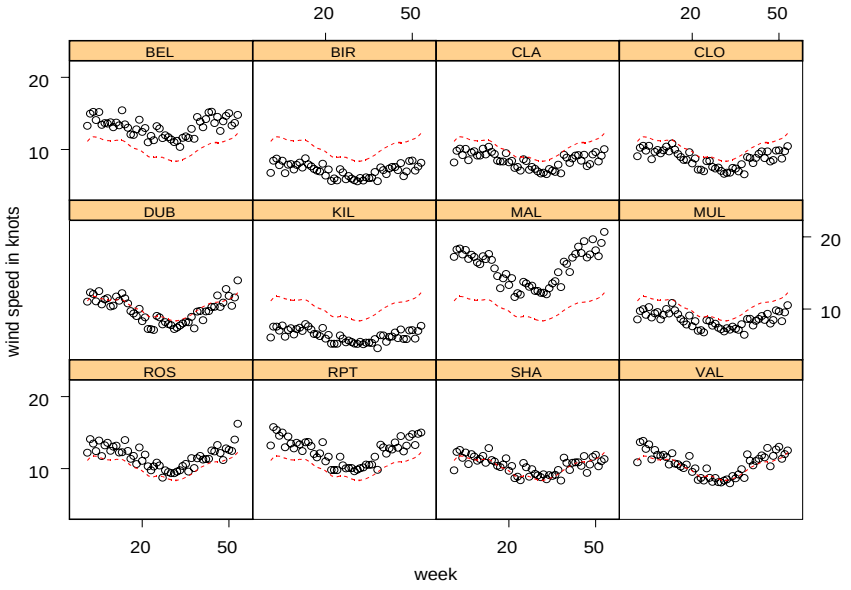
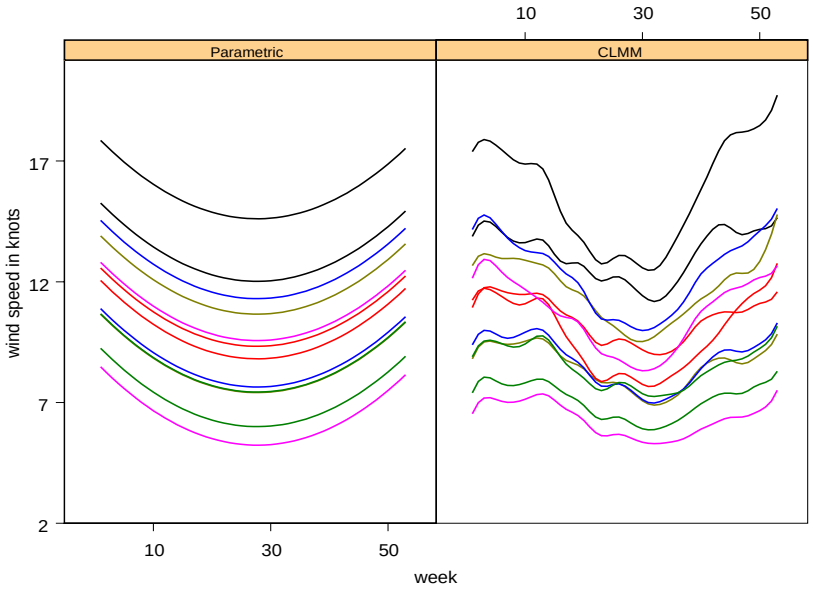


Figure 6.3: Plot of CLMM and Parametric Cluster Specific Fits (h=0.05)



bandwidth of 0.05. The population average is the dotted line and the cluster specific fit is the solid line. Figure 6.2 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 6.3 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.

All three bandwidth selectors resulted in identical bandwidths. In addition, the conditional and marginal local population average curves look similar. The conditional and marginal local mixed models will be analyzed in more detail in the simulation study of Chapter 8.

6.8 Summary

This chapter presented two nonparametric mixed models. The conditional local mixed model, obtaining pointwise fits by weighting the conditional variance-covariance matrix, offers the user population average and cluster specific curves. The marginal local mixed model, which placed weights around the marginal variance-covariance matrix, produces a pointwise population average curve. Three bandwidth selectors for the mixed models were presented and the local estimation methods were applied to a real data set.

The local models provided fits that were far superior to the parametric fits in terms of bias reduction. However, the fits tended to be quite variable. Chapter 7 develops mixed model robust regression based upon the work by Einsporn and Birch (1993) and Mays, Birch, and Starnes (2001). Mixed model robust regression will combine the parametric and nonparametric fits in the hope of obtaining an estimate of the regression function with less bias and variance than the parametric and nonparametric estimates, respectively.

Chapter 7

Semiparametric Estimation for the Mixed Model

Chapters 6 and 7 provided parametric and nonparametric regression methods for the mixed model. The parametric linear mixed model will result in a smooth curve, but one with considerable bias if the model is misspecified. The nonparametric model may have less bias than the parametric model, but the fits may be too variable for small bandwidths. A method is to be developed that combines the smoothness of the parametric regression curve with the low bias of the local method. The method described is the focus of this chapter. An example of this method on a data set will also be provided.

7.1 Mixed Model Robust Regression

Recall from Chapter 4 that Einsporn and Birch (1993) developed a semiparametric regression method for the fixed effects model called Model Robust Regression 1 (MRR1). The rationale behind MRR1 is that the user has some knowledge about the underlying model from which the data had been generated, but this model fails over a portion of the data. The user can specify a nonparametric model, but using the local model exclusively results in loss of information about the model. The nonparametric model also has a tendency to produce fits with large variances. Thus a combination of the two fits may be advantageous; the nonparametric model would "correct" for irregularities from the parametric model while retaining some information about the true model contained in the parametric model. Recall

that the MRR1 fit was a convex combination of the two fits and was defined in Chapter 4 as

$$\hat{\mathbf{Y}}^{\mathsf{MRR1}} = (1-\lambda)\hat{\mathbf{Y}}^{\mathsf{P}} + \lambda\hat{\mathbf{Y}}^{\mathsf{NP}},$$

where $\hat{\mathbf{Y}}^{\mathsf{P}}$ is the parametric fit and $\hat{\mathbf{Y}}^{\mathsf{NP}}$ is the nonparametric fit. The mixing parameter λ is a value between 0 and 1.

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 mixed model. Specifically, the MMRR fit is

$$\hat{\mathbf{Y}}^{\mathsf{MMRR}} = (1-\lambda)\hat{\mathbf{Y}}^{\mathsf{P}} + \lambda\hat{\mathbf{Y}}^{\mathsf{NP}}, \tag{7.1}$$

where $\hat{\mathbf{Y}}^{\mathsf{P}}$ is the fit from the parametric linear mixed model and $\hat{\mathbf{Y}}^{\mathsf{NP}}$ is a local mixed model fit. The mixing parameter is an element between 0 and 1. A value of λ=1 produces an MMRR fit equal to the nonparametric fit; λ=0 results in a 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 fit parametric fit in combination with the cluster specific conditional local mixed model fit.

The mixed model robust fits should fall between the nonparametric and parametric fits since it is a convex combination of them and $0 \leq \lambda \leq 1$. Interestingly, for moderate values of the mixing parameter λ, the MMRR estimate of the regression function retains the same shape as the nonparametric regression curve, but tends to be smoother. This smoothness results from the parametric component.

It is expected that the MMRR fit will have bias that is less than or equal to the bias of the parametric fit. The MMRR fit should be less variable than the nonparametric fits unless λ=1. It is desired that the MMRR estimates have a small mean square error

for misspecified models. Such results would be consistent with those found in Einsporn and Birch (1993) and Mays, Birch, and Starnes (2001) for the fixed effects model. A simulation study in Chapter 8 investigates the approximate integrated mean square error in fitting parametric, nonparametric, and mixed model robust regression models for varying degrees of model misspecification.

7.2 Theoretical Bias, Variance, and MSE Formulas

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

$$\begin{aligned} E\sum(\hat{y}_0 - g(x_0))^2 &= E\sum(\hat{y}_0 + E(\hat{y}_0) - E(\hat{y}_0) - g(x_0))^2 \\ &= Var(\hat{y}_0) + [Bias(\hat{y}_0)]^2. \end{aligned}$$

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. Note that the theoretical mean square error formulas cannot be applied in the real world; regression analysis would be inappropriate if the user knew the true mean function.

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. For population average estimation evaluated at the design points, the bias formula can be written as

$$Bias(\hat{\mathbf{Y}}_{PA}) = -(\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f}) \tag{7.2}$$

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, $\mathbf{f} = E(\mathbf{Y}) - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\epsilon}$ is the misspecified portion, and the smoother matrix $\mathbf{H}$ equals

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

for the parametric mixed model,

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

for the conditional local mixed model, and

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

for the marginal local mixed model. The row vector $\tilde{\mathbf{x}}_k'$ is the k^{th} row of $\tilde{\mathbf{X}}$. Notice that the bias formula for the parametric linear mixed model simplifies to

$$\begin{aligned} -(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{P}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f}) &= -\mathbf{X}\boldsymbol{\beta} + \mathbf{X}(\mathbf{X}'\mathbf{V}^{\text{-}1}\mathbf{X})^{\text{-}1}\mathbf{X}'\mathbf{V}^{\text{-}1}\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{P}})\mathbf{f} \\ &= -(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{P}})\mathbf{f}. \end{aligned}$$

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

$$\mathrm{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{P}}) = \mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}\,\prime}, \tag{7.3}$$

$$\mathrm{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{C}}) = \mathbf{H}_{\mathsf{PA}}^{\mathsf{C}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{C}\,\prime}, \tag{7.4}$$

$$\mathrm{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{M}}) = \mathbf{H}_{\mathsf{PA}}^{\mathsf{M}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{M}\,\prime}. \tag{7.5}$$

The parametric variance-covariance matrix for population average estimation can be simplified:

$$\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}\,\prime} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{\text{-}1}\mathbf{X})^{\text{-}1}\mathbf{X}'\mathbf{V}^{\text{-}1}\mathbf{V}\mathbf{V}^{\text{-}1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{\text{-}1}\mathbf{X})^{\text{-}1}\mathbf{X}' = \mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}\mathbf{V}.$$

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 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 found by conditioning on the random effects for a fixed true mean function. The cluster specific bias formula for estimation at the design points is

$$\text{Bias}(\hat{\mathbf{Y}}_{\mathsf{CS}}|\mathbf{b}) = -(\mathbf{I}-\mathbf{H})(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b}+\mathbf{f}) \tag{7.6}$$

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

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

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

respectively. The row vectors $\mathbf{i}_k'$ and $\tilde{\mathbf{z}}_k'$ are the k^{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

$$\text{Var}(\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{CS}}|\mathbf{b}) = \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{R}\mathbf{H}^{\mathsf{P}}{}_{\mathsf{CS}}{}' \tag{7.7}$$

$$\text{Var}(\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{CS}}|\mathbf{b}) = \mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}\mathbf{R}\mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}{}'. \tag{7.8}$$

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 derivations of the population average and cluster specific bias and variance formulas for the parametric, CLMM, and MLMM models and fixed true mean function appear in Appendices H – L.

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, 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

$$\text{Bias}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}) = -\lambda(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{C}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}})\mathbf{f} \tag{7.9}$$

$$\text{Bias}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,M}}) = -\lambda(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{M}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}})\mathbf{f} \tag{7.10}$$

$$\text{Bias}(\hat{\mathbf{Y}}_{\mathsf{CS}}^{\mathsf{MMRR}}|\mathbf{b}) = -\lambda(\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{C}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{Z}\mathbf{b} - (\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{f}, \tag{7.11}$$

where

$$\begin{aligned}
\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}} &= (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}} + \lambda\mathbf{H}_{\mathsf{PA}}^{\mathsf{C}}, \\
\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}} &= (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}} + \lambda\mathbf{H}_{\mathsf{PA}}^{\mathsf{M}}, \text{ and} \\
\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}} &= (1-\lambda)\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}} + \lambda\mathbf{H}_{\mathsf{CS}}^{\mathsf{C}}
\end{aligned}$$

are the mixed model robust regression smoother matrices for population average (using CLMM and MLMM) estimation and cluster specific prediction. Notice that all three model robust procedures have bias expressions that are similar; the cluster specific model has an additional term for random effects, and of course the smoother matrices in each expression differ. 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

$$\text{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}) = \lambda\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{C}\,\prime} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}\,\prime} \tag{7.12}$$

$$\text{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,M}}) = \lambda\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{M}\,\prime} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}\,\prime} \tag{7.13}$$

$$\text{Var}(\hat{\mathbf{Y}}_{\mathsf{CS}}^{\mathsf{MMRR}}) = \lambda\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{C}\,\prime} + (1-\lambda)\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}\,\prime}. \tag{7.14}$$

Again, all of the variance expressions for the model robust methods follow a similar format, the difference being the smoother matrices and values for λ in the population average setting. Notice that the variance formulas given above can be expressed in terms of covariances. For example,

$$\text{Cov}[\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}, \hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{C}}] = \text{Cov}[\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{Y}, \mathbf{H}_{\mathsf{PA}}^{\mathsf{C}}\mathbf{Y}] = \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{C}\,\prime},$$

and

$$\text{Cov}[\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}, \hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{P}}] = \text{Cov}[\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{Y}, \mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}\mathbf{Y}] = \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}\ \prime}.$$

Thus,

$$\text{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}) = \lambda\text{Cov}[\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}, \hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{C}}] + (1-\lambda)\text{Cov}[\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}, \hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{P}}].$$

Similarly, expressions for $\text{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,M}})$ and $\text{Var}(\hat{\mathbf{Y}}_{\mathsf{CS}}^{\mathsf{MMRR}})$ in terms of covariances may be found. The MMRR fits have λ on the nonparametric fit and $(1-\lambda)$ on the parametric fit; likewise, the MMRR variances have λ on the covariances between the MMRR and nonparametric fits, and $(1-\lambda)$ on the covariances between the MMRR and parametric fits. The mean square errors for the model robust fits can be found in the same fashion as the parametric and nonparametric mixed models. Derivations of the bias and variance formulas for the mixed model robust regression models for a fixed true mean function can be found in Appendices M, N, and O.

Mean square prediction formulas can also be developed for the unconditional true mean function. Again, we assume that the variance-covariance matrices $\mathbf{V}$, $\mathbf{B}$, and $\mathbf{R}$ are known. The mean square prediction error is again the variance plus the squared bias; however, the true mean function is also random and contains additional variation. The cluster specific mean square prediction error with a random true mean function is

$$\text{MSE} = \mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}\ \prime} + (\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})\mathbf{Z}\mathbf{B}\mathbf{Z}'(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})' + (\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})'(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})'. \tag{7.15}$$

The smoother matrix $\mathbf{H}$ equals $\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}, \mathbf{H}_{\mathsf{CS}}^{\mathsf{C}}$, and $\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}$ for the cluster specific parametric, conditional local, and mixed model robust mean square prediction errors. Notice that the variance term in (7.15) is $\mathbf{H}\mathbf{R}\mathbf{H}' + (\mathbf{I}-\mathbf{H})\mathbf{Z}\mathbf{B}\mathbf{Z}'(\mathbf{I}-\mathbf{H})'$. The variance term for the cluster specific mean square formulas when the true mean function was fixed was $\mathbf{H}\mathbf{R}\mathbf{H}'$. Thus, there is additional variation $((\mathbf{I}-\mathbf{H})\mathbf{Z}\mathbf{B}\mathbf{Z}'(\mathbf{I}-\mathbf{H})')$ due to the conditional expectation of the difference between the fit and the true mean function. Therefore, the variance has increased.

The squared bias, on the other hand, has decreased. The squared bias term in (7.15) is $(\mathbf{I}-\mathbf{H})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})'(\mathbf{I}-\mathbf{H})'$, whereas the squared bias term in the cluster specific

mean square formulas conditioned on the random effects is $(\mathbf{I}-\mathbf{H})(\mathbf{X}\boldsymbol{\beta}+\mathbf{Zb}+\mathbf{f})(\mathbf{X}\boldsymbol{\beta}+\mathbf{Zb}+\mathbf{f})'(\mathbf{I}-\mathbf{H})'$. Thus, the unconditional cluster specific mean square prediction will be larger than the conditional cluster specific mean square prediction if the increase in variance is more than the decrease in bias. The unconditional cluster specific mean square prediction will be smaller than the conditional cluster specific mean square prediction if the increase in variance is less than the decrease in bias. Derivations of the mse formula for the cluster specific unconditional true mean function can be found in Appendix P.

7.3 Estimation of the Mixing Parameter

The mixing parameter, λ, is a indicator of the degree of parametric model misspecification. If λ is close to zero, the parametric model provides an adequate fit to the data and little of the local fit is needed. If λ is close to one, the parametric fit is poor and the nonparametric fit should be used. The mixing parameter is an unknown quantity and must be estimated from the data. Conditioning on the bandwidth h, notice that

$$\begin{aligned}\hat{\mathbf{Y}}^{\mathsf{MMRR}} &= (1-\lambda)\hat{\mathbf{Y}}^{\mathsf{P}}+\lambda\hat{\mathbf{Y}}^{\mathsf{NP}}\\ \hat{\mathbf{Y}}^{\mathsf{MMRR}} &= \hat{\mathbf{Y}}^{\mathsf{P}}+\lambda(\hat{\mathbf{Y}}^{\mathsf{NP}}-\hat{\mathbf{Y}}^{\mathsf{P}})\\ (\hat{\mathbf{Y}}^{\mathsf{MMRR}}-\hat{\mathbf{Y}}^{\mathsf{P}}) &= \lambda(\hat{\mathbf{Y}}^{\mathsf{NP}}-\hat{\mathbf{Y}}^{\mathsf{P}}).\end{aligned}$$

Thus we see that λ is the "slope" parameter in a no-intercept model with a response equal to $(\hat{\mathbf{Y}}^{\mathsf{MMRR}}-\hat{\mathbf{Y}}^{\mathsf{P}})$ and the "explanatory" variable $(\hat{\mathbf{Y}}^{\mathsf{NP}}-\hat{\mathbf{Y}}^{\mathsf{P}})$. The least square estimate of the slope, $\hat{\lambda}$, may be estimated as

$$\hat{\lambda}=\frac{(\hat{\mathbf{Y}}^{\mathsf{NP}}-\hat{\mathbf{Y}}^{\mathsf{P}})'(\hat{\mathbf{Y}}^{\mathsf{MMRR}}-\hat{\mathbf{Y}}^{\mathsf{P}})}{(\hat{\mathbf{Y}}^{\mathsf{NP}}-\hat{\mathbf{Y}}^{\mathsf{P}})'(\hat{\mathbf{Y}}^{\mathsf{NP}}-\hat{\mathbf{Y}}^{\mathsf{P}})}.$$

This is similar to the mixing parameter estimate

$$\hat{\lambda}=\frac{(\hat{\mathbf{Y}}^{\mathsf{NP}}_{\mathsf{i},\text{-}\mathsf{i}}-\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{i},\text{-}\mathsf{i}})'(\mathbf{Y}-\hat{\mathbf{Y}}^{\mathsf{P}})}{(\hat{\mathbf{Y}}^{\mathsf{NP}}-\hat{\mathbf{Y}}^{\mathsf{P}})'(\hat{\mathbf{Y}}^{\mathsf{NP}}-\hat{\mathbf{Y}}^{\mathsf{P}})} \tag{7.16}$$

given in Mays, Birch, and Starnes (2001). In equation (7.16), $\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{i},\text{-}\mathsf{i}}$ and $\hat{\mathbf{Y}}^{\mathsf{NP}}_{\mathsf{i},\text{-}\mathsf{i}}$ are the parametric and nonparametric estimates of the mean response at $\mathrm{x_i}$ computed without the point

(x_i,y_i). Burman and Chaudhuri (1992) suggest the substitution of $\hat{\mathbf{Y}}^{P}_{i,-i}$ and $\hat{\mathbf{Y}}^{NP}_{i,-i}$ in (7.16) 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. Expression (7.16) is an estimate of the theoretically optimal mixing parameter. This optimal mixing parameter is given in the following section.

The mixing parameter for mixed model robust regression is similar to the expression given above. For the fixed effects model of Mays, Birch, and Starnes (2001), the estimates of the mean response at x_i computed without the point (x_i,y_i) appear in $\hat{\lambda}$. For the cluster correlated mixed model, parametric and nonparametric fits for the i^{th} cluster with the i^{th} cluster deleted will be used in $\hat{\lambda}$. This is consistent with the work in Chapter 6, where the bandwidth selectors PRESS, PRESS*, and PRESS** also used fits with a cluster deleted. Additional questions arise, however, in the mixed model case. Specifically, whether the fits used in the estimate of λ be population average or cluster specific fits and whether an inverse variance-covariance matrix should be used as weight.

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 $\hat{\lambda}$ 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 the MMRR fit using the conditional local population average fits will be the same $\hat{\lambda}$ used in the MMRR fits using the conditional local cluster specific fits. That is, the estimate used in the research is

$$\hat{\lambda} = \frac{(\hat{\mathbf{Y}}^{NP}_{i,-i} - \hat{\mathbf{Y}}^{P}_{i,-i})'(\mathbf{Y} - \hat{\mathbf{Y}}^{P}_{CS})}{(\hat{\mathbf{Y}}^{NP}_{CS} - \hat{\mathbf{Y}}^{P}_{CS})'(\hat{\mathbf{Y}}^{NP}_{CS} - \hat{\mathbf{Y}}^{P}_{CS})},$$

rather than

$$\hat{\lambda} = \frac{(\hat{\mathbf{Y}}^{NP}_{i,-i} - \hat{\mathbf{Y}}^{P}_{i,-i})'(\mathbf{Y} - \hat{\mathbf{Y}}^{P}_{PA})}{(\hat{\mathbf{Y}}^{NP}_{PA} - \hat{\mathbf{Y}}^{P}_{PA})'(\hat{\mathbf{Y}}^{NP}_{PA} - \hat{\mathbf{Y}}^{P}_{CS})}.$$

Thus, the mean square error for the MMRR estimate using the CLMM population average fits may not be minimized. It may be the case that λ may work well for the cluster specific

fits, but poorly for the population average. An estimate of λ that works best for MMRR with CLMM population average fits may be one that uses population average fits throughout $\hat{\lambda}$, as above. Because we are interested primarily in cluster specific fits for CLMM, however, $\hat{\lambda}$ using 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.

The next section provides the asymptotic theory for the optimal estimator of λ. Convergence rates of the optimal estimator will also be derived.

7.4 Asymptotic Theory for the Mixing Parameter

We can write our population average model as

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

and the cluster specific model for the i^{th} clusters as

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

for i=1,...,s and j=1,$\cdots$,n_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 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 $s \to \infty$ for fixed values of the regressor. The asymptotic theory for the cluster specific model would concern $n \to \infty$, not $s \to \infty$. As the data are correlated, asymptotic theory for the cluster specific model is complicated, and is beyond the current scope of this research.

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 $\boldsymbol{\theta} = \left[\theta(x_{11}), \cdots, \theta(x_{sn_s})\right]'$ 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

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

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

$$\langle \mathbf{h}_1, \mathbf{h}_2 \rangle = n^{-1} \sum_{i=1}^{n} h_1(x_{ij}) h_2(x_{ij}),$$

where h_1 and h_2 are two functions of x_{ij} and z_{ij}. The norm is similarly defined as

$$\|\mathbf{h}_1\|^2 = \langle \mathbf{h}_1, \mathbf{h}_1 \rangle.$$

Define distances δ_s and γ_s as

$$\delta_s = \{\inf\|\theta - f(\beta)\| : \beta \in \Re^d\}$$

and

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

The first measure is the smallest distance between the parametric fit and the true model. If the infimum is unique, this value will be denoted as β^*, so that the distance measure becomes

$$\delta_s = \|\theta - f(\beta^*)\|.$$

The subscript s denotes the fact that this distance measure is dependent upon the number of clusters; note, however, that $s \to \infty$, 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. The subscript s on γ_s^2 indicates that this measure

also depends upon the the number of clusters. 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

$$\|(1-\lambda)\hat{f} + \lambda\hat{g} - \theta\|.$$

The value of λ that minimizes this distance is the theoretically optimal mixing parameter. Because this distance is a 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

$$\lambda^* = \frac{\langle \hat{f} - \hat{g}, \theta - \hat{f} \rangle}{\|\hat{f} - \hat{g}\|^2} = \frac{\langle \hat{f} - \hat{g}, \theta \rangle - \langle \hat{f} - \hat{g}, \hat{f} \rangle}{\|\hat{f} - \hat{g}\|^2}. \tag{7.17}$$

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

$$\hat{\lambda}^* = \frac{\langle \hat{f}_{-\mathsf{i}} - \hat{g}_{-\mathsf{i}}, \mathrm{Y} - \hat{f} \rangle}{\|\hat{f} - \hat{g}\|^2}, \tag{7.18}$$

where $\hat{f}_{-\mathsf{i}}$ and $\hat{g}_{-\mathsf{i}}$ are the parametric and nonparametric estimates obtained 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 $\hat{f}_{-\mathsf{i}}$ and $\hat{g}_{-\mathsf{i}}$ 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:

$$A1.\ \|f(\hat{\beta}, .) - f(\beta^*, .)\| = \mathrm{O}_{\mathsf{p}}(\pi)$$

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

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

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 $\|\hat{g} - \theta\| = O_p(\gamma_s)$. 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,

$$\|\hat{f} - \hat{g}\| = \begin{cases} O_p(1), & \text{if } \lim_{s\to\infty} \delta_s \neq 0 \\ O_p(\gamma_s), & \text{if } \delta_s = 0. \end{cases}$$

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

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

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

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

Lemma 1 gives the convergence rates of the distance between the parametric and the nonparametric estimate. Recall that δ_s is zero if the parametric estimate is correct, and the $\lim_{s\to\infty} \delta_s$ 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 when 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. Lemmas 1 and 2 and Theorem 1, along with their proofs, are given in Burman and Chaudhuri (1992) with λ as the multiplier of the parametric portion and $1-\lambda$ on the nonparametric fit. The proofs of these results, with λ as the multiplier of the nonparametric fit, are given in Appendix Q. Further discussion about the convergence rates of the optimal mixing parameter may be found in Burman and Chaudhuri (1992), Starnes (1999), and Mays, Birch, and Starnes (2001). Asymptotic notational definitions may be found in Bishop, Fienberg, and Holland (1975).

Asymptotic results are needed for the asymptotically optimal data driven mixing parameter of (7.18). 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 will be considered future work.

7.5 An Example of Mixed Model Robust Regression

As in the previous chapters, the wind speed data set will be used in the mixed model robust regression analysis. The parametric linear mixed model was a quadratic model with a random intercept. The local model was a local linear mixed model with a random intercept. Our mixed model robust regression estimate is

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

where $\hat{\mathbf{Y}}^{\mathsf{P}}$ is the parametric linear mixed model fit and $\hat{\mathbf{Y}}^{\mathsf{NP}}$ 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.

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. The local models all assume independence for the between and within-cluster variation. The data are still marginally correlated in the local models.

The bandwidth selectors PRESS, PRESS*, and PRESS** were used for the local models. For the wind speed data set, PRESS, PRESS*, and PRESS** chose a bandwidth of 0.05 for both the conditional and marginal local mixed models. 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. In addition, notice that both estimates of lambda are fairly large. This is consistent with the results of Chapters 5 and 6. 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.

Population average and cluster specific curves for MMRR appear in Figures 7.1 through 7.5. The data are not included in Figures 7.1–7.4 in order to focus the attention on the estimates of the regression curve; the reader is referred to Figures 5.1, 6.1, and 6.2 that do contain the data. Figures 7.1 and 7.2 are plots of the parametric, local, and mixed model robust regression population average curves using CLMM and MLMM with h=0.05, respectively. Figure 7.3 is a comparison of the two PA MMRR curves. Figure 7.4 is a plot of the parametric, conditional local, and mixed model robust regression cluster specific curves using h=0.05, and Figure 7.5 examines the cluster specific curves for station MAL;

Figure 7.1: MMRR using CLMM (Plot of Population Average Curve)

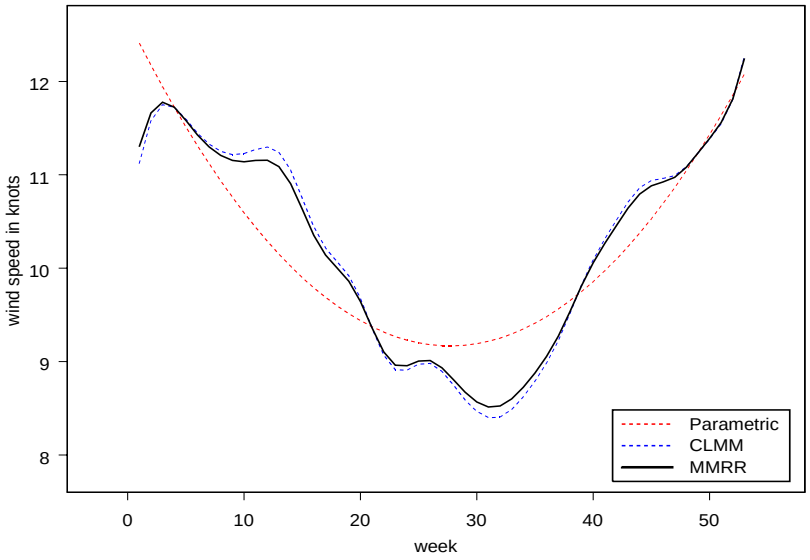


Figure 7.2: MMRR using MLMM (Plot of Population Average Curve)

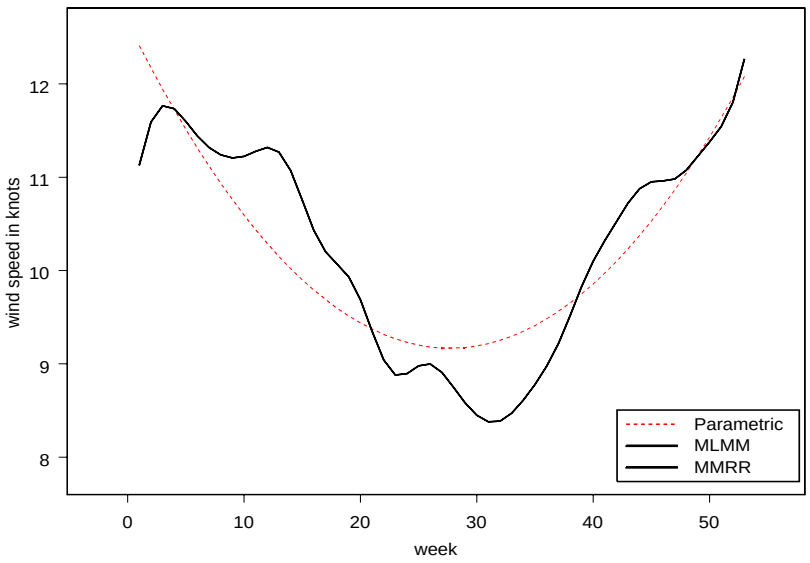


Figure 7.3: Comparison of Population Average MMRR using CLMM and MLMM

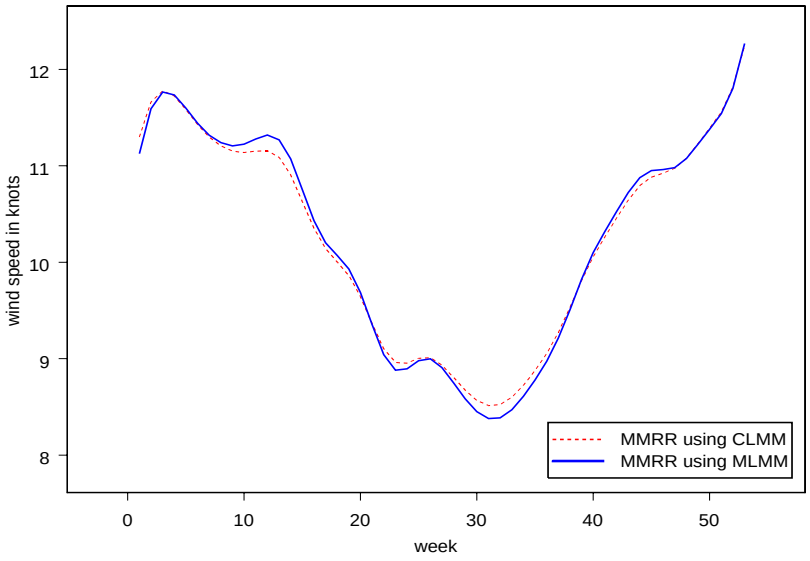


Figure 7.4: MMRR using CLMM (Plot of Cluster Specific Curves by Station)

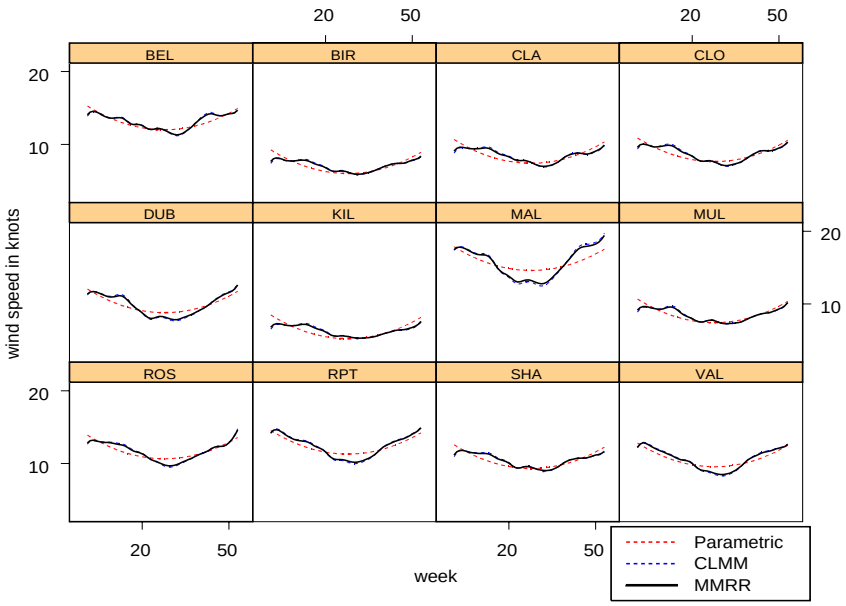
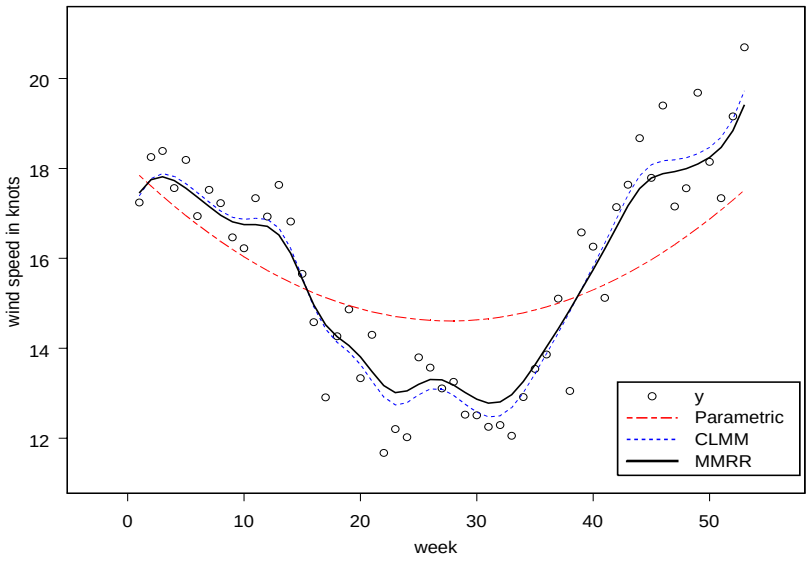


Figure 7.5: Cluster Specific Curves for cluster MAL



although this station has a trend similar to the other stations, it appears to behave oddly.

Obviously, the mixed model robust fits lie between the local and parametric fits. For example, in Figure 7.5, the CLMM fit is slightly smaller in midyear than the mixed model robust fit. Notice that the MMRR fits retain a shape similiar to that given by the local fit. This is true for moderate to large values of the mixing parameter, in the cases where the nonparametric and the parametric fits differ significantly. All three bandwidth selectors choose the same bandwidth in this example, and the population average marginal local and conditional local mixed models provided similar fits. Notice, however, that the PA MMRR curve using CLMM is smoother than the PA MMRR curve using MLMM. As will be seen in Chapter 8, the two mixed model robust fits will not be identical for every data set, and the bandwidth selector used in the analysis will depend upon the inference space (population average or cluster specific) that the user specifies.

7.6 Summary

In Chapter 7, the semiparametric version of the mixed model, mixed model robust regression, was presented. The MMRR method is the mixed model extension of the work by Mays, Birch, and Starnes (2001) for the fixed effects model. The MMRR estimate is a convex combination of the linear parametric mixed model and a local mixed model. An estimate of the mixing parameter was established, along with asymptotic theory on the theoretically optimal mixing parameter. The MMRR regression technique was applied to an adapted portion of the wind speed data set from Haslett and Raftery (1989), and resulted in MMRR fits that were virtually identical to the local fits. Chapter 8 is a comparison of the mean square error of the parametric, local, and MMRR fits for mixed models for varying amounts of model misspecification. It is expected that the MMRR fits will have a small MSE in comparison to the parametric and local methods for a variety of model misspecification amounts.

Chapter 8

Simulation Study Results

The previous chapters have presented the parametric, nonparametric, and model robust modeling techniques for the fixed effect and mixed model. As of yet, the techniques have only been applied to a single data set, namely the wind speed data set. Criteria, such as the mean square error, are needed to compare the parametric, nonparametric, and model robust methods. The Monte-Carlo simulation study will investigate a number of issues. A comparison of the three bandwidth selectors (PRESS, PRESS*, and PRESS**) is performed to compare the appropriateness of the selectors for the two local models. The study will be carried out with varying degrees of model misspecification, variance-covariance structures, and cluster sizes. Simulated mean square errors for each of the methods will be calculated from the simulated datasets. The population average and cluster specific models will be evaluated separately, and the "winner", for a given amount of misspecification, variance structure, bandwidth selector, and cluster size, is the method that minimizes the mean square error over all population and cluster specific methods, respectively. The average bandwidth and mixing parameters may be obtained for each combination and compared to an optimal value. Finally, the univariate distributions of the mixing parameter and bandwidth for differing amounts of model misspecification are desired; rough estimates may be obtained for the bandwidths and mixing parameters from the simulation.

8.1 Description of the Study

A Monte-Carlo simulation generates a specified number of data sets and calculates a quantity of interest from each. In the example studied here, the model is similar to the model used in Mays, Birch, and Starnes (2001), except that in this work the cluster correlated, random coefficient case is considered. The data are generated from the cluster specific model

$$Y_{ij} = (2 + b_{i1})(X_j - 5.5)^2 + (5 + b_{i2})X_j + \gamma \left[10\sin\left(\frac{\pi(X_j - 1)}{2.25}\right)\right] + \epsilon_{ij}, \tag{8.1}$$

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

The population average model can be expressed as

$$Y_{ij} = 2(X_j - 5.5)^2 + 5X_j + \gamma \left[10\sin\left(\frac{\pi(X_j - 1)}{2.25}\right)\right] + \epsilon_{ij}. \tag{8.2}$$

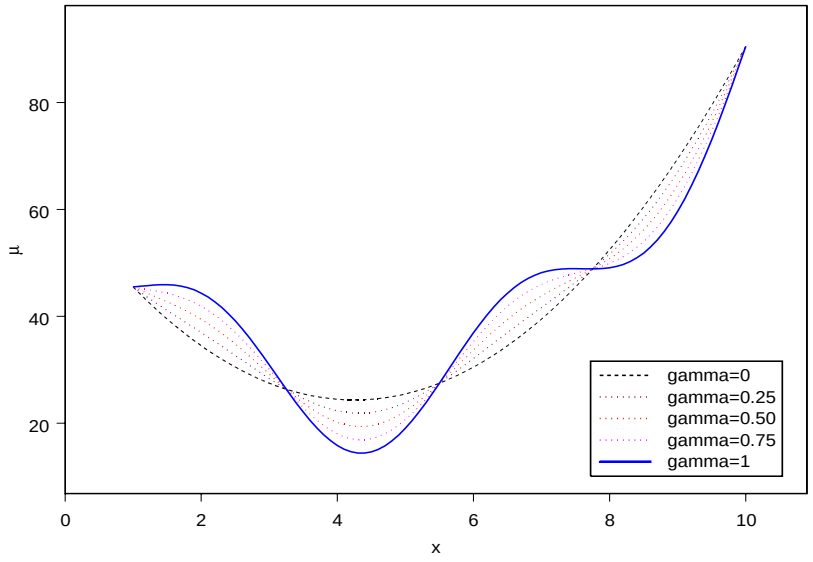
The parameter γ is the misspecification parameter. That is, the user will assume that the data are from the quadratic model

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

but the true model is the one given in (8.1); the trigonometric component of (8.1) serves as the misspecification. When γ is zero, the user's model equals the true model, and there is no model misspecification, otherwise the model has been misspecified. The degree of misspecification increases with γ. Values of γ equal to 0, 0.25, 0.50, 0.75, and 1 will be used in the study. A plot of the population average models using the γ values from the study is given in Figure 8.1. The smooth parabola (dashed line, γ=0) occurs when the user's model equals the true model. The solid curve represents the most misspecification in the population average, when γ=1. The large disparity between the γ=0 and γ=1 models should be reflected in the MSE results from the simulation study.

Three different variance-covariance structure for the random errors ϵ_{ij} were considered. The first variance-covariance structure considered was independence, with the variance of

Figure 8.1: Plot of Population Average Underlying Models (γ is misspecification parameter)



the errors equal to 16. Thus, $\mathbf{R} = \sigma^2\mathbf{I} = (16)\mathbf{I}$. An autoregressive, lag-one model (AR(1)) with ρ=0.20 and σ^2=16, and

$$\text{Cov}[\epsilon_{ij}, \epsilon_{ik}] = \sigma^2\rho^{|x_j - x_k|} = (16)(0.20)^{|x_j - x_k|}$$
$$\text{Cov}[\epsilon_{ij}, \epsilon_{kl}] = \ 0 \ \forall \ i \neq k$$

was the second covariance structure. The last covariance structure considered was the AR(1) model with ρ=0.80 and σ^2=16. Thus, the three variance-covariance structures cover three correlation ranges: uncorrelated (independence case), low correlation (AR(1) with ρ=0.20), and the high correlation case (AR(1) with ρ=0.80).

It is assumed that there is no 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 with an AR(1) structure for $\mathbf{R}$.

The local model used in the conditional local and marginal local mixed model analyses is the local linear mixed model with a random intercept, a simple model that can be very flexible. The local models use an independence structure to model the within-cluster

variation; the CLMM, for example, has a variance-covariance structure with

$$\begin{aligned} \mathrm{Var}[Y_{ij}] &= \frac{\sigma^2}{k_{ii}} + \sigma^2_{b_0} \\ \mathrm{Cov}[Y_{ij}, Y_{kl}] &= \begin{cases} \sigma^2_{b_0} & \text{if i=k } \forall \text{ j,l} \\ 0 & \text{if i}\neq\text{k.} \end{cases} \end{aligned}$$

where k_{ii} is the i^{th} kernel weight of $\mathbf{K_0}$.

The bandwidth selectors used in the study were PRESS, PRESS*, and PRESS** (discussed in Ch. 3). The bandwidths were found for each dataset using the Golden Section Search Method (see Press et al., (1989)). The brackets used in the search were 0.05, 0.15, and 0.30. A value of 0.05 was practically a lower limit; we were unable to choose a smaller h because of a lack of information to fit the mixed models at small bandwidths. The bandwidth 0.30 was chosen because it appeared to be an "upper bound" for our study; preliminary studies produced bandwidths less than 0.30 for each bandwidth selector in every situation. Thus these values were selected to minimize the distance covered by the search method. The mixing parameter λ was found using formula (7.16). Because no bounds are imposed by the mixing parameter formula, if $\tilde{\lambda}$ was the solution of the minimization problem, then the estimate was

$$\hat{\lambda} = \begin{cases} 0 & \text{if } \tilde{\lambda} \leq 0 \\ \tilde{\lambda} & \text{if } 0 < \lambda \leq 1 \\ 1 & \text{if } \tilde{\lambda} > 1. \end{cases}$$

Both the bandwidth and the mixing parameter were found by summing over the design points. Using the bandwidth and mixing parameter for a given dataset, the integrated mean square error 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), conditional local mixed model (population average and cluster specific), the marginal local mixed model (population average), and the mixed model robust regression models (population average using the conditional local mixed model, population average using the marginal local mixed model, and cluster specific using the conditional local mixed model).

We are interested in the approximate integrated MSE (INTMSE) as a function of cluster size, of correlation, and of γ. To keep the number of scenarios reasonably manageable s=5 and s=20 clusters per data set are examined over 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 subtantial computing resources. To examine the practically feasible number of needed simulation runs that also provides sufficient precision of Monte-Carlo averages, we examined standard errors of Monte-Carlo mean square errors. Figures 8.2 and 8.3 are plots of the standard errors of the average MSE values versus the number of runs for the case of five clusters, independence structure, and γ=1 using PRESS as the bandwidth selector. Each line represents a standard error of the average INTMSE for a different procedure. Figure 8.2 is for population average estimation, and Figure 8.3 is for cluster specific prediction. As the number of runs increases, the standard error decreases and levels off. Obviously, 500 or more runs could be advantageous, but computing time is excessive at this point. Hence, we decided on 250 runs, acknowledging that this will retain Monte-Carlo variability higher than usually desired, but that still allows a comparison of the methods.

8.2 Simulation Study

8.2.1 Bandwidth Study

A small simulation study to compare the bandwidth selectors was performed. The purpose was to determine if one of the bandwidth selectors could be eliminated in future studies (less computation), and to get a feel for which selector is appropriate for a given estimation technique. The study used the model in (8.1) with the data obtained at the ten regressor locations for five clusters. The within-cluster and between-cluster variance-covariance structures are independent. The approximate integrated mean square errors appear in Table 8.1. The values in bold will be discussed in section 8.2.2 when comparing the parametric, local, and model robust methods across γ.

Figure 8.2: Plot of Standard Error of average INTMSE versus the number of Monte Carlo runs (Population Average)

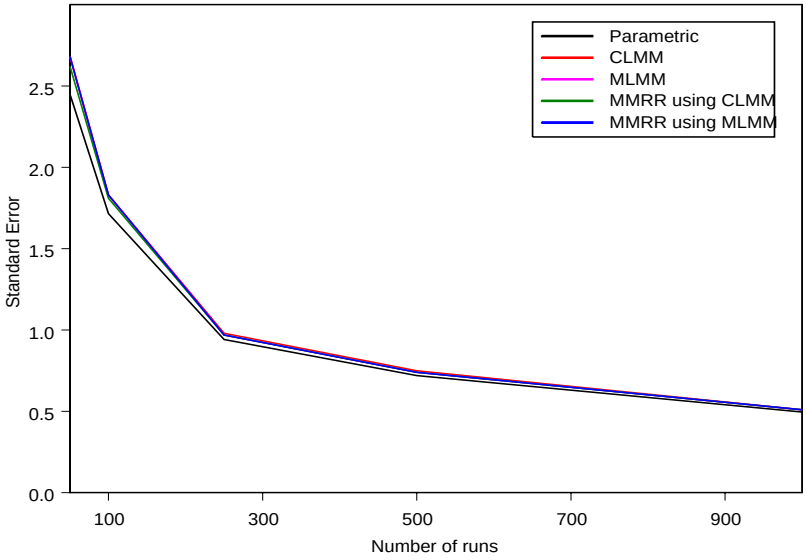


Figure 8.3: Plot of Standard Error of average INTMSE versus the number of Monte Carlo runs (Cluster Specific)

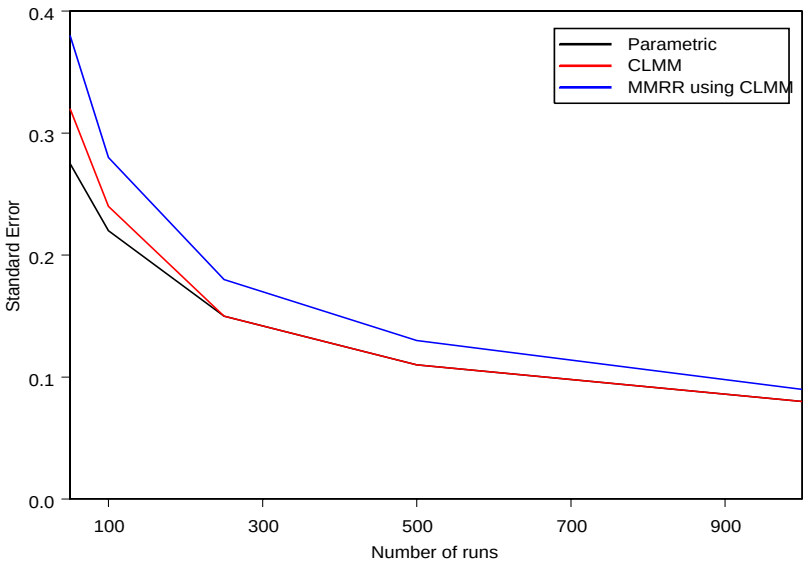


Table 8.1: Simulated INTMSE using Independence (10 regressor locations and 5 clusters)

γ	selector	PA Para-metric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Para-metric	CS CLMM	CS MMRR CLMM
0.00	PRESS	**13.83**	15.59	15.76	13.84	14.00	**2.65**	8.41	2.73
0.00	PRESS*	**13.83**	17.88	16.18	13.88	13.97	**2.65**	8.58	2.71
0.00	PRESS**	**13.83**	16.09	15.98	13.87	13.95	**2.65**	7.54	2.73
0.25	PRESS	16.75	15.78	16.03	15.81	**15.60**	5.62	8.98	**4.99**
0.25	PRESS*	16.75	18.45	17.11	16.28	**16.27**	5.62	9.19	**5.21**
0.25	PRESS**	16.75	16.36	16.73	**15.94**	16.06	5.62	8.06	**4.99**
0.50	PRESS	25.51	**16.19**	**16.19**	18.21	16.33	14.63	10.27	**8.48**
0.50	PRESS*	25.51	19.54	**19.47**	20.08	19.51	14.63	10.51	**9.65**
0.50	PRESS**	25.51	**16.77**	17.75	18.28	17.78	14.63	9.20	**8.35**
0.75	PRESS	40.11	16.62	**16.57**	19.68	16.69	29.87	11.53	**11.25**
0.75	PRESS*	40.11	**20.26**	20.74	21.30	20.74	29.87	**11.76**	12.06
0.75	PRESS**	40.11	**17.26**	17.78	19.35	17.78	29.87	**10.48**	10.67
1.00	PRESS	60.56	17.23	**17.14**	20.97	17.23	51.35	**12.67**	13.57
1.00	PRESS*	60.56	**20.40**	20.62	21.30	20.62	51.35	**12.67**	13.08
1.00	PRESS**	60.56	**17.69**	18.04	20.33	18.04	51.35	**11.86**	12.68

Each row of a table corresponds to a value of γ. Each column corresponds to a regression method; the first five MSE columns (section one) are population average (PA) results, while the remaining three (section two) are cluster specific (CS) results. Each section progresses through columns for the parametric, the local, and the mixed model robust methods. The column selector indicates which bandwidth selector (PRESS, PRESS*, or PRESS**) was used in a given row.

Two important conclusions can be drawn from Table 8.1. Notice that in cellwise comparisons, virtually all of the INTMSE values using PRESS* as a bandwidth selector are larger than the INTMSE values using PRESS and PRESS**. This suggests that PRESS* does not perform as well as the other two selectors in terms of minimizing the integrated mean square error. Thus for the remainder of this work, only PRESS and PRESS** will be utilized.

The INTMSE values for estimation of the population average appear to be the small-

est when PRESS is the bandwidth selector; this is suggested in a comparison between the PA CLMM and PA MLMM columns for selector=PRESS and selector=PRESS**. This suggests that if a user is interested in population average estimation, they should use PRESS as a bandwidth selector. This is consistent with results in Clark (2002), where PRESS was the bandwidth selector of choice; his work was concerned estimation of population averages only.

The opposite is true for cluster specific prediction. The INTMSE values are smallest when PRESS** is the selector, as is gathered from the CS CLMM columns for selector=PRESS and selector=PRESS**. If interest is in cluster specific predictions, PRESS** appears to be the bandwidth selector of choice. This finding is consistent with conclusions made in Mays, Birch, and Starnes (2001). One can think of their data structure as consisting of a single cluster; Mays, Birch, and Starnes (2001) show that PRESS** is superior to PRESS and PRESS* for the fixed effects model with independent data. The conclusions made about the appropriateness of the bandwidth selectors for a given prediction preference will be verified through additional simulation studies and the simulated optimal bandwidths later on in this chapter.

In the results that follow, PA and CS results for both PRESS and PRESS** will still be shown. These results will demonstrate the consistency of our findings across scenarios and to highlight changes in results.

8.2.2 Varying Cluster Size and Correlation Structure

The INTMSE values for five clusters assuming a within-cluster independence structure were given in the previous section. It is expected that as the number of clusters increase, the INTMSE for the local population average should decrease since more observations at a given 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 x_0 only, and resulting in a local population average fit that connects the mean responses at each value of the regressor. The INTMSE values for the local cluster specific fits should be unaffected by the addition

Table 8.2: Simulated INTMSE using PRESS and Independence (10 regressor locations and 20 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**3.37**	4.08	4.08	**3.37**	3.43	**2.49**	10.27	**2.49**
0.25	6.30	4.18	**4.16**	5.42	**4.16**	5.42	10.83	**4.81**
0.50	15.06	4.43	**4.42**	7.81	4.45	14.35	12.73	**8.81**
0.75	29.67	4.86	**4.84**	9.25	4.93	29.57	13.64	**12.10**
1.00	50.12	5.47	**5.45**	10.52	5.51	51.42	**14.24**	14.64

Table 8.3: Simulated INTMSE using PRESS** and Independence (10 regressor locations and 20 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**3.37**	4.16	4.50	3.38	3.45	**2.49**	9.56	**2.49**
0.25	6.30	**4.19**	4.43	5.42	4.40	5.42	10.73	**4.81**
0.50	15.06	**4.48**	4.52	7.50	4.53	14.35	11.04	**8.40**
0.75	29.67	4.92	**4.86**	8.96	**4.86**	29.57	12.84	**11.69**
1.00	50.12	**5.47**	5.48	10.52	5.51	51.42	**14.24**	14.64

of clusters.

In this work, five clusters will be considered our "small" number of clusters, and twenty clusters our "large" number of clusters. Tables 8.2 and 8.3 give the simulated INTMSE values for the independence case for twenty clusters using PRESS and PRESS**, respectively, as the bandwidth selectors. A population average and a cluster specific MSE are highlighted in each row; this bolded value is the minimum population average and cluster specific INTMSE for the given γ value. As expected, the INTMSE decreases for the population average estimation; this decrease is substantial for population average estimation. There is a marginal decrease for the parametric cluster specific prediction. The conditional local cluster specific INTMSE values tend to be larger for the 20 cluster case.

The question now arises– which local method should we use for population average estimation? The population average fits should be chosen via PRESS. The marginal local mixed model is clearly the winner, as the population average INTMSE values for the conditional local mixed model are typically always larger than those for the marginal local mixed model (Tables 8.1 and 8.2). Initial results from the independence cases suggest that the marginal local mixed model should be used for population average estimation, while the conditional local mixed model is utilized for cluster specific prediction.

The cross-over points are the values of γ at which the minimum INTMSE value switch from parametric to model robust estimation, and from model robust to local estimation. It is interesting to note that the initial cross-over point (from parametric to model robust estimation) occurs earlier (a smaller value of γ) for the population average than for cluster specific estimation in the five cluster case. For twenty clusters, it is not apparent from the tables that the model robust method ever wins for the population average; the initial population average cross-over point must occur somewhere between γ=0 and γ=0.25. Thus a search over a finer grid of γ values was performed in order to ascertain where the initial cross-over point occurs for population average estimation. Tables 8.4 through 8.7 give the INTMSE values for γ values between 0 and 1, with a finer grid between 0 and 0.30, for five and twenty clusters using PRESS and PRESS**, respectively. Plots of the parametric, local, and model robust INTMSE values versus γ over the finer grid (from γ=0 to 0.30 by 0.05) appear as Figures 8.4 through 8.7. The plots are for population average estimation or cluster specific prediction for s=5 and s=20. Each plot gives the results using PRESS and PRESS** as the bandwidth selector. Both the tables and plots give a more accurate picture of the initial cross-over. For both population average and cluster specific prediction, the minimum INTMSE value reverts 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=0.25 using PRESS, whereas the second cluster-

Table 8.4: Cross-Over Points for Mixed Model Robust Regression using PRESS (10 regressor locations and 5 clusters) Independence Case

γ	PA Parm.	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parm.	CS CLMM	CS MMRR CLMM
0.00	**13.83**	15.59	15.76	13.84	14.00	**2.65**	8.41	2.73
0.05	13.95	15.60	15.81	**13.94**	14.08	**2.76**	8.44	2.84
0.10	14.30	15.63	15.82	**14.23**	14.36	**3.12**	8.51	3.16
0.15	14.88	15.69	15.87	**14.69**	14.78	3.71	8.63	**3.66**
0.20	15.70	15.73	15.96	15.24	**15.23**	4.55	8.79	**4.29**
0.25	16.75	15.78	16.03	15.81	**15.60**	5.62	8.98	**4.99**
0.30	18.03	15.88	16.03	16.40	**15.87**	6.93	9.18	**5.71**
0.50	25.51	**16.19**	**16.19**	18.21	16.33	14.63	10.27	**8.48**
0.75	40.11	16.62	**16.57**	19.68	16.69	29.87	11.53	**11.25**
1.00	60.56	17.23	**17.14**	20.97	17.23	51.35	**12.67**	13.57

Table 8.5: Cross-Over Points for Mixed Model Robust Regression using PRESS** (10 regressor locations and 5 clusters) Independence Case

γ	PA Parm.	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parm.	CS CLMM	CS MMRR CLMM
0.00	**13.83**	16.09	15.98	13.87	13.95	**2.65**	7.54	2.73
0.05	**13.95**	16.10	16.00	13.96	14.06	**2.77**	7.55	2.84
0.10	14.30	16.15	16.06	**14.27**	14.36	**3.12**	7.62	3.16
0.15	14.88	16.19	16.24	**14.74**	14.87	3.71	7.73	**3.66**
0.20	15.70	16.25	16.48	**15.31**	15.46	4.55	7.88	**4.29**
0.25	16.75	16.36	16.73	**15.94**	16.06	5.62	8.06	**4.99**
0.30	18.03	**16.45**	17.02	16.54	16.65	6.93	8.28	**5.73**
0.50	25.51	**16.77**	17.75	18.28	17.78	14.63	9.20	**8.35**
0.75	40.11	**17.26**	17.78	19.35	17.78	29.87	**10.48**	10.67
1.00	60.56	**17.69**	18.04	20.33	18.04	51.35	**11.86**	12.68

Table 8.6: Cross-Over Points for Mixed Model Robust Regression using PRESS (10 regressor locations and 20 clusters) Independence Case

γ	PA Parm.	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parm.	CS CLMM	CS MMRR CLMM
0.00	**3.37**	4.08	4.08	**3.37**	3.43	**2.49**	10.7	**2.49**
0.05	**3.49**	4.08	4.08	**3.49**	3.56	**2.60**	10.30	2.61
0.10	3.84	4.09	4.09	3.81	**3.79**	2.95	10.45	**2.94**
0.15	4.43	4.12	4.11	4.28	**3.97**	3.54	10.58	**3.45**
0.20	5.24	4.15	4.12	4.84	**4.08**	4.36	10.72	**4.09**
0.25	6.30	4.18	**4.16**	5.42	**4.16**	5.42	10.83	**4.81**
0.30	7.58	4.22	**4.19**	5.98	4.22	6.72	11.08	**5.59**
0.50	15.06	4.43	**4.42**	7.81	4.45	14.35	12.73	**8.81**
0.75	29.67	4.86	**4.84**	9.25	4.93	29.57	13.64	**12.10**
1.00	50.12	5.47	**5.45**	10.52	5.51	51.42	**14.24**	14.64

Table 8.7: Cross-Over Points for Mixed Model Robust Regression using PRESS** (10 regressor locations and 20 clusters) Independence Case

γ	PA Parm.	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parm.	CS CLMM	CS MMRR CLMM
0.00	**3.37**	4.16	4.50	3.38	3.45	**2.49**	9.56	**2.49**
0.05	**3.49**	4.16	4.51	**3.49**	3.57	**2.60**	9.62	2.61
0.10	3.84	4.17	4.51	**3.81**	3.86	2.95	9.79	**2.94**
0.15	4.43	4.17	4.50	4.28	**4.15**	3.54	10.17	**3.44**
0.20	5.24	**4.16**	4.45	4.84	4.33	4.36	10.51	**4.08**
0.25	6.30	**4.19**	4.43	5.42	4.40	5.42	10.73	**4.81**
0.30	7.58	**4.22**	4.42	5.96	4.43	6.72	10.80	**5.56**
0.50	15.06	**4.48**	4.52	7.50	4.53	14.35	11.04	**8.40**
0.75	29.67	4.92	**4.86**	8.96	4.86	29.57	12.84	**11.69**
1.00	50.12	**5.47**	5.48	10.52	5.51	51.42	**14.24**	14.64

Figure 8.4: Plot of INTMSE versus γ (Population Average for 5 clusters)

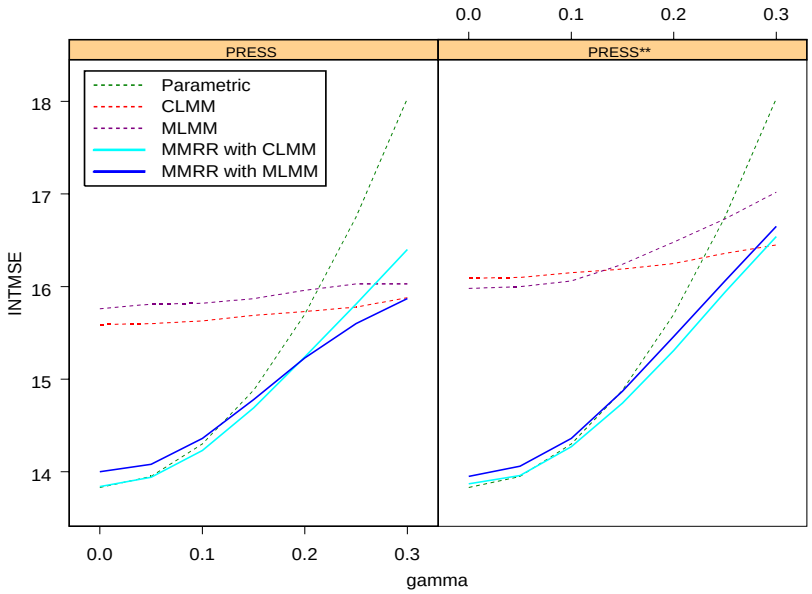


Figure 8.5: Plot of INTMSE versus γ (Cluster Specific for 5 clusters)

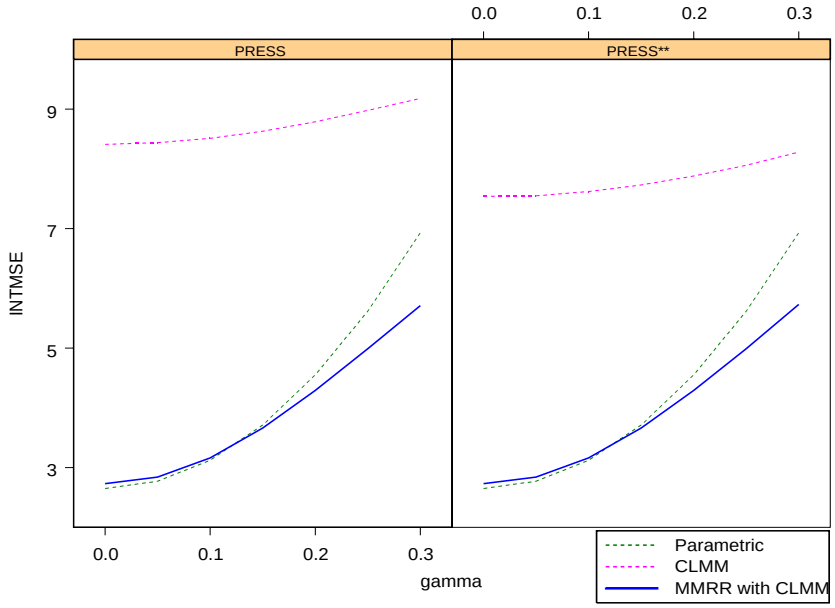


Figure 8.6: Plot of INTMSE versus γ (Population Average for 20 clusters)

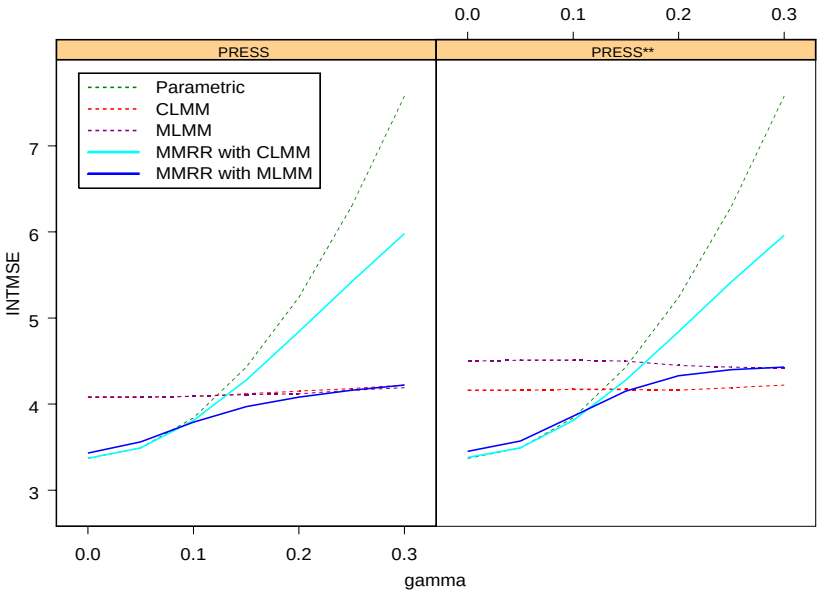
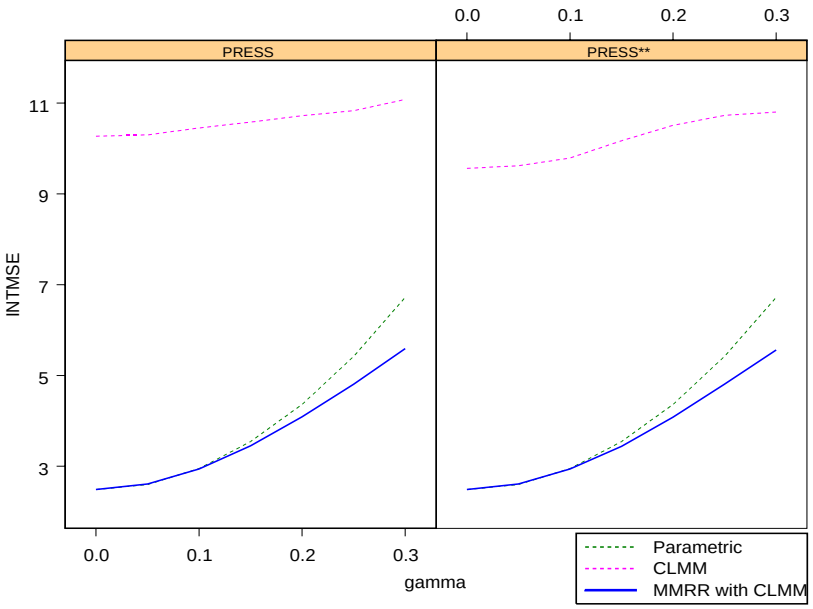


Figure 8.7: Plot of INTMSE versus γ (Cluster Specific for 20 clusters)



specific cross-over point occurs for large misspecification– a γ value between 0.75 and 1.0 (from Tables 8.2 and 8.3). Interestingly, when using PRESS** as a bandwidth selector, the conditional local mixed model and the mixed model robust methods using CLMM do not minimize the population average INTMSE values uniformly. That is, for twenty clusters, MLMM had the smallest INTMSE value for γ=0.75 and MMRR using MLMM obtained the minimum INTMSE value for γ=0.15. This may be due to Monte Carlo variability; however, interesting results in the CLMM versus MLMM debate for the two bandwidth selectors emerge in the correlated errors cases.

Notice that the cross-overs occur earlier when s is large. For example, when using PRESS, the second cross-over point was between γ=0.30 and γ=0.50 for five clusters; the second cross-over point for twenty clusters fell between γ=0.20 and γ=0.25. This pattern occurs throughout this research, and is consistent with the results of Clark (2002).

Two studies were performed using an AR(1) structure on the random errors with a low correlation (ρ=0.20) and a high correlation (ρ=0.80). As in the independence case, the random effects were assumed to be independent, so $\mathbf{B} = \sigma_b^2\mathbf{I}$. The local model is again assuming independence for the within-cluster variation with a random intercept.

One concern in the correlated data case was whether the misspecification term influenced the estimate of ρ. Recall that the user's parametric model, which assumes that the data are from the true variance-covariance structure (in this case, an AR(1) structure), is misspecified when γ differs from zero. How does the misspecification influence 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. We can, however, look at the estimates of ρ under varying misspecification in a Monte-Carlo study. Five hundred data sets were generated for different values of γ, and the average estimate of ρ from the parametric analysis over the five hundred datasets were calculated. The data were generated from an AR(1) process using ρ=0, ρ=0.10, ρ=0.20, ρ=0.33, ρ=0.80, and ρ=0.90. This simulation used 10 regressor locations and 20 clusters. (Table 8.8). At γ=0, when

Table 8.8: Average Estimate of ρ from Parametric Estimation (10 regressor locations, 20 clusters, and 500 iterations)

γ	ρ=0	ρ=0.10	ρ=0.20	ρ=0.33	ρ=0.80	ρ=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

the model is correctly specified, the estimate $\hat{\rho}$ is 0.19 when the true correlation coefficient is 0.20, and $\hat{\rho}$ is 0.79 when the true correlation coefficient is 0.80. These values are fairly close to the true values, which indicates that we are indeed simulating our data from an AR(1) process with the desired value of ρ. The estimates of ρ are both smaller than the true values; this was expected, because although REML estimation has less bias than ML, it still provides estimates that are negatively biased. It is interesting to note that when the true value of ρ is 0.20, as the amount of mean model misspecification increases, the estimate of ρ increases. For extreme model misspecification (γ=1), $\hat{\rho}$ is 0.28, when the true correlation coefficient is 0.20. When the true value of ρ is 0.80, however, the estimate of the correlation coefficient decreases as γ increases. For extreme model misspecification (γ=1), $\hat{\rho}$ is 0.36, when the true correlation coefficient is 0.80. The estimate is less than half of the true value. This essentially says that highly (weakly) correlated data appear less (more) correlated for large model misspecification for our simulation model. Thus, without question, model misspecification unduly influences the estimate of ρ.

We can look at this phenomenon more closely. Table 8.9 provides us with a subset of the data generated from the ρ=0.20 case. This is one of five clusters from one dataset.

Table 8.9: One Cluster from ρ=0.20

X	μ	ϵ	f $\gamma = 0$	f $\gamma = 0.25$	f $\gamma = 0.5$	f $\gamma = 0.75$	f $\gamma = 1$	sign ϵ	sign ϵ + f γ=1
1	61.34	1.01	0	0	0	0	0	+	+
2	53.85	4.84	0	2.46	4.92	7.39	9.85	+	+
3	35.65	-8.17	0	0.86	1.71	2.57	3.42	-	-
4	17.37	0.43	0	-2.17	-4.33	-6.50	-8.66	+	-
5	18.97	-0.88	0	-1.61	-3.21	-4.82	-6.43	-	-
6	36.77	-4.84	0	1.61	3.21	4.82	6.43	-	+
7	49.51	-1.66	0	2.17	4.33	6.50	8.66	-	+
8	53.51	3.13	0	-0.86	-1.71	-2.57	-3.42	+	-
9	68.73	-4.94	0	-2.46	-4.92	-7.39	-9.85	-	-
10	105.804	2.59227	0	0	0	0	0	+	+

The first column gives the value of the regressor, the second column is the conditional mean E[Y|**b**] (denoted by μ), the third column is the realization of the random error term. The remaining columns are the amount of the misspecification determined by the expression

$$\text{f} = 10\gamma \sin(\tfrac{\pi(X-1)}{2.25}).$$

for γ=0, 0.25, 0.50, 0.75, and 1. The f columns differ in the amount of misspecification. Notice that as the misspecification increases, the f increases in magnitude and retains the same sign. At x=1, x=2, and x=3, ϵ and ϵ + f have the same sign. At x=4, however, ϵ + f remains negative, while ϵ is positive. Notice that there are longer runs (more values that have the same sign in succession) for ϵ + f; this has the effect of increasing the lag-one correlation. A plot of the data versus x as a function of γ appears in Figure 8.8. As you can see, the misspecification "realigns" the data; for a given value of the regressor, the response (Y = μ + f + ϵ) for γ=1 is always larger than the response at γ=0 if f if positive and is smaller than the response at γ=0 if f is negative. A similar argument for the ρ=0.80 case holds to show that the estimate of ρ decreases as the misspecification increases.

Figure 8.9 is a plot of the misspecification portion versus x for γ=1. The sinusoidal nature of f creates repeating runs of positive, then negative, values of the misspecification. These runs appear to be the culprit as to why the correlation increases for small ρ and

Figure 8.8: Plot of Data for Varying γ

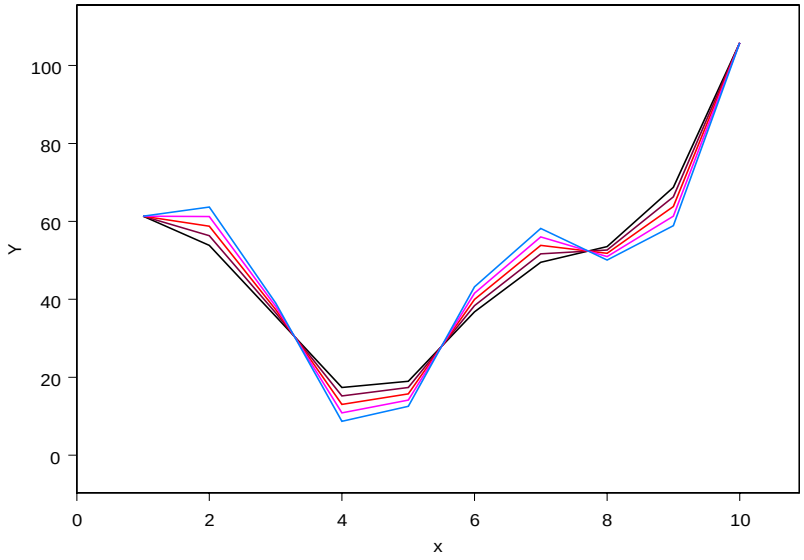


Figure 8.9: Plot of f versus x for γ=1

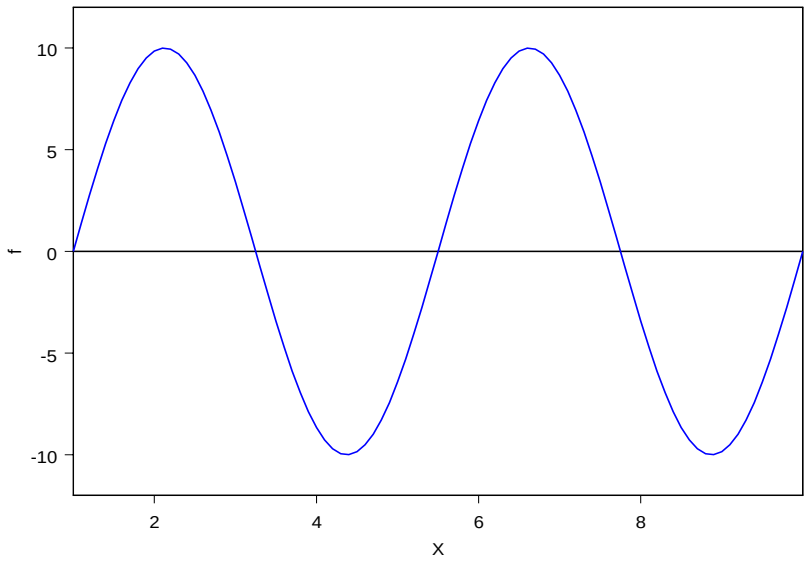


Table 8.10: Simulated MSE using PRESS and AR(1) with rho=0.20 (10 regressor locations and 5 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**13.77**	15.66	15.71	13.78	13.91	**3.74**	10.11	3.78
0.25	16.73	15.81	15.94	15.86	**15.57**	6.74	10.62	**6.02**
0.50	25.53	16.10	**16.09**	18.49	16.28	16.00	11.77	**9.56**
0.75	40.18	16.48	**16.43**	20.24	16.59	31.81	12.85	**12.54**
1.00	60.99	17.06	**16.99**	21.87	17.14	54.00	**13.96**	15.18

Table 8.11: Simulated MSE using PRESS** and AR(1) with rho=0.20 (10 regressor locations and 5 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**13.77**	16.22	16.25	13.79	13.91	**3.74**	9.04	3.77
0.25	16.73	16.38	16.94	**15.99**	16.17	6.74	9.55	**6.05**
0.50	25.53	**16.77**	17.71	18.69	17.77	16.00	10.65	**9.56**
0.75	40.18	**17.18**	17.65	20.15	17.65	31.81	**11.82**	12.12
1.00	60.99	**17.57**	17.81	21.30	17.81	54.01	**13.05**	14.34

decreases for large ρ as f increases. It appears that the estimates do converge to some value. The value of ρ=1/3 provided estimates that were close to the true value, regardless of f. This may be an indication that the value that the estimates are converging to is close to 1/3.

The INTMSE values for the correlated data cases appear in Tables 8.10 – 8.17. Tables 8.10 through 8.13 are for the AR(1) case with ρ=0.20. The first two tables represent cases with s=5, with Table 8.10 using PRESS and Tables 8.11 using PRESS** as bandwidth selectors. The remaining two tables are for s=20, with Table 8.12 and Table 8.13 using PRESS and PRESS**, respectively. Tables 8.14 through 8.17 are for the AR(1) case with ρ=0.80. These tables follow the same layout as the tables for the AR(1) with ρ=0.20 case. As in the independence case, the first five columns are for population average estimation

Table 8.12: Simulated MSE using PRESS and AR(1) with rho=0.20 (10 regressor locations and 20 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**3.42**	3.99	3.97	**3.42**	3.46	**3.34**	11.74	**3.34**
0.25	6.34	4.07	**4.06**	5.42	4.08	6.34	12.23	**5.67**
0.50	15.12	**4.31**	**4.31**	7.95	4.36	15.74	13.95	**9.75**
0.75	29.77	4.74	**4.72**	9.87	4.81	32.31	14.44	**13.25**
1.00	50.28	5.33	**5.31**	11.88	5.37	55.81	**15.03**	16.30

Table 8.13: Simulated MSE using PRESS** and AR(1) with rho=0.20 (10 regressor locations and 20 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**3.42**	4.06	4.29	**3.42**	3.44	**3.34**	10.72	**3.34**
0.25	6.34	**4.09**	4.24	5.42	4.26	6.34	11.80	**5.66**
0.50	15.12	4.36	**4.35**	7.66	4.36	15.74	12.20	**9.36**
0.75	29.77	4.77	**4.73**	9.72	**4.73**	32.31	13.98	**13.04**
1.00	50.28	5.33	**5.32**	11.88	5.35	55.81	**15.03**	16.30

Table 8.14: Simulated MSE using PRESS and AR(1) with rho=0.80 (10 regressor locations and 5 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**14.26**	16.25	16.31	14.37	14.80	**6.48**	15.16	6.63
0.25	17.30	16.35	16.33	16.59	**16.29**	9.41	15.34	**8.97**
0.50	26.24	16.61	**16.41**	18.81	16.83	18.60	15.74	**12.52**
0.75	40.96	17.03	**17.01**	20.27	17.26	34.04	16.26	**15.17**
1.00	61.50	17.64	**17.61**	21.80	17.87	55.93	**16.87**	17.42

Table 8.15: Simulated MSE using PRESS** and AR(1) with rho=0.80 (10 regressor locations and 5 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**14.26**	17.07	17.18	14.44	14.89	**6.48**	14.54	6.60
0.25	17.30	17.21	17.80	**16.97**	17.27	9.41	14.80	**9.14**
0.50	26.24	**17.51**	18.34	19.62	18.48	18.60	15.35	**13.11**
0.75	40.96	**17.71**	18.18	20.44	18.20	34.04	15.90	**15.38**
1.00	61.50	**18.06**	18.21	21.45	18.21	55.93	**16.52**	17.13

Table 8.16: Simulated MSE using PRESS and AR(1) with rho=0.80 (10 regressor locations and 20 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**3.68**	4.25	4.25	**3.68**	3.75	**5.78**	15.63	**5.78**
0.25	6.63	4.33	**4.32**	5.60	4.33	8.85	15.71	**8.39**
0.50	15.46	4.58	**4.57**	7.34	4.62	18.13	15.97	**12.34**
0.75	30.13	5.01	**4.99**	8.79	5.03	33.89	16.39	**15.10**
1.00	50.66	5.62	**5.58**	10.72	5.60	56.55	**17.00**	17.56

Table 8.17: Simulated MSE using PRESS** and AR(1) with rho=0.80 (10 regressor locations and 20 clusters)

γ	PA Parametric	PA CLMM	PA MLMM	PA MMRR CLMM	PA MMRR MLMM	CS Parametric	CS CLMM	CS MMRR CLMM
0.00	**3.68**	4.36	4.62	**3.68**	3.72	**5.78**	14.67	**5.78**
0.25	6.63	**4.38**	4.49	5.62	4.48	8.85	15.04	**8.40**
0.50	15.46	4.64	**4.59**	7.29	**4.59**	18.13	15.39	**12.28**
0.75	30.13	5.01	**4.99**	8.78	5.00	33.89	16.37	**15.09**
1.00	50.66	5.62	**5.58**	10.72	5.50	56.55	**17.00**	17.56

and the last three columns are for cluster specific prediction, with the minimum population average and cluster specific INTMSE value in bold face for each value of γ.

There are similarities between the independence case and the correlated error cases. In the five cluster cases, it appears that the marginal local mixed model has lower INTMSE values when PRESS is the bandwidth selector, and the conditional local mixed model has the minimum INTMSE values when PRESS** is the bandwidth selector.

Notice that population average MLMM and cluster specific CLMM model robust methods 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 local values. As γ increases from zero to one, the INTMSE values for the mixed model robust procedures are either the minimum values (for low to moderate model misspecification), or are close in value to the "winning" INTMSE values.

The one exception to the competitiveness of the model robust methods is the population average CLMM model robust method. Although it performs well cases of low model misspecification (γ=0 and 0.25), it performs poorly for moderate to large model misspecification when compared to the population average MLMM model robust method, especially for s=20 clusters. This issue will be discussed in more detail in the next subsection.

The pattern of the cross-over points in γ across the covariance structures also appears to be similar. The population average cross-over points all seem to occur between γ=0 and γ=0.25, and between γ=0.25 and γ=0.50 for five clusters, and earlier for s=20 clusters. This suggests that population average mixed model robust regression works well for small amounts of model misspecification. The model robust procedures work extremely well for cluster specific prediction; the initial cross-over point is usually between γ=0 and γ=0.25, while the second cross-over points occur between γ=0.75 and γ=1. Thus, the cluster specific mixed model robust regression method generally works well for all levels of misspecification.

There are some key differences between the independent and correlated cases. On

average, the INTMSE values increase as the correlation increases. For example, in the five cluster case, the INTMSE value for the population average parametric model at γ=0 for ρ=0.20 and ρ=0.80 are 13.77 and 14.26, respectively. (The INTMSE value for ρ=0.20 is smaller than the INTMSE value for the uncorrelated case; this is due in part to different seeds used in generating the dataset and Monte-Carlo variability). The mean square error is the sum of the squared bias plus the variance of the fit when conditioned on the values of the random effects. In the γ=0 case, the bias is zero, so the mean square error is the variance of the fit. Thus, as the correlation increases, the variance of the fits must increase. If the sample size n remains fixed, and ρ increases, the effective sample size decreases (like n decreasing) so the variance must increase as ρ increases. In the parametric case for population average estimation at the data points, $\text{Var}(\hat{\mathbf{Y}}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ if the true variance-covariance matrix $\mathbf{V}$ is known. Taking the trace of this matrix gives us the sum of the variances over all of the data points. We can divide this sum by the number of fits to obtain the average variance (and hence the mean square error). For estimation at x=1 to x=10 by units of 0.20 with five clusters, the INTMSE values for the parametric population average are 12.81, 12.87, and 13.46 for the uncorrelated and AR(1) with ρ equal to 0.20 and 0.80, respectively.

Recall that using PRESS as the bandwidth selector as opposed to PRESS** resulted in smaller INTMSE values for the population average. In virtually all of the five cluster cases, the marginal local mixed model outperformed the conditional local mixed model when the local mixed model was the method that minimized mean square error (for "large" values of γ). In addition, we note the interesting fact for s=20, namely, that the marginal local mixed model appears to have the smallest INTMSE value on average over different amounts of γ, regardless of the bandwidth selector. In Tables 8.12 and 8.13, for example, the marginal local mixed model has the smallest INTMSE value for γ=0.50, 0.75, and 1 for both PRESS and PRESS**. When γ=0, however, the parametric or the model robust do better than the local methods. There is only one case where the local method is the winner and the conditional local mixed model outperforms the marginal local mixed model; this is when PRESS** is the bandwidth selector and γ=0.25. This verifies our observation that the marginal local mixed

model is superior to the conditional local mixed model for population average estimation.

8.2.3 Average Bandwidth and Mixing Parameter

Also of interest is the average bandwidth and mixing parameter for different amounts of misspecification, cluster size, and covariance structures (with different amounts of correlation). It is of course expected that, as γ increases, the estimate of λ would increase; more misspecification indicates a greater need for the nonparametric fit. For γ=0, the estimate of λ should be close to zero, as the model has been correctly specified and the model robust fit should equal the parametric fit. As for the bandwidth, it is expected that h will be larger using PRESS** than with PRESS because of the penalty functions. This was the case with the work of Mays, Birch, and Starnes (2001) for the fixed effects case.

Across the two cluster sizes, it is expected that the λ values will be about the same. The bandwidth, on the other hand, should be smaller for the twenty cluster cases. More data at the point of estimation will put more weight on the points at $\tilde{x}_0$ since the weights must sum to one across a dataset (and the sum of the weights for a given cluster to to the inverse of the number of clusters if every cluster has observations at the same values of the regressor).

Across the different variance-covariance structures and differing amounts of correlation, it is again expected that the values of λ will remain relatively constant for a given bandwidth and γ. The bandwidths for the correlated data should be larger. The bandwidths should be smallest for the independence case and largest for the AR(1) with ρ=0.80 case.

Tables 8.18 – 8.20 provide the bandwidth and mixing parameter results for the independence, AR(1) with ρ=0.20, and AR(1) with ρ=0.80 cases. The first column gives the degree of misspecification. The second column denotes whether the row contains the average bandwidth ($\bar{h}$) or average mixing parameter ($\bar{\lambda}$) for the given level of γ. Columns three and four denote whether the conditional or marginal local mixed model was used for five clusters using PRESS as a bandwidth selector; columns five and six are for CLMM and MLMM models with twenty clusters using PRESS. Columns seven through ten are denoted similarly,

Table 8.18: Average Bandwidth and λ from Simulations (Independence Case)

γ		PRESS CLMM clust=5	PRESS MLMM clust=5	PRESS CLMM clust=20	PRESS MLMM clust=20	PRESS** CLMM clust=5	PRESS** MLMM clust=5	PRESS** CLMM clust=20	PRESS** MLMM clust=20
0.00	$\bar{\lambda}$	0.07	0.18	0.01	0.11	0.07	0.15	0.01	0.09
	$\bar{h}$	0.09	0.13	0.07	0.08	0.11	0.19	0.07	0.13
0.25	$\bar{\lambda}$	0.25	0.55	0.20	0.84	0.27	0.49	0.20	0.88
	$\bar{h}$	0.09	0.11	0.07	0.06	0.11	0.17	0.07	0.10
0.50	$\bar{\lambda}$	0.53	0.91	0.46	0.98	0.61	0.95	0.50	1.00
	$\bar{h}$	0.07	0.08	0.05	0.06	0.09	0.13	0.07	0.07
0.75	$\bar{\lambda}$	0.69	0.97	0.63	0.98	0.78	1.00	0.66	1.00
	$\bar{h}$	0.07	0.06	0.05	0.05	0.08	0.10	0.06	0.07
1.00	$\bar{\lambda}$	0.77	0.99	0.73	0.99	0.84	1.00	0.73	1.00
	$\bar{h}$	0.06	0.06	0.05	0.05	0.07	0.08	0.05	0.06

Table 8.19: Average Bandwidth and λ from Simulations (AR(1) with rho=0.20 Case)

γ		PRESS CLMM clust=5	PRESS MLMM clust=5	PRESS CLMM clust=20	PRESS MLMM clust=20	PRESS** CLMM clust=5	PRESS** MLMM clust=5	PRESS** CLMM clust=20	PRESS** MLMM clust=20
0.00	$\bar{\lambda}$	0.04	0.14	0.01	0.12	0.05	0.12	0.01	0.09
	$\bar{h}$	0.09	0.12	0.07	0.07	0.12	0.19	0.08	0.13
0.25	$\bar{\lambda}$	0.22	0.52	0.20	0.85	0.23	0.46	0.20	0.88
	$\bar{h}$	0.08	0.10	0.06	0.07	0.11	0.18	0.07	0.09
0.50	$\bar{\lambda}$	0.49	0.90	0.45	0.97	0.56	0.94	0.48	1.00
	$\bar{h}$	0.07	0.08	0.05	0.05	0.09	0.13	0.07	0.07
0.75	$\bar{\lambda}$	0.65	0.97	0.60	0.98	0.73	1.00	0.61	1.00
	$\bar{h}$	0.06	0.07	0.05	0.05	0.08	0.10	0.05	0.07
1.00	$\bar{\lambda}$	0.73	0.98	0.68	0.99	0.79	1.00	0.68	1.00
	$\bar{h}$	0.06	0.06	0.05	0.05	0.07	0.08	0.05	0.05

Table 8.20: Average Bandwidth and λ from Simulations (AR(1) with rho=0.80 Case)

γ		PRESS CLMM clust=5	PRESS MLMM clust=5	PRESS CLMM clust=20	PRESS MLMM clust=20	PRESS** CLMM clust=5	PRESS** MLMM clust=5	PRESS** CLMM clust=20	PRESS** MLMM clust=20
0.00	$\bar{\lambda}$	0.05	0.14	0.01	0.16	0.06	0.20	0.01	0.09
	$\bar{h}$	0.06	0.07	0.05	0.05	0.12	0.18	0.08	0.14
0.25	$\bar{\lambda}$	0.24	0.58	0.24	0.87	0.21	0.44	0.24	0.92
	$\bar{h}$	0.06	0.06	0.05	0.05	0.11	0.17	0.07	0.09
0.50	$\bar{\lambda}$	0.54	0.88	0.54	0.97	0.58	0.95	0.56	1.00
	$\bar{h}$	0.05	0.05	0.05	0.05	0.09	0.13	0.06	0.07
0.75	$\bar{\lambda}$	0.69	0.95	0.68	0.99	0.76	1.00	0.68	1.00
	$\bar{h}$	0.05	0.05	0.05	0.05	0.08	0.09	0.05	0.06
1.00	$\bar{\lambda}$	0.76	0.97	0.74	1.00	0.81	1.00	0.74	1.00
	$\bar{h}$	0.05	0.05	0.05	0.05	0.07	0.08	0.05	0.05

with PRESS** being used in these scenarios.

In all three tables, the estimate of λ ($\bar{\lambda}$) is performing as expected. When γ equals zero, the parametric method is correctly specified and the mixed model robust fit should equal the parametric fit. The estimate $\bar{\lambda}$ is then "close" to zero for every combination of cluster size, bandwidth selector, and covariance structure. As the number of clusters increase, the average value of λ approaches the desired value; it approaches zero for γ=0 and one for γ=1. The average value of λ is quite close to zero for the conditional local mixed model, but is larger for the marginal local mixed model.

As the amount of misspecification increases, so does the value of $\bar{\lambda}$. The average λ value is approximately one for the marginal local mixed model; however, the conditional local mixed model does not use quite as much of the local fit when γ=1. The values of $\bar{\lambda}$ for the conditional local mixed model in fact vary from 0.68 to 0.84 across the three values of ρ. This indicates that the estimate of λ, which uses cluster specific fits in its calculation, does not depend upon the local model as much as the population average does for misspecified models. The discrepancy of the $\bar{\lambda}$ values for the conditional and marginal local mixed model may explain why the model robust mixed model for the population average using CLMM performs

poorly in comparison to the population average model robust mixed model using MLMM. It appears that $\bar{\lambda}$ for population average mixed model robust regression using CLMM does not choose enough of the local fit, thereby increasing the mean square error by increasing the bias. This pattern a topic of discussion later in the chapter.

The bandwidth values tend to be larger, on average, when PRESS** is the bandwidth selector for every local method, cluster size, and covariance combination, as expected. For example, in the AR(1) ρ=0.20 case and γ=0, PRESS chose $\bar{h}$=0.088, while PRESS** chose $\bar{h}$=0.116 for the conditional local mixed model for five clusters. This is consistent with the work of Mays, Birch, and Starnes (2001).

The λ values appear to be similar across the cluster sizes over the different levels of γ. For example, the estimates of λ for PRESS, CLMM using independence are close to zero when γ=0, in the 0.20 – 0.25 range when γ=0.25, in the 0.45 – 0.55 range when γ=0.50, etc. Notice, however, it does appear that on average, the twenty cluster case selects an estimate of the mixing parameter that is slightly smaller than $\bar{\lambda}$ for five clusters.

The average bandwidth decreases as the amount of misspecification increases, as expected. This is true regardless of cluster size, local method, or covariance structure. As γ increases, the true underlying curve is less smooth, and a smaller bandwidth is needed to reflect the mean structure.

Across the different variance-covariance structures, the estimates of λ are about the same over the values of γ. For example, $\bar{\lambda}$ is 0.07, 0.04, and 0.05 for five cluster, CLMM estimation using PRESS for independence and AR(1) with a ρ of 0.20 and 0.80, respectively. There was no reason to expect that the estimates of the mixing parameter should differ for different variance-covariance structures.

For both local models, $\bar{h}$ does decrease as the number of clusters increase. Again, this is intuitive; the more information (data) is available at $\tilde{x}_0$, the less the fit is dependent on points at regressor locations far from $\tilde{x}_0$, and therefore a smaller bandwidth will suffice.

One would expect that the average bandwidth $\bar{h}$ would increase as the correlation increases. As the correlation increases, the observations with $\tilde{x}$ close to the point of estimation

$\tilde{x}_0$ are more positively correlated with the observations at $\tilde{x}_0$. A a wider "window" appears necessary to capture equivalent amounts information, as compared to the independence case. This wider window is obtained through an increase in the bandwidth.

However, just the opposite is found in our simulation study. For example, for five clusters and the marginal local mixed model using PRESS as the bandwidth selector, $\bar{h}$ decreases from 0.126 to 0.121 to 0.068 as one moves from the uncorrelated to the AR(1) cases with ρ=0.20 and 0.80, respectively. One way that we can verify the decreasing bandwidth trend is to compare $\bar{h}$ to the simulated optimal h. Recall that in Tables 8.18 – 8.20, the bandwidth is chosen to minimize the PRESS and PRESS** statistics using a search method. The simulated optimal value bandwidth for a given level of misspecification and correlation structure is found by calculating the AVEMSE for each simulated dataset over a variety of bandwidths. The AVEMSE values are then averaged over the number of datasets for each bandwidth value and the optimal bandwidth (h_{opt}) is 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 value as above. The simulated optimal mixing parameter is the value of λ corresponding to the smallest average AVEMSE value for the given γ and ρ.

The simulated optimal bandwidth and mixing parameters are given in Tables 8.21 and 8.22, respectively. Histograms of $\bar{h}$ and $\bar{\lambda}$ are shown in the next section for varying levels of γ. For the conditional local mixed model, the values of h_{opt} do not vary as the cluster size changes. It was expected that the bandwidths would stay about the same for the conditional local mixed model, as they are based on quantities that are cluster specific in nature (and is thus invariant to the number of clusters). The simulated optimal bandwidth does decrease as ρ increases; for example, $\bar{h}$ drops from 0.12 for low to no correlation to 0.10 for the high correlation case with no misspecification.

The reason behind the trend of decreasing bandwidth for increasing correlation may be due in part to the mixed model nature of the data and to the marginal covariance structure used in the local methods. For example, the conditional local mixed model assumes an

Table 8.21: Simulated Optimal Bandwidth h_{opt}

correlation	model	cluster	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 8.22: Simulated Optimal Mixing Parameter λ_{opt}

correlation	model	cluster	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 8.23: Estimates of σ^2 and σ_b^2

h	Variance	$\tilde{x}$=1	$\tilde{x}$=2	$\tilde{x}$=3	$\tilde{x}$=4	$\tilde{x}$=5	$\tilde{x}$=6	$\tilde{x}$=7	$\tilde{x}$=8	$\tilde{x}$=9	$\tilde{x}$=10
0.12	$\hat{\sigma}_b^2$	169.27	89.04	34.98	19.11	21.13	27.15	37.30	65.79	134.01	223.53
	$\hat{\sigma}^2$	0.27	0.44	0.38	0.31	0.26	0.27	0.31	0.40	0.44	0.27
0.10	$\hat{\sigma}_b^2$	184.86	91.28	35.52	20.18	22.82	28.74	38.49	66.26	136.81	240.45
	$\hat{\sigma}^2$	0.20	0.35	0.27	0.22	0.20	0.20	0.22	0.29	0.35	0.20
0.08	$\hat{\sigma}_b^2$	203.08	93.01	37.04	22.20	25.59	31.37	40.77	67.63	138.84	260.08
	$\hat{\sigma}^2$	0.12	0.23	0.17	0.15	0.13	0.13	0.14	0.18	0.22	0.12
0.06	$\hat{\sigma}_b^2$	220.75	95.35	39.78	25.39	29.64	35.27	44.36	70.16	141.36	279.01
	$\hat{\sigma}^2$	0.03	0.07	0.05	0.04	0.04	0.04	0.04	0.05	0.07	0.03

independent structure on $\mathbf{R}$ and $\mathbf{B}$, whereas the true model contains an AR(1) structure on $\mathbf{R}$ in the correlated cases. Suppose that $\tilde{\mathbf{R}} = \sigma^2\mathbf{I}$ and $\tilde{\mathbf{B}} = \sigma_b^2\mathbf{I}$ is known and fixed. The structure on $\mathbf{V}_0^*$ has $\frac{\sigma^2}{k_{ii,0}} + \sigma_b^2$ on the i^{th} diagonal and σ_b^2 on the off-diagonal cells. Thus the covariance term is actually a variance (the between-cluster variance) that is specific to the mixed model with a random intercept. A Monte-Carlo study, based on 250 runs, was conducted to see how the estimates of σ^2 and σ_b^2 behave as the bandwidth changes for given levels of ρ and γ over the design points. Table 8.23 contains the results.

Notice that $\hat{\sigma}^2$ decreases and $\hat{\sigma}_b^2$ increases, on average, as the bandwidth decreases. The covariance, $\hat{\sigma}_b^2$, is increasing. This has the effect of increasing the correlation for those points associated with regressor values close to the point of estimation $\tilde{x}_0$ and decreasing the correlation for those further away. It is possible that a smaller bandwidth is needed as ρ increases to imitate the true variance-covariance matrix.

The optimal bandwidths of Table 8.21 can be compared to the simulated bandwidths given in Tables 8.18 through 8.20.The simulated optimal bandwidths, however, do suggest that the bandwidth selectors are performing adequately and that PRESS is appropriate for MLMM and PRESS** for CLMM. For example, the simulated bandwidths for CLMM, PRESS**, s=5, and ρ=0.20 are 0.12, 0.11, 0.09, 0.08, and 0.07 for γ=0, 0.25, 0.50, 0.75, and 1. The optimal bandwidths for CLMM, s=5 and ρ=0.20 are 0.12, 0.11, 0.10, 0.09, and

0.08 for γ=0, 0.25, 0.50, 0.75, and 1. There are a few exceptions. In the cases where γ=0 for independence and AR(1) with ρ=0.20, the simulated optimal bandwidth is much larger than the one given by PRESS for the marginal local mixed model. If you were to look at the profile of the simulated AVEMSE values over bandwidth for these cases, a local minimum does occur around 0.09. Thus, the bandwidth selectors, when using the search method, may be finding a local rather than the global minimum.

There are some similarities among the average estimates of the mixing parameter (found by the formula using the simulated data conditioned on h) and the simulated optimal mixing parameters conditioned on h (Table 8.22). Both λ_{opt} and $\bar{\lambda}$ are "close" to zero and increase to one as γ increases. For the marginal local mixed model with twenty clusters, λ_{opt} and $\bar{\lambda}$ are very close, except in the case where γ=0. In many cases, λ_{opt} and $\bar{\lambda}$ are within 0.10 of each other, with the majority of the differences between λ_{opt} and $\bar{\lambda}$ less than or equal to 0.05.

However, there are some instances where λ_{opt} and $\bar{\lambda}$ are strikingly different. For example, in the γ=0 case for the marginal local mixed model with five clusters, λ_{opt} is zero, while $\bar{\lambda}$=0.181 when using PRESS as the bandwidth selector. For the conditional local mixed model with twenty clusters, λ_{opt} is 0.91 and $\bar{\lambda}$=0.655 when using PRESS** as the bandwidth selector in the γ=0.75 case. In fact, we can generalize the pattern seen here. It appears that the formula for $\hat{\lambda}$ works well in every case for the marginal local mixed model except when γ=0. There is a tendency for $\bar{\lambda}$ to be too large when γ=0, as in the example given above. The mixing parameter estimate is using too much of the nonparametric fit.

The opposite is true for the conditional local mixed model. The formulas work well for small amounts of misspecification; however the simulated optimal mixing parameter and the estimate of the mixing parameter found by the formula differ significantly for γ=0.50, 0.75, and 1. In fact, $\bar{\lambda}$, on average, is too small for large misspecification. This is equivalent to saying that for misspecified models, the conditional local mixed model does not use enough of the nonparametric fit; it relies too heavily on the parametric part.

Two alterations to the mixing parameter formula could be considered. One is to

incorporate variance-covariance matrices in the formulas for $\hat{\lambda}$, or

$$\hat{\lambda}^{\mathsf{M}} = \frac{(\hat{\mathbf{Y}}^{\mathsf{NP}}_{\mathsf{i},\text{-}\mathsf{i}} - \hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{i},\text{-}\mathsf{i}})'\mathbf{V}^{\text{-}1}(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{\mathsf{P}})}{(\hat{\mathbf{Y}}^{\mathsf{NP}} - \hat{\mathbf{Y}}^{\mathsf{P}})'\mathbf{V}^{\text{-}1}(\hat{\mathbf{Y}}^{\mathsf{NP}} - \hat{\mathbf{Y}}^{\mathsf{P}})} \tag{8.3}$$

and

$$\hat{\lambda}^{\mathsf{C}} = \frac{(\hat{\mathbf{Y}}^{\mathsf{NP}}_{\mathsf{i},\text{-}\mathsf{i}} - \hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{i},\text{-}\mathsf{i}})'\mathbf{R}^{\text{-}1}(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{\mathsf{P}})}{(\hat{\mathbf{Y}}^{\mathsf{NP}} - \hat{\mathbf{Y}}^{\mathsf{P}})'\mathbf{R}^{\text{-}1}(\hat{\mathbf{Y}}^{\mathsf{NP}} - \hat{\mathbf{Y}}^{\mathsf{P}})}, \tag{8.4}$$

where the fits in $\hat{\lambda}^{\mathsf{M}}$ are population average and the fits in $\hat{\lambda}^{\mathsf{C}}$ are cluster specific. Another option for the mixing parameter formula is to change the delete cluster portion of the formula for the conditional local mixed model. Because $\hat{\mathbf{b}}_{\mathsf{i},\text{-}\mathsf{i}} = \mathbf{0}$, the delete cluster fits in the formula for the cluster specific model are really the population average delete cluster fits. It may be that an adjustment is needed to those delete cluster fits, which in turn may result in estimates of λ that are close to the simulated optimal mixing parameter, or to use the fits from the entire dataset in place of the delete cluster fits. These issues are left for further study.

8.2.4 Estimates of the Distributions of $\hat{\mathbf{h}}$ and $\hat{\lambda}$

The distributions of the bandwidth and mixing parameter for a given level of misspecification are also of interest. A crude estimate of the distribution may be found using a histogram. For the five cluster, independence case, the estimates of the mixing parameter and bandwidth for each simulated data set for each level of γ were obtained, and histograms were constructed. Each trellis plot is labelled to identify the local method and the bandwidth selector used.

Figures 8.10 and 8.11 are histograms for $\bar{\mathrm{h}}$ and $\bar{\lambda}$, respectively, for γ=0; Figures 8.12 and 8.13 are plots for $\bar{\mathrm{h}}$ and $\bar{\lambda}$ when γ=0.25, and so forth. In Figures 8.17 and 8.19, there is no plot for $\bar{\lambda}$ in the marginal local mixed model using PRESS– virtually all of the estimates were one in those cases.

In the histograms for $\bar{\mathrm{h}}$, it is obvious that the variance decreases as γ increases. For instance, the bandwidths range from 0.05 to 0.24 for the marginal local mixed model (using PRESS) for γ=0; the bandwidths vary from 0.05 to 0.10 for MLMM (using PRESS) for

Figure 8.10: Histogram of h for γ=0

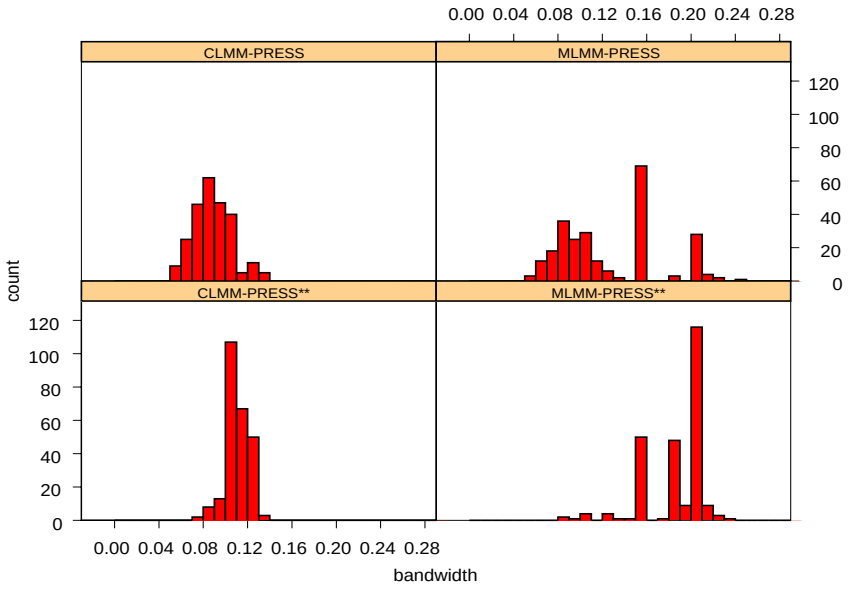


Figure 8.11: Histogram of λ for γ=0

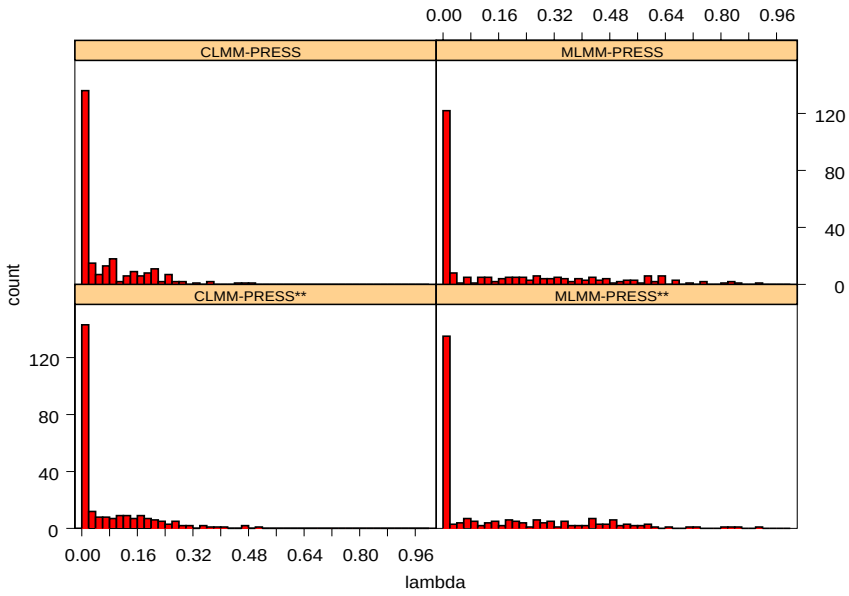


Figure 8.12: Histogram of h for γ=0.25

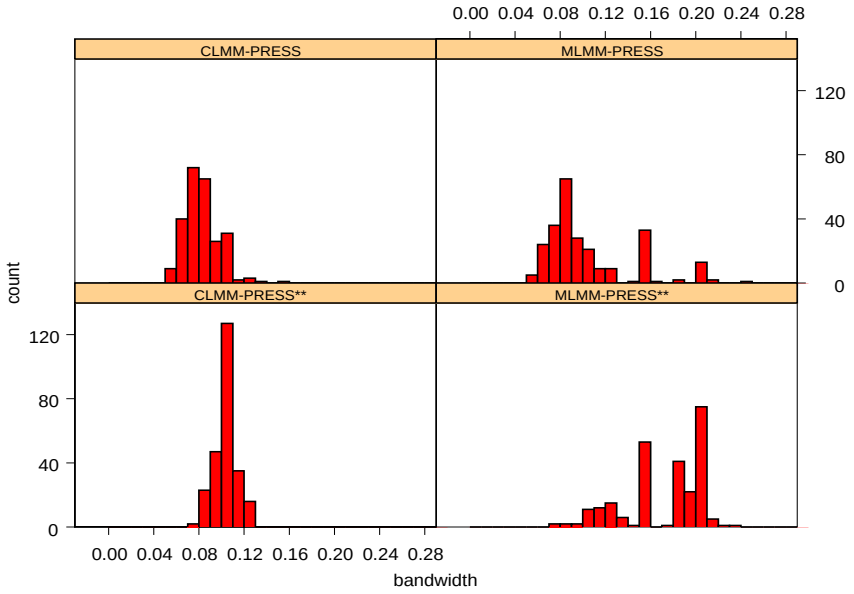


Figure 8.13: Histogram of λ for γ=0.25

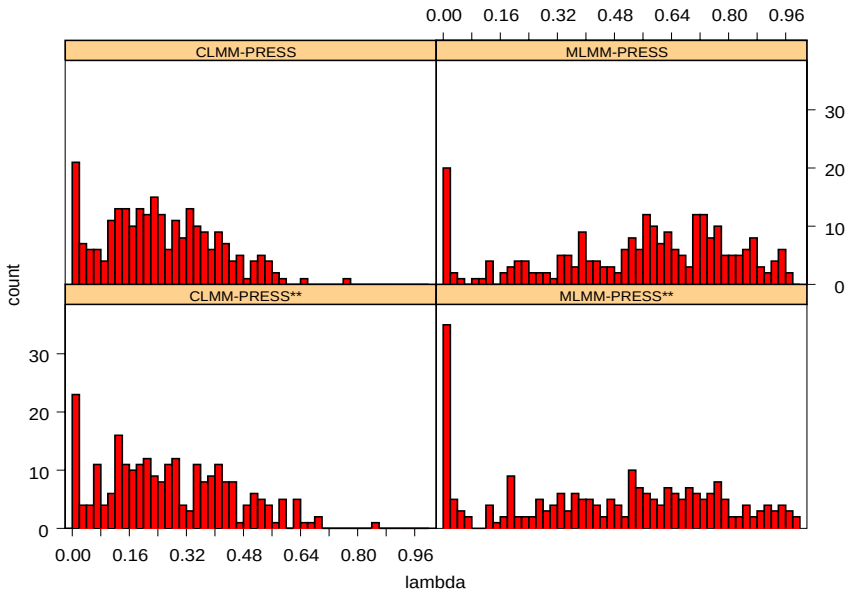


Figure 8.14: Histogram of h for γ=0.50

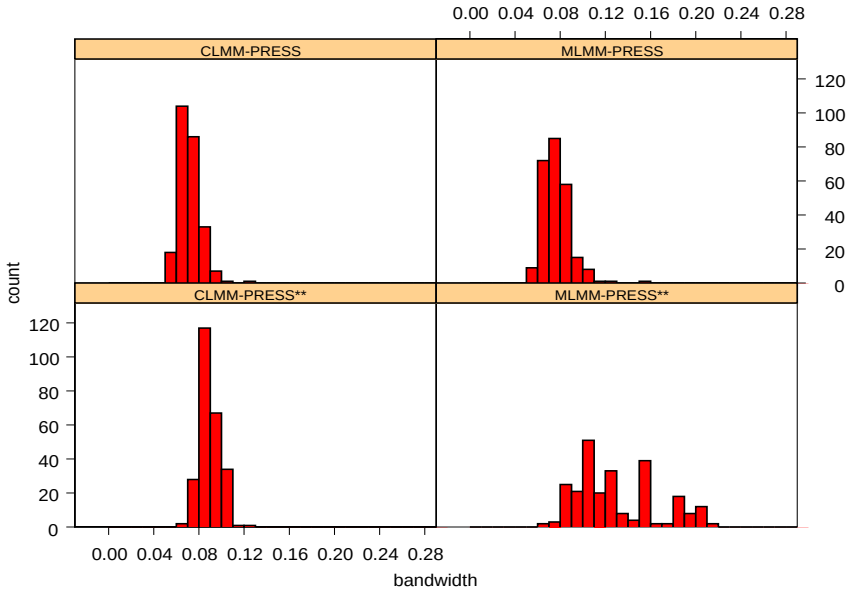


Figure 8.15: Histogram of λ for γ=0.50

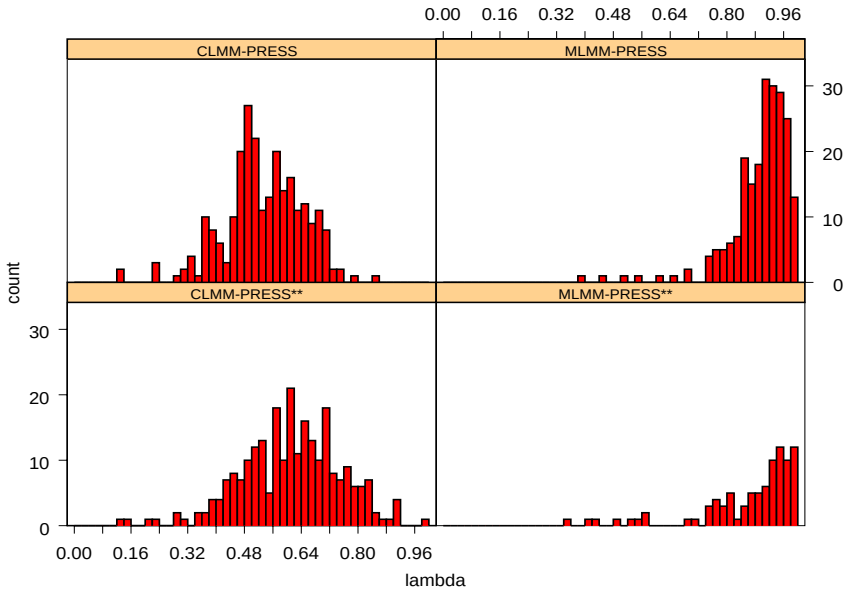


Figure 8.16: Histogram of h for γ=0.75

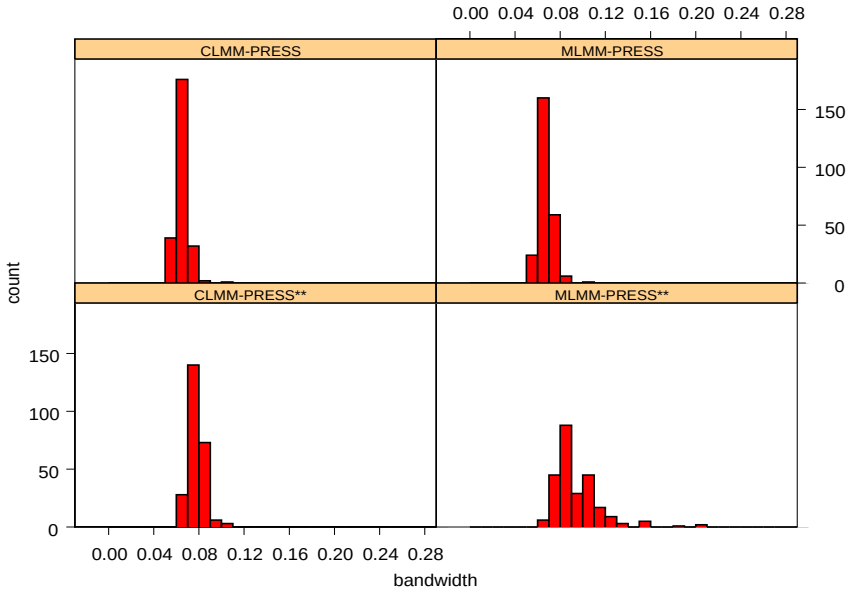


Figure 8.17: Histogram of λ for γ=0.75

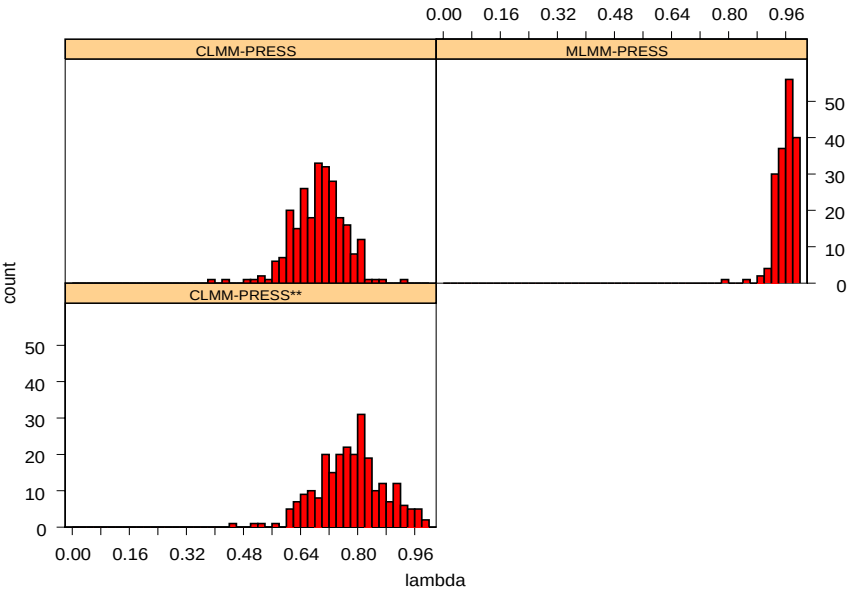


Figure 8.18: Histogram of h for γ=1

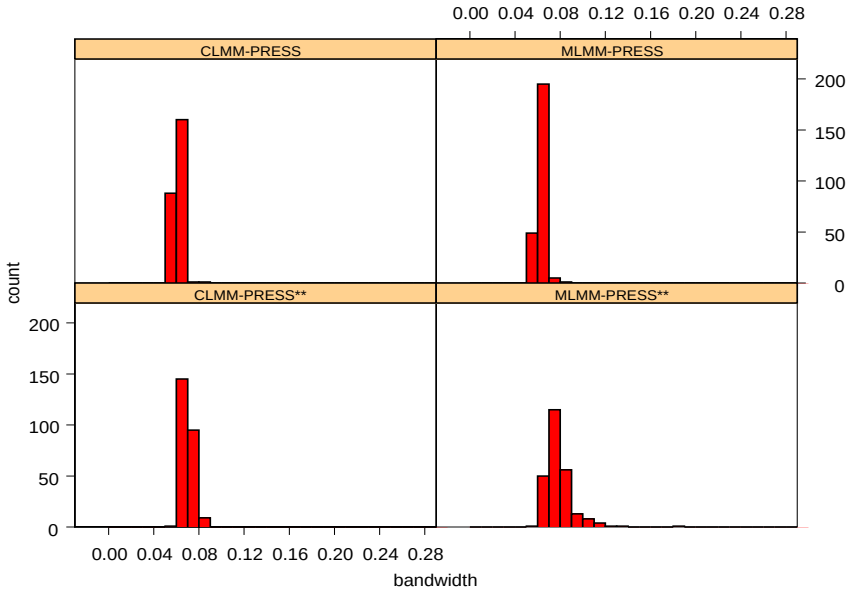
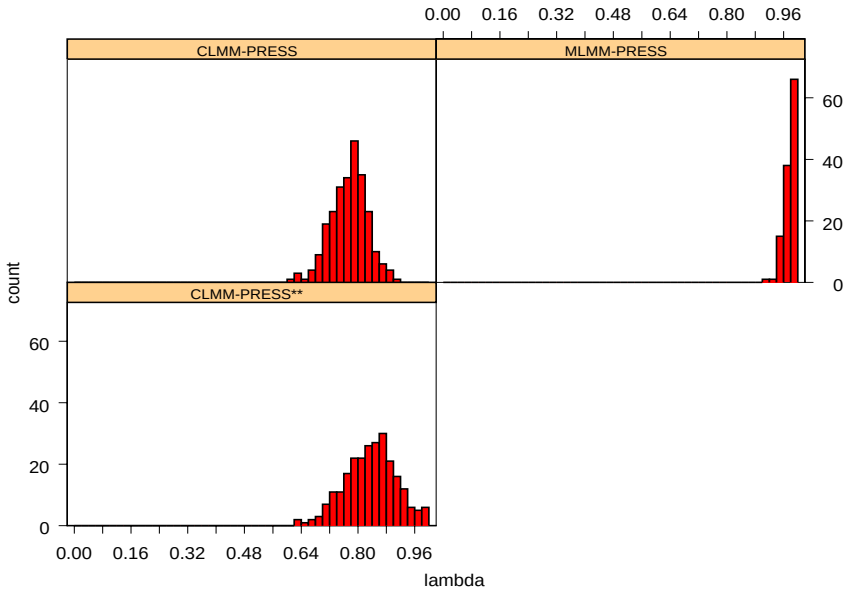


Figure 8.19: Histogram of λ for γ=1



the largest amount of misspecification. The same pattern is observed for the conditional local mixed model. In comparison between the two local methods, there tends to be more variation in $\bar{h}$ for the marginal local mixed model than for the conditional local mixed model.

Notice that there tends to be more variation in $\bar{h}$ for PRESS rather than PRESS**. Recall that PRESS** and PRESS are the bandwidth selectors for the conditional and marginal local mixed models, respectively, which again suggests more variation in MLMM. This trend was noted in Clark (2002); he found some cases where $\bar{h}$ chosen by PRESS was closer to $\bar{h}_{\mathsf{opt}}$ than $\bar{h}$ chosen by PRESS**, yet the AVEMSE was smallest when using $\bar{h}$ from PRESS**.

As for the estimates of the mixing parameter, the distributions are skewed when γ=0 because $\bar{\lambda}$ should be zero. For low to moderate misspecification, there is tremendous variability; in the marginal local mixed model, the estimates take on values all across the range when γ=0.25. The distribution for the marginal local mixed model is again skewed for large γ, since the value of γ should be 1. This is not the case for the conditional local mixed model, in which the distribution appears to be bell-shaped for large misspecification. It is interesting to notice that the mixing parameter estimates when γ=1 for the conditional local mixed model (using PRESS**) are almost exclusively less than 0.96, the value given by λ_{opt}.

8.3 Summary

In Chapter 8, a simulation study was performed to compare the bandwidths selectors, the estimates of the mixing parameters, and the approximate integrated mean square errors of the parametric, local, and mixed model robust models. Both population average and cluster specific inference were performed for varying amounts of model misspecification, correlation structures, cluster sizes, and bandwidth selectors. Mixed model robust regression for cluster specific and population average using MLMM were competitive with the parametric and local models over all levels of misspecification. For small to moderate misspecification, MMRR was the consistently the "winner", often having the smaller INTMSE values than the parametric and local methods.

The average bandwidth and mixing parameter estimates were also obtained for the different combinations. These values were compared to the simulated optimal bandwidth selectors to determine if the bandwidth selectors and mixing parameters formulas were performing as desired.

Chapter 9 will give a brief summary of the main results obtained in this work. Future work in the areas of local mixed models and mixed model robust regression will also be presented in the next chapter.

Chapter 9

Summary, Conclusions, and Outlook on Future Research

This chapter summarizes the important local and model robust mixed model results from the previous chapters. Because the work presented here is ongoing research, directions for future research pertaining to the local and model robust models will also be given.

9.1 Conclusions

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 or local cubic mixed model with a random intercept at each $\tilde{x}_0$ value will suffice.

PRESS should be used as the bandwidth selector for population average estimation. Our simulation study shows that the approximate INTMSE values are smaller for the population average when using PRESS over PRESS* and PRESS**. Conversely, PRESS** is the bandwidth selector of choice for cluster specific prediction because the cluster specific INTMSE values are smallest for PRESS**. These conclusion 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.

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 will 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. There was a tendency for the marginal local mixed model to include too much of the nonparametric fit when the parametric model was correctly specified, however.

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 correlated data is not easy to work with. Our intuitions, often based upon prior work with independent data, were 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.

9.2 Future Research

The work presented here is a start in applying model robust ideas to linear mixed models. Much additional work needs to be done. The following are suggestions for future work in these areas.

9.2.1 Bandwidth Selectors and Estimates of λ

Three bandwidth selectors were developed for the mixed model in Chapter 6. Although the simulated optimal bandwidths are close to those chosen by PRESS and PRESS** in the simulation, more work on bandwidth selection is needed. Improvements could possibly be made, perhaps to the penalty terms or the delete cluster fits. Alternate selectors could be considered entirely, such as plug-in methods or rule of thumb bandwidth selectors. These selectors, although less refined than PRESS and PRESS**, would be computationally simpler and would make for faster results.

The mixing parameter can also be refined. Suggestions on how to modify the formula are given in Chapter 8, and include the incorporation of a variance-covariance matrix, changes

to the delete cluster fits to have a delete cluster BLUP not equal to zero, or to use the fit from the entire dataset rather than the delete cluster fits in the estimation of λ.

Additional work may allow λ to vary by cluster, rather than having one λ for the entire data set. Currently, the mixing parameter is estimated conditioned on the value of the bandwidth. Therefore, joint estimation of the bandwidth and mixing parameter would also be of interest.

9.2.2 Asymptotic Theory

Convergence rates for the theoretically optimal mixing parameter were given in Chapter 7. However, convergence rates for the data driven theoretically optimal mixing parameter have yet to be developed. Whereas the asymptotic theory for the theoretically optimal mixing parameter is just an extension of Burman and Chaudhuri (1992) and Starnes (1999), it is a different matter for the data driven theoretically optimal mixing parameter because the proof given in the references above use Whittle's inequality. This inequality is based upon independent data. Our data is of course correlated, so future work on the asymptotic theory for the data driven estimate must take into account the fact that the data are correlated.

Asymptotic theory for the local methods must be considered. Because the local methods are new, the nonparametric asymptotic theory has yet to be developed. In Chapter 7, we have generically denoted the convergence rate of the nonparametric estimate as γ_s. Convergence rates of the nonparametric estimates for both the local models need to be developed, as do the asymptotic results for the cluster specicific model.

9.2.3 Multiple Regressors and Diagnostics

The work presented here has been for one regressor. Of course, there are many instances in which the researcher is interested in using more than one regressor. Virtually all of the past work in model robust regression has been in the single regressor case, but extensions to multiple regression are vital. The difficulty that arises is the extension of the local methods to nonparametric multiple regression. The method would be to fit locally weighted mixed

model multiple regressions with multivariate kernels.

Confidence intervals for the mean response and prediction intervals for a future observation are also desired, along with multicollinearity and outlier diagnostics. Work on diagnostics by Hurtado-Rodriguez (1993) and Hilden-Minton (1995) for mixed models has been completed, but extensions of their work for mixed model robust regression are needed.

9.2.4 Messy Data

The work presented here considered "nice" datasets– those with no missing observations, no data sparseness, equally spaced values of the regressor, and clusters having the same regressor values (balanced, rectangular data). Messy datasets, as opposed to nice datasets are often seen in the biological and social sciences. It is expected that our local and mixed model robust regression methods can accomodate reasonable departures from "perfect" datasets. For example, the models that we have defined can accomodate unequal spacing and different sample sizes per cluster. However, as with most nonparametric procedures, sparse data and datasets with many missing observations pose a problem. It is expected that the local mixed model (and hence the model robust methods) will not be an exception. Most likely, the local and model robust mixed models will have bias difficulties, will fail to converge, etc. It would be interesting to see how our methods perform with such data.

9.2.5 Alternative Misspecification

The model misspecification considered in this work is that of the model matrices. There are a number of other ways in which a model can be incorrect, including the variance-covariance structure or effect classification. That is, what would happen to our mixed model robust fits if the variance-covariance structure specified in the parametric estimation was incorrect?

Also of interest is the effect of having a random term in the misspecification term in the model. Our simulation did not considered the case where f (the misspecification term) included a random effect; additional simulation studies with such a misspecification term would be insightful.

9.2.6 The MRR2 Estimate for the Mixed Model

The model robust mixed model in this work is based upon the MRR1 estimate of Einsporn and Birch (1993). The MRR2 estimate of Mays, Birch, and Starnes (2001), combines a parametric fit with a fraction of the nonparametric fit of the residuals. An MRR2 extension to the mixed model model would be of interest. Much of the work for mixed model robust regression based upon MRR2 would simply be a generalization of the work of Mays, Birch, and Starnes (2001), coupled with the results given in this work. Mays, Birch, and Starnes (2001) found that the MMR2 estimate performed slightly better than the MRR1 estimate, although the two methods were competitive. It would be interesting to compare the mean square error values of the two mixed model robust methods.

9.2.7 Nonnormal Data

Our research assumes that the random effects and random errors are normally distributed. Although normality is a common assumption, it would be interesting to extend this work to the generalized linear mixed model (GLMM). This work would be a combination of the research presented here (for the linear mixed model) and the work of Clark (2002), who worked with generalized estimating equations in the fixed effects model.

Glossary

	Acronym or Term	Description
Symbols		
	$\aleph$	X-space
	γ	Misspecification Parameter
	λ	Mixing Parameter
	ρ	Correlation Parameter
A		
	ASE	Average Squared Error
	AVEMSE	Average Mean Square Error
B		
	BLUE	Best Linear Unbiased Estimator
	BLUP	Best Linear Unbiased Predictor
C		
	CISE	Conditional Integrated Squared Error
	CLMM (or C)	Conditional Local Mixed Model
	CS	Cluster Specific
	CV	Cross-Validation
E		
	EGLS	Estimated Generalized Least Squares
G		
	GEE	Generalized Estimating Equations
	GLMM	Generalized Linear Mixed Model
	GLS	Generalized Least Squares

H

h	Bandwidth
H	smoother matrix

I

ISE	Integrated Squared Error
INTMSE	Integrated Mean Square Error

K

Ker	Kernel Regression

L

LLR	Local Linear Regression
LPR	Local Polynomial Regression

M

ML	Maximum Likelihood
MLMM (or M)	Marginal Local Mixed Model
MMRR,C (or MMRR,CLMM)	Mixed Model Robust Regression using the Conditional Local Mixed Model
MMRR,M (or MMRR,MLMM)	Mixed Model Robust Regression using the Marginal Local Mixed Model
MMRR	Mixed Model Robust Regression
MRR1	Model Robust Regression 1
MRR2	Model Robust Regression 2
MSE	Mean Square Error

N

n	Sample Size
n_i	Sample Size for the i^{th} sample
NP	Nonparametric (Local)
NPMLE	Nonparametric Maximum Likelihood Estimation

O

OLS	Ordinary Least Squares
opt	Optimal (as in optimal bandwidth or optimal mixing parameter)

P

P	Parametric
PA	Population Average
PLR	Partial Linear Regression
PR	Prediction Recursion Method
PRESS	Prediction Error Sum of Squares (used as a bandwidth selector)
PRESS*	Penalized PRESS statistic (used as a bandwidth selector)
PRESS**	Penalized PRESS statistic (used as a bandwidth selector)

R

REML	Restricted Maximum Likelihood

S

s	Number of Clusters
SNPMLE	Smooth Nonparametric Maximum Likelihood Estimation

U

UMVU	Uniform Minimum Variance Unbiased (Estimator)

Appendix A

Wind Speed Data Set

Week	BEL	BIR	CLA	CLO	DUB	KIL	MAL	MUL	ROS	RPT	SHA	VAL
1	13.29	6.78	8.20	9.07	11.03	6.07	17.24	8.57	12.20	13.22	9.78	10.89
2	14.97	8.45	9.83	10.23	12.32	7.58	18.25	9.75	14.10	15.75	12.34	13.67
3	15.21	8.75	10.13	10.50	12.08	7.56	18.39	9.99	13.48	15.38	12.60	13.85
4	14.10	7.83	9.24	9.86	11.09	7.01	17.56	9.23	12.42	14.55	11.49	12.73
5	15.19	8.44	10.10	10.50	12.49	7.74	18.19	10.27	13.92	15.02	12.23	13.35
6	13.43	6.73	8.53	8.66	10.66	6.19	16.94	8.83	11.75	12.97	10.64	11.27
7	13.67	7.89	9.56	9.66	11.34	7.12	17.52	9.40	13.25	14.45	12.02	12.58
8	13.58	8.00	9.77	9.93	11.51	7.36	17.23	9.61	13.57	13.47	11.69	11.92
9	13.79	7.24	9.17	9.47	10.36	6.49	16.46	8.58	12.48	12.81	11.11	11.78
10	13.08	7.87	9.18	9.88	10.48	7.23	16.23	9.21	13.05	13.59	11.35	11.92
11	13.73	8.18	10.13	10.41	11.74	7.41	17.34	10.00	13.20	13.37	11.79	11.60
12	13.35	7.50	9.31	9.74	11.16	6.97	16.93	9.39	12.24	12.43	10.81	10.56
13	15.42	8.77	10.38	10.85	12.21	7.93	17.63	10.85	12.28	13.65	12.86	12.15
14	13.50	7.94	9.66	9.97	11.50	7.62	16.82	9.84	713.96	13.70	11.18	10.72
15	13.02	7.69	9.38	9.38	10.80	7.19	15.66	9.32	12.45	13.13	11.00	10.62
16	12.09	7.30	8.44	8.98	9.73	6.56	14.58	8.65	11.37	12.05	10.40	10.30
17	12.02	7.089	8.36	8.57	9.38	6.51	12.91	8.27	11.77	11.56	10.21	10.08
18	12.81	6.96	8.35	8.60	9.01	6.27	14.27	7.85	10.64	12.11	10.55	10.59
19	14.12	8.04	9.50	9.63	10.03	7.41	14.86	9.11	12.95	13.68	11.44	11.78
20	12.44	6.40	8.22	8.06	8.33	5.92	13.33	7.61	11.08	11.02	9.69	9.55
21	12.97	7.22	8.43	8.75	8.87	6.42	14.30	8.32	11.94	11.65	10.38	10.03
22	11.00	5.66	7.48	7.20	7.23	5.19	11.67	6.99	10.33	9.81	8.70	8.45
23	11.90	5.98	7.60	7.20	7.22	5.19	12.20	7.09	9.78	9.83	8.88	8.73
24	11.29	5.71	7.09	6.99	7.10	5.20	12.02	6.78	10.30	9.81	8.41	8.35
25	13.24	7.27	8.65	8.37	9.04	6.34	13.80	8.44	10.84	11.67	10.81	10.03

Adapted from 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.

Week	BEL	BIR	CLA	CLO	DUB	KIL	MAL	MUL	ROS	RPT	SHA	VAL
26	12.91	6.85	8.40	8.43	8.85	5.92	13.57	8.35	10.44	10.47	10.21	8.74
27	11.59	5.88	7.17	7.59	7.88	5.40	13.11	7.71	8.76	10.04	8.88	8.16
28	11.99	6.33	7.59	7.38	8.20	5.64	13.25	7.96	9.76	10.11	9.93	8.71
29	11.70	5.96	7.33	7.48	7.94	5.44	12.52	7.43	9.54	10.08	9.28	8.14
30	11.40	5.75	6.97	7.04	7.77	5.21	12.51	7.21	9.38	9.65	9.09	8.09
31	11.03	5.60	6.67	6.62	7.22	5.16	12.25	6.91	9.38	9.87	8.82	8.36
32	11.19	6.04	6.80	6.81	7.45	5.46	12.29	7.34	9.60	10.06	9.22	8.50
33	10.35	5.70	6.58	6.81	7.74	5.12	12.05	7.15	9.77	10.17	8.54	7.98
34	11.62	6.14	7.30	7.46	8.03	5.37	12.91	7.56	10.37	10.57	9.10	8.97
35	11.83	6.03	7.08	7.36	8.19	5.37	13.54	7.27	10.66	10.53	9.00	8.68
36	11.61	6.10	6.88	6.97	8.16	5.32	13.86	7.14	9.61	10.57	9.14	9.20
37	12.87	6.86	7.91	7.96	8.93	5.84	15.10	7.90	11.48	11.65	9.82	9.93
38	11.48	5.61	6.69	6.55	7.33	4.58	13.05	6.44	10.08	9.83	8.35	8.72
39	14.50	7.51	9.27	8.94	9.66	6.37	16.58	8.65	11.45	13.31	11.53	11.99
40	13.87	7.16	8.87	8.83	9.74	6.34	16.26	8.67	11.78	12.96	10.83	11.10
41	13.08	6.58	8.13	8.12	8.46	5.46	15.12	7.73	11.26	12.09	9.81	10.47
42	14.21	7.38	8.76	8.82	9.68	6.16	17.14	8.37	11.32	12.81	10.83	11.07
43	15.15	7.58	9.17	9.41	9.73	6.20	17.64	8.65	11.37	12.67	10.90	11.36
44	15.23	7.56	9.24	9.75	10.28	6.74	18.68	9.02	12.51	13.63	10.90	11.83
45	13.66	7.15	8.44	8.80	10.33	5.98	17.79	8.51	12.40	12.91	10.39	11.57
46	14.54	8.11	9.29	9.74	11.89	7.18	19.40	9.33	13.29	14.50	11.77	12.88
47	12.58	6.30	7.70	8.35	10.30	5.87	17.15	8.00	12.17	12.37	9.40	10.32
48	13.91	6.97	8.00	8.54	10.88	5.84	17.56	8.47	11.19	13.19	10.55	11.78
49	14.68	8.39	9.32	9.86	12.79	7.03	19.69	9.83	12.82	14.41	11.63	12.67
50	15.02	8.47	9.64	9.89	11.82	7.09	18.15	9.67	12.62	14.81	11.93	13.02
51	13.33	7.07	8.35	8.72	10.47	5.92	17.34	8.33	12.44	13.30	10.32	11.38
52	13.70	7.64	9.15	9.76	11.62	6.98	19.16	9.56	14.03	14.82	11.04	12.01
53	14.80	8.16	10.04	10.47	13.97	7.68	20.70	10.55	16.24	15.01	11.34	12.53

Appendix B

B.1 Derivation of Mixed Model Equations for the Parametric Linear Mixed Model using Henderson's joint likelihood of b and $\boldsymbol{\epsilon}$.

We take the derivatives of log f with respect to $\boldsymbol{\beta}$ and $\mathbf{b}$.

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\epsilon})}{\partial \mathbf{b}} = \frac{\partial(-\frac{1}{2}[\mathbf{b}'\mathbf{B}\mathbf{b} + (\mathbf{Y}' - \boldsymbol{\beta}'\mathbf{X}' - \mathbf{b}'\mathbf{Z}')\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})])}{\partial \mathbf{b}}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\epsilon})}{\partial \boldsymbol{\beta}} = \frac{\partial(-\frac{1}{2}[\mathbf{b}'\mathbf{B}\mathbf{b} + (\mathbf{Y}' - \boldsymbol{\beta}'\mathbf{X}' - \mathbf{b}'\mathbf{Z}')\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})])}{\partial \boldsymbol{\beta}}$$

Setting the derivatives equal to zero, we obtain

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\epsilon})}{\partial \mathbf{b}} = -\frac{1}{2}(2\mathbf{B}^{-1}\mathbf{b} - 2\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}) = \mathbf{0}$$

$$\mathbf{B}^{-1}\mathbf{b} - \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} = \mathbf{0}$$

$$(\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})\mathbf{b} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\epsilon})}{\partial \boldsymbol{\beta}} = -\frac{1}{2}(-2\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} + 2\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + 2\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}) = \mathbf{0}$$

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} = \mathbf{0}$$

$$\mathbf{X}'\mathbf{R}^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) = \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y}.$$

We can rewrite this in matrix form to obtain the mixed model equations

$$\begin{bmatrix} \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & (\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}) \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}.$$

B.2 Derivation of $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ for the Parametric Linear Mixed Model using Henderson's joint likelihood of $\mathbf{b}$ and $\boldsymbol{\epsilon}$.

We solve the first equation for $\hat{\mathbf{b}}$

$$\hat{\mathbf{b}} = (\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Now,

$$\begin{aligned}
&(\mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B})(\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}) \\
&= \mathbf{B}\mathbf{B}^{-1} + \mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{B}^{-1} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \\
&= \mathbf{I} + \mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \\
&= \mathbf{I} + \mathbf{B}\mathbf{Z}'(\mathbf{R}^{-1} - \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1})\mathbf{Z} \\
&= \mathbf{I} + \mathbf{B}\mathbf{Z}'(\mathbf{R}^{-1} - \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1} - \mathbf{V}^{-1}\mathbf{R}\mathbf{R}^{-1} + \mathbf{V}^{-1}\mathbf{R}\mathbf{R}^{-1})\mathbf{Z} \\
&= \mathbf{I} + \mathbf{B}\mathbf{Z}'(\mathbf{R}^{-1} - \mathbf{V}^{-1} - \mathbf{V}^{-1}(\mathbf{Z}\mathbf{B}\mathbf{Z}' + \mathbf{R})\mathbf{R}^{-1} + \mathbf{V}^{-1})\mathbf{Z} \\
&= \mathbf{I} + \mathbf{B}\mathbf{Z}'(\mathbf{R}^{-1} - \mathbf{V}^{-1}\mathbf{V}\mathbf{R}^{-1})\mathbf{Z} \\
&= \mathbf{I} + \mathbf{B}\mathbf{Z}'(\mathbf{R}^{-1} - \mathbf{R}^{-1})\mathbf{Z} = \mathbf{I}.
\end{aligned}$$

So,

$$(\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})^{-1} = (\mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}).$$

Thus we have

$$\begin{aligned}
\hat{\mathbf{b}} =&(\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= (\mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B})\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} - \mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} + \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{B}\mathbf{Z}'(\mathbf{I} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}')\mathbf{R}^{-1}\mathbf{Y} - \mathbf{B}\mathbf{Z}'(\mathbf{I} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}')\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}.
\end{aligned}$$

Now,

$$(\mathbf{I} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}') = (\mathbf{I} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}' - \mathbf{V}^{-1}\mathbf{R} + \mathbf{V}^{-1}\mathbf{R})$$

$$= (\mathbf{I} - \mathbf{V}^{-1}(\mathbf{Z}\mathbf{B}\mathbf{Z}' + \mathbf{R}) + \mathbf{V}^{-1}\mathbf{R}) = (\mathbf{I} - \mathbf{V}^{-1}\mathbf{V} + \mathbf{V}^{-1}\mathbf{R}) = \mathbf{V}^{-1}\mathbf{R}.$$

Substituting in, we obtain the predictor

$$\hat{\mathbf{b}} = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{Y} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

$$= \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Plugging the predictor of $\mathbf{b}$ into the second equation, we obtain

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} = \mathbf{X}'\mathbf{R}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}})$$

$$= \mathbf{X}'\mathbf{R}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}(\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}))$$

$$= \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$
$$+ \mathbf{X}'\mathbf{R}^{-1}\mathbf{R}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \mathbf{X}'\mathbf{R}^{-1}\mathbf{R}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}(\mathbf{Z}\mathbf{B}\mathbf{Z}' + \mathbf{R})\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$
$$- \mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{V}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} - \mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Therefore,

$$\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

Appendix C

C.1 Derivation of Mixed Model Equations for the Conditional Local Mixed Model using Henderson's joint likelihood of $\mathbf{b}_0$ and $\boldsymbol{\epsilon}_0$.

We take the derivatives of log f with respect to $\boldsymbol{\beta}_0$ and $\mathbf{b}_0$.

$$\frac{\partial \log \text{ f}(\mathbf{b}_0, \boldsymbol{\epsilon}_0)}{\partial \mathbf{b}_0} = \frac{\partial(-\frac{1}{2}[\mathbf{b}_0'\tilde{\mathbf{B}}\mathbf{b}_0 + (\mathbf{Y}' - \boldsymbol{\beta}_0'\tilde{\mathbf{X}}' - \mathbf{b}_0'\tilde{\mathbf{Z}}')\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \tilde{\mathbf{Z}}\mathbf{b}_0)])}{\partial \mathbf{b}_0}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}_0, \boldsymbol{\epsilon}_0)}{\partial \boldsymbol{\beta}_0} = \frac{\partial(-\frac{1}{2}[\mathbf{b}_0'\tilde{\mathbf{B}}\mathbf{b}_0 + (\mathbf{Y}' - \boldsymbol{\beta}_0'\tilde{\mathbf{X}}' - \mathbf{b}_0'\tilde{\mathbf{Z}}')\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \tilde{\mathbf{Z}}\mathbf{b}_0)])}{\partial \boldsymbol{\beta}_0}$$

Setting the derivatives equal to zero, we obtain

$$\frac{\partial \log \text{ f}(\mathbf{b}_0, \boldsymbol{\epsilon}_0)}{\partial \mathbf{b}_0} = -\frac{1}{2}(2\tilde{\mathbf{B}}^{-1}\mathbf{b}_0 - 2\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0) + 2\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\tilde{\mathbf{B}}^{-1}\mathbf{b}_0 - \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0) + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$(\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}})\mathbf{b}_0 + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 = \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}_0, \boldsymbol{\epsilon}_0)}{\partial \boldsymbol{\beta}_0} = -\frac{1}{2}(-2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \tilde{\mathbf{Z}}\mathbf{b}_0) = \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}.$$

We can rewrite this in matrix form to obtain the conditional local mixed model equations

$$\begin{bmatrix} \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & (\tilde{\mathbf{B}}^{-1}+\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}) \\ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\mathbf{b}}_0 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} \\ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} \end{bmatrix}.$$

C.2 Derivation of $\hat{\boldsymbol{\beta}}_0$ and $\hat{\mathbf{b}}_0$ for the Conditional Local Mixed Model using Henderson's joint likelihood of $\mathbf{b}_0$ and $\boldsymbol{\epsilon}_0$.

We solve the first equation for $\hat{\mathbf{b}}_0$

$$\hat{\mathbf{b}}_0 = (\tilde{\mathbf{B}}^{-1}+\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0).$$

Now,

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

So,

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

Thus we have

$$\begin{aligned}
\hat{\mathbf{b}}_0 &= (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\
&= (\tilde{\mathbf{B}} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}})\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\
&= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 \\
&\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 \\
&= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0.
\end{aligned}$$

Now,

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

Substituting in, we obtain the predictor

$$\begin{aligned}
\hat{\mathbf{b}}_0^{\mathsf{C}} &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 \\
&= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 = \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0).
\end{aligned}$$

Plugging the predictor of $\mathbf{b}_0$ into the second equation, we obtain

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

Therefore,

$$\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) = \mathbf{0}$$

$$\hat{\boldsymbol{\beta}}_0^{\mathsf{C}} = (\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{Y}.$$

Appendix D

D.1 Derivation of Mixed Model Equations for the Marginal Local Mixed Model using Henderson's joint likelihood of $\mathbf{b}_0$ and $\boldsymbol{\epsilon}_0$.

We take the derivatives of log f with respect to $\boldsymbol{\beta}_0$ and $\mathbf{b}_0$.

$$\frac{\partial \log \text{ f}(\mathbf{b}_0, \boldsymbol{\epsilon}_0)}{\partial \mathbf{b}_0} = \frac{\partial(-\frac{1}{2}[\mathbf{b}_0'\tilde{\mathbf{B}}\mathbf{b}_0 + (\mathbf{Y}' - \boldsymbol{\beta}_0'\tilde{\mathbf{X}}' - \mathbf{b}_0'\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}})\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0)])}{\partial \mathbf{b}_0}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}_0, \boldsymbol{\epsilon}_0)}{\partial \boldsymbol{\beta}_0} = \frac{\partial(-\frac{1}{2}[\mathbf{b}_0'\tilde{\mathbf{B}}\mathbf{b}_0 + (\mathbf{Y}' - \boldsymbol{\beta}_0'\tilde{\mathbf{X}}' - \mathbf{b}_0'\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}})\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0)])}{\partial \boldsymbol{\beta}_0}$$

Setting the derivatives equal to zero, we obtain

$$\frac{\partial \log \text{ f}(\mathbf{b}_0, \boldsymbol{\epsilon}_0)}{\partial \mathbf{b}_0} = -\frac{1}{2}(2\tilde{\mathbf{B}}^{-1}\mathbf{b}_0 - 2\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0) + 2\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\tilde{\mathbf{B}}^{-1}\mathbf{b}_0 - \tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0) + \tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$(\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}})\mathbf{b}_0 + \tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 = \tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}_0, \boldsymbol{\epsilon}_0)}{\partial \boldsymbol{\beta}_0} = -\frac{1}{2}(-2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}.$$

We can rewrite this in matrix form to obtain the mixed model equations

$$\begin{bmatrix} \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & (\tilde{\mathbf{B}}^{-1}+\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}) \\ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\mathbf{b}}_0 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} \\ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} \end{bmatrix}.$$

D.2 Derivation of $\hat{\boldsymbol{\beta}}_0$ and $\hat{\mathbf{b}}_0$ for the Marginal Local Mixed Model using Henderson's joint likelihood of $\mathbf{b}_0$ and $\boldsymbol{\epsilon}_0$.

We solve the first equation for $\hat{\mathbf{b}}_0$

$$\hat{\mathbf{b}}_0 = (\tilde{\mathbf{B}}^{-1}+\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0).$$

Now,

$$(\tilde{\mathbf{B}}-\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}})(\tilde{\mathbf{B}}^{-1}+\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}) = \mathbf{I}.$$

in a proof similar to that given in Appendix B. So,

$$(\tilde{\mathbf{B}}^{-1}+\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}})^{-1} = (\tilde{\mathbf{B}}-\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}).$$

Thus we have

$$\begin{aligned} \hat{\mathbf{b}}_0 &= (\tilde{\mathbf{B}}^{-1}+\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\ &= (\tilde{\mathbf{B}}-\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}})\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I}-\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I}-\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0. \end{aligned}$$

Now,

$$(\mathbf{I} - \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}') = (\mathbf{I} - \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}' - \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}} + \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}})$$

$$= (\mathbf{I} - \tilde{\mathbf{V}}^{-1}(\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}' + \tilde{\mathbf{R}}) + \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}})$$

$$= (\mathbf{I} - \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{V}} + \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}}) = \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}}.$$

Substituting in, we obtain the predictor

$$\begin{aligned}\hat{\mathbf{b}}_0^{\mathsf{M}} &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 \\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{V}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) \\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0).\end{aligned}$$

Plugging the predictor of $\mathbf{b}_0$ into the second equation, we obtain

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

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

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

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

$$
\begin{aligned}
&= \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{V}_0^{**}\mathbf{V}_0^{**-1}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) - \tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0)\\
&+ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0
\end{aligned}
$$

$$
\begin{aligned}
&= \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) - \tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0)\\
&+ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0
\end{aligned}
$$

$$
\begin{aligned}
&= \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0\\
&- \tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) + \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0.
\end{aligned}
$$

Therefore,

$$\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0) = \mathbf{0}$$

$$\hat{\boldsymbol{\beta}}_0^{\mathsf{M}} = (\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\mathbf{Y}.$$

Appendix E

E.1 Derivation of Delete Cluster Mixed Model Equations for the Parametric Linear Mixed Model using Henderson's joint likelihood of b, $\boldsymbol{\phi}$, and $\boldsymbol{\epsilon}$. (Hurtado-Rodriguez, 1993)

We take the derivatives of log f with respect to $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, and $\mathbf{b}$.

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\epsilon})}{\partial \mathbf{b}} = \frac{\partial(-\frac{1}{2}[\mathbf{b}'\mathbf{B}\mathbf{b} + (\mathbf{Y}' - \boldsymbol{\beta}'\mathbf{X}' - \boldsymbol{\phi}'\mathbf{U}' - \mathbf{b}'\mathbf{Z}')\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\phi} - \mathbf{Z}\mathbf{b})])}{\partial \mathbf{b}}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\epsilon})}{\partial \boldsymbol{\beta}} = \frac{\partial(-\frac{1}{2}[\mathbf{b}'\mathbf{B}\mathbf{b} + (\mathbf{Y}' - \boldsymbol{\beta}'\mathbf{X}' - \boldsymbol{\phi}'\mathbf{U}' - \mathbf{b}'\mathbf{Z}')\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\phi} - \mathbf{Z}\mathbf{b})])}{\partial \boldsymbol{\beta}}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\beta})}{\partial \boldsymbol{\phi}} = \frac{\partial(-\frac{1}{2}[\mathbf{b}'\mathbf{B}\mathbf{b} + (\mathbf{Y}' - \boldsymbol{\beta}'\mathbf{X}' - \boldsymbol{\phi}'\mathbf{U}' - \mathbf{b}'\mathbf{Z}')\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\phi} - \mathbf{Z}\mathbf{b})])}{\partial \boldsymbol{\phi}}$$

Setting the derivatives equal to zero, we obtain

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\epsilon})}{\partial \mathbf{b}} = -\frac{1}{2}(2\mathbf{B}^{-1}\mathbf{b} - 2\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\phi}) + \mathbf{2Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}) = \mathbf{0}$$

$$\mathbf{B}^{-1}\mathbf{b} - \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\phi}) + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} = \mathbf{0}$$

$$(\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})\mathbf{b} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{U}\boldsymbol{\phi} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y}$$

$$\frac{\partial \log \text{ f}(\mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\epsilon})}{\partial \boldsymbol{\beta}} = -\frac{1}{2}(-2\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} + 2\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + 2\mathbf{X}'\mathbf{R}^{-1}\mathbf{U}\boldsymbol{\phi} + 2\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}) = \mathbf{0}$$

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{U}\boldsymbol{\phi} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} = \mathbf{0}$$

$$\mathbf{X}'\mathbf{R}^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\phi} + \mathbf{Z}\mathbf{b}) = \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y}$$

$$\frac{\partial \log f(\mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}} = -\frac{1}{2}(-2\mathbf{U}'\mathbf{R}^{-1}\mathbf{Y} + 2\mathbf{U}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + 2\mathbf{U}'\mathbf{R}^{-1}\mathbf{U}\boldsymbol{\phi} + 2\mathbf{U}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}) = \mathbf{0}$$

$$\mathbf{U}'\mathbf{R}^{-1}\mathbf{Y} - \mathbf{U}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{U}'\mathbf{R}^{-1}\mathbf{U}\boldsymbol{\phi} - \mathbf{U}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{b} = \mathbf{0}$$

$$\mathbf{U}'\mathbf{R}^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\phi} + \mathbf{Z}\mathbf{b}) = \mathbf{U}'\mathbf{R}^{-1}\mathbf{Y}.$$

The delete cluster mixed model equations for the parametric linear mixed model are given as

$$\begin{bmatrix} \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & (\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}) & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{U} \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{U} \\ \mathbf{U}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{U}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{U}'\mathbf{R}^{-1}\mathbf{U} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \\ \hat{\boldsymbol{\phi}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{U}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}.$$

E.2 Derivation of $\hat{\boldsymbol{\beta}}_{-i}, \hat{\boldsymbol{\phi}}_{-i}$ and $\hat{\mathbf{b}}_{-i}$ for the Parametric Linear Mixed Model using Henderson's joint likelihood of $\boldsymbol{\beta}, \boldsymbol{\phi}$ and $\mathbf{b}$.

We solve the first equation for $\hat{\mathbf{b}}$

$$\hat{\mathbf{b}} = (\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{U}\hat{\boldsymbol{\phi}}).$$

Now,

$$(\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})^{-1} = (\mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}) \text{ from Appendix B.}$$

So,

$$\hat{\mathbf{b}} = (\mathbf{B}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{U}\hat{\boldsymbol{\phi}})$$

$$= (\mathbf{B} - \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B})\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{U}\hat{\boldsymbol{\phi}})$$

$$\begin{aligned} &= \mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} - \mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}} \\ &- \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} + \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} \\ &+ \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}} \end{aligned}$$

$$\begin{aligned} &= \mathbf{B}\mathbf{Z}'(\mathbf{I} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}')\mathbf{R}^{-1}\mathbf{Y} - \mathbf{B}\mathbf{Z}'(\mathbf{I} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}')\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} \\ &- \mathbf{B}\mathbf{Z}'(\mathbf{I} - \mathbf{V}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}')\mathbf{R}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}} \end{aligned}$$

Now,

$$(\mathbf{I} - \mathbf{V}^{-1}\mathbf{ZBZ}') = (\mathbf{V}^{-1}\mathbf{V} - \mathbf{V}^{-1}\mathbf{ZBZ}') = \mathbf{V}^{-1}(\mathbf{V} - \mathbf{ZBZ}') = \mathbf{V}^{-1}\mathbf{R}.$$

So,

$$\hat{\mathbf{b}} = \mathbf{BZ}'\mathbf{V}^{-1}\mathbf{RR}^{-1}\mathbf{Y} - \mathbf{BZ}'\mathbf{V}^{-1}\mathbf{RR}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{BZ}'\mathbf{V}^{-1}\mathbf{RR}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}}$$

$$= \mathbf{BZ}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{U}\hat{\boldsymbol{\phi}}).$$

Using the second delete cluster mixed model equation, we obtain

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} = \mathbf{X}'\mathbf{R}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{ZBZ}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{U}\hat{\boldsymbol{\phi}}) + \mathbf{U}\hat{\boldsymbol{\phi}})$$

$$= \mathbf{X}'\mathbf{R}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{ZBZ}'\mathbf{V}^{-1}\mathbf{Y} - \mathbf{ZBZ}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{ZBZ}'\mathbf{V}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}} + \mathbf{U}\hat{\boldsymbol{\phi}})$$

and

$$\mathbf{X}'\mathbf{R}^{-1}((\mathbf{I} - \mathbf{ZBZ}'\mathbf{V}^{-1})\mathbf{X}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{ZBZ}'\mathbf{V}^{-1})\mathbf{U}\hat{\boldsymbol{\phi}}) = \mathbf{X}'\mathbf{R}^{-1}(\mathbf{I} - \mathbf{ZBZ}'\mathbf{V}^{-1})\mathbf{Y}.$$

Now

$$(\mathbf{I} - \mathbf{ZBZ}'\mathbf{V}^{-1}) = (\mathbf{I} + \mathbf{RV}^{-1} - \mathbf{RV}^{-1} - \mathbf{ZBZ}'\mathbf{V}^{-1})$$

$$= (\mathbf{I} + \mathbf{RV}^{-1} - (\mathbf{R} + \mathbf{ZBZ}')\mathbf{V}^{-1})$$

$$= (\mathbf{I} + \mathbf{RV}^{-1} - \mathbf{VV}^{-1})$$

$$= \mathbf{RV}^{-1}.$$

Plugging this into the equation, we find

$$\mathbf{X}'\mathbf{R}^{-1}(\mathbf{R}\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{R}\mathbf{V}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}}) = \mathbf{X}'\mathbf{R}^{-1}\mathbf{R}\mathbf{V}^{-1}\mathbf{Y}$$

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{V}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}).$$

Now,

$$\begin{aligned}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{U}\boldsymbol{\phi}) &= (\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}) \\ &\quad + \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\phi}) + \mathbf{U}\boldsymbol{\phi})) \\ &= (\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}) \\ &\quad + \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}) - \mathbf{U}\boldsymbol{\phi}) + \mathbf{U}\boldsymbol{\phi})) \\ &= (\mathbf{I} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1})(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}) \\ &\quad - (\mathbf{I} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1})(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\phi})) \\ &\quad + \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Y} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\phi} + \mathbf{U}\boldsymbol{\phi}.\end{aligned}$$

Recall that $(\mathbf{I} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}) = \mathbf{R}\mathbf{V}^{-1}$. So we have

$$\begin{aligned}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{U}\boldsymbol{\phi}) &= \mathbf{R}\mathbf{V}^{-1}(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}) \\ &\quad - \mathbf{R}\mathbf{V}^{-1}(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\phi}) \\ &\quad + \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Y} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\phi} + \mathbf{U}\boldsymbol{\phi}.\end{aligned}$$

So, using the third equation,

$$\mathbf{U}'\mathbf{R}^{-1}\mathbf{Y} = \mathbf{U}'\mathbf{R}^{-1}\mathbf{R}\mathbf{V}^{-1}(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}) - \mathbf{R}\mathbf{V}^{-1}(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\phi})$$

$$+ \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Y} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\phi} + \mathbf{U}\boldsymbol{\phi}$$

$$= \mathbf{U}'\mathbf{V}^{-1}(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}) + \mathbf{U}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Y}$$
$$- \mathbf{U}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\phi} + \mathbf{U}'\mathbf{R}^{-1}\mathbf{U}\boldsymbol{\phi}$$

$$= \mathbf{U}'\mathbf{V}^{-1}(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}) + \mathbf{U}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Y}$$
$$+ \mathbf{U}'\mathbf{R}^{-1}(\mathbf{I} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1})\mathbf{U}\boldsymbol{\phi}$$

$$= \mathbf{U}'\mathbf{V}^{-1}(\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}) + \mathbf{U}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Y}$$
$$+ \mathbf{U}'\mathbf{V}^{-1}\mathbf{U}\boldsymbol{\phi}$$

Notice that

$$\mathbf{U}'\mathbf{R}^{-1}(\mathbf{I} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1})\mathbf{Y} = \mathbf{U}'\mathbf{V}^{-1}\mathbf{Y}.$$

Let $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. Then

$$\begin{aligned}
\mathbf{U}'\mathbf{V}^{-1}\mathbf{Y} &= \mathbf{U}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}) + \mathbf{U}'\mathbf{V}^{-1}\mathbf{U}\hat{\boldsymbol{\phi}} \\
\mathbf{0} &= \mathbf{U}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}) - \mathbf{U}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}) \\
\mathbf{0} &= \mathbf{U}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}) \\
\mathbf{0} &= \mathbf{U}'\mathbf{P}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}) \\
\hat{\boldsymbol{\phi}}_{\text{-i}} &= (\mathbf{U}'\mathbf{P}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P}\mathbf{Y}.
\end{aligned}$$

Back substitution gives us

$$\hat{\boldsymbol{\beta}}_{-\mathsf{i}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}_{-\mathsf{i}})$$

$$= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P})\mathbf{Y}$$

and

$$\hat{\mathbf{b}}_{-\mathsf{i}} = \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{-\mathsf{i}} - \mathbf{U}\hat{\boldsymbol{\phi}}_{-\mathsf{i}})$$

$$= \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P})\mathbf{Y}$$
$$- \mathbf{U}(\mathbf{U}'\mathbf{P}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P}\mathbf{Y})$$

$$= \mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}((\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{Y}$$
$$- (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{U}(\mathbf{U}'\mathbf{P}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P}\mathbf{Y})$$

$$= \mathbf{B}\mathbf{Z}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P})\mathbf{Y}$$

$$= \mathbf{B}\mathbf{Z}'\mathbf{P}(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}\mathbf{U})^{-1}\mathbf{U}'\mathbf{P})\mathbf{Y}.$$

Appendix F

F.1 Derivation of Delete Cluster Mixed Model Equations for the Conditional Local Mixed Model using Henderson's joint likelihood of $\mathbf{b}_0$, $\boldsymbol{\phi}_0$, and $\boldsymbol{\epsilon}_0$.

We take the derivatives of log f with respect to $\boldsymbol{\beta}_0$, $\boldsymbol{\phi}_0$, and $\mathbf{b}_0$.

$$\frac{\partial \log f(b_0, \phi_0, \epsilon_0)}{\partial b_0} = \frac{\partial(-\frac{1}{2}[b_0' \tilde{B} b_0 + (Y' - \beta_0' \tilde{X}' - \phi_0' U' - b_0' \tilde{Z}') K_0^{\frac{1}{2}} \tilde{R}^{-1} K_0^{\frac{1}{2}} (Y - \tilde{X}\beta_0 - U\phi_0 - \tilde{Z} b_0)])}{\partial b_0}$$

$$\frac{\partial \log f(b_0, \phi_0, \epsilon_0)}{\partial \beta_0} = \frac{\partial(-\frac{1}{2}[b_0' \tilde{B} b_0 + (Y' - \beta_0' \tilde{X}' - \phi_0' U' - b_0' \tilde{Z}') K_0^{\frac{1}{2}} \tilde{R}^{-1} K_0^{\frac{1}{2}} (Y - \tilde{X}\beta_0 - U\phi_0 - \tilde{Z} b_0)])}{\partial \beta_0}$$

$$\frac{\partial \log f(b_0, \phi_0, \beta_0)}{\partial \phi_0} = \frac{\partial(-\frac{1}{2}[b_0' \tilde{B} b_0 + (Y' - \beta_0' \tilde{X}' - \phi_0' U' - b_0' \tilde{Z}') K_0^{\frac{1}{2}} \tilde{R}^{-1} K_0^{\frac{1}{2}} (Y - \tilde{X}\beta_0 - U\phi_0 - \tilde{Z} b_0)])}{\partial \phi_0}$$

Setting the derivatives equal to zero, we obtain

$$\frac{\partial \log f(\mathbf{b}_0, \boldsymbol{\phi}_0, \boldsymbol{\epsilon}_0)}{\partial \mathbf{b}_0} = -\frac{1}{2}(2\tilde{\mathbf{B}}^{-1}\mathbf{b}_0 - 2\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{U}\boldsymbol{\phi}_0) + \mathbf{2}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\tilde{\mathbf{B}}^{-1}\mathbf{b}_0 - \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{U}\boldsymbol{\phi}_0) + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$\begin{aligned}&(\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}})\mathbf{b}_0 + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 \\ &+ \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 = \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}\end{aligned}$$

$$\frac{\partial \log f(\mathbf{b}_0, \boldsymbol{\phi}_0, \boldsymbol{\epsilon}_0)}{\partial \boldsymbol{\beta}_0} = -\frac{1}{2}(-2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \mathbf{U}\boldsymbol{\phi}_0 + \tilde{\mathbf{Z}}\mathbf{b}_0) = \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}$$

$$\frac{\partial \log f(\mathbf{b}_0, \boldsymbol{\phi}_0, \boldsymbol{\epsilon}_0)}{\partial \boldsymbol{\phi}_0} = -\frac{1}{2}(-2\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + 2\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + 2\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 + 2\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 - \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \mathbf{U}\boldsymbol{\phi}_0 + \tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}.$$

The delete cluster mixed model equations for the conditional local linear mixed model are given as

$$\begin{bmatrix} \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}) & \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U} \\ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}} & \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U} \\ \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}} & \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\mathbf{b}}_0 \\ \hat{\boldsymbol{\phi}}_0 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} \\ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} \\ \mathbf{U}'\tilde{\mathbf{R}}^{-1}\mathbf{Y} \end{bmatrix}.$$

F.2 Derivation of $\hat{\boldsymbol{\beta}}_{0,\text{-i}}, \hat{\boldsymbol{\phi}}_{0,\text{-i}}$ and $\hat{\mathbf{b}}_{0,\text{-i}}$ for the Conditional Local Mixed Model.

We solve the first equation for $\hat{\mathbf{b}}_0$

$$\hat{\mathbf{b}}_0 = (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0).$$

Now,

$$(\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}})^{-1} = (\tilde{\mathbf{B}} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}) \text{ from Appendix C.}$$

So,

$$\begin{aligned}\hat{\mathbf{b}}_0 &= (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0)\\ &= (\tilde{\mathbf{B}} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}})\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0)\\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\hat{\boldsymbol{\phi}}_0\\ &\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0\\ &\quad + \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\hat{\boldsymbol{\phi}}_0\\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}\\ &\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0\\ &\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\hat{\boldsymbol{\phi}}_0\end{aligned}$$

Now,

$$(\mathbf{I} - \mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}') = (\mathbf{V}_0^{*-1}\mathbf{V}_0^{*} - \mathbf{V}_0^{*-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}') = \mathbf{V}_0^{*-1}(\mathbf{V}_0^{*} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}') = \mathbf{V}_0^{*-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{\frac{1}{2}}.$$

So,

$$\begin{aligned}\hat{\mathbf{b}}_0 &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0\\ &\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\hat{\boldsymbol{\phi}}_0\\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0).\end{aligned}$$

Using the second delete cluster mixed model equation, we obtain

$$\begin{aligned}\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} &= \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0) + \mathbf{U}\hat{\boldsymbol{\phi}}_0)\\ &= \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{Y} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1}\mathbf{U}\hat{\boldsymbol{\phi}}_0 + \mathbf{U}\hat{\boldsymbol{\phi}}_0)\end{aligned}$$

and

$$\begin{aligned}&\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}((\mathbf{I}-\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1})\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0+(\mathbf{I}-\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1})\mathbf{U}\hat{\boldsymbol{\phi}}_0)\\&=\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{I}-\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*-1})\mathbf{Y}.\end{aligned}$$

Now

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

Plugging this into the equation, we find

$$\begin{aligned}&\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{*-1}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0+\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{*-1}\mathbf{U}\hat{\boldsymbol{\phi}}_0)\\&=\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{*-1}\mathbf{Y}\\&\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0+\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{U}\hat{\boldsymbol{\phi}}_0=\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\mathbf{Y}\\&\hat{\boldsymbol{\beta}}_0=(\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*-1}(\mathbf{Y}-\mathbf{U}\hat{\boldsymbol{\phi}}_0).\end{aligned}$$

Now,

$$\begin{aligned}(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \tilde{\mathbf{Z}}\mathbf{b}_0 + \mathbf{U}\boldsymbol{\phi}_0) &= (\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}_0) \\ &\quad + \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{U}\boldsymbol{\phi}_0) + \mathbf{U}\boldsymbol{\phi}_0)) \\ \\ &= (\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}_0) \\ &\quad + \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}_0) \\ &\quad - \mathbf{U}\boldsymbol{\phi}_0) + \mathbf{U}\boldsymbol{\phi}_0)) \\ \\ &= (\mathbf{I} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1})(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y}) \\ &\quad - (\mathbf{I} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1})(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\mathbf{U}\boldsymbol{\phi}_0)) \\ &\quad + \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{U}\boldsymbol{\phi}_0 + \mathbf{U}\boldsymbol{\phi}_0.\end{aligned}$$

Recall that $(\mathbf{I} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}) = \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{*\text{-}1}$. So we have

$$\begin{aligned}(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \tilde{\mathbf{Z}}\mathbf{b}_0 + \mathbf{U}\boldsymbol{\phi}_0) &= \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{*\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y}) \\ &\quad - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{*\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\mathbf{U}\boldsymbol{\phi}_0) \\ &\quad + \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{U}\boldsymbol{\phi}_0 + \mathbf{U}\boldsymbol{\phi}_0.\end{aligned}$$

So, using the third equation,

$$\begin{aligned}\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} &= \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{*\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y}) \\ &\quad - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{*\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\mathbf{U}\boldsymbol{\phi}_0) \\ &\quad + \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{U}\boldsymbol{\phi}_0 + \mathbf{U}\boldsymbol{\phi}_0 \\ \\ &= \mathbf{U}'\mathbf{V}_0^{*\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}_0) \\ &\quad + \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y} \\ &\quad - \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{U}\boldsymbol{\phi}_0 + \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{U}'\mathbf{V}_0^{*\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y}-\mathbf{U}\boldsymbol{\phi}_0)\\
&+ \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y}\\
&+ \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{I}-\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1})\mathbf{U}\boldsymbol{\phi}_0
\end{aligned}
$$

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

Notice that

$$
\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{I}-\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1})\mathbf{Y} = \mathbf{U}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y}.
$$

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

$$
\begin{aligned}
\mathbf{U}'\mathbf{V}_0^{*\text{-}1}\mathbf{Y} &= \mathbf{U}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y}-\mathbf{U}\boldsymbol{\phi}_0) + \mathbf{U}'\mathbf{V}_0^{*\text{-}1}\mathbf{U}\hat{\boldsymbol{\phi}}_0\\
\mathbf{0} &= \mathbf{U}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y}-\mathbf{U}\hat{\boldsymbol{\phi}}_0) - \mathbf{U}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y}-\mathbf{U}\boldsymbol{\phi}_0)\\
\mathbf{0} &= \mathbf{U}'(\mathbf{V}_0^{*\text{-}1} - \mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1})(\mathbf{Y}-\mathbf{U}\hat{\boldsymbol{\phi}}_0)\\
\mathbf{0} &= \mathbf{U}'\mathbf{P}^*(\mathbf{Y}-\mathbf{U}\hat{\boldsymbol{\phi}}_0)\\
\hat{\boldsymbol{\phi}}_{0,\text{-i}}^{\mathsf{C}} &= (\mathbf{U}'\mathbf{P}^*\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^*\mathbf{Y}.
\end{aligned}
$$

Back substitution gives us

$$\hat{\boldsymbol{\beta}}^{\mathsf{C}}_{0,\text{-i}} = (\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}})$$

$$= (\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}^*\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^*)\mathbf{Y}$$

and

$$\hat{\mathbf{b}}^{\mathsf{C}}_{0,\text{-i}} = \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{0,\text{-i}} - \mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}})$$

$$= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}^*\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^*)\mathbf{Y}$$
$$- \mathbf{U}(\mathbf{U}'\mathbf{P}^*\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^*\mathbf{Y})$$

$$= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{*\text{-}1}((\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1})\mathbf{Y}$$
$$- (\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1})\mathbf{U}(\mathbf{U}'\mathbf{P}^*\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^*\mathbf{Y})$$

$$= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{V}_0^{*\text{-}1} - \mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{*\text{-}1})(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}^*\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^*)\mathbf{Y}$$

$$= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{P}^*(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}^*\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^*)\mathbf{Y}.$$

Appendix G

G.1 Derivation of Delete Cluster Mixed Model Equations for the Marginal Local Mixed Model using Henderson's joint likelihood of $\mathbf{b}_0$, $\boldsymbol{\phi}_0$, and $\boldsymbol{\epsilon}_0$.

We take the derivatives of log f with respect to $\boldsymbol{\beta}_0$, $\boldsymbol{\phi}_0$, and $\mathbf{b}_0$.

$$\frac{\partial \log f(b_0, \phi_0, \epsilon_0)}{\partial b_0} = \frac{\partial(-\frac{1}{2}[b_0' \tilde{B} b_0 + (Y' - \beta_0' \tilde{X}' - \phi_0' U' - b_0' \tilde{Z}' K_0^{-\frac{1}{2}}) K_0^{\frac{1}{2}} \tilde{R}^{-1} K_0^{\frac{1}{2}} (Y - \tilde{X}\beta_0 - U\phi_0 - K_0^{-\frac{1}{2}} \tilde{Z} b_0)])}{\partial b_0}$$

$$\frac{\partial \log f(b_0, \phi_0, \epsilon_0)}{\partial \beta_0} = \frac{\partial(-\frac{1}{2}[b_0' \tilde{B} b_0 + (Y' - \beta_0' \tilde{X}' - \phi_0' U' - b_0' \tilde{Z}' K_0^{-\frac{1}{2}}) K_0^{\frac{1}{2}} \tilde{R}^{-1} K_0^{\frac{1}{2}} (Y - \tilde{X}\beta_0 - U\phi_0 - K_0^{-\frac{1}{2}} \tilde{Z} b_0)])}{\partial \beta_0}$$

$$\frac{\partial \log f(b_0, \phi_0, \beta_0)}{\partial \phi_0} = \frac{\partial(-\frac{1}{2}[b_0' \tilde{B} b_0 + (Y' - \beta_0' \tilde{X}' - \phi_0' U' - b_0' \tilde{Z}' K_0^{-\frac{1}{2}}) K_0^{\frac{1}{2}} \tilde{R}^{-1} K_0^{\frac{1}{2}} (Y - \tilde{X}\beta_0 - U\phi_0 - K_0^{-\frac{1}{2}} \tilde{Z} b_0)])}{\partial \phi_0}$$

Setting the derivatives equal to zero, we obtain

$$\frac{\partial \log f(\mathbf{b}_0, \boldsymbol{\phi}_0, \boldsymbol{\epsilon}_0)}{\partial \mathbf{b}_0} = -\frac{1}{2}(2\tilde{\mathbf{B}}^{-1}\mathbf{b}_0 - 2\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{U}\boldsymbol{\phi}_0) + 2\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\tilde{\mathbf{B}}^{-1}\mathbf{b}_0 - \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{U}\boldsymbol{\phi}_0) + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$(\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}})\mathbf{b}_0 + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 = \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}$$

$$\frac{\partial \log f(\mathbf{b}_0, \boldsymbol{\phi}_0, \boldsymbol{\epsilon}_0)}{\partial \boldsymbol{\beta}_0} = -\frac{1}{2}(-2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 + 2\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 - \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \mathbf{U}\boldsymbol{\phi}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}$$

$$\frac{\partial \log f(\mathbf{b}_0, \boldsymbol{\phi}_0, \boldsymbol{\epsilon}_0)}{\partial \boldsymbol{\phi}_0} = -\frac{1}{2}(-2\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + 2\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + 2\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 + 2\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{0}$$

$$\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0 - \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}\mathbf{b}_0 = \mathbf{0}$$

$$\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \mathbf{U}\boldsymbol{\phi}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0) = \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}.$$

The delete cluster mixed model equations for the marginal local linear mixed model are given as

$$\begin{bmatrix} \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}}) & \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U} \\ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}} & \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U} \\ \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}} & \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}} & \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\mathbf{b}}_0 \\ \hat{\boldsymbol{\phi}}_0 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} \\ \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} \\ \mathbf{U}'\tilde{\mathbf{R}}^{-1}\mathbf{Y} \end{bmatrix}.$$

G.2 Derivation of $\hat{\boldsymbol{\beta}}_{0,-i}, \hat{\boldsymbol{\phi}}_{0,-i}$ and $\hat{\mathbf{b}}_{0,-i}$ for the Marginal Local Mixed Model.

We solve the first equation for $\hat{\mathbf{b}}_0$

$$\hat{\mathbf{b}}_0 = (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0).$$

Now,

$$(\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}})^{-1} = (\tilde{\mathbf{B}} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}).$$

So,

$$\begin{aligned}\hat{\mathbf{b}}_0 &= (\tilde{\mathbf{B}}^{-1} + \tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0)\\ &= (\tilde{\mathbf{B}} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}})\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0)\\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\hat{\boldsymbol{\phi}}_0\\ &\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} + \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0\\ &\quad + \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{V}_0^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\hat{\boldsymbol{\phi}}_0\\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{V}_0^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y}\\ &\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{V}_0^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0\\ &\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'(\mathbf{I} - \mathbf{V}_0^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}')\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\hat{\boldsymbol{\phi}}_0\end{aligned}$$

Now,

$$(\mathbf{I} - \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}') = (\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{V}} - \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}') = \tilde{\mathbf{V}}^{-1}(\tilde{\mathbf{V}} - \tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}') = \tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}}.$$

So,

$$\begin{aligned}\hat{\mathbf{b}}_0 &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0\\ &\quad - \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{R}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\hat{\boldsymbol{\phi}}_0\\ &= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0).\end{aligned}$$

Using the second delete cluster mixed model equation, we obtain

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} = \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 - \mathbf{U}\hat{\boldsymbol{\phi}}_0) + \mathbf{U}\hat{\boldsymbol{\phi}}_0)$$

$$= \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}\mathbf{Y} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0$$
$$- \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}\mathbf{U}\hat{\boldsymbol{\phi}}_0 + \mathbf{U}\hat{\boldsymbol{\phi}}_0)$$

and

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}((\mathbf{I} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1})\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + (\mathbf{I} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1})\mathbf{U}\hat{\boldsymbol{\phi}}_0)$$
$$= \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{I} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1})\mathbf{Y}.$$

Now

$$(\mathbf{I} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}) = (\mathbf{I} + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1})$$

$$= (\mathbf{I} + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1} - (\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}} + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}})\mathbf{V}_0^{**-1})$$

$$= (\mathbf{I} + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1} - \mathbf{V}_0^{**}\mathbf{V}_0^{**-1})$$

$$= \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}.$$

Plugging this into the equation, we find

$$\tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}\mathbf{U}\hat{\boldsymbol{\phi}}_0)$$
$$= \tilde{\mathbf{X}}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**-1}\mathbf{Y}$$

$$\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_0 + \tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\mathbf{U}\hat{\boldsymbol{\phi}}_0 = \tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\mathbf{Y}$$

$$\hat{\boldsymbol{\beta}}_0 = (\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**-1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}_0).$$

Now,

$$
\begin{aligned}
(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 + \mathbf{U}\boldsymbol{\phi}_0) &= (\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}_0) \\
&\quad + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_0 - \mathbf{U}\boldsymbol{\phi}_0) + \mathbf{U}\boldsymbol{\phi}_0)) \\
&= (\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}_0) \\
&\quad + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}_0) \\
&\quad - \mathbf{U}\boldsymbol{\phi}_0) + \mathbf{U}\boldsymbol{\phi}_0)) \\
&= (\mathbf{I} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1})(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{Y}) \\
&\quad - (\mathbf{I} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1})(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{U}\boldsymbol{\phi}_0)) \\
&\quad + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{Y} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{U}\boldsymbol{\phi}_0 + \mathbf{U}\boldsymbol{\phi}_0.
\end{aligned}
$$

Recall that $(\mathbf{I} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}) = \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}$. So we have

$$
\begin{aligned}
(\tilde{\mathbf{X}}\boldsymbol{\beta}_0 + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{b}_0 + \mathbf{U}\boldsymbol{\phi}_0) &= \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{Y}) \\
&\quad - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{U}\boldsymbol{\phi}_0) \\
&\quad + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{Y} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{U}\boldsymbol{\phi}_0 + \mathbf{U}\boldsymbol{\phi}_0.
\end{aligned}
$$

So, using the third equation,

$$
\begin{aligned}
\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{Y} &= \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{Y}) \\
&\quad - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{R}}\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\mathbf{U}\boldsymbol{\phi}_0) \\
&\quad + \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{Y} - \mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{U}\boldsymbol{\phi}_0 + \mathbf{U}\boldsymbol{\phi}_0) \\
&= \mathbf{U}'\mathbf{V}_0^{**\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \mathbf{U}\boldsymbol{\phi}_0) \\
&\quad + \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{Y} \\
&\quad - \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{U}\boldsymbol{\phi}_0 + \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}\mathbf{U}\boldsymbol{\phi}_0
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbf{U}'\mathbf{V}_0^{**\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y}-\mathbf{U}\boldsymbol{\phi}_0) \\
&+ \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}\mathbf{Y} \\
&+ \mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{\text{-}1}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{I}-\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1})\mathbf{U}\boldsymbol{\phi}_0 \\
\\
&= \mathbf{U}'\mathbf{V}_0^{**\text{-}1}(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y}-\mathbf{U}\boldsymbol{\phi}_0) \\
&+ \mathbf{U}'\mathbf{V}_0^{**\text{-}1}\mathbf{U}\boldsymbol{\phi}_0
\end{aligned}
$$

Notice that

$$
\mathbf{U}'\mathbf{K}_0^{\frac{1}{2}}\tilde{\mathbf{R}}^{-\frac{1}{2}}\mathbf{K}_0^{\frac{1}{2}}(\mathbf{I}-\mathbf{K}_0^{-\frac{1}{2}}\tilde{\mathbf{Z}}\tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1})\mathbf{Y} = \mathbf{U}'\mathbf{V}_0^{**\text{-}1}.
$$

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

$$
\begin{aligned}
\mathbf{U}'\mathbf{V}_0^{**\text{-}1}\mathbf{Y} &= \mathbf{U}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y}-\mathbf{U}\boldsymbol{\phi}_0) + \mathbf{U}'\mathbf{V}_0^{**\text{-}1}\mathbf{U}\hat{\boldsymbol{\phi}}_0 \\
\mathbf{0} &= \mathbf{U}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y}-\mathbf{U}\hat{\boldsymbol{\phi}}_0) - \mathbf{U}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y}-\mathbf{U}\boldsymbol{\phi}_0) \\
\mathbf{0} &= \mathbf{U}'(\mathbf{V}_0^{**\text{-}1} - \mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1})(\mathbf{Y}-\mathbf{U}\hat{\boldsymbol{\phi}}_0) \\
\mathbf{0} &= \mathbf{U}'\mathbf{P}^{**}(\mathbf{Y}-\mathbf{U}\hat{\boldsymbol{\phi}}_0) \\
\hat{\boldsymbol{\phi}}_{0,\text{-i}}^{\mathsf{M}} &= (\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^{**}\mathbf{Y}.
\end{aligned}
$$

Back substitution gives us

$$\hat{\boldsymbol{\beta}}^{\mathsf{M}}_{0,\text{-i}} = (\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}})$$

$$= (\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^{**})\mathbf{Y}$$

and

$$\hat{\mathbf{b}}^{\mathsf{M}}_{0,\text{-i}} = \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{0,\text{-i}} - \mathbf{U}\hat{\boldsymbol{\phi}}_{0,\text{-i}})$$

$$= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}(\mathbf{Y} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^{**})\mathbf{Y}$$
$$- \mathbf{U}(\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^{**}\mathbf{Y})$$

$$= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{V}_0^{**\text{-}1}((\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1})\mathbf{Y}$$
$$- (\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1})\mathbf{U}(\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^{**}\mathbf{Y})$$

$$= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}(\mathbf{V}_0^{**\text{-}1} - \mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1}\tilde{\mathbf{X}})^{\text{-}1}\tilde{\mathbf{X}}'\mathbf{V}_0^{**\text{-}1})(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^{**})\mathbf{Y}$$

$$= \tilde{\mathbf{B}}\tilde{\mathbf{Z}}'\mathbf{K}_0^{-\frac{1}{2}}\mathbf{P}^{**}(\mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{P}^{**}\mathbf{U})^{\text{-}1}\mathbf{U}'\mathbf{P}^{**})\mathbf{Y}.$$

Appendix H

The Population Average Parametric Mixed Model

All results given here are conditioned on the random effects $\mathbf{b}$ for fixed. known $\mathbf{V}$.
The true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\epsilon}$, where Var($\boldsymbol{\epsilon}$)=$\mathbf{R}$ and Var($\mathbf{Y}$)=$\mathbf{V}$.
The parametric fit is $\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{PA}} = \mathbf{H}^{\mathsf{P}}_{\mathsf{PA}}\mathbf{Y}$, where $\mathbf{H}^{\mathsf{P}}_{\mathsf{PA}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$.

$$
\begin{aligned}
\text{Bias}(\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{PA}}) &= \text{E}(\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{PA}}) - \text{E}(\mathbf{Y}) \\
&= \text{E}(\mathbf{H}^{\mathsf{P}}_{\mathsf{PA}}\mathbf{Y}) - \text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{P}}_{\mathsf{PA}})\text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{P}}_{\mathsf{PA}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f}) \\
&= -(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f}) \\
&= -(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{f} \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{P}}_{\mathsf{PA}})\mathbf{f}.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{PA}}) &= \mathrm{Var}(\mathbf{H}^{\mathsf{P}}_{\mathsf{PA}}\mathbf{Y}) \\
&= \mathbf{H}^{\mathsf{P}}_{\mathsf{PA}}\mathrm{Var}(\mathbf{Y})\mathbf{H}^{\mathsf{P}}_{\mathsf{PA}}{}' \\
&= \mathbf{H}^{\mathsf{P}}_{\mathsf{PA}}\mathbf{V}\mathbf{H}^{\mathsf{P}}_{\mathsf{PA}}{}' \\
&= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{H}^{\mathsf{P}}_{\mathsf{PA}}\mathbf{V}.
\end{aligned}
$$

Appendix I

The Population Average Local Mixed Model
The Conditional Local Mixed Model

All results given here are conditioned on the random effects $\mathbf{b}$ for a fixed bandwidth h and known $\mathbf{V}$.
The true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\epsilon}$, where $\text{Var}(\boldsymbol{\epsilon})$=$\mathbf{R}$ and $\text{Var}(\mathbf{Y})$=$\mathbf{V}$.
The conditional local fit is $\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{PA}} = \mathbf{H}^{\mathsf{C}}_{\mathsf{PA}}\mathbf{Y}$, where $\mathbf{H}^{\mathsf{C}}_{\mathsf{PA}}$ is given in Chapter 7.

$$\begin{aligned}
\text{Bias}(\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{PA}}) &= \text{E}(\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{PA}}) - \text{E}(\mathbf{Y}) \\
&= \text{E}(\mathbf{H}^{\mathsf{C}}_{\mathsf{PA}}\mathbf{Y}) - \text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{C}}_{\mathsf{PA}})\text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{C}}_{\mathsf{PA}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f})
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{PA}}) &= \text{Var}(\mathbf{H}^{\mathsf{C}}_{\mathsf{PA}}\mathbf{Y}) \\
&= \mathbf{H}^{\mathsf{C}}_{\mathsf{PA}}\text{Var}(\mathbf{Y})\mathbf{H}^{\mathsf{C}}_{\mathsf{PA}}{}' \\
&= \mathbf{H}^{\mathsf{C}}_{\mathsf{PA}}\mathbf{V}\mathbf{H}^{\mathsf{C}}_{\mathsf{PA}}{}'.
\end{aligned}$$

Appendix J

The Population Average Local Mixed Model
The Marginal Local Mixed Model

All results given here are conditioned on the random effects $\mathbf{b}$ for a fixed bandwidth h and known $\mathbf{V}$.
The true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\epsilon}$, where $\text{Var}(\boldsymbol{\epsilon})$=$\mathbf{R}$ and $\text{Var}(\mathbf{Y})$=$\mathbf{V}$.
The marginal local fit is $\hat{\mathbf{Y}}^{\mathsf{M}}_{\mathsf{PA}} = \mathbf{H}^{\mathsf{M}}_{\mathsf{PA}}\mathbf{Y}$, where $\mathbf{H}^{\mathsf{M}}_{\mathsf{PA}}$ is given in Chapter 7.

$$\begin{aligned}
\text{Bias}(\hat{\mathbf{Y}}^{\mathsf{M}}_{\mathsf{PA}}) &= \text{E}(\hat{\mathbf{Y}}^{\mathsf{M}}_{\mathsf{PA}}) - \text{E}(\mathbf{Y}) \\
&= \text{E}(\mathbf{H}^{\mathsf{M}}_{\mathsf{PA}}\mathbf{Y}) - \text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{M}}_{\mathsf{PA}})\text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{M}}_{\mathsf{PA}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f})
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\mathbf{Y}}^{\mathsf{M}}_{\mathsf{PA}}) &= \text{Var}(\mathbf{H}^{\mathsf{M}}_{\mathsf{PA}}\mathbf{Y}) \\
&= \mathbf{H}^{\mathsf{M}}_{\mathsf{PA}}\text{Var}(\mathbf{Y})\mathbf{H}^{\mathsf{M}\,\prime}_{\mathsf{PA}} \\
&= \mathbf{H}^{\mathsf{M}}_{\mathsf{PA}}\mathbf{V}\mathbf{H}^{\mathsf{M}\,\prime}_{\mathsf{PA}}.
\end{aligned}$$

Appendix K

The Cluster Specific Parametric Mixed Model

All results given here are conditioned on the random effects $\mathbf{b}$ for a fixed bandwidth h and known $\mathbf{V}$.
The true model is $\mathbf{Y}|\mathbf{b} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{f} + \boldsymbol{\epsilon}$, where $\text{Var}(\boldsymbol{\epsilon})$=$\mathbf{R}$ and $\text{Var}(\mathbf{Y}|\mathbf{b})$=$\mathbf{R}$.
The parametric model fit is $\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{CS}} = \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y}$, where
$\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}} = (\mathbf{I} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1})\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}$.

$$\begin{aligned}
\text{Bias}(\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{CS}}|\mathbf{b}) &= \text{E}(\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{CS}}|\mathbf{b}) - \text{E}(\mathbf{Y}|\mathbf{b}) \\
&= \text{E}(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y}|\mathbf{b}) - \text{E}(\mathbf{Y}|\mathbf{b}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}})\text{E}(\mathbf{Y}|\mathbf{b}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{f}).
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{CS}}|\mathbf{b}) &= \text{Var}(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y}|\mathbf{b}) \\
&= \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\text{Var}(\mathbf{Y}|\mathbf{b})\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}{}' \\
&= \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{R}\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}{}'.
\end{aligned}$$

Appendix L

The Cluster Specific Local Mixed Model
The Conditional Local Mixed Model

All results given here are conditioned on the random effects $\mathbf{b}$ for a fixed bandwidth h and known $\mathbf{V}$.
The true model is $\mathbf{Y}|\mathbf{b} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{f} + \boldsymbol{\epsilon}$, where $\text{Var}(\boldsymbol{\epsilon})$=$\mathbf{R}$ and $\text{Var}(\mathbf{Y}|\mathbf{b})$=$\mathbf{R}$.
The conditional local model fit is $\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{CS}} = \mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}\mathbf{Y}$, where $\mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}$ is given in Chapter 7.

$$\begin{aligned}
\text{Bias}(\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{CS}}|\mathbf{b}) &= \text{E}(\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{CS}}|\mathbf{b}) - \text{E}(\mathbf{Y}|\mathbf{b}) \\
&= \text{E}(\mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}\mathbf{Y}|\mathbf{b}) - \text{E}(\mathbf{Y}|\mathbf{b}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{C}}_{\mathsf{CS}})\text{E}(\mathbf{Y}|\mathbf{b}) \\
&= -(\mathbf{I} - \mathbf{H}^{\mathsf{C}}_{\mathsf{CS}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{f}).
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\mathbf{Y}}^{\mathsf{C}}_{\mathsf{CS}}|\mathbf{b}) &= \text{Var}(\mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}\mathbf{Y}|\mathbf{b}) \\
&= \mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}\text{Var}(\mathbf{Y}|\mathbf{b})\mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}{}' \\
&= \mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}\mathbf{R}\mathbf{H}^{\mathsf{C}}_{\mathsf{CS}}{}'.
\end{aligned}$$

Appendix M

The Population Average Model Robust Mixed Model Using the Conditional Local Mixed Model

All results given here are for a fixed bandwidth h, λ, and known $\mathbf{V}$.
The true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\epsilon}$, where Var($\boldsymbol{\epsilon}$)=$\mathbf{R}$ and Var($\mathbf{Y}$)=$\mathbf{V}$.
The model robust mixed model fit using CLMM is
$\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}} = [\lambda \mathbf{H}_{\mathsf{PA}}^{\mathsf{C}} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}]\mathbf{Y} = \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{Y}$.

$$
\begin{aligned}
\text{Bias}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}) &= \text{E}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}) - \text{E}(\mathbf{Y}) \\
&= \text{E}(\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{Y}) - \text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}})\text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}})\mathbf{f} \\
&= -(\mathbf{I} - [\lambda \mathbf{H}_{\mathsf{PA}}^{\mathsf{C}} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}])\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}})\mathbf{f} \\
&= -\lambda(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{C}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}})\mathbf{f}.
\end{aligned}
$$

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,C}}) &= \mathrm{Var}(\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{Y}) \\
&= \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathrm{Var}(\mathbf{Y})\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}\prime} \\
&= \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}\prime} \\
&= \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{V}[\lambda\mathbf{H}_{\mathsf{PA}}^{\mathsf{C}} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}]\prime \\
&= \lambda\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{C}}{}' + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,C}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}{}'.
\end{aligned}$$

Appendix N

The Population Average Model Robust Mixed Model Using the Marginal Local Mixed Model

All results given here are for a fixed bandwidth h, λ, and known $\mathbf{V}$.
The true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\epsilon}$, where Var($\boldsymbol{\epsilon}$)=$\mathbf{R}$ and Var($\mathbf{Y}$)=$\mathbf{V}$.
The model robust mixed model fit using MLMM is
$\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,M}} = [\lambda \mathbf{H}_{\mathsf{PA}}^{\mathsf{M}} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}]\mathbf{Y} = \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{Y}$.

$$
\begin{aligned}
\text{Bias}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,M}}) &= \text{E}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,M}}) - \text{E}(\mathbf{Y}) \\
&= \text{E}(\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{Y}) - \text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}})\text{E}(\mathbf{Y}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{f}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}})\mathbf{f} \\
&= -(\mathbf{I} - [\lambda \mathbf{H}_{\mathsf{PA}}^{\mathsf{M}} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}])\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}})\mathbf{f} \\
&= -\lambda(\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{M}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}})\mathbf{f}.
\end{aligned}
$$

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{Y}}_{\mathsf{PA}}^{\mathsf{MMRR,M}}) &= \mathrm{Var}(\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{Y}) \\
&= \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathrm{Var}(\mathbf{Y})\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}\prime} \\
&= \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}\prime} \\
&= \mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{V}[\lambda\mathbf{H}_{\mathsf{PA}}^{\mathsf{M}} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}}]' \\
&= \lambda\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{M}\ \prime} + (1-\lambda)\mathbf{H}_{\mathsf{PA}}^{\mathsf{MMRR,M}}\mathbf{V}\mathbf{H}_{\mathsf{PA}}^{\mathsf{P}\ \prime}.
\end{aligned}$$

Appendix O

The Cluster Specific Model Robust Mixed Model

All results given here are conditioned on the random effects for a fixed h, λ, and known $\mathbf{V}$.
The true model is $\mathbf{Y}|\mathbf{b} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{f} + \boldsymbol{\epsilon}$, where Var($\boldsymbol{\epsilon}$)=$\mathbf{R}$ and Var($\mathbf{Y}$)=$\mathbf{V}$.
The model robust mixed model fit using CLMM is
$\hat{\mathbf{Y}}_{\mathsf{CS}}^{\mathsf{MMRR,C}}|\mathbf{b} = [\lambda\mathbf{H}_{\mathsf{CS}}^{\mathsf{C}} + (1-\lambda)\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}]\mathbf{Y} = \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{Y}$.

$$
\begin{aligned}
\text{Bias}(\hat{\mathbf{Y}}_{\mathsf{CS}}^{\mathsf{MMRR}}|\mathbf{b}) &= \text{E}(\hat{\mathbf{Y}}_{\mathsf{CS}}^{\mathsf{MMRR}}|\mathbf{b}) - \text{E}(\mathbf{Y}|\mathbf{b}) \\
&= \text{E}(\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{Y}|\mathbf{b}) - \text{E}(\mathbf{Y}|\mathbf{b}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\text{E}(\mathbf{Y}|\mathbf{b}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{f}) \\
&= -(\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{Z}\mathbf{b} - (\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{f} \\
&= -(\mathbf{I} - [\lambda\mathbf{H}_{\mathsf{CS}}^{\mathsf{C}} + (1-\lambda)\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}])\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{Z}\mathbf{b} - (\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{f} \\
&= -\lambda(\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{C}})\mathbf{X}\boldsymbol{\beta} - (\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{Z}\mathbf{b} - (\mathbf{I} - \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}})\mathbf{f}.
\end{aligned}
$$

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{Y}}_{\mathsf{CS}}^{\mathsf{MMRR}}|\mathbf{b}) &= \mathrm{Var}(\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{Y}|\mathbf{b}) \\
&= \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathrm{Var}(\mathbf{Y}|\mathbf{b})\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}\prime} \\
&= \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}\prime} \\
&= \mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{R}[\lambda\mathbf{H}_{\mathsf{CS}}^{\mathsf{C}} + (1-\lambda)\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}]' \\
&= \lambda\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{C}\ \prime} + (1-\lambda)\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}\ \prime}.
\end{aligned}$$

Appendix P

The Cluster Specific Parametric Mixed Model

All results are derived with $\mathbf{b}$ random for a fixed bandwidth h, λ, and known $\mathbf{V}$.
The true model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{f} + \boldsymbol{\epsilon}$, where $\text{Var}(\boldsymbol{\epsilon})$=$\mathbf{R}$ and $\text{Var}(\mathbf{Y})$=$\mathbf{V}$.
The parametric model fit is $\hat{\mathbf{Y}}^{\mathsf{P}}_{\mathsf{CS}} = \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y}$, where
$\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}} = (\mathbf{I} - \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1})\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{Z}\mathbf{B}\mathbf{Z}'\mathbf{V}^{-1}$.

$$\begin{aligned}
\text{MSE} &= \text{E}[(\hat{\mathbf{Y}} - \text{E}(\mathbf{Y}|\mathbf{b}))(\hat{\mathbf{Y}} - \text{E}(\mathbf{Y}|\mathbf{b}))'] \\
&= \text{E}[(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - \text{E}(\mathbf{Y}|\mathbf{b}))(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - \text{E}(\mathbf{Y}|\mathbf{b}))'] \\
&= \text{E}[(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon))(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon))'] \\
&= \text{E}[(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon) - \text{E}[\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon)] + \text{E}[\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon)]) \\
&\quad (\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon) + \text{E}[\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon)] - \text{E}[\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon)])'] \\
&= \text{Var}(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon)) + \text{E}[(\text{E}(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon)))(\text{E}(\mathbf{H}^{\mathsf{P}}_{\mathsf{CS}}\mathbf{Y} - (\mathbf{Y} - \epsilon)))'] \\
&= \text{Var}(-(\mathbf{I} - \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}})\mathbf{Y} + \epsilon) + \text{E}[(\text{E}(-(\mathbf{I} - \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}})\mathbf{Y} + \epsilon))(\text{E}(-(\mathbf{I} - \mathbf{H}^{\mathsf{P}}_{\mathsf{CS}})\mathbf{Y} + \epsilon))'].
\end{aligned}$$

Now,

$$\begin{aligned}\mathrm{Var}(-(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})\mathbf{Y}+\epsilon) &= \mathrm{Var}(\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}\mathbf{Y}-\mathbf{Z}\mathbf{b}) \\ &= \mathrm{E}[\mathrm{Var}[\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}\mathbf{Y}-\mathbf{Z}\mathbf{b}|\mathbf{b}]]+\mathrm{Var}[\mathrm{E}(\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}\mathbf{Y}-\mathbf{Z}\mathbf{b}|\mathbf{b})] \\ &= \mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}\prime}+\mathrm{Var}[-(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})\mathbf{Z}\mathbf{b}] \\ &= \mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}\prime}+(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})\mathbf{Z}\mathbf{B}\mathbf{Z}'(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})'.\end{aligned}$$

The bias portion can be expressed as

$$\begin{aligned}&\mathrm{E}[(\mathrm{E}(-(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})\mathbf{Y}+\epsilon))(\mathrm{E}(-(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})\mathbf{Y}+\epsilon))'] \\ &= \mathrm{E}[(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})\mathrm{E}(\mathbf{Y}))((\mathrm{E}(\mathbf{Y}))')(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})] \\ &= \mathrm{E}[(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})'(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})] \\ &= (\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})'(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}).\end{aligned}$$

So the unconditional cluster specific mean square error for the parametric mixed model is

$$\mathrm{MSE} = \mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}\mathbf{R}\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}\prime}+(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})\mathbf{Z}\mathbf{B}\mathbf{Z}'(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})'+(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})(\mathbf{X}\boldsymbol{\beta}+\mathbf{f})'(\mathbf{I}-\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}).$$

The mean square error formula given above is for the cluster specific parametric mixed model. Similar mean square error formulas for the cluster specific CLMM and MMRR models may be found by replacing $\mathbf{H}_{\mathsf{CS}}^{\mathsf{P}}$ with $\mathbf{H}_{\mathsf{CS}}^{\mathsf{CLMM}}$ and $\mathbf{H}_{\mathsf{CS}}^{\mathsf{MMRR}}$, respectively.

Appendix Q

Proof of Lemma 1

(from Burman and Chaudhuri, 1992):

$$\begin{aligned}\|\hat{f}-\hat{g}\| &= \|\hat{f}-f-\hat{g}+\theta+f-\theta\| \\ &= \|(\hat{f}-f)-(\hat{g}-\theta)+(f-\theta)\| \\ &= \|\hat{f}-f\|^2+\|\hat{g}-\theta\|^2+\|f-\theta\|^2-2\langle\hat{f}-f,\hat{g}-\theta\rangle \\ &+ 2\langle\hat{f}-f,f-\theta\rangle-2\langle\hat{g}-\theta,f-\theta\rangle.\end{aligned}$$

Now,

$$\|\hat{f}-f\|^2 = \mathrm{O_p}(\pi)$$

$$\|\hat{g}-\theta\|^2 = \mathrm{O_p}(\gamma_s^2).$$

by assumptions A1 and A2. By the Cauchy-Schwartz Inequality,

$$\begin{aligned}|\langle\hat{f}-\mathrm{f},\hat{g}-\theta\rangle| &\leq \|\hat{f}-f\|\|\hat{g}-\theta\| &= \mathrm{O_p}(\pi\gamma_s)\\ |\langle\hat{f}-\mathrm{f},f-\theta\rangle| &\leq \|\hat{f}-f\|\|f-\theta\| &= \mathrm{O_p}(\pi\delta_s)\\ |\langle\hat{g}-\theta,f-\theta\rangle| &\leq \|\hat{g}-\theta\|\|f-\theta\| &= \mathrm{O_p}(\gamma_s\delta_s).\end{aligned}$$

Notice that

$$\begin{aligned}\|\hat{g}-\theta\| &\quad \text{dominates} \quad \|\hat{f}-f\|\\ \|\hat{g}-\theta\| &\quad \text{dominates} \quad \langle\hat{f}-f,\hat{g}-\theta\rangle\\ \langle\hat{g}-\theta,f-\theta\rangle &\quad \text{dominates} \quad \langle\hat{f}-f,f-\theta\rangle.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\hat{f} - \hat{g}\|^2 &= O_p(\gamma_s^2) + \|f - \theta\|^2 + O_p(\gamma_s)\|f - \theta\| \\ &= O_p(\gamma_s^2) + O_p(\delta_s^2) + O_p(\gamma_s)O_p(\delta_s).\end{aligned}$$

If $\lim_{s\to\infty} \delta_s \neq 0$, then

$$\|\hat{f} - \hat{g}\|^2 = O_p(1).$$

If δ_s=0, then

$$\|\hat{f} - \hat{g}\|^2 = O_p(\gamma_s^2).$$

QED.

Proof of Lemma 2

(from Burman and Chaudhuri, 1992):

Consider the case $\lim_{s\to\infty} \delta_s \neq 0$.

Now,

$$|1-\lambda^*| = |\langle \hat{g}-\theta, \hat{g}-\hat{f}\rangle| \|\hat{f}-\hat{g}\|^{-2}.$$

But

$$\begin{aligned}
|\langle \hat{g}-\theta, \hat{g}-\hat{f}\rangle| &\leq |\langle \hat{g}-\theta, \theta-f\rangle| + \|\hat{g}-\theta\|^2 + |\langle \hat{g}-\theta, \hat{f}-f\rangle| \\
&\leq \|\hat{g}-\theta\|\|\theta-f\| + \|\hat{g}-\theta\|^2 + \|\hat{g}-\theta\|\|\hat{f}-f\| \\
&= O_p(\gamma_s)\delta_s + O_p(\gamma_s^2) + O_p(\gamma_s\pi) \\
&= O_p(\gamma_s)\delta_s + O_p(\gamma_s^2).
\end{aligned}$$

Since $\lim_{s\to\infty} \delta_s \neq 0$, then $|\langle \hat{g}-\theta, \hat{g}-\hat{f}\rangle| = O_p(\gamma_s)$. Then, using Lemma 1, we can conclude that

$$|1-\lambda^*| = |\langle \hat{g}-\theta, \hat{g}-\hat{f}\rangle| \|\hat{f}-\hat{g}\|^{-2} = O_p(\gamma_s)$$

.

So $\lambda^* = O_p(\gamma_s)$.

Consider the case $\delta_s = 0$. Now,

$$\begin{aligned}
|\lambda^*| &= |\langle \hat{f}-\theta, \hat{g}-\hat{f}\rangle| \|\hat{f}-\hat{g}\|^{-2} \\
&\leq \|\hat{f}-\theta\|\|\hat{g}-\hat{f}\|\|\hat{f}-\hat{g}\|^{-2} \\
&\leq \{\|\hat{f}-f\|\|f-\theta\|\}\|\hat{f}-\hat{g}\|^{-1} \\
&= \{O_p(\pi) + \delta_s\}\|\hat{f}-\hat{g}\|^{-1}
\end{aligned}$$

Using Lemma 1 and the fact that $\delta_s = 0$ gives us,

$$|\lambda^*| = O_p(\gamma_s^{-1}\pi).$$

QED.

Proof of Theorem 1

(from Burman and Chaudhuri, 1992):

$$\begin{aligned}\|(1-\lambda^*)\hat{f}+\lambda^*\hat{g}-\theta\| &\le |1-\lambda^*|\|\hat{f}-\theta\|+|\lambda^*|\|\hat{g}-\theta\| \\ &\le |1-\lambda^*|\{\|\hat{f}-f\|+\|f-\theta\|\}+|\lambda^*|\|\hat{g}-\theta\| \\ &= |1-\lambda^*|\{O_p(\pi)+\delta_s\}+|\lambda^*|O_p(\gamma_s)\end{aligned}$$

If $\lim_{s\to\infty}\delta_s \neq 0$, we have

$$\begin{aligned}\|(1-\lambda^*)\hat{f}+\lambda^*\hat{g}-\theta\| &= O_p(\gamma_s)\{O_p(\pi)+\delta_s\}+O_p(\gamma_s^2) \\ &= O_p(\gamma_s).\end{aligned}$$

If $\delta_s = 0$, we have

$$\begin{aligned}\|(1-\lambda^*)\hat{f}+\lambda^*\hat{g}-\theta\| &= O_p(\pi\gamma_s^{-1})\{O_p(\pi)+0\}+O_p(\pi) \\ &= O_p(\pi).\end{aligned}$$

QED.

Appendix R

SAS code for the weekly wind speed dataset for local estimation at the data points.

```
/* ********************** for the local models****************************** */
/* The data set wind contains three variables– the cluster (denoted as cluster),
the explanatory variable (denoted as week), and the response (denoted as Y).*/


options nonotes; /*this option turns off the SAS log*/
proc sort data=wind out=wind;
    by cluster week;
run;
macro getband(h,xno,options=);
    proc datasets memtype=data lib=work nolist;
        delete allweights CLMMC CLMMM MLMMM;
    run;
    %do i=1 %to &xno;
        /*Obtaining the weights for each x0 value (x0=1 to &xno) */
        proc datasets memtype=data lib=work nolist;
            delete weights;
        run;
        data differences;
            set wind;
            X0=&i;
            diff=ABS(week-X0);
            weight = exp(-(diff/&h/(&xno-1))**2);
        run;
        proc means data=differences noprint;
        /*creating a dataset called weightsum that contains the sum of the weights*/
            var weight;
            output out=weightsum sum=sum;
        run;
```

```
    data differences;
        set differences;
        if _n_ = 1 then merge weightsum(keep=sum);
        w = weight/sum;
        z=1/(sqrt(w));
    run;
    data weights;
        merge differences wind;
    run;
    proc append base=allweights data=weights force;
    /*stacking the dataset for each set of weights*/
    run;
%end;
proc sort data=allweights;
    by x0;
run;
/* The conditional local mixed model */
/*In PROC MIXED, the outpm option gives the PA fits, the outp*/
/*option gives the CS fits, the sub=command defines the subject,*/
/*vi=1 gives the inverse of V for the first cluster, and the type=*/
/*option defines the variance covariance structure. The type=*/
/*option on the random statement is for B, and on the*/
/*repeated statement to model R.*/
proc mixed data=allweights;
    by x0;
    class cluster;
    model Y = week/ outp=CLMMC outpm=CLMMM solution;
    weight w;
    random int / sub=cluster solution vi=1 type=vc;
    repeated / sub=cluster type=vc;
run;
/* The marginal local mixed model */
proc mixed data=allweights;
    by x0;
    class cluster;
    model Y = week /outpm=MLMMM solution;
    weight w;
    random z / sub=cluster solution vi=1 type=vc;
    repeated / sub=cluster type=vc;
run;
ods listing; /*this turns on the printing in the SAS output window */
```

```
    data CLMMC;
    /*The conditional local mixed model cluster specific fits */
    /*(keeping those values where x=x0),where pred is the CS fit*/
        set CLMMC;
        where week=x0;
        keep cluster week y pred;
        rename pred=CLMMCS;
    run;
    data CLMMM;
     /*The conditional local mixed model population average fits */
    /*(keeping those values where x=x0),where pred is the PA fit*/
        set CLMMM;
        where week=x0;
        keep cluster week y pred;
        rename pred=CLMMPA;
    run;
    data MLMMM;
    /*The marginal local mixed model population average fits */
    /*(keeping those values where x=x0),where pred is the PA fit*/
        set MLMMM;
        where week=x0;
        keep cluster week y pred;
        rename pred=MLMMPA;
    run;
    proc print data=CLMMC;
    run;
    proc print data=CLMMM;
    run;
    proc print data=MLMMM;
    run;
%mend getband;
%getband(h=0.05,xno=53);
```

Bibliography

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

Assaid, C. (1997) "Outlier Resistant Model Robust Regression", Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Bishop, Y., Fienberg, S., and Holland, P. (1975), Discrete Multivariate Analysis: Theory and Practice. Cambridge : MIT Press.

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, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Cleveland, W. (1979), "Robust locally weighted regression and smoothing scatter plots", Journal of the American Statistical Association, **74**, 829-836.

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

Deaton, M.L., Reynolds, M.R., and Myers, R.H. (1983), "Estimation and hypothesis testing in regression in the presence of nonhomogeneous error variances", Communications in Statistics B12(1): 45-66.

Einsporn, R. (1987), "HATLINK: A Link Between Least Squares Regression and Nonparametric Curve Estimation", Ph.D. dissertation, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Einsporn, R., 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. (1992), "Variable bandwidth and local linear regression smoothers", Annals of Statistics, **20**, 2008-2036.

Fan, J., and Gijbels, I. (1995), "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation", Journal of the Royal Statistical Society B, **57**, 371-394.

Fan, J., and Gijbels, I. (1996), Monographs on Statistics and Applied Probability 66, Local Polynomial Modeling and Its Applications. London : Chapman and Hall.

Fan, Y.Q., and Ullah, A. (1999), "Asymptotic normality of a combined regression estimator", Journal of Multivariate Analysis, **71**, 2, 191-240.

Gasser, T. and Müller, H. (1979), "Kernel estimation of regression functions", in Smoothing Techniques for Curve Estimation, eds. Gasser and Rosenblatt. Heidelberg : Springer-Verlag.

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.

Harville, D.A. (1976), "Extensions of the Gauss-Markoff theorem to include the estimation of the random effects", Annals of Statistics, **4**, 384-395.

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, J. (1995), "Multilevel Diagnostics in Mixed and Hierarchical 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. (1978), "Nonparametric maximum likelihood estimation of a mixing distribution", Journal of the American Statistical Association, 73, 364, 805-811.

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

Lin, X.H., and Carroll, R.J. (2001), "Semiparametric regression for clustered data Using generalized estimating equations", Journal of the American Statistical Association, **96**, 1045-1056.

Magder, L.S., and Zeger, S.L. (1996), "A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians", Journal of the American Statistical Association, 91, 435, 1141-1151.

Mays, J.E. (1995), "Model Robust Regression: Combining Parametric, Nonparametric, and Semiparametric Methods", Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

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.

Myers, R.H. (1990), Classical and Modern Regression with Applications, second edition. Boston, MA : PWS-KENT.

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

Newton, M.A., and Zheng, Y.L. (1999), "A recursive algorithm for nonparametric analysis with missing data", Biometrika, **86**, 15 - 26.

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

Priestley, M., and Chao, M. (1972), "Nonparametric function fitting", Journal of the Royal Statistical Society B, **34**, 384-392.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1989), Numerical Recipes in Pascal : the Art of Scientific Computing. Cambridge : Cambridge University Press.

Robinson, P.M. (1988), "Root-N-consistent semiparametric regression", Econometrica, **56**, 931 - 954.

Robinson, T. (1997) "Dual Model Robust Regression", Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Ruppert, D., Sheather, S.J., and Wand, M.P. (1995), "An effective bandwidth selector for local least squares regression", Journal of the American Statistical Association, **90**, 1257-1270.

Schabenberger, O., and Pierce, F. (2001), Contemporary Statistical Models in the Plant and Soil Sciences. Boca-Raton, FL : CRC Press.

Searle, S.R. (1971), Linear Models. New York : Wiley and Sons.

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

Starnes, B.A. (1999), "Asymptotic Results for Model Robust Regression", Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

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

Tao, H., Palta, M., Yandell, B.S., and Newton, M.A. (1999), "An estimation technique for the semi-parametric mixed effects model", Biometrics, **55**, 102 - 110.

Ullah and Vinod (1993), "General nonparametric regression estimation and testing in econometrics", Handbook of Statistics, **11**.

Whittle, P. (1960) "Bounds for the moments of linear and quadratic forms in independent variables", Theory of Probability and its Applications, **3**, 302 - 305.

Wooldridge, J. M. (1992), "A test for functional form against nonparametric alternative", Econometric Theory, **8**, 452 - 475.

Vita

Megan Janet Tuttle Waterman was born January 22, 1975 in Auburn, New York to Earl and Constance Tuttle. She was a 1993 honors graduate from Marcellus High School in Marcellus, New York. She earned a Bachelor of Arts degree Magna Cum Laude in Mathematics and Economics from Nazareth College of Rochester in Rochester, New York in December, 1996. She continued her studies in the Statistics Department at Virginia Polytechnic Institute and State University where she received a Master of Science degree in Statistics in December of 1998. In May through July of 2000, she was an intern through the Lewis' Educational Research Collaborative Internship Program– a joint program through the Ohio Aerospace Institute and NASA Glenn Research Center in Cleveland, Ohio. Megan completed her Ph.D. degree in Statistics in May, 2002 and has accepted a job as a mathematician intern in Maryland.